ABSTRACT

Title: Classifying Attitude by Topic Aspect
for English and Chinese Document Collections

Yejun Wu, Doctor of Philosophy, 2008

Dissertation directed by: Professor Douglas W. Oard
College of Information Studies &
Institute for Advanced Computer Studies, UMCP

The goal of this dissertation is to explore the design of tools to help users make sense of subjective information in English and Chinese by comparing attitudes on aspects of a topic in English and Chinese document collections. This involves two coupled challenges: topic aspect focus and attitude characterization. The topic aspect focus is specified by using information retrieval techniques to obtain documents on a topic that are of interest to a user and then allowing the user to designate a few segments of those documents to serve as examples for aspects that she wishes to see characterized. A novel feature of this work is that the examples can be drawn from documents in two languages (English and Chinese). A bilingual aspect classifier which applies monolingual and cross-language classification techniques is used to assemble automatically a large set of document segments on those same aspects. A test collection was designed for aspect classification by annotating consecutive sentences in documents from the Topic Detection and Tracking collections as aspect instances. Experiments show that classification effectiveness can often be increased by using training examples from both languages.

Attitude characterization is achieved by classifiers which determine the subjectivity and polarity of document segments. Sentence attitude classification is the focus of the experiments in the dissertation because the best presently available test collection for Chinese attitude classification (the NTCIR-6 Chinese Opinion Analysis Pilot Task) is focused on sentence-level classification. A large Chinese sentiment lexicon was constructed by leveraging existing Chinese and English lexical resources, and an existing character-based approach for estimating the semantic orientation of other Chinese words was extended. A shallow linguistic analysis approach was adopted to classify the subjectivity and polarity of a sentence. Using the

large sentiment lexicon with appropriate handling of negation, and leveraging sentence subjectivity density, sentence positivity and negativity, the resulting sentence attitude classifier was more effective than the best previously reported systems.

Classifying Attitude by Topic Aspect for English and Chinese Document Collections

by

Yejun Wu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2008

Advisory Commmittee:

Professor Douglas W. Oard
Professor Dagobert Soergel
Professor Marilyn D. White
Professor John E. Newhagen
Professor Jimmy Lin

DEDICATION

To my parents. without whom this would not have been possible.

ACKNOWLEDGMENTS

Many people have contributed to my success in this endeavor. My great and sincere gratitude is due to my advisor, Dr. Douglas W. Oard, who has been inspiring, encouraging, challenging, and mentoring me throughout the dissertation research. Without his dedication of numerous hours of mentoring, it would have been impossible for me to succeed. Dr. Oard has not only guided me through my dissertation study, but also has had great influence on shaping my philosophy of structuring, connecting, and presenting research ideas.

Many other people contributed to the success of my dissertation as well. I am grateful to all the members of my dissertation committee: Dr. Dagobert Soergel, Dr. Marilyn D. White, Dr. John Newhagen, and Dr. Jimmy Lin, who met several times to challenge and enrich the ideas in the dissertation, and have provided valuable advice and comments. Special thanks go to Dr. John Newhagen for helping me connect to social psychology study of attitude. Colleagues in the Laboratory for Computational Linguistics and Information Processing (CLIP) at the University of Maryland Institute for Advanced Computer Studies (UMIACS) have been very helpful for discussing ideas and providing computing resources. Particular thanks go to Dr. Jianqiang Wang (now at the State University of New York at Buffalo) for providing his bi-directional Chinese and English translation probability tables and for explaining the general issues in cross-language information retrieval, Dr. Philip

Resnik and Michael Subotin for providing and tutoring Chinese processing tools (e.g., parser and segmenter), Dr. Gina-Anne Levow (now at University of Chicago) for providing Chinese processing resources (e.g., stopwords and lexicons), Dr. Bonnie Dorr and Dr. Amy Weinberg for guiding me into the world of computational linguistics, and Mark Damrongsri for setting up hardware and software.

Many professors at the College of Information Studies (CLIS) have also contributed to the success of my graduate study. I would like to greatly thank my earlier academic advisor, Dr. Robert B. Allen (now at Drexel University) who encouraged and financially sponsored me to take courses at the Department of Computer Science, and provided me with research experience in digital libraries so that I prepared myself for later graduate study, my earliest academic advisor Dr. Eileen Abels for understanding and supporting my pursuing a different research area from hers, Dr. Ann Prentice, former dean of CLIS, who made my attending the University of Maryland possible by financially sponsoring my beginning stage of graduate study, Dr. Diane Barlow for providing me with graduate assistantships for the earlier stage of my graduate study.

Several research projects at UMIACS have provided research assistantship for my studies. They have not only been helpful financially but also created an interdisciplinary, cross-department research environment to develop research ideas and to learn to cooperate to attack big research problems. Among them, the Joint Institute for Knowledge Discovery (JIKD) project has enriched my research skills in information retrieval; the Global Autonomous Language Exploitation (GALE) project has directly sponsored my dissertation research. Computing resources have been

provided by UMIACS and CLIS; their support is gratefully acknowledged.

I would like to thank the ACM SIGIR travel grant, which allowed me to attend SIGIR 2006 to discuss the dissertation ideas with Professor Bruce Croft and Dr. Dave Hawking, enhancing my confidence in bilingual aspect classification. My thanks also go to Japan's National Institute of Informatics (NII) NTCIR travel grant and University of Maryland Goldhaber travel grant, which sponsored my trip to the NTICR-6 Workshop, enriching my research ideas for Chinese attitude classification.

I owe my deepest thanks to my family. My parents and parents-in-law have been very supportive for my life and graduate study in a distant country. My wife was taking big family responsibilities in my first two years of graduate study when she took care of our little daughter, and she has accompanied me throughout my later stage of graduate study. I owe much to my parents-in-law who took care of our daughter when both my wife and I have been studying at the University of Maryland. I also thank my younger sister and brother who held family responsibilities when I was not available for help. Without their spiritual and logistic support, my graduate study in the United States would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

"If men are to be precluded from offering their sentiments on a matter which may involve the most serious and alarming consequences... reason is of no use to us."—George Washington[1] (First President of the United States, 1732-1799)[2]

"Attitude is a little thing that makes a big difference."—Winston Churchill (British Orator, and Prime Minister during World War II, 1874-1965)[3]

"A wise person makes his own decisions, a weak one obeys public opinion."—Chinese proverb[4]

Attitudes are a crucial aspect of human society, and they influence the affairs of persons, organizations, and nations. From the above quotes, it is clear that the importance of attitudes is perceived across cultures. Often we seek to make sense of both our own and relevant others' attitudes before we take actions.

[1] http://dictionary.reference.com/search?q=opinion

[2] http://en.wikipedia.org/wiki/George_Washington (last visited on October 8, 2006).

[3] http://en.thinkexist.com/quotes/with/keyword/attitude/ (last visited on October 8, 2006.

[4] http://www.sccs.swarthmore.edu/users/01/kyla/quotations/prov.html (last visited on October 8, 2006).

Attitudes are tendencies to like or dislike specific attitude objects [136], and so they have orientation, valence, or polarity. “Attitude” evokes several related concepts, such as opinion, sentiment, affect, emotion and mood, all of which have been studied in social psychology. I[5] do not differentiate between attitude, opinion, sentiment, affect, emotion, and mood in this dissertation, instead I focus on the operational task of automatically measuring the semantic orientation of whatever combination of these concepts I find expressed in text. So I use “attitude” in this dissertation as an inclusive term for the set of expressions that exhibit positive, negative, neutral, or ambivalent tendencies. This dissertation aims to create components for an information system that users could employ to compare attitudes toward the aspects of a topic in news articles across multiple languages. The components are evaluated with intrinsic measures using fixed test collections; integrating these components into a complete interactive system remains for future work. This chapter addresses motivations, the design space, research questions, contributions, and the organization of the remainder of the dissertation.

1.1 Motivations

This dissertation is motivated by a desire to characterize attitudes about aspects of a topic. Before addressing this goal, I address the needs for attitude, how the

[5] Whether to use “we” or “I” is a matter of writing style. ”We” and ”I” are actually used interchangeably in the dissertation based on context. Many people, indicated in the Acknowledgements, contributed to the ideas discussed in the dissertation, but the experiments and writing are entirely my work. The term “the investigator” also refers to me.

needs are served today, our desire to automate the way we serve the needs, and the sources we can use to mine attitude.

1.1.1 Attitude: Needs and Service

Needs for Attitude

Why do we care about other people's attitudes? Many attitudes serve to help us understand our world and to make sense of occurrences around us. "This does not imply that they provide a factually truthful picture of the world — merely one that is meaningful and understandable to the particular individual who holds them" [132](p.76). Attitudes represent people's views and feelings about the issues concerned and have implications for potential behavior [51].

Attitudes can be used to predict behavior with considerable success under appropriate conditions [135]. Attitudes guide behavior in two ways. When the situation both motivates the individual to consider his or her action carefully and allows the individual the opportunity for conscious consideration, attitudes guide behavior through deliberation about the implications of one's attitudes for a given course of action (this is the theory of reasoned action) [158]. In other cases, attitudes act as an association in memory between the attitude object and one's evaluation of that object — the attitude must be activated from memory when the individual sees the attitude object if the attitude is to exert any influence (this is Fazio's Attitude-to-Behavior Process Model) [158]. "Attitudes that are accessible from memory can guide an individual's behavior in a satisfying direction without the individual having to engage in conscious, deliberative reasoning" [158](p. 90).

Whereas attitudes generally guide behavior, emotions particularly do this by changing our motivations [45]. Emotion influences cognition by short-circuiting cognitive processing. "Our feelings provide us with information. We use our awareness of our feelings to make evaluative judgments and decisions, based on how we feel" [45](p. 137). Emotion thus helps us to organize our behavior (although negative moods may disorganize behavior) [45]. Emotion also has interpersonal functions — "emotional expressions serve to inform others about the expressers' intentions and motives, and function to motivate various actions of the part of the perceiver" [45](p. 139).

Examples of Needs for Information About Attitudes

Since attitude can help to guide behavior and emotion may influence motivation, people in government, commercial, and political domains are eager to collect and understand other people's attitudes and emotions toward specific objects of concern. People may have different information seeking tasks concerning attitudes, as in the following examples:

- Political candidates may wish to know both aggregate attitudes regarding their candidacy and which groups of people like/dislike specific positions the candidate has taken.

- Policy makers may want to know the attitudes expressed to different audiences by institutional stakeholders (e.g., foreign governments) on an issue.

- Advertisers may want to measure changes in aggregate attitudes after an advertisement is delivered to a targeted population.

- Individuals may want to know a certain celebrity's attitudes about an issue, or to find people who share their attitudes with whom they might have a discussion, or to find people who disagree with their attitudes who they might thus seek to persuade.

- Scholars in various disciplines study attitudes. Survey researchers study public opinions.

- Journalists use other people's attitudes to frame stories.

How the Needs are Served

Survey research analyzes and measures opinions and attitudes expressed in large-scale surveys. Opinion polls or surveys, such as the Gallop Poll,[6] are often used to examine public opinions toward social issues and to measure consumer behavior. Psychometricians, e.g., Guttman, Guilford, and Likert, developed the systematic attitude measurement techniques that quantify the strength of a person's conviction in an opinion [175]. The goal of many surveys is to learn about the distribution and correlates of attitudes in a population by collecting reports from a representative sample [155]. Here are three typical example public opinion survey questions created

[6] Gallop Poll, http://poll.gallup.com/ (last visited on October 10, 2006).

by the PEW Research Center:[7]

1. Now thinking about our country, overall, are you satisfied or dissatisfied with the way things are going in our country today?

2. Please tell me if you have a very favorable, somewhat favorable, somewhat unfavorable, or very unfavorable opinion of the United States.

3. In your opinion, would it be a good thing or a bad thing if the European Union becomes as powerful as the U.S.?

Historians ask factual questions and opinion questions when doing their research. For instance, some American historians study "popular history: what the middle class think about the history and culture of their nation" [19]. Darwin scholars would like to know, among the set of Darwin's personal letters, in which letter Darwin vigorously defended his theory of evolution[8]. To answer these questions, they need to find relevant information objects that express attitudes about their topics.

"Content analysis is any research technique for making inferences by systematically and objectively identifying specified characteristics within text" [166](p. 5). It has been primarily associated with research in the field of journalism and empirical political science, and has been undertaken in a variety of other disciplines, such as

[7] The PEW Research Center, http://people-press.org/reports/ (last visited on October 10, 2006).

[8] The example is from Daniel Cohen of the Center for History and New Media at George Mason University in February, 2006.

anthropology, education, history, literature, philology, psychiatry, psychology, and sociology [166]. To meet a need of finding negative attitudes toward an object, Media Tenor, an international media analysis company "offers current and up-to-date research on the basis of Media Tenor comprehensive media content analysis with particular emphasis on issues such as Customer Service Representative, Brand Value, Country Images, Reputation Management, CEO-Benchmarks and Election campaigning."[9] Their content analysis is conducted manually.

Since both survey and manual content analysis for examining public opinions are expensive, being able to automatically identify opinions and attitudes in text should be of interest to many users. Computer scientists have only recently started to try to answer questions about opinions, such as those listed in the OpQA collection [167]:

Topic: Kyoto

- Q: Are the Japanese unanimous in their opinion of Bush's position on the Kyoto Protocol?
- Q: How do European Union countries feel about the US opposition to the Kyoto protocol?

Although people communicate their opinions and attitudes via various media (such as radio, television, Internet, lectures, talks) and people seek opinions and attitudes through various venues (such as interviews, surveys, or reading blogs), the questions of who seeks and how people seek and use information about attitudes have not yet been systematically studied in library and information science.

[9] http://www.mediatenor.com/ (last visited on October 8, 2008).

Library users do seek information about opinions and attitudes. Although people seek factual information such as "tell me the governing party of Canada," they also seek subjective, evaluative information such as [160]:

(1) Locate a brief discussion of the merits of the paperback novel.

(2) Locate an evaluative biography of the late British poet A.E. Houseman.

(3) Find an annotated, evaluative list of books on word usage.

(4) Find evaluative remarks on recent books in the field of information science.

So people seek opinions and attitudes from multiple sources, such as books, personal letters, news articles, or by directly surveying people. Next we addresses the sources of information about attitude in this research.

1.1.2 Sources of Information About Attitude

Attitudes are often reported in formal documents, such as news articles and legal documents, they are often expressed directly in informal documents, such as personal letters, discussion boards, mailing lists, emails, Usenet, blogs (or weblogs), and conversational speech. Some sources (e.g., personal letters and emails) are generally not available for research on automated attitude analysis due to privacy issues. In this study, we therefore consider news articles and blogs as candidate sources, and ultimately settled on news articles for experiments in order to leverage investments in existing test collections.

News

“News” is a research topic in journalism just as “relevance” is in information science. There are many definitions of news. Charles A. Dana who ran the New York Sun from 1869 to 1897 defined news as “anything that interests a large part of the community and has never been brought to its attention before” [117](p. 56). Two general guidelines abstracted from the various definitions are [117]:

- News is information about a break from the normal flow of events, an interruption in the expected.
- News is information people need to make sound decisions about their lives.

News is a “selective account of reality” [116]. News is “not merely a neutral reflection of events or record of public debates,” but “a social production” that stems from journalists’ news decision making (e.g. selecting, editing, presenting news elements) under specific individual, cultural or social circumstances [23](p. 101). The main factors that seem to influence a journalist’s news decisions are: “news factors,” “institutional objectives” (e.g., economic, cultural, political or ideological objectives), and “the subjective beliefs of journalists” (i.e., journalists’ predispositions toward an issue or an actor) [33](p. 135).

Despite differences across countries, “journalists from different countries share many news values, and even the characteristics of them are quite identical worldwide” [186](p456). Zhong [211] studied news decisions of a total of 60 U.S. journalists and 60 Chinese journalists, and found that both groups used psychologically salient news elements (i.e., information with some generally agreed-upon news values) in their stories more often than they used cultural or ideological elements.

Reporters and editors generally strive for balance and fairness. Balance means that views from all related sides are covered. Fairness means honesty, straightforwardness ahead of flashiness, and reporting relevant and important facts [117]. Unfair and unbalanced journalism might be described as failures in objectivity. Objectivity in journalism means that "the news story is free of the reporter's opinion and feelings, that it contains facts and that the account is by an impartial and independent observer" [117](p. 46).

Lack of balance and the absence of fairness are often inadvertent. "Since writing is as much as act of the unconscious as it is the conscious use of controlled and disciplined intelligence, the feelings of reporters crop up now and then" [117](p. 44).

We can therefore imagine three sources of attitudes in news articles. One is news stories that unconsciously (or consciously) report the journalists' attitudes. Henley et al. did content analysis of newspaper articles for biases toward anti-gay attacks by analyzing the frequency and specificity of using nominal referents to the violence. They found a significant difference in the attitudes between two major US newspapers — the Washington Post and the San Francisco Chronicle (1988-94) — in reporting anti-gay attacks [68]. The Washington Post (addressing a less gay-friendly community) strongly distinguished its treatment of crimes against lesbians and gay men from that of crimes against straights, whereas the Chronicle (directed at a more gay-friendly target audience) made no significant distinction in its reporting of violence against gays and straights. “The difference is due to bias and not something intrinsic to the crimes themselves (e.g. that crimes against straights are more viciously carried out)” [68](p. 95).

A second source of attitude would be news stories that explicitly report the attitudes of all involved parties of an issue. The third source would be intentionally opinionated coverage, such as commentary, editorials, Op-Eds, and news analysis.

In countries with a free press, reporters typically try to cover all sides of an issue and to eliminate bias, but many countries also have operated state-run news organizations, which present the government’s views. Therefore news articles may report attitudes of governments, news organizations, or individuals.

Blogs

Blogs (or weblogs), defined as “frequently modified web pages in which dated entries (posts) are listed in reverse chronological sequence”, are becoming an increasingly popular form of communication on the World Wide Web [70](p. 1). Blog types, grouped according to their primary purpose, include filter blogs (in which authors link to and comment on the contents of other web sites), personal journals, and knowledge blogs (k-logs) [70]. Educators and business people see blogs as environments for knowledge sharing; blogs created for this purpose within an organization are sometimes called knowledge logs [50, 69, 141]. Most research has focused on filter blogs and personal journals, characterizing a blog as “a Website that contains an online personal journal with reflections, comments and often hyperlinks provided by the writer" [187]. These online diaries come in many shapes and styles, ranging from "people willing to sharing their views, pictures and links, to companies interested in another way of reaching their customers" [8]. According to Sifry's Alerts,[10] the blog analysis firm Technorati now tracks over 35.3 million blogs; the blogosphere is doubling in size every 6 months; it is now over 60 times bigger than it was 3 years ago; and on average, a new blog is created every second of every day.

Blogs are a type of computer-mediated communication. Computer-mediated communication (CMC) has been reported to be "significantly higher than face-to-face (FtF) communication on certain types of hostile or profane speech acts, leading to characterizations of CMC as uninhibited and depersonalized" [182]. CMC language is more egocentric than FtF speech [182]. "CMC may represent a new resource for eliciting emotionally rich, relationship-oriented verbal interaction among emotionally disturbed adolescents" [212](p. 228).

Blogs are being hailed as fundamentally different from other Internet communication protocols (e.g., email, Web), and as possessing a socially-transformative, democratizing potential [69]. Journalists see blogs as alternative sources of news and public opinion [96]. "Blogs tend to be impressionistic, telegraphic, raw, honest, individualistic, highly opinionated and passionate, often striking an emotional chord" [97]. Private individuals create blogs as a vehicle for self-expression and self-empowerment [9, 69]. Therefore the blogosphere is a huge information space of

[10] http://www.sifry.com/alerts/archives/000432.html (last visited on October 10, 2006).

unstructured informal text in which attitudes are embedded.

The blogosphere is multilingual, and deeply international. According to Sifry's Alerts,[11] English, while being the language of the majority of early bloggers, had fallen to less than a third of all blog posts by April 2006. Japanese and Chinese language blogging had grown significantly. Chinese language blogging, while continuing to grow on an absolute basis, had begun to decline as an overall percentage of the posts that Technorati tracked between October 2005 and April 2006. The ten languages with the greatest number of posts tracked by Technorati in April 2006 were: Japanese (37%), English (31%), Chinese (15%), Spanish (3%), Italian (2%), Russian (2%), French (2%), Portuguese (2%), Dutch (1%), and German (1%). Korean and Persian languages were undersampled in that analysis.

From this overview, we can conclude that news articles would be a good source for filtered, institutional, and governmental attitudes, whereas blogs would be a good source for personal attitudes, and that both types of sources are multilingual. Each type of source is interesting in its own right, and the combinations could well be more interesting than either one alone.

1.2 An Ultimate Goal

When evaluating an issue, people may simply express an overall attitude (such as good or bad, like or dislike, positive or negative, pro or con), or they may express opinions about several aspects of the issue. This happens in our daily life when we

[11] http://www.sifry.com/alerts/archives/000433.html, published on April 17, 2006 (last visited October 10. 2006).

evaluate products, organizations, persons, issues, academic courses, etc. Although both genres of evaluations are important, evaluating aspects provides more information. The notions of topic, aspect, and attitude are operationally defined as follows.

Topic. A topic is an attitude object, which can be a specific entity (such as a person, an organization, a location, a product, a concept or an issue) or an event. In this dissertation, **topic** and **issue** are used interchangeably to refer to attitude objects.

Aspect. An aspect[12] refers to the subtopics of the topic on which attitude holders have expressed their evaluation. This includes the attributes, parts, or facets of the topic. So we mean "aspect" in the same sense as was used in the TREC[13] Interactive Track [198]. We care about the essential and specific aspects of a topic. For instance, when evaluating the issue of abortion, its aspects may include the ethics, the health impacts of abortion, etc. We do not care about the general aspects of a topic, such as the temporal and spatial aspects of an event.

Attitude. Attitude here is operationally defined as the combination of a set of related psychological concepts (i.e., attitude, opinion, sentiment, affect, emotion) that have semantic orientations, so it is an affective evaluation that the attitude holder uses to comment on the topic or its aspects.

Words, sentences, documents, and even Chinese characters may all be used to

[12] In linguistics, "aspect" often denotes "grammatical aspect;" we consistently mean "topic aspect."

[13] Text Retrieval Conference, organized by National Institute for Standards and Technology.

express attitudes. “Semantic orientation” has several synonyms — valence, polarity, attitude tendency. In this dissertation, we use polarity as a basic level concept for the set of concepts. When addressing attitude at word level, we use semantic orientation; when addressing attitude at sentence level, passage level, or document level, we use polarity; when addressing attitude of Chinese characters, we use attitude tendency. When introducing other researchers’ works, the original terms in their work, such as sentiment tendency for Chinese characters, were used to avoid confusions.

Attitude can be explicitly expressed with explicit evaluative orientation or be implicit in the style of expression (such as sarcasm). This research focuses solely on explicit expression of attitude. Automatically making an inference of evaluative orientation from humorous, sarcastic, or implied sentiments is itself an artificial intelligence research topic.

We seek to help users compare attitudes toward the same aspects of a topic across languages. Such a system will need to:

- search the blogosphere and newswire collections in Chinese and English,
- identify aspects of the topic,
- detect the polarity of attitudes toward those aspects in each language, and
- align the aspects across languages for the purpose of comparing the attitudes.

A display of system results we envision (what Russell calls a representation design [148]) is shown in Figure 1.1. A system structure for a general solution is described in Figure 1.2. Due to the text genre difference between news and blogs,

Topic:	Chinese Space Program			
Time Period:	Jan. 1, 2005 - Oct. 1, 2006			
	English (US)		**Chinese (China)**	
Source	News	Blog	News	Blog
Aspect 1: Pride	+	+	++	++
Aspect 2: Technology	−	−	+	+−
Aspect 3: Cost	−	−		0
Aspect 4: Impact on USA	−	−−		+

++: highly positive;

+: somewhat positive;

−: somewhat negative;

−−: highly negative;

+−: ambivalent (both highly positive and highly negative);

0: neutral (both weakly positive and weakly negative);

blank: no attitude (either not subjective or not addressed).

Figure 1.1: An example of the ultimate goal.

system components built for news might not work well for blogs, so the system components are genre-specific. The system has the following key components.

Component 1: **Document segment splitters for aspect analysis**. We want to find an appropriate text granularity for the analysis of an aspect of a topic. We can imagine that a whole document is very likely to address multiple aspects of the topic, but an appropriate document segment is more likely to address only one aspect. Although it can be a dissertation topic to identify optimal segment size for aspect analysis, we simplified the task for this study by tuning existing software to extract suitable segments (or passages).

Component 2: **Search engines for searching Chinese and English news and blogs**. If the end user can read and write both Chinese and English, two search engines must be built, one for each language. Document segments (or passages) are the natural granularity for indexing since a full document may address multiple aspects of a topic.

Component 3: **Classifiers for detecting document segments that address specific aspects of a topic**. This is an aspect classification problem with aspect alignment between two languages. There are two approaches - unsupervised aspect clustering and supervised aspect classification. Unsupervised aspect clustering involves two steps. The first step is unsupervised clustering, in which a system automatically clusters document segments. When clusters are automatically generated, either the system recommends a label for each cluster, or the user examines the clusters and labels them. The second step is aligning the aspect clusters across two languages. The second approach is supervised text classification in which the user

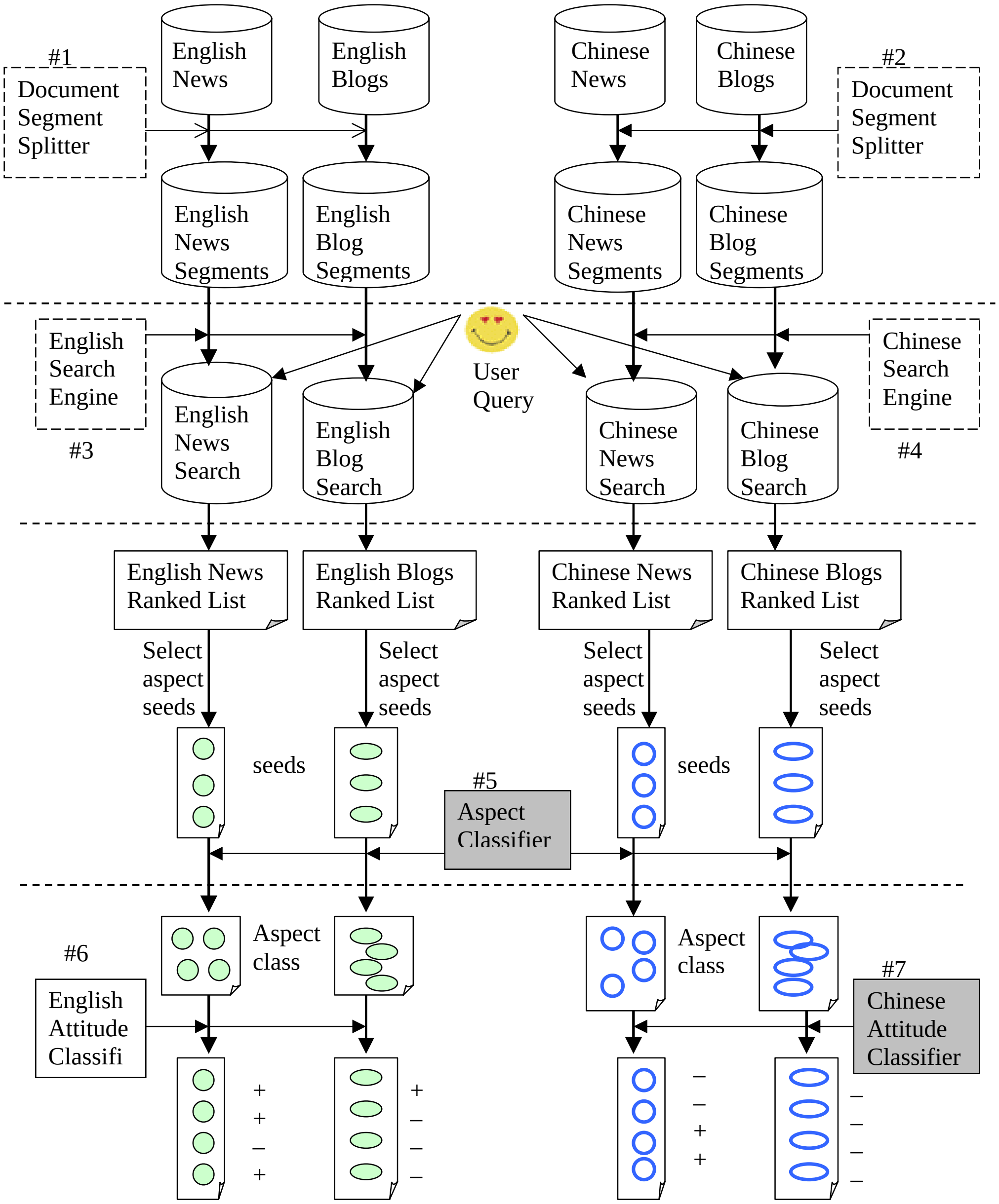


Figure 1.2: An architecture for a general solution.

System components depicted with rectangles are numbered.

Dashed rectangles (#1, #2, #3, #4): built for news in this study.

Solid rectangle without background (#6): built for blogs in an earlier study.

Solid rectangles with grey background (#5, #7): main foci of this study.

provides some seed (or training) document segments about a specified aspect in two languages, then the system classifies the remaining segments. An aspect class generated in this way can be easily labeled by the user when the training segments are designated and the resulting classes are already aligned across the two languages. Methods for this component are elaborated in Chapter 3 (bilingual aspect classification).

Component 4. **An English attitude classifier for detecting attitudes expressed in English document segments**. Attitude classifiers based on linguistic analysis have been the focus of extensive research as described in the literature review (Chapter 2). We have built a simple English attitude classifier for blogs, the results of which are recorded elsewhere [127], but that is not a focus of our present research.

Component 5. **A Chinese attitude classifier for detecting attitudes in Chinese document segments**. Attitude classifiers based on automated linguistic analysis are built to do this. Techniques for this component will be elaborated in Chapter 4 (Chinese attitude classification).

To facilitate better understanding of the system capabilities when the whole system is built, a use case is described in the next section, followed by a narrower research goal for this study.

1.3 A Use Case

The use case here describes, step by step, the main user actions, the system operations in response to the user actions, and the output of system operations (see Table 1.1).

Step	Operations	Descriptions
1	System operation	Process English and Chinese news and blogs into document segments, index English document segments; index Chinese document segments
	System output	a user interface for English search engine, a user interface for Chinese search engine.
2	User action	Issue an English query (or two queries, one for searching news, the other for searching blogs) about a topic.
	System operation	Process the query and retrieve 2 sets of relevant document segments (with news and blogs separated).
	System output	Present a list of English news articles and a list of English blogs.
3	User action	Examine the sets of English news and blogs, identify an aspect of the topic and select several seed segments that are about that aspect, and label the aspect. Repeat as needed for additional aspects of the same topic.
	System operation	For each aspect, identify additional English document segments on that aspect.

	System output	For each aspect, many news and blog segments that address that aspect.
4	User action	Issue a Chinese query (or two queries, one for searching news, the other for searching blogs) about the same topic.
	System operation	Process the query and retrieve 2 sets of relevant document segments (with news and blogs separated).
	System output	Present a list of Chinese news articles and a list of Chinese blogs.
5	User action	Examine the sets of Chinese news and blogs, and, for each aspect identified in the English segments, select several seed segments.
	System operation	For each aspect, identify additional Chinese document segments on that aspect.
	System output	For each aspect, many news and blog segments that address that aspect.
6	System operation	Perform English and Chinese attitude classification.
	System output	Summary of attitude categories for each aspect, presented in a representation like Figure 1.1.

Table 1.1: A use case.

The assumed user's task is to seek the attitudes of several aspects of a subject.

The user is assumed to be able to read and write in Chinese and English and to have the information literacy of searching information in search engines. In the use case, the user examines English news and blogs (at step 3) before she examines Chinese news and blogs (at step 5), but this is not the enforced order. The user can take any order at her will.

Our ultimate goal could be expanded further to include more languages (e.g., Spanish), more sources (e.g., transcribed speech from talk shows) or more automation (e.g., automatically segregating blogs from different demographic groups). But the ultimate goal outlined here is already too large for one dissertation. We have, therefore, elected to focus on news for now, and to leave the corresponding work with blogs for future work. The reason we focus on news is that we have news collections with pre-cleared intellectual property rights and news test collections for both bilingual aspect classification and Chinese attitude classification, which are introduced in Chapter 3 and Chapter 4.

1.4 Research Questions

In the previous sections, we introduced English and Chinese aspect classification and Chinese attitude classification. This section addresses the research questions involving these two tasks. The research questions are narrowly defined so that experiments can be designed to answer them.

Bilingual Aspect Classification

Context. A bilingual user who can read and write in English and Chinese is interested in finding text about aspects of a topic from two document collections in two languages. She will search for documents on the topic in the two collections and browse the two sets of retrieved documents to identify the aspects of the topic which interest her. She will define an aspect instance by selecting a segment (or portion) of a document. She cannot hope to read all retrieved documents, so she will instead read several retrieved documents and select a few (e.g., 2-4) aspect instances for each aspect from the two collections, relying on a classification system that could help her to find the remaining aspect instances.

Assumption. The granularity of a text unit appropriate for defining aspect instances is important for aspect classification. Since a whole document may address multiple aspects of a topic, we assume document segments are appropriate as aspect instances, assuming that one document segment addresses one aspect. A document segment is composed of consecutive sentences.

The instances the user selected for an aspect serve as training examples for the aspect classification system. Since we have only a few training examples for each aspect in each language, it may be useful to train an aspect classification system using the training examples in the two languages. So we will ask, can we use the training examples in the two languages to improve classification performance?

Research question. Does adding training examples in a second language (in the same aspect) help aspect classification using training examples in the first language only?

We define the first language (or native language) as **main language**, and the second language (or foreign language) as **supporting language**, then the research question can be rephrased as: does adding supporting-language training examples help aspect classification using main-language training examples only? Whether a language is a main language or a supporting language depends on the user's choice. In this study, we deal with English and Chinese only.

A more general research question we can ask is: given a small number of training examples, each of several languages, does combining the training examples from multiple languages improve classification effectiveness using training examples for a single language only?

Research method. To answer this question, we designed a test collection for aspect classification by annotating consecutive sentences in documents from the Topic Detection and Tracking evaluations (TDT3 and TDT4) as aspect instances. We tried three variants of the k-Nearest-Neighbor (kNN) technique for aspect classification, using document segments generated as a fixed partition by TextTiling. The automatically generated document segments were mapped onto the human-annotated segments for evaluation purpose. We mapped the document segment vectors between English and Chinese using a statistical translation technique, and reduced the dimensionality of the term space through local Latent Semantic Analysis (LSA).

Our experiment results are very encouraging. Experiments show that when few training examples are available in either language, classification using training examples from both languages can often achieve higher effectiveness than using training examples from just one language. Results of this study are presented in Chapter 3.

Chinese Attitude Classification

Context. In our ultimate goal (introduced in Section 1.2), the task for Chinese attitude classification is to identify the attitudes expressed about an aspect in a document segment. The best presently available test collection for Chinese attitude classification (the NTCIR-6[14] Chinese Opinion Analysis Pilot Task test collection) is, however, focused on sentence level classification. Here we have elected to focus on sentence-level attitude classification due to two reasons. One is that, although it is unknown whether human beings aggregate segment-level attitude from sentence-level attitude, we can use segment-level attitude to generate segment-level attitude as a system design approach. So sentence-level attitude classification is useful. The other reason is that designing a test collection for other kinds of segments can be both time-consuming and expensive.

Assumption. Since a segment is composed of sentences, we expect that sentence-level classification can be usefully aggregated to perform segment-level attitude classification, at least as a system design approach.

Research Questions. There are two main research questions for Chinese sentence attitude classification:

[14] NTCIR stands for Japan's National Institute for Informatics Test Collection for Information Retrieval

- How best to detect Chinese sentences in which an attitude is expressed (i.e., “opinionated” or “subjective” sentences)?

- How best to classify the polarity of opinionated Chinese sentences?

Research Method. We model the attitude expressed in a sentence as an aggregate of the attitude expressed by the semantic orientations of the words in the sentence, so evidence about the semantic orientation of Chinese words is needed. A Chinese sentiment lexicon was constructed using four lexical resources:

- First, we acquired Chinese prior-polarity positive and negative lexicons from National Taiwan University (NTU) [93].

- Second, we automatically translated Wilson and Wiebe’s English prior-polarity subjectivity lexicon [197] into Chinese and manually pruned the translated lexicon.

- Third, we extracted words with annotated polarity from four sets of training documents for the NTCIR-6 Opinion Analysis Pilot Task.

- Fourth, we rekeyed a dictionary of positive Chinese terms [205] and a dictionary of negative Chinese terms [159] using Traditional Chinese characters (since the test collection is in Traditional Chinese).

- For words not in the lexicon, we extended Ku et al.’s character-based approach for computing the prior polarity of unknown Chinese words [93].

We adopted a shallow linguistic analysis approach to classifying the subjectivity and polarity of a sentence. When aggregating a sentence’s subjectivity and polarity from its words, a negation mechanism and the following features were taken into consideration:

- sentence subjectivity density,
- sentence aggregated polarity,
- sentence positivity, and
- sentence negativity.

We tested three negation mechanisms for a subjective word:

- *1-word adjacency negation*: the negation is applied to the word immediately following it;
- *2-word adjacency negation*: the negation is applied to the two words immediately following it;
- dependency-based negation: the dependency relationship between a negation word and its related words in a sentence is applied based on dependency parsing.

By combining various features and adjusting thresholds for subjectivity and aggregated polarity, we created 12 systems for subjectivity classification and 21 systems for polarity classification. Using our largest sentiment lexicon with the "2-word adjacency negation" approach, and leveraging sentence subjectivity density, sentence positivity and negativity, we have created systems more effective than the best previously reported systems. Experimental results for this study are presented in Chapter 4.

1.5 Contributions

The contributions of this dissertation include conceptual contributions, technical contributions, and the resources created for relevant research communities.

- Conceptual contributions:
  - Defining a novel grand challenge problem that can guide the development of specific technologies for aspect and attitude classification.
  - Enriching the research space for topical classification by studying topical classification at a lower level of granularity (i.e., aspects) across languages.
  - Drawing on social psychology theories of attitude activation and presenting experiment results consistent with those theories.
- Technical contributions:
  - Demonstrating improved aspect classification effectiveness from integrating cross-language text classification techniques.
  - Demonstrating improvements over existing character-based approaches to semantic orientation classification for Chinese words.
  - Demonstrating Chinese sentence subjectivity and polarity classification techniques that are more effective than the best previously reported results.
- Resource contributions:

  - Developing a new test collection for bilingual and cross-language aspect classification.
  - Creating a large Chinese lexicon for use in attitude classification.

1.6 Organization of the Dissertation

The dissertation is organized as follows. Chapter 2 introduces social psychological studies on attitude and related concepts, then reviews previous research work on topical classification, subjectivity analysis, English attitude classification, and Chinese attitude classification. Chapter 3 describes bilingual aspect classification, including the methodology, test collection design, experimental design, experimental results, findings and discussion. Chapter 4 describes Chinese sentence attitude classification, including the methodology, experimental design, experimental results, findings, and discussion. Chapter 5 concludes the dissertation with an overview of the results, a discussion of limitations, and an outline of future work.

Chapter 2

Background and Related Work

This chapter reviews previous research on topic-based classification, introduces social psychology studies of attitude and related concepts, and then reviews English and Chinese attitude classification research. Topic-based classification techniques include clustering and automatic text classification. English attitude classification includes general approaches to subjectivity analysis and specific techniques for classifying words, sentences and documents. Chinese attitude classification includes approaches for classifying Chinese words and sentences.

2.1 Document Clustering and Automated Text Classification

Aspect classification is a topical classification task in which aspects are like any other topics with a finer granularity, and in which the aspects of a broader topic may be related in some way. There are two approaches for topic-based classification: clustering, which is typically an unsupervised learning approach, and classification, which is typically a supervised learning approach.

2.1.1 Document Clustering

Clustering is a statistical technique for multivariate analysis that assigns items to automatically created groups based on a calculation of the degree of association between items and groups. The formed groups should have a high degree of association between members of the same group and a low degree of association between members of different groups. Here, the data set of our interest is document set, so we are dealing with document clustering or classification.

In order to cluster documents into groups, some means of quantifying the degree of association between them is required. This may be a distance measure, or a measure of similarity or dissimilarity. The determination of inter-document similarity depends on both the document representation (e.g., the weights assigned to the indexing terms characterizing each document), and the similarity function that is chosen [140]. If documents are represented using the vector-space model, each document, d, is represented as a vector, **d**, in the term-space,

$$\mathbf{d} = (w_1, w_2, ..., w_n),$$

where w_i is the weight of the i^{th} term in the document, which is usually computed based on Term Frequency (TF) and Inverse Document Frequency (IDF).

In document clustering, the function that computes the similarity of two document vectors is usually normalized by the document length. Examples include the Dice coefficient, Jaccard coefficient, and cosine coefficient [140]. The cosine coefficient is the most commonly used, which is defined as [164]:

$$cosine(\mathbf{d}_1, \mathbf{d}_2) = \frac{\mathbf{d}_1 \bullet \mathbf{d}_2}{|\mathbf{d}_1||\mathbf{d}_2|}$$

where $\bullet$ indicates the vector dot product and $|\mathbf{d}|$ is the length of vector **d**. Given a set of documents, S, the **centroid** vector **c** of S is defined as:

$$\mathbf{c} = \frac{1}{|S|} \sum_{d \in S} \mathbf{d},$$

which is the vector obtained by averaging the weights of the various terms present in the document set S. Note that a centroid almost never corresponds to an actual data point. Similarly, the similarity between a document and a centroid vector, and between two centroid vectors can be computed using the cosine coefficient, that is,

$$cosine(\mathbf{d}, \mathbf{c}) = \frac{\mathbf{d} \bullet \mathbf{c}}{|\mathbf{d}||\mathbf{c}|},$$

$$cosine(\mathbf{c}_1, \mathbf{c}_2) = \frac{\mathbf{c}_1 \bullet \mathbf{c}_2}{|\mathbf{c}_1||\mathbf{c}_2|}$$

The two main types of clustering methods are *non-hierarchical* (or partitioning), which divides a document set of N items into K clusters, and the *hierarchical*, which produces a nested document set in which pairs of documents or clusters are successively linked until every document in the document set is connected [140]. When no overlap is allowed, the result is called a partition.

The non-hierarchical methods can create a one-level (un-nested) partitioning of the document set based on the idea that a cluster is represented by its centroid. While there are many partitioning techniques, the K-means algorithm is widely used in document clustering, which is briefly described below [164]:

1. Select K document points as the initial cluster centroids;
2. Assign all documents to their closest centroids;
3. Recompute the centroid of each cluster;

4. Repeat step 2 and 3 until the centroid of every cluster does not change.

The hierarchical methods typically produce a nested sequence of partitions of the document set, with a single, all-inclusive cluster at the top and singleton clusters of individual document points at the bottom. These can be graphically displayed as a tree or a dendrogram. There are two basic approaches to hierarchical clustering: *agglomerative*, which starts with all the document points as individual clusters and, at each step, merges the most similar pair of clusters; and *divisive*, which starts with one, all-inclusive cluster and, at each step, splits a cluster until only singleton clusters of individual document points remain [164].

In the Topic Detection and Tracking (TDT)[1] *topic detection* task, a system performs unsupervised clustering of the incoming news stream, forming clusters without reference to an initial set of on-topic seed documents [54]. A topic tracking system is controlled by a threshold of confidence score for whether a news story should be clustered with the seed documents. Franz et al. defined a document-document similarity function based on a symmetrized version of the Okapi term weighting formula, then used the similarity function to compute the similarity score of a document with a cluster represented by its centroid [54].

While document clustering is sometimes referred to as automatic document classification, this is not strictly accurate since the classes formed are not known prior to clustering [140]. The strength of clustering lies in the fact that it does not require training data; a weakness is that it is hard to evaluate the clustering performance.

[1] http://www.nist.gov/speech/tests/tdt/tasks/ (last visited on October 22, 2008).

The next section introduces automatic text classification, which does require training data.

2.1.2 Automated Text Classification

An Overview

The goal of text classification is to classify the topic or theme of a document [113]. Automated text classification (TC) is a supervised learning task, defined as automatically assigning pre-defined category labels to documents [203]. It is a well studied task, with many effective techniques. There are two approaches to this task: knowledge engineering and machine learning. In the knowledge engineering approach, a classifier is built manually by domain experts. This can yield reasonably good results and is widely used in practice, but the cost of manual class profile construction and maintenance can be quite high. In the research community, the dominant approach is based on machine learning techniques, in which a general inductive process automatically builds a classifier by learning the characteristics of the categories from a set of pre-classified documents [156]. This can reach a similar accuracy as the knowledge engineering approach when suitable training data is available [6].

In text domains, effective feature selection is essential to make the learning task efficient and more accurate. The purpose of feature selection is to reduce the dimensionality of the term space since high dimensionality of the term space may result in the overfitting of a classifier to the training data. Forman [53] presented an empirical comparison of 12 feature selection methods (e.g., Chi-Squared, Information Gain, Odds Ratio, Probability Ratio, Document Frequency, Bi-Normal Separation, Power, Accuracy, F_1-measure) evaluated on a benchmark of 229 binary text classification problem instances. The overall feature selection procedure is to score each potential feature (usually words) according to a particular feature selection method, and then to take the best k features. Scoring involves counting the occurrences of a feature in training examples (positive examples and negative examples separately), and then computing a function of these. Forman used a variety of classifiers, including Naive Bayes, C4.5, logistic regression, and Support Vector Machine with a linear kernel (each using the WEKA open-source implementation with default parameters). The study shows that for multiple objective functions (accuracy, F-measure, precision, and recall), Bi-Normal Separation is the top single choice although for optimizing precision, Information Gain yields the best result most often. Their study did not include k-Nearest-Neighbor (kNN) for classification and latent semantic analysis (LSA) for feature selection.

Yang and Pedersen studied five feature selection methods for aggressive dimensionality reduction: term selection based on document frequency (DF), information gain (IG), mutual information (MI), a χ^2 test (CIII), and term strength (TS) [204]. Using the kNN and Linear Least Squares Fit mapping (LLSF) classification techniques, they found IG and CIII most effective in aggressive term removal without losing categorization accuracy. Strong correlations were found between the DF, IG and CIII scores of a term, indicating the importance of common terms in text classification. They also found that DF thresholding, the simplest method with the lowest cost in computation could reliably replace IG or CIII when the computations of those measure were expensive.

Popular techniques for text classification include probabilistic classifiers (e.g, Naive Bayes classifiers), decision tree classifiers, regression methods (e.g., Linear Least-Square Fit), on-line (filtering) methods (e.g., perceptron), the Rocchio method, neural networks, example-based classifiers (e.g., k-Nearest Neighbors), Support Vector Machines, Bayesian inference networks, genetic algorithms, and maximum entropy modelling [156].

Yang and Liu [203] conducted a controlled study on five well-known text classification methods: support vector machine (SVM), a k-Nearest Neighbor (kNN) classifier, a neural network (NNet) approach, the Linear Least-Square Fit (LLSF) mapping, and a Naive Bayes (NB) classifier. Their results show that SVM, kNN, and LLSF significantly outperform NNet and NB when the number of positive training examples per category are small (less than 10), and that all the methods perform comparably when the categories are sufficiently common (over 300 examples).

In automatic text classification, cross-validation is often used to assess classification effectiveness. Basically it is a resampling method to estimate the prediction error [56]. The idea of cross-validation is to partition a dataset into two mutually exclusive subsets: a training set and a test set. The training set is used to build the model, and the test set is used to test the model. In K-fold (or K-round) cross-validation, the dataset is partitioned into K mutually exclusive subsets. Of the K subsets, a single subset is retained as the test set for testing the model which was constructed on the K-1 training subsets. This process is repeated K times (folds or rounds), with each subset used exactly once as the test data. The K results then can be combined to produce a single estimate of the classification performance [90]. Because the training and test data are drawn from the same collection, cross-validation introduces some risk of reporting overly optimistic results, but it can nevertheless be useful early in the development process when only a single annotated collection is available.

Cross-language Text Classification

In monolingual text classification, both training and test data are in the same language. Cross-language text classification arises when training data are not in the same language, but rather in some other language. Cross-language text classification is a new area in text classification. There have been only a few studies on this issue since 1999.

In 1999, Topic Detection and Tracking (TDT) research was extended from English and Chinese [185]. In *topic tracking*, a system is given several (e.g., 1-4) initial seed documents and asked to monitor the incoming news stream for further documents on the same topic [54], so this is a text classification task. Researchers performed monolingual and cross-language text classification tasks. Systems were evaluated with a normalized cost measure which was combined from miss and false alarm. When trained on English data and tested on English data, the best system obtained a normalized cost of 0.077; when tested on Chinese data, its cost was 0.111. When trained on Chinese data and tested on Chinese data, it obtained a normalized cost of 0.080; when tested on English data, its cost was 0.115 [185]. So the effectiveness of cross-language classifiers was worse than monolingual classifiers.

Bel et al. [6] studied an English and Spanish bilingual classification task for the International Labor Organization (ILO) corpus, which had 12 categories. The ILO corpus consisted of 2,165 English documents and 1,590 Spanish documents. They evaluated two classification methods — the Rocchio method, which computes a class profile as the centroid of the training documents, and the Winnow method which, like SVM, computes an optimal linear separator in the term space between positive and negative training examples. They used monoligual classification as their baseline, which yielded a micro-averaged F_1 measure of 0.865 for English and 0.790 for Spanish, both achieved by the Winnow classifier which performed better than Rocchio. They studied two approaches — a *poly-lingual approach* in which both English and Spanish training and test data were available, and *cross-lingual approach* in which training examples were available in one language. Using the *poly-lingual approach*, in which a single classifier was built from a set of training documents in both languages, their Winnow classifier achieved an accuracy of 0.811 for English and Spanish (mixed), which was worse than their monolingual English classifier but better than their monolingual Spanish classifier. For the *cross-lingual approach*, they used two translation methods — *terminology translation* and *profile translation*. For *terminology translation*, they constructed a terminology for each class, and translated all the domain terms using an aligned corpus. When trained on English and tested on pseudo-English (Spanish translated into English), their Winnow classifier achieved an accuracy of 0.792; when trained on Spanish and tested on pseudo-Spanish (English translated into Spanish), it achieved a lower accuracy of 0.618 due to insufficient training data. Both were worse than their monolingual classifiers. For *profile translation*, they extracted a reduced vocabulary of 150 best terms in classifying all English documents with Winnow. They then trained a classifier on all English documents using only the reduced vocabulary and tested on all Spanish documents, translating only the Spanish terms having a translation in the reduced vocabulary. This method achieved an accuracy of 0.724, which was reported as “not bad” even though it was worse than their monolingual approach.

Rigutini et al. [143] studied English and Italian cross-language text classification in which training data were available in English and the documents to be classified were in Italian. They used a Naive Bayes classifier to classify English and Italian newsgroups messages of three categories: *Hardware*, *Auto* and *Sports*. English training data (1,000 messages for each category) were translated into Italian using *Office Translator Idiomax*, a plug-in for the *Office XP* suite. Their Italian monolingual baseline achieved both average recall and average precision of 0.94 across the three categories. Their cross-language classifier was created using Expectation-Maximation (EM), with English training data (translated into Italian) used to initialize the EM iteration on the unlabeled Italian documents. Once the Italian documents were labeled, these documents were used to train an Italian classifier. This approach achieved both average recall and average precision of 0.91 across the three categories. The cross-language classifier performed slightly worse than monolingual classifier, probably due to the quality of their translated Italian training data.

Gliozzo and Strapparava [59] investigated English and Italian cross-language text classification by using comparable corpora and bilingual dictionaries. The comparable corpus consisted of 32,354 Italian and 27,821 English news in four fixed categories, which covered same topics in the same period of time. The comparable corpus was used for Latent Semantic Analysis (LSA) which exploits the presence of common words among different languages in the term-by-document matrix to create a space in which documents in both languages were represented. To augment the number of common words between English and Italian in the matrix, two multilingual resources were exploited: MultiWordNet and the Collins English-Italian bilingual dictionary. The MultiWordNet was used to augment each document with the synset-ids of all the words, whereas the Collins dictionary was used to translate the words in the documents. Using the SVM classification method with a BoW kernel, the monolingual classifiers achieved an F_1 measure of 0.95 for English, and 0.92 for Italian. Using the comparable corpora only, the performance of cross-language classifier was 0.66 for training on English and testing on Italian, and 0.55 for training on Italian and testing on English. Using both the comparable corpora and the bilingual dictionary, the F_1 measure reached 0.88 for training on English (or Italian) and testing on Italian (or English), worse than their monolingual classifier (with F_1 =0.95 for English and 0.92 for Italian).

Olsson et al. [130] classified Czech documents using English training data. They translated Czech document vectors into English document vectors using a probabilistic dictionary which contained conditional word-translation probabilities for 46,150 word translation pairs. Their concept label kNN classifier ($k = 20$) achieved precision of 0.40, which is 73% of the precision of a corresponding monolingual classifier (with P=0.55).

We are interested in applying cross-language text classification techniques to classifying document segments into aspects, rather than classifying documents into topics. We are not aware of previous work at that level of topic aspect and text segment granularity. We are also interested in using training data in two languages at the same time, and in cases where relatively little training data is available.

2.2 Social Psychology Studies of Attitude and Related Concepts

Attitude,[2] opinion, belief, value, and habit are a set of related concepts [132]. According to Merriam-Webster Online Dictionary,[3] opinion, view, belief, conviction, persuasion, sentiment are synonyms that mean "a judgment one holds as true." We introduce some of the main concepts here.

2.2.1 Attitude

Definitions and Models

There are lay conceptualizations of attitude and more precise formulations in social psychology [112]. Standard dictionary definitions embrace the diversity of lay views. According to the Pocket Oxford Dictionary [171], an attitude is: "1: an opinion or way of thinking; behavior reflecting this (don't like his attitude). 2: bodily posture; pose. 3: position of an aircraft etc. relative to given points."

[2] Throughout this dissertation, we consistently refer to "attitude" to avoid unhelpful variations in our use of terminology. Since the remainder of the chapter introduces related concepts, we use each author's terminology where appropriate in this chapter.

[3] http://www.m-w.com/dictionary/opinion (last visited on June 20, 2006).

According to the Merriam-Webster Online Dictionary, the first of these (or focus in this dissertation), attitude has the following senses: “1 a: a mental position with regard to a fact or state; b: a feeling or emotion toward a fact or state. 2: an organismic state of readiness to respond in a characteristic way to a stimulus (as an object, concept, or situation). 3 a : a negative or hostile state of mind; b: a cocky or arrogant manner.”

The social psychology definition of attitude is that attitudes are tendencies to like or dislike specific attitude objects [136]. Attitudes are “general and enduring favorable or unfavorable feelings about, evaluative categorizations of, and action predispositions toward stimuli”[16](p. 401). The attitude objects (or stimuli) can be anything that has the potential to be evaluated favorably or unfavorably, including specific persons (e.g., Marilyn Monroe, Bill Clinton), social groups (e.g., lawyers, Iraqis), policy decisions (e.g., raising taxes, reducing NASA spending), personal action decisions (e.g., having an abortion, taking a Yoga course), abstract concepts (e.g., democracy, free trade), or consumer products (e.g., Honda sedans, Canon digital cameras) [158]. The attitudinal orientations toward these attitude objects are subjective rather than objective, because they reflect how a person sees an object and not necessarily how the object exists in reality [112].

There are two main theoretical viewpoints about the essential nature of attitudes—the older three-component model and the newer separate-entities viewpoint. The three-component model holds that an attitude is a single entity composed of three components: a cognitive component, an affective (emotional) component, and a behavioral component [132]. The *cognitive component* refers to beliefs about an attitude object. For instance, an individual may believe that the free trade policy is beneficial to the prosperity of the United States. The *affective component* refers to feelings or emotions associated with an attitude object. For instance, an individual may indicate that she does not like the free trade policy because she might lose her job some day when foreign countries become more competitive. The *behavioral component* refers to past behaviors associated with the attitude object [62]. For instance, a person might possess a positive attitude toward the free trade policy due to having signed a petition in favor of this issue. According to the tri-component model, attitudes are global evaluations of attitude objects that are derived from beliefs, feelings, and past behaviors regarding the attitude object. In general, people who have positive attitude toward an attitude object should often exhibit beliefs, feelings, and behaviors that are favorable toward the object, whereas people who have negative attitudes toward an attitude object should often exhibit beliefs, feelings, and behaviors that are unfavorable toward the object [35, 111].

The separate-entities viewpoint holds that the three components described above are distinct, separate entities which may or may not be related, depending on the particular situation, and that the term "attitude" should be reserved solely for the affective dimension, indicating an evaluation toward an attitude object [132].

The homeostasis model of attitudes holds that an individual's system of attitudes functions homologously. Individuals do not have the time, energy, or ability to access and review all of the contents of the psychological objects encountered every day. "An attitude toward the psychological object provides a rapid, cognitively inexpensive heuristic for deriving meaning from, imparting predictability to, and deriving behavioral guidelines for dealing with a complex world" [18](p. 335). This model emphasizes the perspective that attitudes are evaluative heuristics for guiding organismic-environmental transactions. "The central property of attitudes that reflect the behavioral organization of interest is the individual's stable and global positive/negative responses (evaluations) toward the psychological object" [18](p. 338).

From the above viewpoints, attitude indicates a stable and enduring evaluation toward an attitude object and has a polarity (e.g., positive, negative). Let's now take a look at the measurement of attitudes since we want to automatically compute attitudes.

The Measurement of Attitude

The bipolar rating scale (a bipolar continuum ranging from favorable through neutral to unfavorable) developed by Thurstone [173] was designed as a measure of an individual's favorable or unfavorable potential action toward some attitude object [18]. Likert [102] proposed a summated bipolar scaling procedure in which subjects are asked to indicate whether they strongly agree, agree, do not know, disagree, or strongly disagree in response to attitude stimuli. The assumption underlying the popular bipolar measurement of attitude is that "an attitude is reducible to the net difference between the positive and negative valent processes... This assumption can be expressed as three key principles: (a) an attitude is a joint function of positively and negatively valent activation functions (principle of evaluative activation);

(b) positively and negatively valent activation functions have generally opposing effects on an attitude (principle of opposing evaluative actions); and (c) positively and negatively valent activation functions are reciprocally controlled (principle of reciprocal evaluative activation)" [16](p. 401). Cacioppo and Bernston replaced the third principle with the principle of bivalent modes of evaluative activation, that is, "positively and negatively activation functions can be activated reciprocally (e.g., mutually exclusive and incompatible), uncoupled (e.g., singularly activated), or non-reciprocally (e.g., coactivational or coinhibitory)" [16](p. 402). The introduction of bivalent modes of evaluation activation requires a two-dimensional representation of positive and negative evaluative activation, one dimension for positivity (from low to high) and one for negativity (from low to high). The conceptualization of an attitude represented by the three principles and the bivariate framework can be expressed quantitatively as follows [16]:

$$Attitude = \frac{W_p}{W_p + W_n} P - \frac{W_n}{W_p + W_n} N + I(P, N) \qquad (2.1)$$

where P is the level of positivity activated by an attitude object;

N is the level of negativity activated by an attitude object;

I captures the nonadditive effects (e.g, the potentially mutually inhibitory interaction between positivity and negativity as their mutual activation increases);

W_p is a weighting factor representing the relative attitudinal effect of variations in positivity;

W_n is a weighting factor representing the relative attitudinal effect of variations in negativity.

Incorporating the unidimensional model and bidimensional model of attitude structure [112], attitudes can be (1) favorable, (2) unfavorable, (3) neutral (neither favorable nor unfavorable), (4) ambivalent (i.e., both favorable and unfavorable), and (5) no attitude (as in the situation that an individual has never heard of the attitude object). However, to reflect the principle of bivalent modes of evaluation activation, attitudes are better measured in two dimensions (i.e., positivity and negativity). When an aggregate attitude score is desired, the above attitude equation can be applied. Figure 2.1 roughly depicts a way to aggregate positivity, negativity and their strength into an attitude category for computational purpose.

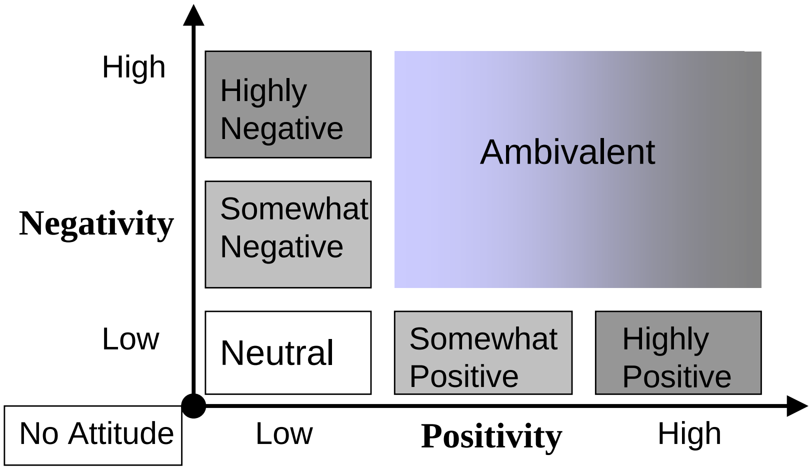


Figure 2.1: Attitude categories described in two dimensions.

2.2.2 Related Concepts

When studying attitude, other related psychological concepts, such as opinion, affect, sentiment, emotion, and mood, are often addressed. Here a brief introduction of these concepts is given, but is not meant to be a thorough comparisons.

Opinion

According to the Merriam-Webster Online Dictionary, opinion has the following senses: “1 a: a view, judgment, or appraisal formed in the mind about a particular matter; b: APPROVAL, ESTEEM. 2 a: belief stronger than impression and less strong than positive knowledge; b: a generally held view. 3 a: a formal expression of judgment or advice by an expert; b: the formal expression (as by a judge, court, or referee) of the legal reasons and principles upon which a legal decision is based.”[4]

Opinions are different from attitudes. “Opinions are equivalent to beliefs, rather than to attitudes” [132]. Beliefs are statements indicating a person’s subjective probability that an object has a particular characteristic, or asserting the truth or falsity of propositions about the object [52]. Opinions “are usually narrower in content or scope than the general evaluative orientation which we call an attitude, and they are primarily cognitive rather than emotion-laden. Another way of putting this is that opinions involve a person’s judgments about the likelihood of events or relationships, whereas attitudes involve a person’s feelings or emotions about objects or events. Thus, in this view, ‘I think this book is interesting’ is an opinion or belief,

[4] http://www.m-w.com/dictionary/opinion (last visited on June 20, 2006).

whereas ‘I like this book’ is an attitude” [132].

Krech and Crutchfield emphasized feelings of being “for or against” the psychological objects and having “positive or negative affect” toward the psychological objects in differentiating attitudes from opinions [18, 91]. So attitudes have polarity (positive, negative, neutral, ambivalent, no attitude) and strength (strong, weak) whereas opinions may or may not. For example, answers to the question “Which country do you think will be the world’s largest economy by 2020?” would be opinions probably without polarity, whereas the statement “this piece of software is difficult to use due to a high learning curve, but I like it since I am used to it” expresses a negative opinion but a positive attitude.

In spite of the difference between opinions and attitudes, the two concepts continue to be used synonymously, particularly in the area of survey research and polling. There *public opinion* is commonly accepted as the shared attitudes *and* opinions of large groups of people who have particular characteristics in common [132], whether this is a community, a society, a specific class or gender, or people from a specific religion or culture.

Sentiment

When talking about attitude and opinion, we very often hear the concept of sentiment. According to the Merriam-Webster Online Dictionary, sentiment has 3 senses: “1 a: an attitude, thought, or judgment prompted by feeling: PREDILECTION; b: a specific view or notion: OPINION; 2 a: EMOTION; b: refined feeling:

delicate sensibility especially as expressed in a work of art; c: emotional idealism; d: a romantic or nostalgic feeling verging on sentimentality; 3 a: an idea colored by emotion; b: the emotional significance of a passage or expression as distinguished from its verbal context."

"Human beings possess dispositions to respond affectively to particular objects or kinds of events. More precisely, we attribute affective dispositions to individuals to account for individual differences in this regard. Such dispositions are called sentiments or emotional attitudes. They are usually referred to as 'likes' or 'dislikes,' or else by emotion words followed by an object name or generic expression" [57](p. 64).

The structure of sentiments can be described in two ways. "First, sentiments consist of cognitive dispositions to appraise an object in a particular way. Sentiments can be understood as cognitive schemas, whose informational content gives rise to the appraisal when meeting the object... Second, sentiments are dispositions to treat the object in a way corresponding to that of the action readiness during the emotions; they constitute latent motivations that become acute on actual or possible confrontation with the relevant object... Sentiments indeed often can be described as desires" [57](p. 65).

So sentiment has both the meanings of opinion and attitude; it has a cognitive and an emotional component, and so is closer to attitude (i.e., its emotion-laden nature) than to opinion (i.e., its cognitive nature).

Affect, Emotion and Mood

Affect, emotion, mood, and sentiment are close psychological concepts. Frijda refers to them as “a varieties of affect” [57]. The Merriam-Webster Online Dictionary defines these concepts synonymously as follows.

Affect has 2 senses: “1: FEELING, AFFECTION; 2 : the conscious subjective aspect of an emotion considered apart from bodily changes; also: a set of observable manifestations of a subjectively experienced emotion.”

Emotion has 3 senses: “1: obsolete: DISTURBANCE; EXCITEMENT; 2a: the affective aspect of consciousness: FEELING; 2b: a state of feeling; 2c: a conscious mental reaction (as anger or fear) subjectively experienced as strong feeling usually directed toward a specific object and typically accompanied by physiological and behavioral changes in the body.”

Mood has 3 senses: “1: a conscious state of mind or predominant emotion: FEELING; 2: archaic: a fit of anger: RAGE; 3a: a prevailing attitude: DISPOSITION; 3b: a receptive state of mind predisposing to action; 3c: a distinctive atmosphere or context.”

Psychologists have tried to identify the properties of these concepts and to differentiate among them. Thurstone defined affect as appetition or aversion that varies in intensity [18, 174]. Frijda referred to affect as pleasant or unpleasant feeling [57]. Zajonc empirically verified the hypothesis of affective independence of cognition, that is, “affect can be aroused without the participation of cognitive processes... provided we mean by ‘cognition’ something more than pure sensory input” [209](p.

262).

Emotion has the following characteristics: automatic appraisal, commonalities in antecedent events, presence in other primates, quick onset, brief duration, unbidden occurrence, and distinctive physiology [44]. The valence of emotion is prominent. A rapid, seemingly automatic response to stimuli as positive or negative is common to many emotional experiences and is very likely universal across cultures [46, 208].

Frijda [57] differentiated emotions from moods. Emotions imply and involve relationships of the person with a particular object. In the states of emotions, affect (pleasant or unpleasant feeling), appraisal (the perception and evaluation of the emotional event with regard to its valence and its relevant properties for dealing with it), and action readiness (action tendencies or impulses to establish or disrupt relationships to the environment) are all object-focused, whereas in the states of moods, these elements lack such a focus [57]. Moods are nonintentional states [57]. "Emotions can be very brief, typically lasting a matter of seconds or at most minutes... Moods last much longer than emotions... Moods last for hours, sometimes for days. If the state endures for weeks or months, however, it is not a mood but more properly identified as an affective disorder" [43](p. 56).

The distinction between emotions and moods is related to object-relatedness. These concepts have intimate and obvious relationships. "Emotion may fade into moods; but moods may give rise to emotions, since they often imply a lowering of threshold for particular emotions. Sentiments may be the precipitate of emotions, and may underlie the emergence of emotions" [57](p. 67).

Computationally Modeling Attitude

The introduction of these psychological concepts is meant to demonstrate that attitudes, opinions, sentiments, affects, emotions, and moods are closely related but not identical. However, we do not plan to automatically compute their differences since it is very difficult to do so. What we can automatically compute is the polarity (or valence) of affective evaluation of objects. Therefore what we plan to automatically measure is the polarity of the combination of these psychological concepts, including attitudes, opinions with polarity, sentiments, affects, and emotions (with moods excluded since they are not object-related). In the field of computational linguistics, a general covering term for subjective evaluation is *private state* [139]. However, *attitude* is used in the title of the dissertation because attitude indicates a stable and enduring evaluation toward an attitude object and has a polarity, and so that term is a useful description we want to characterize.

In the next section we address the regularities of attitude expressions which make it possible for us to detect attitude from linguistic statements.

2.2.3 Regularities of Attitude Expressions

There are regularities in attitude expressions. Eiser observes that the kinds of statements people make apparently to express their personal feelings and beliefs about particular issues often do *not* show the kind of boundless creativity which Chomsky and others argue that the rules of language permit [41]. “On the contrary, many attitude statements appear to be drawn from a rather limited and familiar repertoire, and, while leaving some room for stylistic embellishment, may be somewhat stereotypical in form and content"(p. 24). Eiser observes that many of the attitude statements people make are of a rather special kind of conditioned and/or imitative responses. Eiser assumes that such linguistic behavior is acquired socially from other people, much of it is likely to be shown in similar forms by large numbers of people. "To the extent that attitude statements are socially conditioned/imitative responses, therefore, they are likely to give the impression of widely shared social attitudes" [41](p. 25). Therefore if we could identify the regularities of attitude-bearing words and expressions, we should often be able to identify the attitude language. This is the reason why sentiment detection systems that simply use attitude words (such as dynamic adjectives) are a good baseline. Therefore classifying the sentiment of words and expressions is a good investment.

Eiser also argues that people use a shared linguistic style for expressing particular attitudes and select shared aspects of issues to evaluate [41]. "Although language provides us with a wide choice of ways to express our attitudes, in practice there are social constraints on such choice. We learn what dimensions of description are likely to be acceptable in different social contexts when expressing our viewpoints on a given issue. We acquire, that is, a *shared* linguistic style for expressing particular attitudes, reflecting at least partly a *shared* selectivity in the aspects of the issue to be regarded as salient. Since language is our prime means of communication, it is also the prime route through which attitudes may come to be shared. Language, too, provides us with concepts and sets of categories in terms of which events can be evaluated and represented" [41](p. 73). This leads us to believe that people evaluate and share their attitudes on the aspects of a topic.

2.3 English Sentiment Classification

2.3.1 Subjectivity Analysis

There has been a recent flurry on research in automatic opinion and sentiment recognition, including lexical acquisition, semantic orientation detection for words and phrases, distinguishing subjective from objective language, and detecting subjectivity at document and sentence level. The remainder of this chapter introduces the previous work on English and Chinese sentiment detection.

Subjective language refers to the aspects of language used to express *private states* (i.e., opinions, beliefs, attitudes, evaluations, emotions, rants, allegations, accusations, suspicions, and speculations) in the context of a text or conversation [145, 193]. The purpose of subjectivity analysis is to distinguish subjective language from language used to objectively present factual information. Good clues are needed to perform automatic subjectivity analysis.

According to Weibe et al. [193], although some manually developed resources exist, such as General Inquirer[5] and the *attitude* adverb features in COMLEX [109], there is no comprehensive dictionary of subjective language, and moreover, many words with subjective usages can also be used objectively. Thus, a system "may not merely consult a list of lexical items to accurately identify subjective language, but must disambiguate words, phrases, and sentences in context" [193]. To reflect the

[5] http://www.wjh.harvard.edu/ inquirer/homecat.htm (last visited on June 25, 2006).

ambiguity of subjective language, a *potential subjective element* (PSE) is defined as "a linguistic element that may be used to express subjectivity," and a *subjective element* is defined as "an instance of a potential subjective element, in a particular context, that is indeed subjective in that context" [193].

To judge subjective language in context, Wiebe et al. [193] found that valuable clues of subjectivity include hapax legomena (i.e., unique words), automatically identified subjective collocations (including fixed n-grams and n-grams with placeholders for unique words), and distributionally similar words that share pragmatic usages for expressing subjectivity. To disambiguate whether or not a PSE instance is indeed subjective in context, they found that the density of other potentially subjective expressions in the surrounding context is valuable, (i.e., a sufficient number of other PSE instances nearby imply that a particular PSE instance is more likely to be subjective). To classify documents into subjective versus objective categories, they used the count of the all PSE instances in the document as the single feature (i.e., each document was characterized by the count of all PSE instances), and their k-Nearest-Neighbor classifier (with the distance between two documents defined as the absolute value of the difference between the normalized PSE counts for the two documents) reported positive results.

Wiebe and her colleagues [15, 66, 145, 191, 193] did a series of studies on recognizing subjectivity. They created subjective versus objective classifiers at the sentence level by using a seeding process that utilized known subjective vocabulary to automatically create training data [191]. They found that dynamic adjectives, semantically oriented adjectives, and gradable adjectives are strong predictors of subjectivity. They recognized subjectivity by manual tagging; used a bootstrapping process to learn linguistically rich extraction patterns for subjective expressions; developed subjectivity classifiers using only unannotated texts for training, and created subjective and objective sentence classifiers.

Riloff and Wiebe [145] used high-precision rule-based subjectivity classifiers to automatically identify subjective and objective sentences in unannotated texts. The labeled sentences were then used as a training set to automatically learn linguistically rich extraction patterns for identifying subjective expressions.

Wiebe and Riloff [191] created rule-based sentence-level subjective and objective classifiers (with high precision and low recall) using well-established general subjectivity clues that had been published in the literature as baselines, then applied the rule-based classifiers to unannotated text to automatically generate training data. The training data were used to learn extraction patterns (a set of syntactic templates) associated with subjectivity and objectivity. The extraction patterns were found to be good clues for distinguishing subjective sentences from objective ones, but were not sufficient by themselves. Adding the extraction patterns to the rule-based classifiers increased recall, with a relatively small drops in precision. They then created a naive Bayes classifier using the labeled sentences identified by the rule-based classifiers. A *self-training* approach was used to improve the Bayes classifier by generating a new training set using the classifier itself. Using the new training data increased the recall of the learned subjective and objective patterns substantially with a minor drop in precision.

Bruce and Wiebe [15] found that adjectives were statistically significantly positively correlated with subjective sentences. Quirk et al. [139] posited three semantic scales for adjectives: stative/dynamic, gradable/nongradable, and inherent/nonherent. According to Quirk et al. [139], "adjectives are characteristically stative. Many adjectives, however, can be seen as dynamic. In particular, most adjectives that are susceptible to subjective measurement are capable of being dynamic. Stative and dynamic adjectives differ syntactically in a number of ways. For example, a stative adjective such as *tall* cannot be used with the progressive aspect or with the imperative: *He's being tall. Be tall.* On the other hand, we can use *careful* as a dynamic adjective: *He's being careful. Be careful.* A general semantic feature of dynamic adjectives seems to be that they denote qualities that are thought to be subject to control by the possessor and hence can be restricted temporally"(p. 434). Subjectivity was found by Bruce and Wiebe to be part of the semantics of dynamic adjectives [15].

Wiebe [189] studied sentence subjectivity clues using Lin's [103] method for clustering words according to distributional similarity to identify adjective features, seeded by a small number of manually annotated sentences. They wanted to measure a simple adjective feature, which is the conditional probability that a sentence is subjective given that at least one adjective appears. The feature was learned from a thesaurus automatically created from a news corpus, and the WordNet[6] synonyms. Their approaches achieved a precision from 0.61 to 0.66. The adjective features were further refined with the addition of lexical semantic features of adjectives, specifically polarity and gradability.

[6] http://wordnet.princeton.edu (last visited on June 28, 2006).

Since the presence of one or more adjectives is useful for predicting that a sentence is subjective, Hatzivassiloglou and Wiebe [66] studied the effects of dynamic adjectives, semantically oriented adjectives, and gradable adjectives on a simple subjectivity classifier, and established that they are strong predictors of subjectivity. Moreover, each of these sets of adjectives were found to be better predictors of subjective sentences than the class of adjectives as a whole.

Research in genre classification has included recognition of subjective genres such as *editorials* and objective genres such as *business* or *news* [84, 87, 192, 207]. Yu and Hatzivassiloglou [207] applied a Naive Bayes classifier to separate editorials from regular news stories, and then applied unsupervised, statistical techniques for detecting opinions at sentence level.

2.3.2 Classifying the Semantic Orientation of English Words and Phrases

Much of the previous work on attitude classification focuses on lexical acquisition, identifying positive, or negative words and phrases. Hatzivassiloglou and McKeown [65] presented a supervised learning method for predicting the semantic orientation or polarity of adjectives, which is the direction the word deviates from the norm for its semantic group or lexical field. They proposed and verified a hypothesis that conjunctions (such as *and*, *but*, *or*) constrain the orientation of conjoined adjectives. The conjunctions between adjectives were extracted using a shallow parser in a 21 million word corpus of Wall Street Journal articles. They used the constraints to develop and train a log-linear statistical model that predicts whether two conjoined adjectives are of same or different orientation with 82% accuracy. Combining the constraints across many adjectives, a clustering algorithm separated the adjectives into groups of different orientations with 92% accuracy.

Hatzivassiloglou and McKeown's method [65] for computing the semantic orientation of words is restricted to adjectives; it requires labeled adjectives as training data, and the process is difficult to implement and to analyze theoretically [178]. Turney and Littman [178, 179] presented an unsupervised learning method for inferring semantic orientation of words (including adjectives, adverbs, nouns, and verbs) from semantic association. Seven positive words (good, nice, excellent, positive, fortunate, correct, and superior) and seven negative words (bad, nasty, poor, negative, unfortunate, correct, and superior) were selected intuitively as paradigms for positive and negative semantic orientation due to their lack of sensitivity to context. The semantic orientation of any given word could then be calculated from the strength of its association with the seven positive words, minus the strength of its association with the seven negative words. The magnitude of the difference can be considered as the strength of the semantic orientation. The strength of the semantic association between words was computed using two methods: pointwise mutual information (PMI) and latent semantic analysis. PMI was computed using word co-occurrence data collected from the Web search engine AltaVista (i.e., the hits of a query "*word1* NEAR *word2*" generated by AltaVista) as follows:

$$PMI(word_1, word_2) = log_2(\frac{\frac{1}{N}hits(word_1 NEAR word_2)}{\frac{1}{N}hits(word_1)\frac{1}{N}hits(word_2)}) \tag{2.2}$$

where N is the total number of documents indexed by the search engine.

The online demonstration of LSA[7] was used to compute the strength of association of two words. Their methods were evaluated against a list of 3,596 words (1,614 positive and 1,982 negative) collected from the General Inquirer lexicon [166] after removing the words with multiple senses. The PMI method, given an unlabeled Web training corpus of approximately one hundred billion words, attained an accuracy of 82.8%, which is comparable with Hatzivassiloglou and McKeown's more complex method [65]. PMI requires a large corpus, but it is simple, easy to implement, and is not restricted to adjectives. The LSA method yielded lower accuracy (65%) due to the small ten-million word corpus (i.e., TASA-ALL) that was used to train LSA.

Turney [177] extended the PMI method to compute the semantic orientation of phrases by calculating the mutual information between a given phrase and the word "excellent" minus the mutual information between the given phrase and the word "poor." The semantic orientations of phrases extracted from the reviews collected from Epinions were used to classify reviews of automobiles, banks, movies, and travel destinations to the categories "recommended" and "not recommended."

Following Turney's method [177], Yu and Hatzivassiloglou [207] experimented with seed sets containing 1, 20, 100, and over 600 positive and negative pairs of adjectives. For a given seed set size, the set of positive seeds is denoted as ADJ_p and the set of negative seeds as ADJ_n. They calculated a modified log-likelihood ratio $L(W_i, POS_j)$ for a word W_i with part of speech POS_j (j can be adjective, adverb, noun or verb) as the ratio of its collocation frequency with ADJ_p and ADJ_n within

[7] http://lsa.colorado.edu/ (last visited on January 5, 2008).

	Recall								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
P (seed=1)	0.46	0.47	0.48	0.49					
P (seed=20)	0.76	0.75	0.73	0.71	0.68	0.66	0.65		
P (seed=100)		1.0	0.8	0.78	0.76	0.76	0.75	0.74	0.72

Table 2.1: Precision (P) at each recall level with various number of seeds (roughly adapted from [207]).

a sentence.

$$L(W_i, POS_j) = log(\frac{\frac{Freq(W_i,POS_j,ADJ_p)+\varepsilon}{Freq(W_{all},POS_j,ADJ_p)}}{\frac{Freq(W_i,POS_j,ADJ_n)+\varepsilon}{Freq(W_{all},POS_j,ADJ_n)}}) \tag{2.3}$$

where $Freq(W_{all}, POS_j, ADJ_p)$ represents the frequency of all words W_{all} with part of speech POS_j collocated with ADJ_p; ε is a smoothing constant (set to 0.5).

The Brill's tagger [12] was used to obtain POS. Unlike Turney's, their method does not consult any Web search engine. Evaluating on Hatzivassiloglou and MaKeown's [65] 1,336 manually labeled positive and negative adjectives, they found that both recall and precision increased as the seed set became larger. Table 2.1 roughly displays the precision of each recall level with different seed size.

Kamps and Marx [83] used WordNet [120, 49] to determine the affective or emotive aspects (i.e., semantic orientation) of a word. A word's semantic orientation was calculated based on its semantic distance from "good" compared to its semantic distance from "bad." Semantic distance was calculated based on the path-length between two words in WordNet. The approach has not been evaluated empirically, however.

Gradability or grading is "the semantic property that enables a word to participate in comparative constructs and to accept modifying expressions that act as intensifiers (such as *large*) or diminishers (such as *small*)" [108](p. 271). To distinguish between gradable adjectives (such as *careful*, *accurate*) and non-gradable adjectives (such as *domestic*, *military*), Hatzivassiloglou and Wiebe [66] have developed a trainable log-linear statistical model that takes into account the number of times an adjective has been observed in a form or context indicating gradability relative to the number of times it has been seen in non-gradable context. Their method can extract gradability values quite reliably.

Rather than just learning the prior (out-of-context) semantic orientation for words, some studies have sought to identify the contextual polarity of phrases in which instances of those words occur. For example, Nasukawa, Yi, and colleagues [123, 206] classified the contextual polarity of expressions that refer to a given subject (such as a product). They manually developed high-quality patterns to classify polarity, yielding quite high precision, but very low recall. Wilson et al. [197] used a two-step process that employed machine learning and a variety of features to identify the contextual polarity of subjective expressions. In the first step, they developed classifiers using the BoosTexter AdaBoost.HM [152] machine learning algorithm, then classified each phrase containing a subjectivity clue as neutral or polar. The neutral-polar classifier used 28 features covering word features, modification features, structure features, sentence features, and document features.

In the second step, they developed a polarity classifier using 10 features, including word features (such as word token, word prior polarity) and polarity features (such as polarity shifters) to classify all the phrases marked as polar and to disambiguate their contextual polarity. The 10-feature classifier achieved an accuracy of 65.7%.

2.3.3 Classifying the Attitude of English Sentences

Some previous research focuses on classifying the attitude of sentences-scale. A common approach among these is to create an out-of-context prior-polarity lexicon at first, then to assign a polarity score to a sentence in some way [61, 77, 88, 207].

Yu and Hatzivassiloglou [207] took a two-step approach to classifying the sentiment of sentences as positive, negative, or neutral. First, they classified sentences into opinion and fact. Three approaches were used to find opinion sentences.

The first approach used SIMFINDER [64], a system for measuring sentence similarity based on shared words, phrases, and WordNet synsets, to compute the similarity of a sentence to each sentence in pre-classified opinion or fact documents on the same topic, then averaged those scores and classified the sentence to the category (fact or opinion) for which the average was higher.

The second approach trained a Naive Bayes classifier using the pre-classified opinion and fact documents. The features included words, bigrams, trigrams, and the parts of speech in each sentence, the counts of positive and negative words in the sentence, the counts of parts of speech combined with polarity information (e.g., positive adjectives), the polarity (if any) of the head verb, the main subject, their

	Approaches		
Class	1	2	3
Fact	{1.00, 0.27}	{0.44, 0.50}	{0.44, 0.53}
Opinion	{0.16, 0.64}	{0.88, 0.56}	{0.91, 0.86}

Table 2.2: {Recall, precision} of fact/opinion sentence classification (adapted from [207]).

immediate modifiers, and the average semantic orientation score of the words in the sentence.

The third approach used multiple classifiers, each relying on a different subset of features. The goal was to reduce the training set to the sentences that were most likely to be correctly labeled so as to boost classification accuracy.

Evaluated on manually annotated 400 sentences (as shown in Table 2.2), the first approach achieved high recall and low precision for classifying facts, and low recall and medium precision for classifying opinions. The second approach achieved a higher recall and precision for classifying opinions than for facts. The third approach slightly outperformed the second approach for both facts and opinions.

Once opinion sentences were detected, they used the average per-word log-likelihood scores defined in formula 2.3 to assign an aggregate polarity to the sentences based on a heuristic cutoff threshold learned from the training data. They experimented with different combinations of part-of-speech classes for calculating the aggregate polarity scores; combined evidence from adjectives, adverbs, and verbs achieved the highest accuracy (90%) for sentence polarity.

Grefenstette et al. [61], Hu and Liu [77], and Kim and Hovy [88] considered local negation to reverse polarity, then multiplied or counted the prior polarities of clue instances in the sentence. Kim and Hovy [88] developed a sentiment classifier for words and sentences using thesauri, but their template-based approach needs an annotated corpus for learning, and words in thesauri are not always consistent in semantic orientation.

2.3.4 Classifying the attitude of English Documents

Automatic sentiment analysis has been applied to many applications, including classification of product and movie reviews [26, 123, 133, 178], analysis of product reputations, tracking sentiments toward events, incorporating opinions into question answering and multi-document summarization systems [207], identifying children's emotional expressions in online environments [2], rating the affective content of texts [34], recognition of hostile messages [163], and distinguishing positive from negative reviews [197].

The essential issues in sentiment analysis are to identify how sentiments are expressed in texts and whether the expressions indicate positive (favorable, recommended) or negative (unfavorable, not recommended) opinions toward the subject [123]. Both linguistic and machine learning approaches have been applied. The following is a brief review of these approaches, a summary of which is given in Table 2.3 to facilitate comparisons.

Authors	**Documents**	**Approaches**	**Categories**
Turney [177]	Epinions reviews	Linguistic analysis: extract two-word phrases and calculate their semantic orientations (SOs); aggregate their SOs	positive, negative
Pang et al. [133]	movie reviews	Machine learning: (Naive Bayes, maximum entropy, SVM), unigram model	positive, negative
Mullen and Collier [122]	movie reviews	Machine learning: hybrid SVMs	positive, negative
Kennedy and Inkpen [86]	movie reviews	Linguistic analysis: valence shifters; machine learning: SVM	positive, negative
Spertus [163]	Web feedback and email messages	Machine Learning: C4.5 decision tree; features based on syntax and semantics of sentences	flame, maybe, okay
Das and Chen [24]	stock message boards	Machine learning: couple multiple classifiers by a voting scheme	positive, negative
Durbin et al. [34]	customer emails	Linguistic analysis: sentiment lexicon, valence modifiers	positive, negative
Pang and Lee [134]	movie reviews	Machine learning: multi-class categorization	multi-point scale
Wilson et al. [196]	MPQA corpus	Machine learning: support vector regression,	opinion strength (strong, weak)
Wilson et al. [197]	MPQA corpus	Linguistic feature analysis for contextual polarity	positive, negative, both, neutral
Nasukawa and Yi [123]	Web pages and news articles	Linguistic analysis: extract sentiments for subjects, sentiment lexicon, shallow parser	positive, negative

Table 2.3: A summary of approaches to classifying the attitude of documents.

There have been many studies of classifying reviews. Turney [177] applied a three-step approach to classify Epinions reviews. First a part-of-speech tagger (the Brill tagger) was applied to the reviews and two-word phrases (such as "horrific events") were extracted. The second step was to use Pointwise Mutual Information (formula 2.2) to calculate semantic orientations for the extracted phrases. Finally the average semantic orientation of the extracted phrases was computed as the orientation of the review. That is, if the average was positive, then the review was classified as positive; otherwise, negative.

Pang et al. [133] applied classical text classification techniques to the task of classifying movie reviews as positive or negative. They evaluated three different supervised learning algorithms (i.e., Naive Bayes, maximum entropy, and SVM) and eight different sets of features. The best result was achieved using an SVM with features based on the presence or absence (rather than the frequency) of single words (rather than bigrams) with an accuracy of 83%. Mullen and Collier [122] used SVMs to bring together Pang et al.'s [133] unigram model and lemmatized versions of unigram models, several favorability measures for phrases and adjectives (such as semantic orientation with Turney's [177] PMI and, Kamps and Marx's [83] differentiation with WordNet) and, where available, knowledge of the topic of the text (such as topic proximity). Their hybrid SVMs reached an accuracy of 84.4%, slightly superior to Pang et al.'s.

Kennedy and Inkpen [86] used three types of valence shifters (negations, intensifiers and diminishers) and machine learning algorithms to detect the semantic orientation of movie reviews. More attitude classification for reviews of products, movies, paper peer reviewers, and stock investment are discussed in [21, 24, 77, 105, 121, 142, 149].

Some projects have focused on document-level sentiment classification of text other than reviews. Spertus [163] identified inflammatory texts with a system named Smokey. Smokey builds a 47-element feature vector based on the syntax and semantics of each sentence, combining the vectors for the sentences within each message (in this case, comments that were sent via feedback forms on Web pages). A training set of 720 messages was used by Quinlan's C4.5 decision-tree generator to determine feature based rules that were able to correctly classify 64% of the "flames" and 98% of the non-flames in a separate test set of 460 messages. Das and Chen [24] applied an algorithm that coupled different classifier algorithms together, using a voting scheme to extract small-investor sentiment from stock market message boards.

Durbin et al. [34] constructed a system for affect rating of texts with a particular domain (customer emails) that were written in with multiple languages. First, a few thousand sentiment words were manually collected and rated. Second, a document part-of-speech tagger was run, and individual rated words (i.e., those on a list) were identified. Third, valence modifiers such as "very" or "slightly" were detected, and the ratings of words immediately following were modified appropriately. Finally, an overall rating was assigned to the document. They suspected that their system would perform well on typical customer emails, but a corpus of rated messages were not available. Instead, as an experiment, they evaluated their system on the collection of movie reviews assembled by Pang et al. [133]; the system achieved an accuracy of 63%, well below the best result of 83% obtained by Pang et al [133] who trained various machine learning algorithms on the same dataset. This implies that attitude classifiers are genre-specific.

Devitt and Ahmad explored positive and negative polarity in financial news text, which could be used in a quantitative analysis of news sentiment impact on financial market [32]. They took a cohesion-based text representation approach which built a graph representation of part-of-speech tagged text (without disambiguation) using WordNet [49]. The graph structure was composed of nodes representing concepts in or derived from the text, connected either by relations between those concepts in WordNet (e.g., hyponymy), or derived from the text (e.g. adjacency). The positive and negative polarity of the concepts in the graph were taken from the Sentiwordnet lexicon [48]. They found that the relationship type could productively be used to adjust the polarity of a node. Some types of relationships (e.g., antonymy and hyponymy) deserved higher weights than others. They also found that node specificity calculated on the basis of the graphic structure was useful because "highly specific nodes or concepts may carry more information and, by extension, affective content than less specific ones" [32](p. 987).

Beyond classifying documents into positive and negative categories, Pang and Lee [134] classified movie reviews with respect to a multi-point scale (e.g., one to five "stars"). Wilson and Wiebe have also classified the strength of opinions [196] and contextual polarity of the polar expressions [197].

Instead of classifying a whole document into positive and negative sentiments, Nasukawa and Yi [123] extracted sentiments for specific subjects from a document. Specifically they applied sentiment analysis to text fragments that consisted of a sentence containing a subject term, plus the remainder of that paragraph. The window included at least 5 (and no more than 50) words before and after the subject term. They detected all references to the given subject, and determined sentiment for each of the references. They also applied a shallow syntactic parser and a sentiment lexicon to identify semantic relationships between the sentiment expressions and the subject. Their prototype system achieved high precision (75-95%) but low recall (roughly 20%) for Web pages and news articles. Most of their failures were due to the long and complex sentences.

We have seen attitude classification in many domains or genres—customer emails, product reviews, movie review, news articles, Web pages. Due to text genre difference, attitude classifiers are genre-specific, and so need to be trained from the text genre they work on.

2.4 Attitude Classification for Chinese and Other Languages

Compared with the relatively large amount of previous research on sentiment detection in the English language, the work in other languages is relatively rare. Ku et al. at National Taiwan University (NTU) [93] performed opinion extraction at word, sentence, and document level in news and weblog articles in both English and Traditional Chinese. To learn the semantic orientation of Chinese words, they proposed a character-based word classification algorithm in which a Chinese word's semantic orientation is a function of the Chinese characters it contains. First they collected two sets of sentiment words as seed vocabulary, from General Inquirer (translated into Chinese) and the Chinese Network Sentiment Dictionary [93]. They then enlarged the seed vocabulary by consulting two thesauri: a Chinese synonym dictionary and the Academia Sinica Bilingual Ontological WordNet. The enlarged seed lexicon was split into positive words and negative words. To compute the opinion tendency and strength of an unknown Chinese word, they counted the frequencies of each character of the word in the positive and negative seed vocabularies. They defined *sentiment tendency* of a character as the ratio of a character's frequency count in positive words to its frequency count in the whole seed lexicon, normalized by the total number of characters in both positive and negative words (see below).

$$POS_{c_i} = \frac{P_{c_i} / \sum_{j=1}^{n} P_{c_j}}{P_{c_i} / \sum_{j=1}^{n} P_{c_j} + N_{c_i} / \sum_{j=1}^{m} N_{c_j}} \tag{2.4}$$

$$NEG_{c_i} = \frac{N_{c_i} / \sum_{j=1}^{m} N_{c_j}}{P_{c_i} / \sum_{j=1}^{n} P_{c_j} + N_{c_i} / \sum_{j=1}^{m} N_{c_j}} \tag{2.5}$$

where

- POS_{c_i} and NEG_{c_i} are positive and negative tendency of character c_i respectively,
- P_{c_i} and N_{c_j} are frequency counts of character c_i in positive and negative words respectively,
- n and m denote the total number of unique characters in positive and negative words respectively, and so $\sum_{j=1}^{n} P_{c_j}$ and $\sum_{j=1}^{m} N_{c_j}$ are total number of unique characters in positive and negative words respectively.

The difference between POS_{c_i} and NEG_{c_i} determines the sentiment tendency of character c_i. The semantic orientation of a Chinese word is the average of the sentiment tendency scores of the composing characters. Their sentiment word classification algorithm achieved an average precision of 61.06% at best cases.

For sentiment detection at sentence level, they summed up the semantic orientation scores of all words in the sentence. If a negation operator appears before a sentiment word, the sentiment tendency of the word was reversed. For sentiment detection at document level, they simply summed up the sentiment scores of all the sentences. Evaluated with a small test collection they created using TREC 2003, NTCIR-2 and blog documents, their system achieved a low precision at both the sentence and document levels (11.41% - 40%) due to the fact that sentences not relevant to topics were also evaluated.

To facilitate research on Chinese sentiment classification, the Chinese Opinion Analysis Pilot Task was introduced in the Sixth NTCIR Workshop. There were four subtasks — opinionated sentence detection, opinion holder extraction, relevant sentence detection, and polarity detection, all at sentence level [157]. Here we focus on opinionated sentence detection and polarity detection. Five teams participated in the task. We built the best system for opinionated sentence detection [157, 199], and the Chinese University of Hong Kong (CUHK) created the best system for sentence polarity detection [200]. So we briefly introduce CUHK's approach here.

CUHK proposed that a complete opinion consisted of five components: an opinion holder (which is the governor of an opinion), an opinion object (which is the target of the opinion), an opinion word (that expresses the opinion polarity, e.g., *favorable*), an opinion operator (which is the verb indicating an opinion event, e.g., *emphasize*), and an opinion indicator (which is the word indicating the orientation of an opinion or the orientation trend of multiple opinions, e.g., *but*) [94, 200]. Opinion operators and indicators were semi-automatically learned from the training data of NTCIR-6 and from the documents similar to the training data which were collected from the Web. Opinion word lexicons were initially built from NTU's lexicons [93] and from the Chinese Positive Dictionary and the Chinese Negative Dictionary [159, 205], and then manually classified into context-free opinion words (CFOW) and context-dependent opinion words (CDOW). The CDOW lexicon was further expanded by applying both supervised and unsupervised learning methods to the documents which were crawled from the Web. Their opinion sentence classifier was built using the support vector regression method, which used five features—opinion indicator, operator, opinion words, topic, and named entities (e.g., countries, persons, and organizations). Their classifier achieved an F_1 score of 0.64 for opinionated sentence detection and 0.41 for sentence polarity classification.

The previous work on Chinese sentiment analysis was influenced by English sentiment classification in many ways. The tasks were the same, both working on subjectivity and polarity classification at word, sentence, and document level, and the approaches for building classifiers were almost the same (i.e., lexicon-based linguistic analysis approaches and machine learning approaches).

Applying English attitude classification resources and approaches to other languages has also started. Mihalcea et al. [119] investigated two methods to automatically generate resources for Romanian subjectivity analysis by leveraging English lexicons and manually labeled English corpora. They first translated an English sentiment lexicon to Romanian using a bilingual dictionary and built a rule-based sentence-level subjectivity classifier. They second used an English subjectivity classifier to automatically annotate the English side of an existing English-Romanian parallel corpus, then projected the subjectivity annotations onto the Romanian side of the corpus across the sentence-level alignments available in the corpus, and then used the Romanian annotations to train a subjectivity classifier with machine learning techniques. Their overall results obtained with the corpus projection approach (F-measure of 0.68) were considerably higher than with the lexicon translation approach (F-measure of 0.48). They found that the lexicon translation process led to increased ambiguity and loss of subjectivity [119] because subjectivity is a property associated with word meanings rather than with words [190].

2.5 Summary

In this chapter, we have reviewed the major approaches for document clustering and automatic text classification, social psychology studies of attitude and related concepts, and attitude classification for English and other languages. This review helps us to highlight several gaps between what has been done and what needs to be done.

First, while topic-based classification has been thoroughly studied, and previous studies have established a good set of classification methods, aspect classification with a finer granularity remains largely unexplored.

Second, although cross-language classification has been the focus of some research, using same techniques and cross-language training data together is a new research topic.

Third, much work has been done on English attitude classification whereas far less work has been done on Chinese attitude classification, and much room remains for improvement. Although English attitude classification still remains an interesting research task, advancing the state of art in Chinese attitude classification perhaps offers even greater potential.

Finally, the quantitative conceptualization of attitude in social psychology (Equation 2.1) substantiates the automatic measurement of attitude in computational linguistics. The two dimensional model of attitude (illustrated in Figure 2.1) is an appropriate way of characterizing attitude, which is worth exploring in our study.

In the next chapter, we turn our attention to bilingual aspect classification.

Chapter 3

Bilingual Aspect Classification

In this chapter, we address one of the two major tasks in our ultimate goal—bilingual aspect classification. We address the methods for exploring the research question, test collection design, experiment design, and experiment results.

3.1 Methods

The goal of aspect classification in two languages is to classify English and Chinese document segments (or passages) that are relevant to a topic based on relevance to the aspects of that topic. We apply a classical topic-based classification approach using machine learning methods which require a training data set. Typical automatic classification methods yield greater accuracy when a large set of training data is provided. From our application we assume that the user provides extremely limited amount of training data, so the classification methods here employ what might be called weakly supervised learning.

Here we define *monolingual aspect classification* as an aspect classification task which uses training examples in one language only, and the training examples and test examples are in the same language; we define *bilingual aspect classification* as an aspect classification task which uses training data in two languages.

Bilingual aspect classification involves training examples from two languages—

main language and supporting language. We define *main language* as the language of the test examples, and *supporting language* as the other language that is different from the language of the test examples. For instance, if the task is to classify English document segments into English aspects, English is the main language, and Chinese is the supporting language, so English training examples are main-language training examples, and Chinese training examples are supporting-language training examples.

In order to answer our research question, we need an approach for monolingual aspect classification and an approach for bilingual aspect classification so that we can compare them to see whether bilingual aspect classification can do any better. For any classification experiment, we need a classification method appropriate for our task, a test collection which includes training and test data, and evaluation metrics. In this section, we discuss general approaches to monolingual and bilingual aspect classification, classification methods, and evaluation metrics. Since test collection design is a substantial part of this study, it is addressed in Section 3.2.

3.1.1 Monolingual Aspect Classification

Before we create a classification system that can use supporting-language training examples, we first need to create a classification system that uses at least main-language training examples. The monolingual aspect classification system serves as a baseline. It involves the following steps, as illustrated in Figure 3.1.

(1) A user who can read and write both English and Chinese retrieves a set of

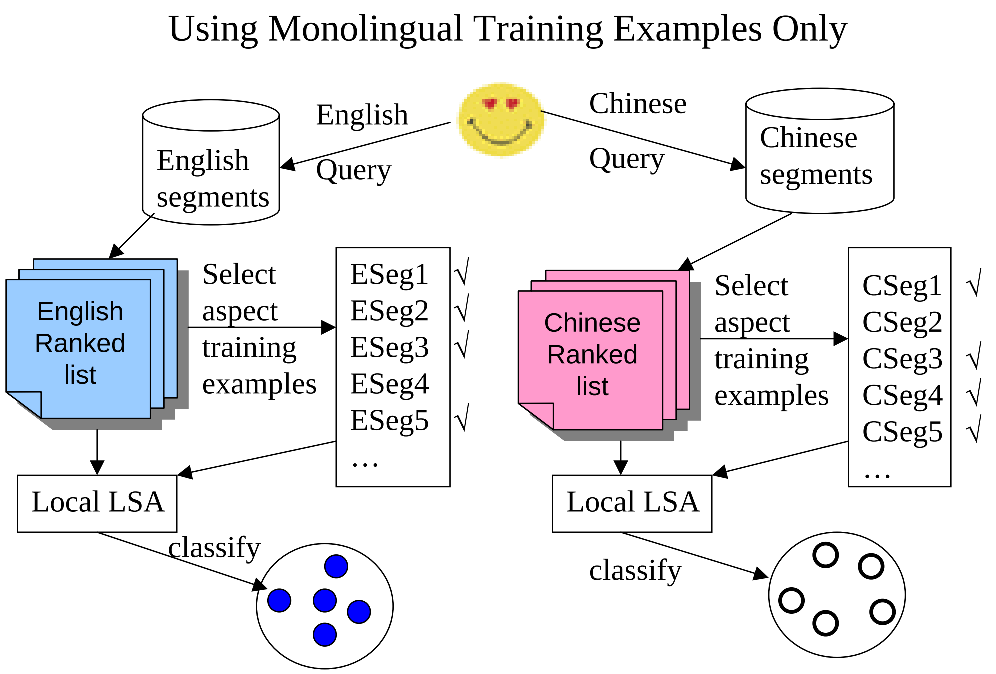


Figure 3.1: A procedure for monolingual aspect classification.

English document segments relevant to a topic from an English collection, and a set of Chinese document segments relevant to the same topic from a Chinese collection. Two information retrieval systems are created for this purpose. The Indri search engine[1] was used to create the two systems.

(2) The user examines the two sets of retrieved document segments and, for each aspect, selects a few (2-4) document segments from the two languages.

(3) Local latent semantic analysis (LSA) is performed on each set of retrieved document segments to reduce the dimensionality of the term space. In the vector space model, documents (and queries) are represented as term vectors in a t-dimensional space (t is the number of terms) [4], which represents both "signal"

[1] http://www.lemurproject.org/indri/ (last visited on January 5, 2008).

(i.e., meaning) and “noise” (from term usage variations). In vector-space approaches to document classification, a notion of distance is defined such that two documents are considered close to the extent that they contain similar terms. LSA reduces the dimensionality of the vector space with semantic information (hopefully) preserved but conflating similar terms towards a “conceptual” representation [125]. For instance, local LSA tends to deemphasize the query terms that generated the topic-specific document collection. The dimensions that are kept are those that explain the most variance. The mathematical basis for LSA is a Singular Value Decomposition (SVD) of the high dimensional term-document matrix, a technique closely related to factor analysis [125]. The SVD represents both terms and documents as vectors in a space of choosable dimensionality [27]. This yields an optimal approximation to the original term-document matrix in the least squares (or L_2 norm) sense. The putative advantages of LSA are the noise reduction and data compaction through the elimination of redundancy [27]. LSA for a large document collection is both computationally expensive and memory intensive, but local LSA is applied to smaller matrices and thus does not suffer from these computational problems. Local LSA is defined as applying SVD to a term-by-document matrix consisting only of the relevant documents relevant to a topic [78].

The SVD decomposes a rectangular matrix of terms by documents (t × d) into three matrices. For example, a t × d matrix of terms and documents X can be decomposed into the product of three other matrices:

$$X = T_0 S_0 D_0^T \tag{3.1}$$

such that T_0 and D_0 are orthonormal matrices of left and right singular vectors and S_0 is the diagonal matrix of singular values. The diagonal elements of S_0 are constructed to be all positive and ordered in decreasing magnitude [27]. By choosing the first k largest singular values in S_0 and setting the remaining smaller ones to zero (and deleting the corresponding columns of T_0 and D_0), we get a matrix $\hat{X}$ which is approximately equal to X, but with rank k.

(4) A classification algorithm takes the manually selected document segments as training examples and identifies which of the unlabeled document segments on that topic best match that aspect (in reduced term space).

The keys to the classification algorithm are a segment-segment similarity function and a threshold for making the classification decision. In the vector space model, a document segment is represented as a vector of term weights, and the similarity of two document segments can be computed as the cosine of the two vectors:

$$sim(D_i, D_j) = \frac{\sum_{k=1}^{t}(w_{ik} * w_{jk})}{\sum_{i=1}^{t} w_{ik}^2 \sum_{j=1}^{t} w_{jk}^2} \tag{3.2}$$

where w_{ik} is the weight of term T_k in document segment D_i,

w_{jk} is the weight of term T_k in document segment D_j,

t is the total number of index terms in the collection.

The term weight w_{ik} is traditionally computed as term frequency (*tf*) multiplied by inverse document frequency (*idf*).

$$w_{ik} = tf_{ik} * idf_k = tf_{ik} * log(\frac{N}{n_k}) \tag{3.3}$$

where w_{ik} is the weight of term T_k in document segment D_i,

tf_{ik} is the frequency of term T_k in document segment D_i,

idf_k is the inverse document segment frequency of term T_k in the

whole document segment collection,

N is the total number of document segments in the collection,

n_k is the total number of document segments that contain term T_k.

A better index term weighting function can lead to a substantial improvement in information retrieval performance. Okapi BM25 term weighting has been shown to be robust and to achieve retrieval effectiveness that is on a par with any other known technologies. Okapi term weighting function is defined as [129]:

$$w_{ik}^{Okapi} = tf_{ik}^{Okapi} * idf_k^{Okapi} \tag{3.4}$$

where

$$tf_{ik}^{Okapi} = \frac{tf_{ik}}{0.5 + 1.5\frac{dl_i}{avdl} + tf_{ik}} \tag{3.5}$$

$$idf_k^{Okapi} = log\frac{N - df_k + 0.5}{df_k + 0.5} \tag{3.6}$$

where dl is the length of document D_i,

$avdl$ is the average document length (DL),

df_k is document frequency (DF) of term T_k.

Okapi term weighting must be applied before local LSA, since after local LSA, the original term-document space becomes a dimension-document space in which there is no way to compute tf and idf.

3.1.2 Bilingual Aspect Classification

Since the user has selected training examples from two languages for the same aspect, the training examples in one language might be used as additional training examples for the other language (if we know how to map them correctly).

Related previous work is the Story Link Detection (SLD) task introduced at TDT'99 [14] which involved determining whether two stories discuss the same topic. However, rather than linking two single documents, we try to link two sets of document segments in two languages. We adopt classical topic-based classification approaches to perform this task. The process involves the following steps, as illustrated in Figure 3.2. The steps are introduced using English as the main language, Chinese as the supporting language, and translating Chinese into English; but it works for the other direction in a similar way.

(1) Once the Chinese aspect training examples are provided by the user, they are translated into English.

(2) Fold in (or map, project) the translated training examples into the original English document segments' LSA space.

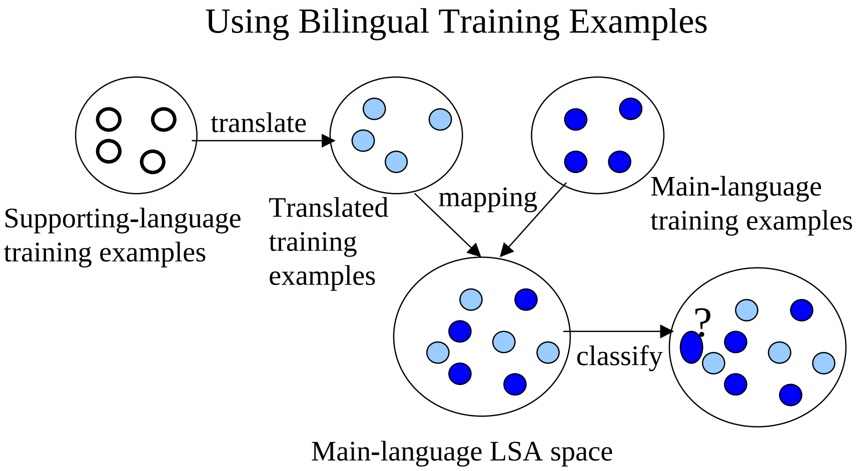

Using Bilingual Training Examples
translate
Supporting-language training examples
Translated training examples
mapping
Main-language training examples
classify
?
Main-language LSA space

Figure 3.2: Translating and folding supporting-language training examples.

(3) This step is optional. Correct the projection of the translated training examples by moving the centroid of these translated segments toward the centroid of the original English training examples.

(4) Classify the unlabeled English segments using the English and Chinese (already translated into English) supporting-language training examples in their English document segments' LSA space.

"Translation" here means mapping term statistics from one language to another, not simply replacing the terms themselves. Due to homonymy, one source-language word can have multiple meanings, and due to target-language synonymy, one meaning can have many suitable translations. So one word in one language could be translated into different words in the other language. For example, in a certain parallel corpus, the Chinese word "*Bai2 Hua4*" might correspond to the English word *Mandarin* with a probability of 0.7, to *Chinese Spoken Language* with a probability of 0.1, to *Modern Chinese Spoken Language* with a probability of 0.05, to *White*

Language with 0.05, etc.

If a translation probability matrix which estimates the probabilities a set of English words will be translated into a set of Chinese words is available, an English segment vector can be translated into a Chinese segment vector by multiplying the segment vector by the translation probability matrix, and then folded into a Chinese LSA space by multiplying the result by the term-by-dimension matrix left singular vector T_0. The same is true for translating and folding a Chinese segment vector. Figure 3.3 illustrates this process.

So all that remains is how to get a bidirectional Chinese/English word translation probability matrix. Translation probabilities can be estimated from parallel corpora, from multilingual dictionaries (when presentation order encodes relative likelihood of general usage), or from the distribution of an attested translation in multiple sources of translation knowledge [184]. In statistical machine translation (MT), translation probabilities are usually learned from parallel corpora. Parallel corpora consist of pairs of documents in two languages that are translations of each other. Such corpora can be aligned at document level, sentence level, or term level. Sentence-aligned parallel corpora are required for statistical MT [183]. With sentence-aligned parallel corpora, the freely available GIZA++ toolkit [128][2] can be used to train translation models. GIZA++ produces a representation of a sparse translation matrix using a three-column table that specifies, for each source-target word pair, the normalized translation probability of the target language word given the source

[2] http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html (last visited on December 20, 2007).

Translation

English document vector — E→ C translation probability matrix — Chinese document vector

$$\begin{array}{cccc} EW_1 & EW_2 & \dots & EW_n \\ \hline ewt_1 & ewt_2 & \dots & ewt_n \end{array} \quad X \quad \begin{array}{c|cccc|} & CW_1 & CW_2 & \dots & CW_m \\ EW_1 & p_{11} & p_{12} & \dots & p_{1m} \\ EW_2 & p_{21} & p_{22} & \dots & p_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ EW_n & p_{n1} & p_{n2} & \dots & p_{nm} \end{array} \quad = \quad \begin{array}{cccc} CW_1 & CW_2 & \dots & CW_m \\ \hline cwt_1 & cwt_2 & \dots & cwt_m \end{array}$$

$$cwt_i = \sum_{j=1}^{n} ewt_i * p_{ji} \quad \text{(i=1 to m)}$$

Folding

Chinese document vector — Chinese Term by Dimension Matrix T_0 — Folded Chinese doc vector

$$\begin{array}{cccc} CW_1 & CW_2 & \dots & CW_m \\ \hline cwt_1 & cwt_2 & \dots & cwt_m \end{array} \quad X \quad \begin{array}{c|cccc|} & Dim_1 & Dim_2 & \dots & Dim_k \\ CW_1 & ct_{11} & ct_{12} & \dots & ct_{1k} \\ CW_2 & ct_{21} & ct_{22} & \dots & ct_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ CW_m & ct_{m1} & ct_{m2} & \dots & ct_{mk} \end{array} \quad = \quad \begin{array}{cccc} Dim_1 & Dim_2 & \dots & Dim_k \\ \hline fwt_1 & fwt_2 & \dots & fwt_k \end{array}$$

$$fwt_i = \sum_{j=1}^{m} cwt_i * wt_{ji} \quad \text{(i=1 to k)}$$

Figure 3.3: Translating an English document vector into Chinese and folding into LSA space.

EW: English Word; CW: Chinese Word;

wt: weight; ewt: English weight; cwt: Chinese weight;

p: probability;

ct: left singular vector values.

language word [183].

The document vectors extracted from the English index of a search engine (i.e., Indri in our experiments) are English word stems and their term weights whereas the translation probability tables were built for Chinese and English words. Although we could train a translation model for English stems, we re-used an existing translated model (see Section 3.3.4), which generated translation probability tables for Chinese and English words. So we need to do some processing of the words and the probabilities.

A word stem could have resulted from multiple word forms, so a Perl script was written to conflate the probabilities of English words into probabilities for the corresponding stems in both translation probability tables. English stopwords were not indexed by Indri, so they were not stemmed but were replaced with "OOV" (Out of Vocabulary) in the translation probability matrices. In the English to Chinese translation probability matrix, if conflation resulted in a summed probability bigger than 1, the probabilities were normalized to 1.

For both monolingual and bilingual aspect classification, we need a classification method that is appropriate for our task. We discuss this issue in the next section.

3.1.3 Classification Methods

Previous studies show that k-Nearest-Neighbor (kNN) and Support Vector Machine (SVM) technologies are among the best text classifiers [203]. Since a topic can have multiple aspects, our classification problem is an m-way multiple class problem. kNN is a natural way for a multiple class problem. However we also briefly investigated SVM because SVM's have demonstrated among the best classification approaches.

kNN vs. SVM

SVM's are a machine learning approach for solving two-class classification problems. The method is defined over a vector space where the problem is to find a decision surface that "best" separates the data points in two classes [203]. A decision surface in a linearly separable space is a hyperplane. The SVM problem is to find the decision surface that maximizes the margin between the data points in a training set.

We did a very small pilot study using SVM and kNN for aspect classification using a single topic. There are two widely used SVM implementations - LibSVM and SVMLight. We chose the LibSVM package since SVMLight requires more extensive manual tuning; LibSVM automatically selects optimal parameters. Since the data points may not be linearly separated, we used a non-linear Radial Basic Function (RBF) kernel. The RBF kernel maps training samples into a higher dimensional space, so it can handle the case when the relation between class labels and attributes is nonlinear [76]. The SVM performed much worse than kNN for that one topic. With only one topic we cannot draw any firm conclusion of course, but there were several reasons that we chose not to further explore to use SVM classifiers:

- It is not our research question whether SVM performs better or worse than kNN for aspect classification. Actually what we need is a classification method that has good performance for our task. The kNN method is a natural choice handling multiple-class problem, and kNN is among the classifiers with highest performance, so we have elected to use kNN for our experiments.

- Aspect classification is in some sense harder than topic classification because all of the aspects are similar to that topic. We suspect that the aspects may not be linearly separable and that even the RBF kernel may not be appropriate for our task. Indeed, it is not clear what kernel function would be appropriate, and customizing an SVM for our specific classification task would therefore be both difficult and time-consuming.

- An SVM generally benefits from training and test data that have approximately balanced number of positive and negative examples. SVM does only binary classification, but our task is a multiple class problem. Although a multiple class problem can be converted to a binary classification problem, our training and test data may not have approximately balanced numbers of positive and negative examples after the conversion. That is, even if a large number of training examples were available, many more negative than positive training examples would be available for each aspect.

- Last but not the least, to achieve a good classification effectiveness, SVM needs lots of training data to learn the decision function that separates two classes, but we have a weakly supervised learning task — only a few training examples are available for each aspect.

kNN and Variants

The k-Nearest-Neighbor (kNN) classifier is a well known instance-based statistical approach which has been studied in pattern recognition and machine learning for over four decades [25] and in text classification for over a decade [5, 188]. kNN is a robust approach to text categorization, ranking among the top-performing classifiers [202]. Here we introduce the classical kNN approach and two variants.

The classical kNN algorithm is very simple: to classify a new object (i.e., a document segment), consult the k training examples that are most similar, where k is an integer, k $\geq$ 1. Each of the k labeled neighbors "votes" for its category. Then count the number of votes each category gets, and assign the category with the maximum number of votes to the new object [113]. In our experiments, the similarity measure was the cosine similarity function (see equation 3.2) with Okapi weights (see equation 3.4, 3.5 and 3.6).

The special case where the category is predicted to be the category of the closest training sample (i.e., when k=1) is called the nearest-neighbor algorithm. kNN for k>1 is more robust than the nearest neighbor method [113]. The best choice of k depends upon the data; generally, larger values of k reduce the effect of noise on the classification, but make boundaries between classes less distinct. A good k can be selected by using empirical techniques (e.g., cross-validation).[3]

In the classical kNN algorithm, the degree of similarity between a test object and training examples is indirectly used (to pick the k nearest neighbors); that is,

[3] see Wikipedia at: http://en.wikipedia.org/wiki/Nearest_neighbor_(pattern_recognition) (last visited on November 20, 2007).

the similarity scores are used to rank the training examples. To make better use of the similarity scores, we introduce two variants of the classical algorithm that directly use the similarity scores.

Franz's algorithm [55], rather than simply assigning the category based on majority voting, sums up the weighted similarity scores as the contribution of the k nearest neighbors to their categories. That is, each category of the k nearest neighbors accumulates the similarity between the new object and the training examples of this category, the category with the maximum sum of similarity scores is assigned to the new object.

A second variant of the classical kNN algorithm was proposed by Yang [202]. Here the conventional M-way classification kNN is adapted to the 2-way classification problem. Since we have multiple aspect classes to work with; when we are working with *Aspect X*, we classify documents either into *Aspect X* or *Non-Aspect X*. Since we have different numbers of positive and negative training examples, we compute a relevance score for each test document as follows [202]:

$$r(x, kp, kn) = \frac{1}{|U_{kp}|} \sum_{y \in U_{kp}} cos(x, y) - \frac{1}{|V_{kn}|} \sum_{z \in V_{kn}} cos(x, z) - Threshold \tag{3.7}$$

where U_{kp} consists of the *kp* nearest neighbors of test document segment vector x among the positive training documents, and V_{kn} consists of the *kn* nearest neighbors of x among the negative training documents. In our aspect classification experiments, we set the threshold to 0 since we do not have any basis for selecting a different value. The basic idea of the formula is to test whether the test document is more similar to the category in consideration or the remaining categories.

In classification, two types of decisions can be made—a *soft* classification decision which the test object can be assigned to more than one classes, and a *hard* classification decision which the test object has to be assigned to only one class. The advantage of the two variants is that if a *soft* classification decision is to be made, the categories with positive relevance values in Yang's variant or with top sums of similarity scores in Franz's variant can be considered as the qualified categories for that test document; if a *hard* classification decision has to be made, the category with the maximum positive relevance value in Yang's variant or the maximum sum of similarity scores in Franz's variant is assigned. In our experiments, we elected to make *hard* classification decisions to make the classifiers easier to build. Soft classification, however, can be useful when users define conceptually overlapped aspects or assign the same or overlapping segments to different aspects as training examples.

We are interested in finding out which kNN approach works best for aspect classification, so we need a metrics to measure the effectiveness of the classifiers created using these approaches.

3.1.4 Evaluation Metrics

Consider a system that is required to make n binary decisions, each of which has exactly one correct answer (either Yes or No, *hard* decisions here). The result of n such decisions can be summarized in a contingency table as shown in Table 3.1.

Given the contingency table, classification effectiveness measures are defined as:

	Yes is Correct	No is Correct	
System-predicted Yes	a	b	a+b
System-predicted No	c	d	c+d
	a+c	b+d	a+b+c+d=n

Table 3.1: Contingency table.

- Precision

$$P = \frac{a}{a+b}$$

- Recall

$$R = \frac{a}{a+c}$$

- F_1 Measure (balanced harmonic mean of precision and recall)

$$F_1 = \frac{2 * P * R}{P + R}$$

- Accuracy

$$ACC = \frac{a+d}{a+b+c+d}$$

We adopted precision, recall, and the F-measure. We did not use accuracy because accuracy is often dominated by d (the count of correctly classifying a truly negative test instance into a negative category) when the negative category is large.

Given x aspects and y document segments for each aspect, there are xy decisions a system needs to make. Given these xy decisions, two ways of computing average effectiveness are available. *Microaveraging* considers all xy decisions as a single group and computes mean precision, recall, and F-measure. *Macroaveraging* computes the mean effectiveness measures separately for each aspect, and then computes the mean of the resulting x effectiveness values [101]. We were not principally interested in the performance of classifying individual document segments, but rather in how well we do on an aspect, so we used the macroaveraged measures; that is, the mean precision, mean recall, and mean F-measure, which we refer to as effectiveness.

Before we are able to build classifiers and measure them, we need training and test data for each aspect, so we need a test collection which is addressed in the next section.

3.2 Test Collection Design

Designing a test collection for this problem turned out not to be an easy task. Several steps are required. First, estimate the number of aspects needed to obtain statistically significance results. Second, prepare the topics and documents for aspect annotation. Third, prepare an annotation procedure approved by the Institutional Review Board (IRB), and then recruit annotators to do the work. Fourth, format the annotation results for use by the aspect classification systems. Finally, assess the inter-annotator agreement and the effect of annotation disagreement on evaluation results. We address all these issues in this section.

3.2.1 Estimating the Number of Topics Needed

Before we start to design a test collection for aspect classification, we need to estimate the number of aspects that need to be created so that we can conduct statistically significant comparisons between two classification systems' effectiveness measures. The two systems work on the same set of aspects, so the statistical test is paired sample test.

As a general statistical exercise, we want to estimate the number of independent samples that are needed to do statistical significance tests. Since the aspects of a topic can be somehow related to each other, they may not be independent samples. However, we can reasonably assume the topics are independent. So we do a statistical exercise here to roughly estimate the number of topics needed.

A quick way to estimate sample size is Cohen's general principle. Assuming precision of aspect classification is normally distributed, according to Cohen's general principle, for $power = 0.8$, $\alpha = 0.05$, the (medium) effect size $d = 0.50$, and a two-tailed paired-sample t-test, total sample size required is $N = 32$ topics [75].

Now we take a closer look at our problem by analyzing a specific case of classifying aspects given 4 training examples and 4 test examples. Assume significance level $\alpha = 0.05$ and mean classification precision $p = 0.60$, we want to test the difference of mean precision between $p1 = 0.60$ (baseline) and $p2 = 0.65$ (alternative system).

Classifying an object into an aspect receives two mutually exclusive outcome—correct or incorrect. Each classification decision is a bernoulli trial, and the distribution of the classification decisions is a binomial distribution. Here $n = 4$ test examples, the probability of making a correct decision is $p = 0.6$, the probability of making a incorrect decision is $q = 0.4$, so

mean (correct decisions) $= np = 2.4$,

standard deviation (correct decisions) $= \sqrt{npq} = 0.9798$ units,

mean (precision) $= 0.6$,

standard deviation (precision) $= \frac{0.9798}{4} = 0.2385$ units.

Using a matched-sample t-test, assuming correlation between baseline precision and alternative precision $\rho = 0.90$, a formula to compute sample size (shown below) can be used to reach N = 36 (see below).

For *power* = 0.8, $\alpha = 0.05$, we know non-centrality power parameter delta $\delta = 2.8$ [75], then effect size

$$d = \frac{p2 - p1}{\rho * \sqrt{2(1-\rho)}} = 0.4688,$$

therefore sample size $N = (\frac{\delta}{d})^2 = 36$.

The estimated sample size for our specific case (36) is close to the estimation made by Cohen's general principle (32). Since it is a rough estimation, and annotation is expensive, we believed we needed at least 32 topics. In order to serve the purpose of classification, we need at least 2 aspect categories within a topic, so we need at least 64 aspects.

3.2.2 Data for Aspect Annotations

To annotate aspects, we need topics, documents, and a definition of document segments. News and blogs in Chinese and English can be collected for our purpose;

however, intellectual property rights considerations make it advantageous to conduct our experiments with news. Fortunately we have English and Chinese news articles from the Topic Detection and Tracking (TDT) collection, which includes pre-annotated topics: the TDT3 collection with 1999 and 2000 evaluation topics, and the TDT4 collection with 2002 and 2003 evaluation topics. TDT3 topics are described in both English and Chinese languages[4] whereas TDT4 topics are described only in English.[5]

The TDT3 and TDT4 collections include news articles and automatically transcribed broadcast news from news sources NYT (New York Times), APW (Associated Press Worldstream), XIN (Xinhua News Agency), ZBN (Zaobao News), VOA (Voice of America), CNN (Cable News Network), NBC (National Broadcasting Company), CBS (Columbia Broadcasting System), and CTS (Chinese Television System), ABC (American Broadcasting Company), MNB (Morning News Beat), and PRI (Public Radio International). TDT3 and TDT4 have also annotated relevant documents for the topics. Most of the English relevant documents are from NYT, APW, and VOA, whereas most of the Chinese relevant documents are from XIN, ZBN, and VOM (VOA Mandarin); adding other sources does not significantly increase the number of relevant documents. So we selected the news documents from NYT, APW, VOA, XIN, ZBN, and VOM for our experiments. The reason we wanted to limit the number of news sources was that each news source had its own format, which complicates document pre-processing. In our test collection, we have

[4] TDT3 topics: http://projects.ldc.upenn.edu/TDT3 (last visited on October 26, 2008.

[5] TDT4 topics: http://projects.ldc.upenn.edu/TDT4 (last visited on October 26, 2008.

33,388 Chinese documents and 37,083 English documents from the selected sources.

3.2.3 Annotation Procedure

When selecting the topics for aspect annotation, we needed a sufficient number of topics (32) and we also needed to select topics with a sufficient number of relevant documents for aspect annotation. We wanted each aspect to have at least 8 document segments so that at least 4 could be used for training and at least 4 for test. We selected 50 topics which had at least 15 relevant documents because we expected that it might be difficult to find at least 8 document segments for at least 2 aspects from fewer than 15 documents. These 50 topics were the maximum number that met that criterion, since some topics might ultimately be rejected if either an insufficient number of aspects or an insufficient number of document segments for an aspect were identified. The 50 topics are listed in Appendix B.

Before annotation started, we had to choose the granularity (i.e., the text unit) for annotation. We could have processed each document into segments and let the annotators work with the document segments directly. This would make the annotation work easier, but it might have made the test collection less useful if the size of a document segment had been poorly chosen. So we decided to divide each document into sentences and defined a document segment as a group of consecutive sentences. The annotators thereby were free to define the length of a segment in any proper manner (e.g., depending on its context).

Chinese sentence boundary detection is relatively easy because Chinese punctuation marks are mostly unambiguous. Chinese periods, question marks, exclamations, and right quotation marks were used to detect sentence boundaries (except for headlines and sub-headlines which usually have no ending punctuation marks). English sentence boundary detection is more complicated because a period is not necessarily a signal of the end of a sentence, so the standard Perl *Sentence* module[6] for sentence splitting was used.

Two graduate students in the Master of Information Management program of the College of Information Studies at the University of Maryland were recruited to annotate the aspects. Both were native speakers of Chinese with good mastery of English. A two-step training session was provided to them before they started the annotation project. In the first step, the investigator explained the concept of aspect. An aspect of a topic was defined as a subtopic and a facet of the topic, and an instance of the aspect was defined as a group of consecutive sentences that addressed the aspect. They were provided with an annotated topic (Topic 30012 Leonid meteor shower) as an example, which was prepared by the investigator. The following aspects were identified for that topic:

1. Aspect 1: reports on what meteors phenomena were observed during the meteors shower,

2. Aspect 2: possible damages caused by the meteors to artificial satellites and communication systems,

[6] http://search.cpan.org/∼shlomoy/Lingua-EN-Sentence-0.25/lib/Lingua/EN/Sentence.pm (last visited on December 20, 2007).

3. Aspect 3: scientific implications of observing the meteors shower,

4. Aspect 4: All Others.

In the second step, the annotators were provided with a topic to do a test run to see whether they understood the process. They asked whether “when,” “where,” and “who” could be defined as aspects of a topic. The investigator explained that these were not the aspects we intended because those were too general, and thus could apply to every news topic. They were told that what we intended were more specific and essential subtopics, such as reasons, consequences, people’s reactions, and influences on our lives. They were instructed not to define too broad aspects that could include multiple subtopics.

The annotation project was split into two phases. In the first phase, each person annotated 25 topics. Provided with the topics and relevant documents in the two languages, they were instructed to identify 2-5 aspects for each topic, and to make an effort to finish annotating each topic within 4 hours. Specifically they were instructed to identify aspects that appeared in both the English and Chinese relevant documents, to provide a brief description of each aspect, to give an example document segment for each defined aspect, and to find at least 8 document segments for each aspect. There was no enforcement of which language the aspects were developed in. In other words, they could identify aspects in Chinese relevant documents, then identify the same aspects in English relevant documents, or vice versa, although the former approach seemed easier because Chinese is their native language.

They were instructed to record the sentence numbers for each document segment. If an aspect appeared in only one news collection (e.g., Chinese) but not in the other (e.g., English), that aspect was not annotated. For examining inter-annotator agreement, in the second phase, each person annotated 5 topics that had already been annotated and were recommended by the other person. The purpose of annotating recommended topics was to speed up the second phase because annotators often recommended the topics they were most confident in their annotations and they provided clear descriptions for the aspects of the topics. This has a potential bias of reporting a higher inter-annotator agreement.

The annotators were instructed to annotate relevant segments about an aspect from as many different documents as possible in order to increase the diversity of document segments for each aspect. The optimal case would be to extract at most one segment about an aspect from one document, but we could not rigidly enforce that restriction if we were to identify at least 8 segments for each aspect. To minimize the chance of extracting multiple document segments from a very small number of documents, we therefore sorted the documents in ascending order of document length. The application for research involving human subjects, informed consent form, and instructions approved by the Institutional Review Board are provided in Appendix A.

The two annotators annotated 176 bilingual aspects for the 50 topics, and 50 *All Others* categories which held non-relevant document segments from the documents where relevant segments had been extracted. The *All Others* categories were incompletely annotated because the annotators focused on annotating their specifically defined aspects and were limited to 4 hours. Recording non-relevant sentence numbers was time-consuming, so they usually recorded fewer than 8 segments for the *All Others* category. The *All Others* categories were not used in our experiments. The annotations were automatically examined to remove non-existent sentence numbers, which the annotators occasionally recorded by mistake.

3.3 Experiment Design

We designed two experiments. Experiment 1 was designed to test three kNN algorithms and five ways of exploiting supporting-language training examples. Experiment 2 then used the best configuration from Experiment 1 to test the effect of varying the number of main and supporting-language training examples on classification effectiveness.

3.3.1 Generating Document Segments

TextTiling is a process for automatically subdividing a text document into multi-paragraph "passages" or subtopic segments that are topically coherent [67]. Before we used it to generate document segments, the documents were preprocessed to strip off XML tags such as <Headline>. The original paragraph boundaries (marked with <P>) were retained by replacing the <P> tags with two newline characters so that Hearst's TextTiling software recognized them. An English period was added to any sentence in the headline and any sentence that did not end with a punctation mark in the document.

Word segmentation for Chinese documents was performed with LDC Segmenter[7] because the translation resources (addressed in Section 3.3.4) were prepared with LDC Segmenter.

Since Hearst's TextTiling software does not work with Chinese documents in UTF8 or GB encoding, Chinese documents were converted to hexadecimal codes. Hexadecimal codes of Chinese punctuation marks were converted into their corresponding English ones. For instance, the hexadecimal code for Chinese empty space (E38080) was replaced with an English empty space.

The TextTiling software has a window size parameter w that needs to be optimized. The parameter defines the length of a text "block." Two text blocks are compared to identify topic shift (or words change). Its default value is 20 words. We ran the program on 2,723 relevant documents (for 23 topics) in the NYT/APE/XIE news collection for the TREC 2003 HARD track[8] 2003 with the w parameter ranging from 2 to 28, generated text tiles, then evaluated them with LDC's passage markings. It turned out that $w = 7$ reached the best retrieval performance measured by R-Precision, which is the precision after R documents have been retrieved, where R is the number of relevant documents for the topic [161]; so this value was adopted for the parameter. Occasionally that parameter setting caused the program to fail to process a Chinese document. When this happened, the parameter was changed to 8, 6, 20, or the length of the entire document, in that order.

[7] http://projects.ldc.upenn.edu/Chinese/LDC_ch.htm (last visited on January 5, 2008).

[8] Text Retrieval Conference, High Accuracy Retrieval from Documents track, a text retrieval task organized by the National Institute of Standards and Technology.

Long tiles are generally not desirable because they might address multiple aspects. So long tiles were further split if certain conditions were met. If a tile had at least 3 sentences and at least 200 words, or had at least 7 sentences and at least 140 words, it was split into segments with a maximum length of 140 words. These parameters were chosen by manually analyzing some of the long tiles and examining the resulted split tiles when different parameters were tried.

3.3.2 Retrieving Document Segments

The English document segments were indexed using Indri.[9] The Porter stemmer was applied and stop words removed during indexing. The Chinese document segments (in hexadecimal codes) were also indexed using Indri. A Chinese stopword list prepared by Gina-Anne Levow [99] was applied.

TDT3 and TDT4 topics have a structure of 7 components: topic number, topic title, seminal event which is composed of What, Who, When, and where, topic explication, rule of interpretation, related articles, and more examples. An example of TDT3 example is provided in Appendix C. Both English and Chinese topics were manually prepared using topic titles, *what* and *who* specifications, and topic explications including "on topic" statements but excluding the notes on "off topic" statements. The 36 English topics were processed into Indri queries with stopwords removed. Indri automatically stems the query words using the porter stemmer.

TDT3 topics 30001-30059 have Chinese versions whereas TDT3 topics 31001-31060 and TDT4 topics 40001-40060 have no Chinese versions. The topics with no

[9] http://www.lemurproject.org/indri/ (last visited on January 5, 2008).

Chinese versions that were used in our experiments were manually translated into Chinese by the investigator with the help of Google translation and English-Chinese dictionaries. If a person or place name had multiple translations, they were all added to the topics. The Chinese topic statements were segmented into words using the LDC Segmenter and then converted into hexadecimal codes.

With the queries and the indexes, we retrieved a ranked list of document segments for each language. Next we experimented with how many document segments we should use to construct the local LSA space.

3.3.3 Dimension Reduction Using LSA

When selecting the number of document segments for building the LSA space, we had two concerns. We needed enough segments so that most of the segments in the gold standard would contribute to the construction of their local LSA space. Although we could have folded in any segment that was not in the LSA space, taking too many segments results in slower SVD computation and less potential for dimensionality reduction. We experimented with taking the top 1,000, 1,500, 2,000, 2,500, and 3,000 segments, and finally chose the top 1,500 Chinese and 2,500 English segments to construct the LSA spaces for those languages, because taking more segments did not considerably increase the number the segments in the gold standard.

Once the number of documents to be retrieved was decided, a term-by-document matrix was constructed (i.e., extracted from the index) for each query. This was the input of the SVD. The outputs we needed were a dimension-by-document matrix D_0, i.e., the document vectors in the LSA space, and a term-by-dimension matrix T_0.

Previous study of the relationship between the number of LSA dimensions retained and mean average precision for retrieval from the Cranfield collection of 1,398 aerospace abstracts showed that retaining 100 dimensions yielded best results [125]. Both the number of abstracts and the length of the abstracts in that experiment were close to our case, so we decided to retain 100 dimensions.

3.3.4 Document Vector Translation and Folding-In

As introduced in Section 3.1.2, the freely available GIZA++ toolkit can be used with sentence-aligned parallel corpora to train translation models and produce translation probabilities for a target language word given a source language word [183]. Fortunately we did not have to actually run GIZA++ to get Chinese-English and English-Chinese translation probability tables because our colleague Jianqiang Wang had prepared the tables we needed for his dissertation [183]. The Chinese-English bidirectional translation probability tables were generated using the Foreign Broadcast Information Service (FBIS) parallel corpus[10]. The word alignment models implemented by GIZA++ are sensitive to the translation direction, so GIZA++ was run twice, one with English as the source language and the other with Chinese as the source language [183]. Wang further improved the English to Chinese transla-

[10] LDC catalog: LDC2003E14, http://projects.ldc.upenn.edu/TIDES/mt2003.html (last visited on January 5, 2008).

tion probabilities by combining the translation probabilities in the two directions and applying statistical synonyms derived from the tables [183]. Specifically, the two directional translation probabilities were used to derive statistical synonyms (translation synsets) for English words and Chinese words. For an English word, the synonyms in its synsets were in Chinese with certain probabilities, and vice versa. An English word and a Chinese word were considered to have the same meaning if each word appeared in one of the other word's translation synsets. The "meaning matching" probability between the two words was estimated by multiplying the probability of the English word in its Chinese translation synsets with the probability of the Chinese word in its English translation synsets. The translation probabilities in the English to Chinese translation probability matrix were then replaced with the meaning matching probabilities, which were ultimately normalized. We directly adopted the resulting translation probability matrices.

We could translate a document segment vector in two ways. One was to directly translate Okapi term weights that was pre-computed using the TF and DF vectors extracted from the Indri index. The other was to extract and translate the TF and DF vectors separately, then to compute the Okapi term weights. We expected that the second way would be better than the first because pre-computed term weights represent the importance of the terms in the source language (e.g., Chinese) collection, and the TF*IDF term weight function is not linear — it rewards rare terms (see Equation 3.3 and 3.6). When a rare term was translated from its source language (e.g., Chinese) to the target language (e.g. English), the resulting term weight could be over-estimated. Estimating the importance of the translated terms

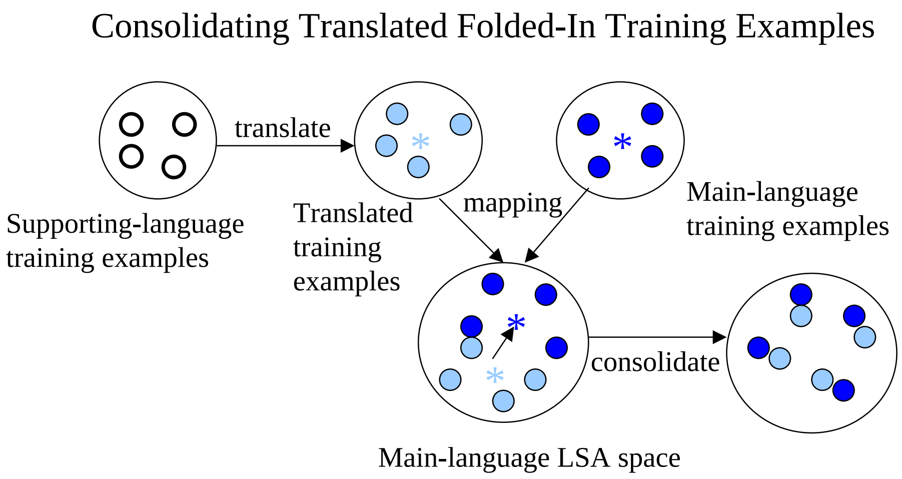

Consolidating Translated Folded-In Training Examples
translate
Supporting-language training examples
Translated training examples
mapping
Main-language training examples
consolidate
Main-language LSA space

Figure 3.4: Consolidating translated folded-in training examples.

in the target language (e.g., English) by translating TF and DF vectors separately before computing TF*IDF can avoid this problem.

3.3.5 Consolidating Translated Folded-In Document Vectors

When supporting-language document vectors are translated and folded into the main-language LSA space, their positions in the space might be systematically distorted by translation errors and projection errors. Adjusting the positions of the folded-in vectors toward the main-language training document vectors so that the centroids of the two sets of training examples overlap might put the folded-in segment vectors in better positions. Figure 3.4 illustrates this process.

3.3.6 Processing the Test Collection for Training and Test

The annotation process resulted in a raw gold standard which could not be directly used for classification experiments because our classification system dealt with machine-generated documents segments, which were generally different from the annotated document segments. We also required that aspects in a topic be mutually exclusive. So the raw gold standard was processed to fit our classification experiments.

Validating and Mapping

Validation was first performed to remove sentences and aspects that were not appropriate for our classification experiments. We used greedy selection to ensure that retained aspects were mutually exclusive. That is, if the same sentence appeared in two or more aspects, the first aspect was kept and every other aspect annotated for that sentence was removed. If an aspect in one language was removed, its corresponding aspect in the other language was also removed. In this way, we guaranteed that no sentence was common to two document segments (or passages) that were annotated with different aspects. If all the aspects of a topic had duplicate sentences, that topic was dropped.

For our classification experiments, we automatically partitioned each document into fixed segments (see Section 3.3.1 above). In the classification training and test process, we needed to know whether a machine-generated document segment was relevant to an aspect, so we had to map the machine-generated segments onto the document segments in the raw gold standard. We did this based on sentence overlap. If at least one sentence in a machine-generated segment was marked as relevant to a certain aspect, this segment might be mapped onto that aspect. Therefore, one machine-generated segment could possibly be mapped onto multiple candidate aspects; when this happened, a single mapping decision was made by a majority voting. That is, if one sentence in the machine-generated segment was mapped to *Aspect 1*, and two sentences were mapped to *Aspect 2*, the decision would be to map it to *Aspect 2*. For example, suppose the raw gold standard had assigned the following sentence S1, S4, S5, S6, S10 and S11 into the following aspects:

- Aspect 1: S1
- Aspect 2: S4, S5, S6
- Aspect 3: S10, S11

Suppose the machine-generated segment was composed of S1, S2, S3, S4 and S5, then S1 was mapped to Aspect 1, S4 and S5 were mapped to *Aspect 2*, so this segment was mapped to *Aspect 2* based on majority voting. When ties were encountered, the mapping outcome was decided arbitrarily by choosing the lowest numbered candidate.

Two Operational Test Collections

Validation and mapping usually decreased the number of document segments for an aspect. Requiring at least 8 document segments per aspect disqualified too many topics and aspects. We therefore required each aspect to have at least 5 segments, so that 4 segments could be used for training and the others for testing. There were a total of 106 bilingual aspects from 36 topics that met this requirement (excluding the *All Others* categories).

Document segments automatically retrieved by using the topic description as a query were used to create LSA spaces in each language. By trial and error, we selected the top 1500 Chinese and 2500 English document segments (see Section 3.3.3). To simplify our experiments, we dropped the document segments that were in the gold standard but were not in the ranked list of selected retrieved segments (although we could have kept them by folding them into the LSA spaces). This made us drop 3 more topics when all (or all but one) of their aspects fell below 5 document segments. Ultimately we used 92 bilingual aspects from 33 topics, including 3 Chinese aspects that could only be used as training data for English aspect classification because each of them had only 4 segments. This is the first version of our operational test collection.

In order to examine the effect of varying the number of training examples on classification effectiveness, a second version of the test collection was created in which we required at least 7 segments for each aspect so that 1-6 segments could be used for training. This version has 40 aspects from 17 topics.

The original 50 topics, the 36 topics after validation and mapping, and the topics for the two operational test collections are all listed in Appendix B.

Inter-annotator Agreement

As mentioned above, in the second phase of the annotation project, each annotator was provided with 5 topics which had already been annotated and recommended by the other annotator. Each annotator was also provided with aspect descriptions and one example for each aspect, and the documents in which segments had been found for that aspect. If the second annotator could not find at least 8 relevant document segments from the provided documents, she could also use any other relevant documents for the topic (which were also provided). This is a semi-open annotation task because the documents for each aspect were provided, but document segments could be freely defined.

When validating the annotations for the inter-annotator agreement study, non-existent sentences were removed, but aspects with duplicate sentences were retained. We could have examined the inter-annotator agreement from these unmapped annotations, but that would give at best indirect information about the effect of disagreements. Therefore we instead measured inter-annotator agreement using machine-generated segments that were mapped onto the annotated aspects. In this way we could directly examine the effect of disagreement on our experiment design.

Inter-assessor agreement was computed across 2 categories: a document segment relevant to or not relevant to an aspect. The average Cohen's kappa was 0.22 for English aspects and 0.50 for Chinese aspects. This seems reasonably good for an semi-open annotation task. The agreement on Chinese aspects was higher than that on English aspects, probably because Chinese was their native language, so they could judge the relevance of Chinese segments more consistently. Since the annotators annotated topics recommended by each other, the inter-annotator agreement reported here can be higher than annotating randomly selected topics.

3.3.7 Experiment 1

Experiment Design

The first experiment was designed to serve two purposes: to test whether adding supporting-language training examples would help to classify main-language segments by aspect, and, if it were helpful, to identify the best way of using the supporting-language training examples. We needed as many aspects as possible, so the first version of the operational test collection was used (in which each aspect has at least 4 segments).

The document segments for each aspect were partitioned into training and test sets using cross-validation. We elected to run a maximum of 70 rounds of cross-validations. If more than 70 combinations were available, only 70 were randomly selected; if fewer than 70 combinations were available, all the combinations were taken. When supporting-language training segments were added together, the language with the greatest number of combinations determined the number of cross-validation rounds.

The parameter k of kNN was automatically selected topic by topic depending on the number of aspects for that topic and the number of available training examples for an aspect. If a topic had m aspects (excluding the *All Others* category), when the number of available training examples in each aspect was no more than 2, k was set to the number of available training examples in the aspect; otherwise, k was set to $2m + 1$. k was set to an odd integer to minimize cases of ties (except when $k = 2$).

When the training data and test data were ready, we ran the aspect classifiers. We have three classification algorithms—classical kNN, Franz's variant, and Yang's variant—and five ways of using supporting-language training data:

- Base: baseline systems using main-language training data only; supporting-language training data were not used.

- Fold: translating supporting-language document segment vectors with pre-computed Okapi term weights, then folding them into main-language local LSA space.

- TrTD: translating the supporting-language document segments' TF and DF vectors separately, then computing their Okapi weights, then folding the vectors with Okapi weights into the main-language local LSA space.

- FoldC: translating the supporting-language document segment vectors with pre-computed Okapi term weights, folding them into main-language local LSA space, then moving them toward the main-language training data so that their centroids meet.

- TrTDC: translating the supporting-language document segments' TF and DF vectors separately, computing their Okapi weights, folding the vectors with Okapi weights into main-language local LSA space, moving them toward the main-language training data so that their centroids meet.

The three classification algorithms were applied to the five ways of using supporting language training data, so there were a total of 15 runs. Based on the results of this experiment, we could find the best algorithm and the best way of exploiting supporting-language training data for Experiment 2.

Results

Experiment 1 tested the effect of adding 4 supporting-language training examples to 4 main-language training examples on classification performance. Table 3.2 shows the comparisons of the three kNN classification algorithms across the five ways of using supporting-language training examples for classifying English aspects. Table 3.3 shows the similar comparisons for classifying Chinese aspects.

Both Table 3.2 and Table 3.3 show that both *FoldC* and *TrTDC* consistently yield lower mean precision, recall, and F-measure than *Fold* and *TrTD* respectively, indicating that moving the supporting-language training examples toward the main-language training examples until their centroids meet does not work well.

There might be three reasons that have resulted in the failure of centroid moving:

- One is that translation and folding the translated training examples into the main-language LSA space at document segment scale introduces less error than centroid moving, so we lose information when we move the centroids of the translated training examples in the LSA space.

- A second reason is that we might have moved the folded-in segment vectors too far away so that more errors have been introduced by centroid moving.

- A third reason is that systematic translation error might not have occurred (some words were correctly translated sometimes, some not, so random error

	Classical (CL)			Franz			Yang		
Run	P	R	F_1 (stdev)	P	R	F_1 (stdev)	P	R	F_1 (stdev)
Base	0.5064	0.5508	0.4952 (0.3285)	0.5357	0.5762	0.5232 (0.3119)	0.5525	0.5919	**0.5359** (0.3162)
Fold	0.5617	0.5931	0.5363 (0.3016)	0.5959	0.6365	**0.5823** (0.2941)	0.5757	0.6244	0.5632 (0.2985)
TrTD	0.5720	0.5947	**0.5394** (0.2991)	0.6144	0.6469	**0.5896** (0.2716)	0.5983	0.6381	**0.5762** (0.3007)
FoldC	0.5183	0.5172	0.4697 (0.2957)	0.5444	0.5424	0.5003 (0.2928)	0.4756	0.4983	0.4413 (0.2941)
TrTDC	0.5298	0.5388	0.4898 (0.3177)	0.5741	0.5646	0.5241 (0.2931)	0.5071	0.5303	0.4727 (0.2924)

Table 3.2: English aspect classification; 4 E and 4 C training examples; mean over 92 aspects from 33 topics. Mean P, R, and F_1 results averaged over 92 English aspects from 33 topics. The mean F_1 is averaged over the F_1 scores of 92 aspects, not computed from mean P and mean R. A bold score is the highest score either in the row or in the column where it is. Stdev stands for standard deviation.

Runs	CLASSICAL (CL)			FRANZ			YANG		
	P	R	F_1 (stdev)	P	R	F_1 (stdev)	P	R	F_1 (stdev)
Base	0.5170	0.5353	0.4922 (0.3054)	0.5442	0.5435	0.5161 (0.3136)	0.6311	0.6209	**0.5940** (0.2880)
Fold	0.5442	0.5435	0.5161 (0.3136)	0.6183	0.6204	0.5877 (0.3037)	0.6268	0.6324	**0.6015** (0.2957)
TrTD	0.5892	0.5760	**0.5529** (0.3168)	0.6369	0.6223	**0.5988** (0.3062)	0.6303	0.6324	**0.6062** (0.3054)
FoldC	0.4803	0.4903	0.4463 (0.2791)	0.5059	0.5235	0.4821 (0.3084)	0.4560	0.4746	0.4322 (0.2988)
TrTDC	0.4539	0.4716	0.4363 (0.2942)	0.5070	0.5176	0.4838 (0.3124)	0.4896	0.5271	0.4732 (0.3120)

Table 3.3: Chinese aspect classification; 4 C and 4 E training examples; mean over 89 aspects from 32 topics. Mean P, R, and F_1 results averaged over 89 Chinese aspects from 32 topics. The mean F_1 is averaged over the F_1 scores of 89 aspects, not computed using mean P and mean R. A bold score is the highest score either in the row or in the column where it is. Stdev stands for standard deviation.

may have occurred), and correcting all the dimensions of the translated training examples in the LSA space might be misleading, and it is hard to detect which dimensions need to be corrected.

Since *FoldC* and *TrTDC* work worse than *Base*, *Fold* and *FoldC*, now we focus on comparisons between *Base*, *Fold*, and *TrTD*. *Fold* and *TrTD* consistently improve classification effectiveness over the baseline, indicating that supporting-language training examples are useful. *TrTD* is better than *Fold*, again confirming that translating TF and DF vectors then computing Okapi term weights is better than translating a vector of pre-computed term weights. The two kNN variants are always better than the classical kNN. There is no consistent difference between *Franz TrTD* and *Yang TrTD*. We therefore arbitrarily elected *Franz TrTD* for Experiment 2.

3.3.8 Experiment 2

Experiment Design

The second experiment was designed to examine the effect on classification effectiveness of varying the number of main- and supporting-language training examples on classification performance. The second version of the test collection was used, in which each aspect has at least 7 segments in both languages. We used 1-6 document segments for training and the remainder for test. Again, a maximum of 70 rounds of cross-validations were performed. Only the best classification algorithm and the best way of applying supporting-language training examples as found by Experiment 1 were used for this experiment. Since fewer than 4 document segments could be used as training examples, the parameter k of kNN was set somewhat differently than in Experiment 1. If an aspect had only 1 or 2 training examples, k was the number of available training examples (i.e., 1. or 2); otherwise, k was set to $2m + 1$ where m was the number of aspects in a topic.

Results and Discussion

Raw experiment results are shown in Table 3.4 and Table 3.5. All of the figures in this section were plotted using the data in these two tables.

Figure 3.5 shows that when only main-language training examples are involved, more training examples generally yields better classification effectiveness. This meets our expectations. Linear regression models for predicting scores from the number of main-language training examples are significant at $p < 0.05$ (the correlation coefficient is 0.841 for English, and 0.968 for Chinese). For Chinese aspect classification, more training examples consistently yield better classification effectiveness, implying that the training examples are well-annotated, as reflected by the Chinese aspect inter-annotator agreement (kappa=0.5).

Notice that, for English aspect classification, when there are at least 4 English training examples are available for an aspect, the classification effectiveness is generally bad—there is a drop when 4 and 6 training examples are provided, and when 5 English training examples are provided, the effectiveness (F=0.6559) is not much higher than the effectiveness when 3 training examples are provided (F=0.6467). There might be two reasons that result in the drop:

k segments	1	2	3	4	5	6	7	8	9	10	11	12
1E+(k-1)C	0.5211	0.5652	0.5910	0.6044	0.5884	0.5944	0.6395					
2E+(k-2)C		0.5588	0.5618	0.5877	0.5865	0.5913	0.5475	0.5797				
3E+(k-3)C			0.6467	0.6366	0.6405	0.6484	0.6199	0.6434	0.6352			
4E+(k-4)C				0.5541	0.6099	0.6692	0.6289	0.6437	0.6341	0.6637		
5E+(k-5)C					0.6559	0.6279	0.6710	0.6585	0.6739	0.6337	0.6741	
6E+(k-6)C						0.6410	0.6580	0.6881	0.6424	0.6496	0.6172	0.6545

Table 3.4: F_1 of classifying English aspects using varying numbers of training examples (k is the total number of main- and supporting-language training examples; E: English, C: Chinese).

k segments	1	2	3	4	5	6
1C+(k-1)E	0.5375	0.6200	0.6018	0.6300	0.6402	0.6180
2C+(k-2)E		0.5656	0.6006	0.5792	0.6366	0.6527
3C+(k-3)E			0.6027	0.6656	0.6931	0.7016
4C+(k-4)E				0.6701	0.7070	0.7078
5C+(k-5)E					0.6900	0.7352
6C+(k-6)E						0.7088
k segments	7	8	9	10	11	12
1E+(k-1)C	0.6143					
2E+(k-2)C	0.6247	0.6267				
3E+(k-3)C	0.6827	0.7085	0.6633			
4E+(k-4)C	0.6922	0.6849	0.7047	0.6941		
5E+(k-5)C	0.7226	0.7433	0.7074	0.7115	0.7308	
6E+(k-6)C	0.6970	0.6361	0.6633	0.6424	0.6719	0.6416

Table 3.5: F_1 of classifying Chinese aspects using varying numbers of training examples (k is the total number of main- and supporting-language training examples; E: English, C: Chinese).

- The main reason is that the English training examples are not consistently good, as reflected by the English aspect inter-annotator agreement (kappa=0.22); that is, some training examples might not have been distinguished well across aspects; for instance, some training examples annotated for one aspect might be equally good for another aspect of the same topic.

- A second reason might be related to cross validation. When more training examples are available, the bad training examples have better chance to be in a round of cross validation. For instance, suppose we have a total of 8 training examples for each aspect, in which 1 example is bad. We run a total of 8 rounds of cross validation. We select 1 example for training, the chance of that bad example is selected is 0.125. If we select 2 examples for training; we run 28 rounds of cross validation. The chance that bad example is selected is 0.25. Now we select 4 examples for training; we run 70 rounds of cross validation; the chance that bad example is selected is 0.5.

Figure 3.6 reinforces the conclusion from Experiment 1 that supporting-language training examples are useful. It shows that when equal numbers of supporting-language and main-language training examples are used, classification effectiveness usually increases (6C and 5E being the exceptions).

Figure 3.6 confirms that supporting-language training examples are helpful, but it cannot tell us whether they are as helpful as an equivalent number of main-language training examples would be. To examine whether main-language training examples are superior to supporting-language ones, Figure 3.7 was plotted. No clear

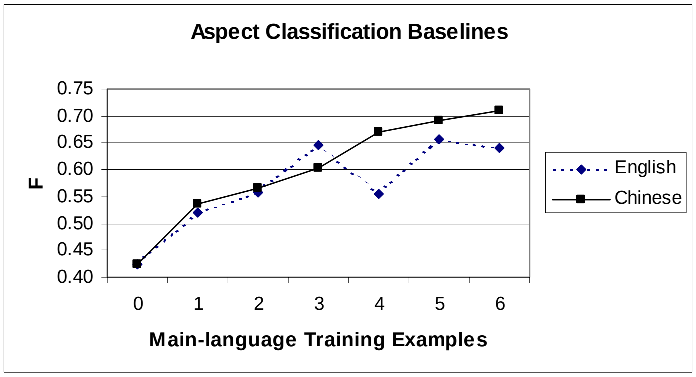


Figure 3.5: Classifying English and Chinese segments using main-language training examples only. When no training examples were provided to 40 aspects from 17 topics, a random classifier would achieve an F of 0.425.

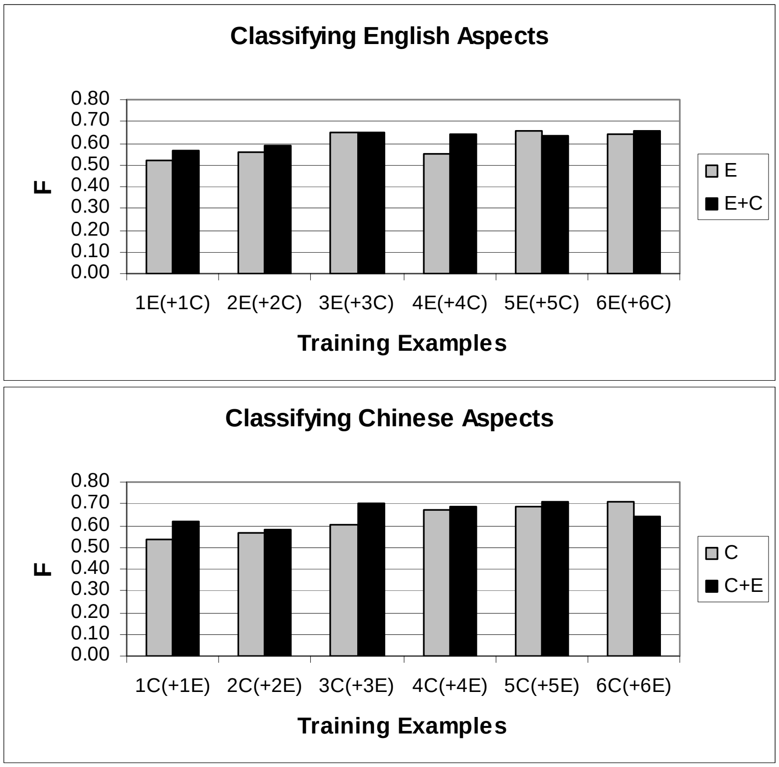


Figure 3.6: Effect of adding equal numbers of supporting-language training examples. Grey bar is monolingual training; black bar is bilingual training.

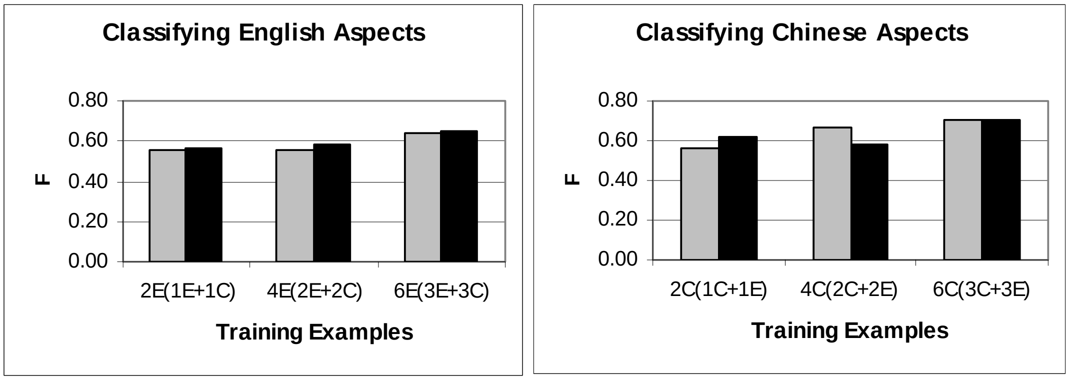


Figure 3.7: Comparing supporting-language training examples with main-language training examples with totals held constant.

trend is evident.

Figure 3.8 shows supplementing varying numbers of main-language training examples with varying numbers of supporting-language ones. Although a bit busy, it generally shows a point of diminishing returns is reached beyond which fluctuations appear random.

When a fixed number of main-language and supporting-language training examples (k=3-9) are available, classification effectiveness is generally positively correlated with the percentage of main-language training examples. Table 3.6 shows that Pearson's r correlation scores are positive (except for one outlier), and some are statistically significant at $p < 0.05$. Figure 3.9 is for $k = 6$. Due to sparse data, the cases $k < 3$ or $k > 9$ are not included in Table 3.6.

From the findings that: (1) for main-language training examples, more training examples yields better classification effectiveness; (2) a point of diminishing returns occurs after a few foreign-language training examples have been added; and (3)

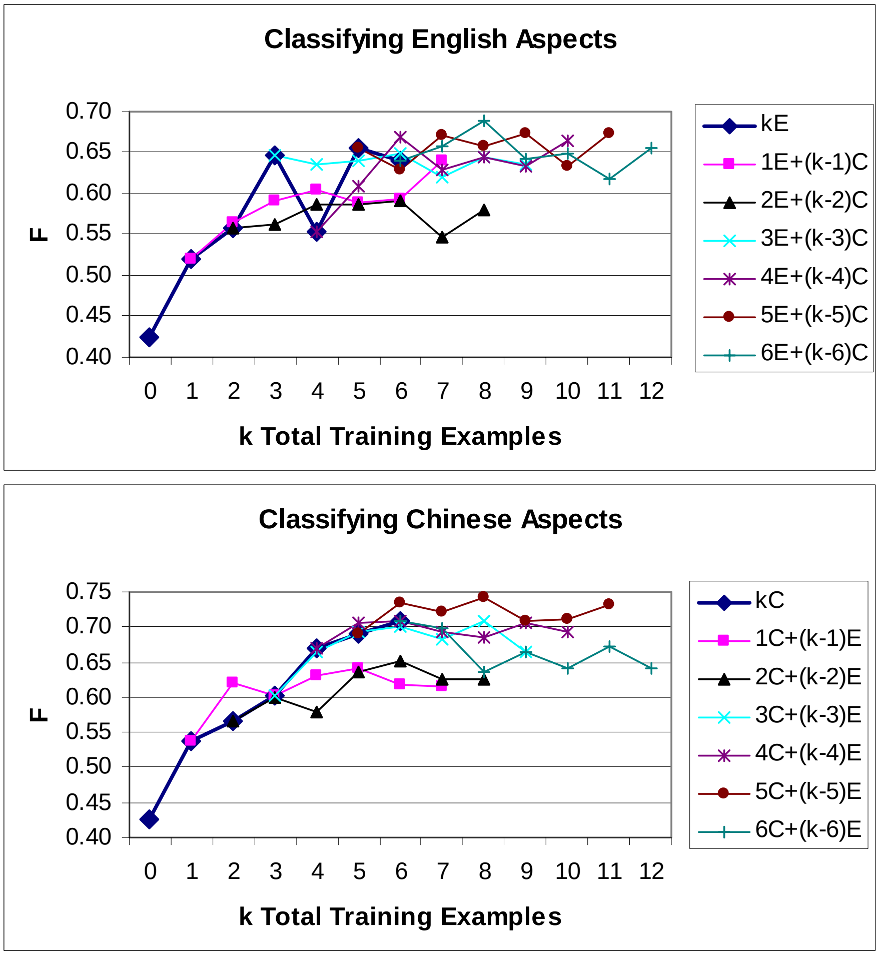


Figure 3.8: Varying numbers of main-language and supporting-language training examples.

k	Correlation		df
	English	Chinese	
3	0.6391	0.4193	1
4	-0.3832	0.6348	2
5	0.8061*	0.8247*	3
6	0.6347	0.8722*	4
7	0.5829	0.8890*	4
8	0.9253*	0.1729	3
9	0.4374	0.0148	2

Table 3.6: Correlation between classification effectiveness and percentage of main-language training examples. * means statistically significant at $p < 0.05$; k is the total number of training examples used; df is degree of freedom.

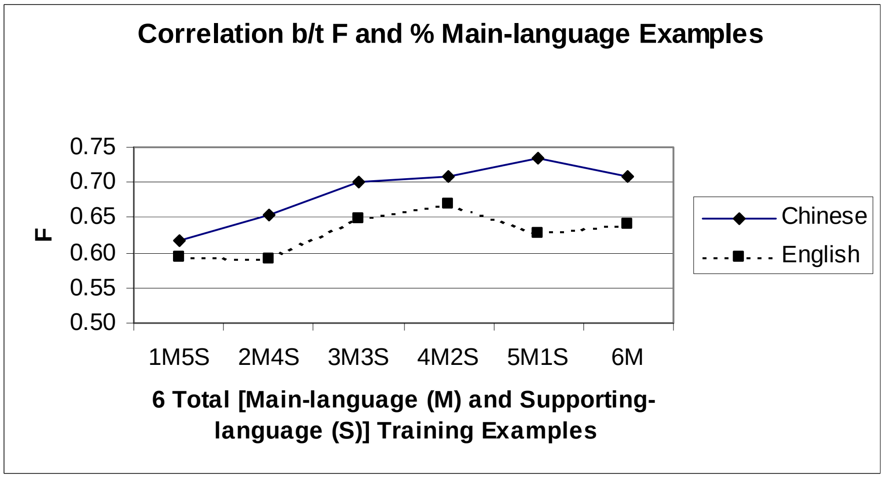


Figure 3.9: Correlation between classification effectiveness and predominance of main-language training examples, given 6 training examples.

classification effectiveness was generally correlated with the percentage of main-language examples, we infer that the usefulness of supporting-language training examples are constrained by:

- annotation errors introduced when annotating the main-language and supporting-language training examples;
- translation errors introduced when translating the supporting-language training examples into main-language;
- projection errors introduced when folding-in the translated training examples into the main-language local LSA space;
- the availability of main-language training examples.

Annotation errors, translation errors, and projection errors are to some extent unavoidable. However, main-language training examples have the problem of annotation errors and availability (or scarcity), whereas supporting-language training examples have the problem of annotation errors, translation errors, and projection errors. When the scarcity of main-language training examples outweighs the problems inherent in using supporting-language training examples, the supporting-language training examples can be still useful.

3.4 Implications for Future Work

Documents on the same topics but in different languages are not rare. Truly polyglot users may be rare, but people who have some knowledge and skills of a supporting language are not rare. The discovery that supporting-language training examples can be useful for main-language aspect classification encourages us to leverage the resources in the two languages to achieve better classification results, especially when the resources (such as labeled training segments) in one language are scarce.

Although we experimented bilingual aspect classification with English and Chinese only, our discovery have encouraged us to expand the experiment design to include more languages in the future.

The idea of moving the translated supporting-language training examples toward the main-language training examples in the LSA space until their centroids meet does not work well in our current experiment. However, translation errors and folding-in errors might have occurred, so it is worth testing the idea of moving the centroid of the translated supporting-language training example a little bit (or half way) toward the centroid of the main-language training examples.

The goal of text classification is to classify the topic or theme of a text document. It can be considered as a special case of information retrieval, in which the query is the text document to be classified and the items to be retrieved are categories. So bilingual aspect classification may have implications for bilingual information retrieval tasks. For instance, the relevant documents identified from one language might be translated and used as relevance feedback examples to get better retrieval results of the other language.

3.5 Summary

In this chapter, we have addressed one of the two major tasks in our ultimate goal—bilingual aspect classification. Specifically we have addressed the methods for answering the research question, test collection design, experiment design, and results and findings.

We have asked whether supporting-language training examples can be useful for main-language aspect classification and found the answer to be yes. In other words, our bilingual aspect classification approach which applies both main-language and supporting-language training examples improves classification effectiveness over our monolingual aspect classification approach which applies main-language training examples only. We used document segments as aspect instances.

For monolingual aspect classification, we retrieved a set of main-language document segments relevant to a topic from the main-language document collection, performed local Latent Semantic Analysis (LSA) to reduce the dimensionality of the term space. For each aspect identified by the user, we used a few main-language document segments provided by the user as training examples to classify the remaining document segments.

For bilingual aspect classification, we retrieved a set of main-language document segments relevant to a topic from the main-language document collection, and a set of supporting-language document segments relevant to the same topic from the supporting-language document collection. We performed main-language local LSA. For each aspect, we have a few training examples from both the main-language and the supporting-language. We translated the supporting-language training examples into main-language, and folded in the translated training examples into the main-language local LSA space, then classified the remaining main-language document segments.

We used k-Nearest-Neighbor (kNN) and its two variants as our classification methods.

- The classical kNN algorithm assigns the category of a new object based on majority voting of the categories of the k nearest neighbors.

- Franz's variant sums up the weighted similarity scores between the new object and the training examples of each category, then assigns the category with the maximum sum of similarity scores to the new object.

- Yang's variant converts a conventional M-way classification kNN into the 2-way classification problem (i.e., *Aspect X* or *Non-Aspect X*), then computes the average similarity between a new object and the training examples in *Aspect X* (or *Non-Aspect X*); if the new object is more similar to the category in consideration (*Aspect X*) than to the remaining categories (*Non-Aspect X*), then the new object is classified into this category.

We tested five ways of using supporting-language training examples:

- Base: baseline systems using main-language training data only; supporting-language training data were not used.

- Fold: translating supporting-language document segment vectors with pre-computed Okapi term weights, then folding them into main-language local LSA space.

- TrTD: translating the supporting-language document segments' TF and DF vectors separately, then computing their Okapi weights, then folding the vectors with Okapi weights into the main-language local LSA space.

- FoldC: translating the supporting-language document segment vectors with pre-computed Okapi term weights, folding them into main-language local LSA space, then moving them toward the main-language training data so that their centroids meet.

- TrTDC: translating the supporting-language document segments' TF and DF vectors separately, computing their Okapi weights, folding the vectors with Okapi weights into main-language local LSA space, moving them toward the main-language training data so that their centroids meet.

To evaluate our classifiers, we designed an aspect classification test collection in which two annotators annotated consecutive sentences in a document as aspect instances for 176 aspects from 50 topics in English and Chinese using the TDT3 and TDT4 news collections. Topically relevant English documents were selected from APW, NYT, and VOA news sources, whereas topically relevant Chinese documents were selected from XIN, ZBN, and VOA news sources.

We processed a document into document segments as aspect instances using a variant of TextTiling, and the document segments were then mapped onto the annotated document segments in the raw gold standard. Two versions of the test collection were created. The first has 92 aspects from 33 English topics, with each aspect having at least 4 document segments. The second has 40 aspects from 17 topics in both languages, with each aspect having at least 7 document segments.

For examining inter-annotator agreement, each annotator annotated 5 topics that had already been annotated and were recommended by the other annotator. We measured inter-annotator agreement using machine-generated segments that were mapped onto the annotated aspects. In this way we could directly examine the effect of disagreement on our experiment design. The average Cohen's kappa was 0.22 for English aspects and 0.50 for Chinese aspects, indicating the annotators have annotated better document segments for Chinese aspects than for English aspects.

Our first experiment aimed to find out the best kNN approach and the best way of using supporting-language training data, so we used the first version of the test collection with the specific case of supplementing 4 main-language training examples with 4 supporting-language training examples. We found that translating the TF and DF vectors of supporting-language training examples, combining them into vectors with Okapi weights, folding them into the main-language LSA space, and then using Franz's kNN variant worked best (although Yang's kNN variant is not statistically significantly different). That is, TrTD with Franz's kNN variant worked best. This confirmed our expectation that translating TF and DF vectors then computing Okapi term weights is better than translating a vector of pre-computed term weights, because pre-computed term weights represent the importance of the terms in the supporting-language collection and could be over-estimated when rare terms are rewarded.

The second experiment used the second version of the test collection to examine the effect of supplementing 1-6 main-language training examples with 0-6 supporting-language training examples. From that experiment we found that:

- For main-language training examples, more training examples yields better classification effectiveness. Chinese aspect classification effectiveness consistently improves when more training examples are provided, whereas English aspect classification effectiveness fluctuates when more than 4 training examples are provided. This might be due to the quality of English training examples as reflected by the relatively lower inter-annotator agreement of English aspect annotations.

- Adding an equivalent number of supporting-language training examples generally helps.

- A point of diminishing returns occurs after a few foreign-language training examples have been added.

- Classification effectiveness was generally correlated with the percentage of main-language examples.

- The usefulness of supporting-language training examples is constrained by the following factors: annotation errors of main-language and supporting-language training examples, translation errors, and projection errors.

Chapter 4

Chinese Attitude Classification

Chinese attitude classification is one of the major tasks in our ultimate goal. This chapter addresses the methods for exploring the research questions for Chinese attitude classification, experiment design, experiment results, and findings.

4.1 Methods

Generally there are two types of document and sentence subjectivity and polarity classification approaches—the machine learning approach that builds classifiers by extracting features from training data, and the linguistic analysis approach that is based on sentiment lexicons and linguistic features.

Previous studies on English document and sentence sentiment classification found that generally the machine learning approach did not work as well as the linguistic analysis approach although a firm conclusion about which approach is better is hard to draw. Furthermore, the machine learning approach requires sufficient amount of training data which are not available to us. Therefore we took the linguistic analysis approach.

A sentence is composed of words structured with syntax. As an engineering simplification, we classify the polarity of a sentence by first classifying the semantic orientation of words in the sentence and then aggregating the sentence polarity from words. A *comprehensive* out-of-context prior-polarity sentiment lexicon is desirable but impossible to get. Fortunately both small Chinese and English sentiment lexicons are available, and we can compose a relatively big Chinese sentiment lexicon by translating the English sentiment lexicon into Chinese, combining them, and expanding the combined lexicon by adding and removing some characters that function as prefix and suffix. For words not in the lexicon, we compute the semantic orientation of the words by extending Ku et al.'s algorithm [93]. The word polarity classifiers are to be evaluated using the lexicon with known polarity.

4.1.1 Lexicon Acquisition and Preparation

Since a sentence is composed of words even though a Chinese sentence does not mark word boundaries, we started sentence attitude classification by examining attitude from the granularity of a word; therefore acquisition of sentiment lexicon and classifying the semantic orientations of words became the first step.

We acquired NTU's Chinese sentiment lexicon which annotated positive and negative words, and extracted sentiment words from the documents for four training topics [199]. We manually rekeyed two books (Chinese Positive Dictionary [159] and Chinese Negative Dictionary [205]) from Simplified Chinese to Traditional Chinese. We also acquired Wilson and Wiebe's English sentiment lexicon [195]. In this section we address how we translated the English sentiment lexicon into Chinese, and created a large lexicon by leveraging all these lexicons. The sentiment words were finally categorized into 9 mutually exclusive categories.

Categories of Sentiment Words

Three types of words play a role in expressing opinions—opinion words, opinion operators, and opinion indicators [200]. An opinion word expresses the polarity of an opinion, such as favorable, unfavorable, and neutral. An opinion operator is a verb indicating an opinion expression, such as *point out*, *claim*, and *criticize*. An opinion indicator is usually a conjunction, an adverb, an adverbial phrase, or a verb which indicates the orientation of an opinion or the orientation trend of multiple opinions. There are four types of typical opinion indicators [200]: (1) negational conjunctions (such as *but*, *however*, *nevertheless*), (2) continual or strengthening conjunctions (such as *and*, *especially*, *furthermore*), (3) adverbs and adverbial phrases directly indicating the polarity of opinionated sentences (such as *unfortunately*, *regrettably*), (4) opinion operators that directly indicate the polarity of the opinionated sentences (such as *criticize*, *praise*).

We created 9 mutually exclusive categories: opinion operators, opinion indicators, intensifiers (mainly adverbs such as *very*), quantifiers (mainly subjective quantity words, such as *a large number of, big mass*), negation characters and words (mainly adverbs), functional negation words (verbs work semantically as negation, such as *quench*, *stop*, *remove*), positive (POS), neutral (NEU), and negative (NEG). There are 4 additional categories which intersect with the positive, negative and neutral categories: words that, depending on context, can be both positive and neutral (POS+NEU), negative and neutral (NEG+NEU), positive and negative (POS+NEG), positive, neutral and negative (POS+NEU+NEG). The 13 categories are shown in Table 4.1.

Lexicon Translation

Wilson and Wiebe's English prior-polarity subjectivity lexicon annotates the semantic orientation (i.e., positive, negative, neutral) of 8221 words [195]. Three English-Chinese lexicon resources were used to translate this English lexicon into Traditional Chinese: Steve Simpson's English-Traditional Chinese dictionary,[1] the ENGLEX English-Simplified Chinese dictionary (automatically converted into Traditional Chinese with some errors), and a messy English-Simplified Chinese dictionary extracted by our colleagues from parallel texts for a machine translation task for the National Institute of Standards and Technology (NIST).

The translated lexicon was manually pruned. The translated lexicon, the NTU's lexicon and the two rekeyed dictionaries were merged and manually put into the 13 categories. The duplicates in each of these sub-lexicons were removed. The 9 mutually exclusive categories were processed to ensure no overlapped words among them. The merged lexicon was further expanded in a way introduced in the next subsection. The number of words in each category is shown in Table 4.1.

Lexicon Expansion

We know that the Chinese language is a character-based language, and lacks the rich morphological variations of words which English has. In a strict sense, the Chinese language lacks the affixes as commonly found in Indo-European languages,

[1] http://www.math.psu.edu/simpson/chinese/ChinText/b5/eng-chi (last visited July 25, 2008).

	NTU	Merged	Expanded
POS	2,816	10,937	26,941
NEG	8,276	17,349	43,848
NEU		3,530	9,175
POS + NEU		929	2,736
NEG + NEU		1,281	3,413
POS + NEG		558	1,373
POS + NEU + NEG		263	863
Operator		761	761
Indicator		847	847
Negation		380	380
Functional negation		191	191
Intensifier		302	302
Quantifier		209	209

Table 4.1: Lexicon size. The merged lexicon is composed of the NTU lexicon, the two rekeyed dictionaries, and the translated lexicon.

and there is no clear distinction between roots and affixes; words are created mainly by compounding [201]. A common way to create compound words is by combining roots and affixes [58] although there is no clear distinction between roots and affixes [201]. We brainstormed some characters and words that function as prefixes and suffixes. Chinese dictionaries usually use these characters and words to explain other words. These characters and words were manually categorized into five rough categories: DE+DI, HAVING, ABILITY, MAKE, and THING. Table 4.2 lists these characters and words and their English glosses.

We expanded the lexicon by adding or removing the characters and words that function as prefixes or suffixes. The expanded words keep their original semantic orientations. Here are the rules for lexicon expansion:

Category	Characters and Words (in Pinyin)	English Gloss
DE+DI	De	adjective suffix (e.g., -ive)
	Di	adverb suffix (e.g., -ly)
HAVING	You3, Ju4	having
	Ke3, Neng2	can, able, capable
ABILITY	Xing4	noun suffix (e.g., -ability)
MAKE	Shi3	verb prefix: make, enable
THING	Ren2, Zhe3, Zhi1 Ren2, Jia1 Huo3	person (of), guy
	Dong1 Xi1, Shi4, Shi4 Wu4, Zhi1 Shi4	thing (of), matter (of)
	Di4 Fang1	place
	Xing2 Wei2, Xing2 Dong4	behavior, action
	Yan2 Xing2	speech and behavior
	Yan2 Ci2	speech and words
	Xiang3 Fa3	idea, thought
	Shou3 Duan4	way, method, means
	Yang4 Zi3	shape, manner
	Gong1 Zuo4	work, task
	Cheng2 Du4	degree, extent
	Dui4 Xiang4	object
	Mu4 Biao1	goal, aim, target

Table 4.2: Chinese characters and words that function as prefixes and suffixes. No efforts were made to create a scientific classification scheme; the categories were created for computational purpose.

(1) if a word ends with ABILITY, or a THING with 1 or 2 characters, remove these suffixes to obtain a word.

(2) if a word begins with MAKE, remove it to obtain a word.

(3) if a word begins with HAVING and ends with DE+DI or ABILITY, remove the first character to obtain a word.

(4) if a word ends with DEDI, exchange De with Di (or vice versa) to create two words, and then remove De and Di to obtain two more words.

(5) if a word does not end with DEDI, add De or Di to obtain 2 words.

The size of the expanded lexicon is also shown in Table 4.1. We observe that the expanded positive (POS), neutral (NEU) and negative (NEG) sub-lexicons are more than twice as big as before expansion.

4.1.2 Classifying Chinese Word Semantic Orientation

Although we have created an expanded lexicon, it still does not include all the opinionated words in the document collection, therefore we need a way to compute the semantic orientation of a word that is not in the lexicon. We have designed two word semantic orientation classifiers for this purpose. For each classifier, we have used the NTU lexicon and the expanded lexicon for training and the NTU lexicon for test.

Word Classifier I: Baseline

Hatzivassiloglou and McKeown [65] defined semantic orientation or polarity as the direction a word deviates from the norm for its semantic group or lexical field. Ku et al. [93] extended this idea to Chinese characters, computing the semantic orientation of a Chinese word as a function of the Chinese characters the word contains. They defined a character's *sentiment tendency* as the ratio of a character's frequency count in positive words to its frequency count in the whole seed lexicon, normalized by the total number of characters in both positive and negative words as follows:

$$POS_{c_i} = \frac{P_{c_i}/\sum_{j=1}^{n} P_{c_j}}{P_{c_i}/\sum_{j=1}^{n} P_{c_j} + N_{c_i}/\sum_{j=1}^{m} N_{c_j}} \tag{4.1}$$

$$NEG_{c_i} = \frac{N_{c_i}/\sum_{j=1}^{m} N_{c_j}}{P_{c_i}/\sum_{j=1}^{n} P_{c_j} + N_{c_i}/\sum_{j=1}^{m} N_{c_j}} \tag{4.2}$$

$$TENDENCY_{c_i} = POS_{c_i} - NEG_{c_i} \tag{4.3}$$

where

- POS_{c_i} and NEG_{c_i} are positive and negative tendency of character c_i respectively,
- P_{c_i} and N_{c_j} are frequency counts of character c_i in positive and negative words respectively,
- n and m denote total number of unique characters in positive and negative

words respectively, and so $\sum_{j=1}^{n} P_{c_j}$ and $\sum_{j=1}^{m} N_{c_j}$ are total number of characters in positive and negative words respectively.

Subtracting NEG_{c_i} from POS_{c_i} yields an estimate of the net attitude $TENDENCY_{c_i}$ expressed by character c_i (which will be positive for positive polarity, and negative for negative polarity). The semantic orientation for a Chinese word is then computed as the average of the estimated net attitude tendency for each character in the word, normalized by the length of the word, so its score is between -1 and 1.

Their approach does not deal with two issues—words with neutral semantic orientation and words with no semantic orientation. We thought a word with neutral semantic orientation could be identified by setting a threshold range around zero for word semantic orientation, but that requires a clean neutral lexicon for training, which is not available to us now. Identifying words with no semantic orientation is a more complex issue which will be discussed in Chapter 5. Next we introduce how we extend this algorithm using negation characters and negation rules.

Word Classifier II: Negation Handling during Training and Classification

In the previous subsection, we have shown that a character's sentiment tendency can be learned from its association with a positive lexicon and a negative lexicon. There are many words in our positive and negative lexicon that contain obvious negation characters (and sometimes negation bigrams). We have collected a negation sub-lexicon as shown in Table 4.1. Figure 4.1 shows some examples of the negation

不 (no)
沒 (no)
無 (no)
未 (not)
不介入 (not involved in)
不受 (not subject to)
不相信 (not believe)
不以為然 (object to)
不再 (not any more)
不復 (not any more)
沒有 (no)
沒那麼 (not that)
從來未 (never ever)
從沒有 (never ever)
無意 (not intend to)
全無 (totally not)

Figure 4.1: Negation characters and bigrams.

characters and bigrams. Removing those characters (or bigrams) often (but not always) yields a word with the opposite polarity. For instance, removing the “Bu4” (not) in “Bu4 Ya3” (not elegant) which has negative semantic orientation, we get “Ya3” (elegant) which has positive semantic orientation. We therefore hand-coded a Perl script which used our negation sub-lexicon to remove words that contain a negation character (or bigram) from its lexicon, stripped off the negation character (or bigrams), and then added the remaining characters back to the lexicon with opposite polarity (if it is not already there). We expect the effect of this processing to be to sharpen the association of individual characters to their sentiment tendency. Using the algorithm of Classifier I and the revised training lexicons, we can build classifier IIa.

In the NTU lexicons, roughly 10 times (and in the expanded lexicons roughly

6 times) as many words with negation characters (or bigrams) were found in the original negative lexicon as in the original positive lexicon. Polarity flipping in the training data alone would enrich the positive examples at the expense of negative examples. This bias can be overcome, however, by detecting negation characters (or bigrams) at classification time, and reversing the sentiment tendency of the characters following the negation character (or bigram). That is, we still used the algorithm of Classifier I, but revised the training data (the positive and negative lexicons) to strengthen the association between characters and their sentiment tendency. When accumulating the word semantic orientation from character sentiment tendency, we accumulated evidence from left to right, reversing the sign of the net polarity for all characters following a negation lexeme, through the end of the word. Characters preceding the negation lexeme were not affected. This classifier is called IIb.

4.1.3 Classifying Chinese Sentence Subjectivity and Polarity

General Framework

We take a shallow linguistic analysis approach for classifying the subjectivity and polarity of a sentence which are aggregated from the subjectivity and semantic orientations of the words in the sentence without considering the syntactic structures governing the words except negation mechanisms. Syntactical analysis for attitude classification is a much more complex issue. Basically we collect the evidences of subjectivity density, positivity, negativity, and aggregated polarity from a sentence,

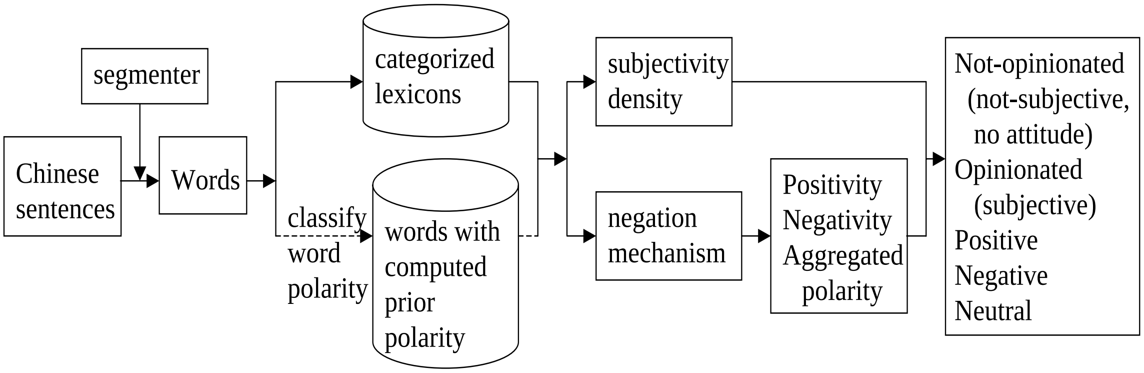


Figure 4.2: General framework for Chinese sentence attitude classification.

then apply heuristic rules to classify sentence subjectivity and polarity. Figure 4.2 describes the general framework of this process.

A sentence is first segmented into words, and its length (i.e., the number of words in the sentence) is calculated. The expanded and categorized lexicon is used to count the number of sentence words found in the lexicon; for the words not found in the lexicon, the word semantic orientation classifier is optionally applied to compute the semantic orientation. A subjectivity density score is computed using the sentence length, the number of words in the lexicon and the number of words optionally computed (with the word semantic orientation classification algorithm) as subjective or having a certain semantic orientation. A negation mechanism is used to accumulate the aggregated polarity of the sentence, and its two components—positivity and negativity. Heuristic rules are then developed to use the evidences of subjectivity density, positivity, negativity, and aggregated polarity to classify sentence subjectivity and polarity. The heuristic rules are described in the experiments in Section 4.4. We will examine:

- the usefulness of the heuristic rules,
- the usefulness of words with a semantic orientation computed with our extended algorithm, and
- the effect of negation mechanisms (introduced below) on the effectiveness of sentence polarity classification.

Negation Mechanisms

Since negation words reverse the semantic orientation of related words, we have studied negation mechanisms. There are two types of negation words: natural negation words (or characters), such as "Bu2" (not) in "Bu2 Huai4" (not bad), and functional negation words, such as "Ting2 Zhi3" (stop) in "Ting2 Zhi3 Fan4 Zui4" (stop crimes). We have designed three negation approaches:

- "1-word adjacency negation": the negation word works on the word immediately following it; that is, if a negation word immediately precedes a subjective word, the polarity of the word is reversed. An example is "Bu2 Huai4" (not bad).
- "2-word adjacency negation": the negation word works on the two words immediately following it; that is, if a negation word is either one or two words immediately preceding a subjective word, the polarity of the word is reversed. An example is "Bu2 Tai4 Huai4" (not too bad).
- "dependency-based negation," which is based on dependency parsing. The

purpose is to capture longer relationship that a negation word involves in. An example is “Bu2 Shi4 Yu1 Chun3 De Cuo4 Wu4” (not a stupid mistake). In this example, what is to be negated is the semantic orientation of “stupid mistake”, not just that of “stupid”.

“In a dependency grammar, one word is the head of a sentence, and all other words are either a dependent of that word, or else dependent on some other word which connects to the headword through a sequence of dependencies” [113](p. 428). To extract the dependency relationship between a negation word and its related words, the Stanford Parser [100] was applied to each sentence that contains a negation word. The parser outputs the governor-dependent relation of two words along with the type of their grammatical relation. A list of these grammatical relation tags is shown in Table 4.3. The goal is to identify a group of words that the negation word works on to the extent that the negation sense is complete. For instance, in the sentence “Ta1 Mei2 Fan4 Zui4” (He did not commit the crime), “not” governs “commit” which governs “crime,” if we extract “not commit,” the negation sense is not complete; what we want to extract is “not commit the crime.”

From the 3343 parsed Chinese sentences which contain negation words, we randomly selected about two dozens for manual analysis. After manually analyzing these parsed Chinese sentences, we deduced the heuristic rules for adding governors and dependents related to a negation word to the dependency relationship:

- if a dependent appears after its governor but not as ccomp (because such a clausal complement relation is usually too far away), this dependent is added;

Tag	Explanation
advmod	adverbial modifier
amod	adjectival modifier
assmod	associative modifier (*)
ccomp	cclausal complement with internal subject
comod	compound modifier (*)
conj	conjunct
mmod	modal modifier (*)
neg	negation modifier
nmod	noun modifier (*)
nsubj	nominal subject
plmod	plural modifier (*)
numod	numeric modifier (*)
rcmod	relative clause modifier

Table 4.3: Grammatical relation tags. The interpretation of the tags with a * is made by the investigator because an official reference is not available from the documentations of the Stanford Parser and Chinese Treebank.

- if the dependent appears before its governor as advmod, amod, nmod, rcmod, comod, numod, mmod, assmod, plmod, neg, or conj, AND the dependent appears after the negation word, this dependent is added;

- if the dependent appears before its governor, but not as ccomp or nsubj, AND the governor appears after the negation word, this governor is added.

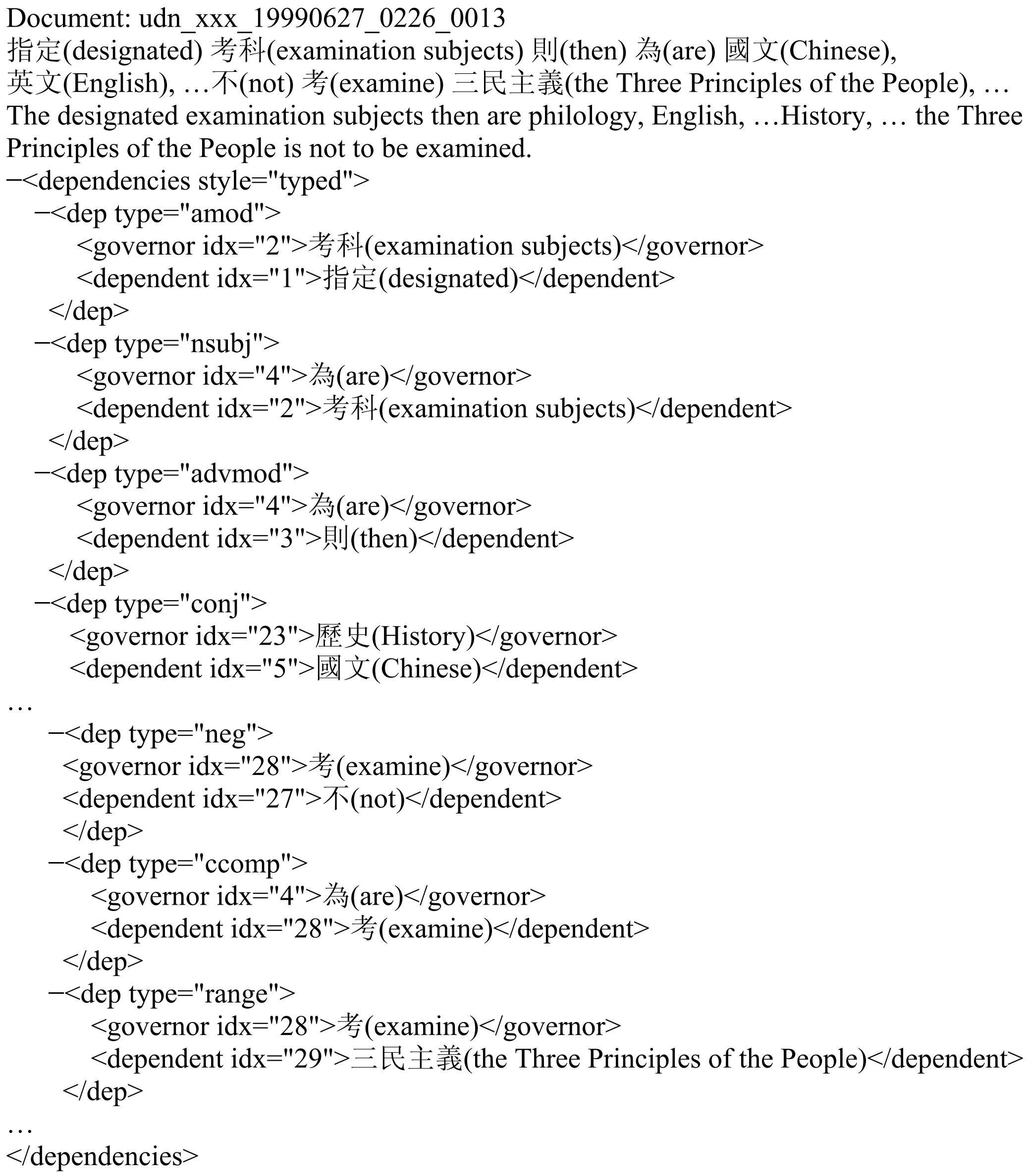
Figure 4.3 shows an example of a parsed Chinese sentence fragment with the negation word "not." The dependency relationship extracted is glossed as "not examine the Three Principles of the People." Figure 4.4 shows an example of a parsed Chinese sentence fragment with the functional negation word "opposes." The dependency relationship extracted is glossed as "opposes to build the 4th nuclear plant."

4.1.4 Evaluation Methods

The Chinese sentence attitude classifiers are evaluated using the NTCIR-6 test collection. Precision, recall, and F_1 measure are used to measure the performance of classifiers. Precision is computed as $\frac{\#SystemCorrect}{\#SystemProposed}$. Recall is computed as $\frac{\#SystemCorrect}{\#OpinionatedSentences}$. F_1 is the balanced harmonic mean of precision and recall, and is defined as $\frac{2PR}{P+R}$.

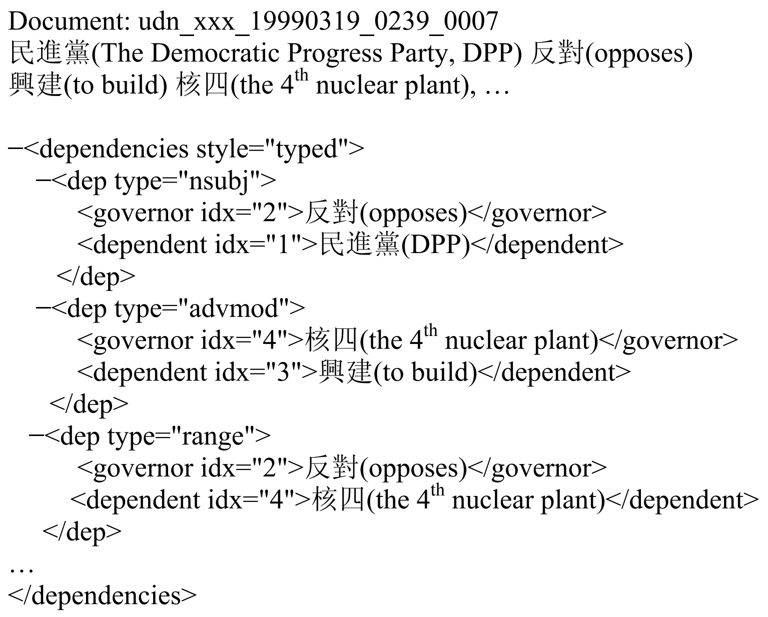
To test the statistical difference between two systems for sentence subjectivity detection and polarity classification, we perform resampling to obtain the standard deviation of precision and recall. The bootstrap resampling method is used. The R statistical program[2] is used to perform bootstrapping, and 2000 samples are taken. The bootstrap distribution of a statistic (i.e., precision and recall), based on many resamples, approximately represents the sampling distribution of the statistic [115]. For precision, resamples are drawn from the sentences proposed by the system as subjective or having a certain polarity. For recall, resamples are drawn from the sentences annotated by the gold standard as subjective and having a certain polarity. The error bar of precision and recall ranges from its mean minus one standard deviation to its mean plus one standard deviation. If the error bars of precision (or recall) of two systems do not touch, or they intersect but neither means of precision (or recall) of two system are inside the error bars of precision (or recall) of each other system, then the precision (or recall) scores of the two systems are statisti-

[2] The R program is available at http://www.r-project.org/ (last visited on October 11, 2008).



```
Document: udn_xxx_19990627_0226_0013
指定(designated) 考科(examination subjects) 則(then) 為(are) 國文(Chinese),
英文(English), …不(not) 考(examine) 三民主義(the Three Principles of the People), …
The designated examination subjects then are philology, English, …History, … the Three
Principles of the People is not to be examined.
−<dependencies style="typed">
  −<dep type="amod">
      <governor idx="2">考科(examination subjects)</governor>
      <dependent idx="1">指定(designated)</dependent>
    </dep>
  −<dep type="nsubj">
      <governor idx="4">為(are)</governor>
      <dependent idx="2">考科(examination subjects)</dependent>
    </dep>
  −<dep type="advmod">
      <governor idx="4">為(are)</governor>
      <dependent idx="3">則(then)</dependent>
    </dep>
  −<dep type="conj">
     <governor idx="23">歷史(History)</governor>
     <dependent idx="5">國文(Chinese)</dependent>
…
    −<dep type="neg">
     <governor idx="28">考(examine)</governor>
     <dependent idx="27">不(not)</dependent>
     </dep>
    −<dep type="ccomp">
       <governor idx="4">為(are)</governor>
       <dependent idx="28">考(examine)</dependent>
     </dep>
    −<dep type="range">
       <governor idx="28">考(examine)</governor>
       <dependent idx="29">三民主義(the Three Principles of the People)</dependent>
     </dep>
…
</dependencies>
```

Figure 4.3: Example of dependency parsing of a sentence fragment with a negation word.



```
Document: udn_xxx_19990319_0239_0007
民進黨(The Democratic Progress Party, DPP) 反對(opposes)
興建(to build) 核四(the 4th nuclear plant), …

−<dependencies style="typed">
   −<dep type="nsubj">
        <governor idx="2">反對(opposes)</governor>
        <dependent idx="1">民進黨(DPP)</dependent>
     </dep>
   −<dep type="advmod">
        <governor idx="4">核四(the 4th nuclear plant)</governor>
        <dependent idx="3">興建(to build)</dependent>
     </dep>
  −<dep type="range">
        <governor idx="2">反對(opposes)</governor>
       <dependent idx="4">核四(the 4th nuclear plant)</dependent>
    </dep>
…
</dependencies>
```

Figure 4.4: Example of dependency parsing of a sentence fragment with a functional negation word.

cally significantly different at 95% confidence; otherwise, they are not statistically significantly different [126]. If both precision and recall of system X are statistically significantly higher than those of system Y, we can infer that the F score of system X is statistically significantly higher than that of system Y. If one of the two statistics (precision and recall) of system X is statistically significantly higher than that of system Y, but the other statistic is statistically significantly lower than or is not statistically significant different from system Y, we are not sure whether their F scores are statistically significantly different.

4.2 Test Collection for Sentence Classification

The National Taiwan University (NTU) created the test collection for opinion analysis in Chinese sentences for NTCIR-6. The collection has 32 topics, 843 documents in Traditional Chinese, 11,907 sentences. When creating the test collection, a pool of seven annotators was recruited to annotate the subjectivity and polarity of the sentences in the documents, with three annotators per topic (and per document). Each annotator judged all the sentences in the documents of a topic if the topic was assigned to her, but a topic could be assigned to three annotators with different combinations. For each topic the inter-annotator agreement between the three annotators was computed. The average agreement (Cohen's kappa score) for each topic ranges from 0.0537 to 0.4065, with an average of 0.2328 [157] which is low, but positive.

NTU created the two gold standards of subjectivity and polarity of each sentence by majority voting. For judging the subjectivity (or opinionatedness), under the *lenient standard*, two of the three annotators must agree on the classification of a sentence, whereas under the *strict standard*, all three annotators must agree [157]. This is the gold standard created by majority voting, so the lenient standard is denoted as *Lenient-M* in this dissertation. We did not use the strict standard (*Strict-M*) in this dissertation due to sparse data.

For judging the polarity, only opinionated sentences are included for evaluation. For the lenient standard (Lenient-M), two of the three annotators must agree that a sentence is opinionated to be included in the evaluation of its polarity. For the strict standard (Strict-M), all the three annotators must agree. The polarity of the sentence, either positive (POS), negative (NEG), or neutral (NEU), is the polarity with the largest number of votes by the annotators. In cases where the polarity of the sentence is ambiguous, POS + NEU is decided as POS, NEG + NEU is decided as NEG, POS + NEG is decided as NEU, and for POS + NEU + NEG, the gold standard is NEU [157].

Among the total of 11,907 sentences, 62% were judged as opinionated according to the lenient standard (25% according to the strict standard) [157]. Four topics (i.e., topic 1, 2, 3, 26) were released to the NTCIR-6 participants for training purpose; 28 topics were used for evaluation [157]. Among the 28 topics, there are 9240 sentences in the *Lenient-M standard*, in which 5453 (or 59.02%) were judged as subjective. Among 5453 subjective sentences, 2470 (26.73% of 9240) were judged as negative, 1209 (13.08%) neutral, and 1744 (19.2% of 9240) positive. In the Strict-M standard, there are 3284 sentences judged with an agreed subjectivity, in which 2048 (or 22.16% of 9240) were judged as subjective. Among the 2048 subjective sentences, 1145 sentences were judged as having an agreed polarity: 613 (6.67% of 9240) were judged as negative, 470 (5.09% of 9240) as positive, and 62 (0.07% of 9240) as neutral. Therefore a lucky random classifier which guesses the most popular category (i.e., subjective and negative) can get a precision of 0.5902 and recall of 1.0 for subjectivity, and a precision of 0.2673 and recall of 0.4530 (2470 out of 5453) for polarity by the Lenient-M standard. By the Strict-M standard, it can get a precision of 0.2216 and recall of 1.0 for subjectivity, and a precision of 0.0666 and recall of 0.2993 (613 out of 2048). Since the test collection in strict standard has a sparse data problem, that is, most sentences are decided as not opinionated, we mainly use the lenient standard for our experiments.

Test collections are usually called gold standards, but sentence attitude annotations are actually the subjective judgements (or opinions) of the annotators, rather than facts. Political and social topics are common in news articles. When judging sentence subjectivity and polarity in news articles, the annotator's political stands on the topics may be also involved in their judgements. For instance, sentence udn_xxx_19990627_0226_0013 (see below) is judged as NEG in the Lenient-M standard, however, the sentence itself can be judged as non-opinionated if no political stands are associated with the Three Principles of the People.

udn_xxx_19990627_0226_0013: The designated examination subjects are Chinese language, English, math for natural science, math for social science, chemistry, physics, biology, geography, and history; the Three Principles of the People is not to be examined, and all the examinees will be grouped according to the current 4-group enrollment plan.

The reason why that sentence is judged as NEG is probably due to the political stands toward the Three principles of the People (Principles of Nationalism, Principles of Democracy, and Principles of People's Livelihood), which were the political creed proposed in 1905 by Sun Yat-sen, founder of the Republic of China. When the Democratic Progress Party which upholds the independence of Taiwan from China came into power, it has been making efforts to remove the concept of China from Taiwan. When the Three principles of the People was no longer examined for high school students, it may have evoked negative political stands among the annotators.

Since the inter-annotator agreement score is low, we naturally question the effectiveness of the test collection for measuring the performance of our systems. Fortunately NTU also released the annotations of each sentence made by the three annotators. This allows us to get three additional standards created by synthetic annotator 1, 2 and 3, denoted as *Lenient-1*, *Lenient-2*, and *Lenient-3* correspondingly. Therefore we can use the four test collections to examine whether the low inter-annotator agreement has any effect on the ranking of our systems.

4.3 Classifying Chinese Word Semantic Orientation Experiments

4.3.1 Experiment Design

We designed two experiments to evaluate the two word semantic orientation classifiers. The first used the NTU lexicon for training and test. In order to assess the word classification effectiveness we adopted a cross-validation strategy. We first randomly partitioned the NTU positive and negative lexicons into 10 positive and 10 negative subsets. We then trained a Word Classifier I on the first 9 partitions and evaluated classification accuracy using the remaining partition. We repeated this process 10 times, each with a different evaluation partition, reporting the average classification accuracy across those 10 runs. We repeated this process for Word Classifier IIa and IIb.

The second experiment used the expanded lexicon for training and the NTU lexicon for test. We thought the expanded lexicon was spoiled during extraction, training, pruning, and manual categorization, and so could not be used as test data.

When preparing the training data, we removed the NTU positive lexicon from the expanded positive lexicon, and the NTU negative lexicon from the expanded negative lexicon; then added the resulted expanded positive lexicon into the 9 partitions of the NTU positive lexicon, and the resulted expanded negative lexicon into the 9 partitions of the NTU negative lexicon. We used the same remaining partitions of the NTU lexicon for test.

4.3.2 Results

Since we do not have an authoritative neutral lexicon to optimize a threshold range for classifying words with neutral semantic orientation, we swept a threshold through 0, ±0.2, ±0.4, ±0.6, and ±0.8. The purpose was to observe how the classifiers behaved with these different thresholds.

Table 4.4 compares word semantic orientation classification results for the two classifiers in the two experiments. Negation handling during training and classification (i.e., classifier IIb) improves F_1 for positive words at every positive threshold over the baseline (i.e. classifier I), and a Wilcoxon matched-pairs signed-ranks test across the 10 folds of cross-validation shows the difference to be statistically significant at threshold of 0 and 0.2 (large improvements are also evident at higher thresholds, but we did not test those for significance). IIb also yields statistically significant improvements for negative words at a threshold of 0, but a rapid drop in recall results in a reduction in F_1 from negation handling for higher negative thresholds.

Experiment	Training	Negation	Positive Threshold					Negative Threshold				
(Classifier)	Lexicon	Handling	0	0.2	0.4	0.6	0.8	0	0.2	0.4	0.6	0.8
1(I)	NTU	No	0.7473	0.7710	0.6602	0.3899	0.1349	0.8926	0.8227	0.6942	0.5006	0.3373
2(IIa)	NTU	No	0.6938	0.7072	0.6734	0.5118	0.2348	0.8500	0.7643	0.6313	0.4292	0.2573
3(IIb)	NTU	Yes	0.8011	**0.8216**	0.7449	0.5210	0.2348	**0.9124**	0.8273	0.6712	0.4340	0.2573
4(I)	Expanded	No	0.7655	0.7806	0.5976	0.3286	0.1180	0.9031	0.8019	0.6461	0.4379	0.2269
5(IIa)	Expanded	No	0.6979	0.7166	0.6522	0.4270	0.1960	0.8562	0.7577	0.6061	0.4045	0.2136
6(IIb)	Expanded	Yes	0.8071	**0.8251**	0.7008	0.4310	0.1962	**0.9161**	0.8123	0.6338	0.4067	0.1946

Table 4.4: Average F_1 of classifying positive and negative words above a threshold.

Negation handling during training (i.e., classifier IIa) performs less well than the baseline on classifying negative words at low thresholds, as might be expected from the disproportional removal of negative words. In classifying positive words at low thresholds (i.e., 0 and 0.2), IIa (P=0.5713, R=0.8835, F_1=0.6938) improves the recall, but hurts the precision over the baseline (P=0.6591, R=0.8638, F_1=0.7473) at the threshold of 0. This indicates that adding training data from the negative lexicon (after simple removal of negation words) has enriched the positive lexicon by increasing the coverage of positive characters, but possibly has also introduced some negative characters. This also suggests that further gain may be available from a more delicate approach of negation handling.

This is a two-way classification task with the average evenly balanced between the two conditions, so random guessing would yield 50% average F_1. All our character-based classifiers do much better than that. Indeed, always guessing positive would yield F_1=0.253, and always guessing negative would yield F_1=0.754, and we substantially outperform both of those simple baselines (at least for low thresholds). To illustrate the results that can be achieved, we obtained precision=0.859, recall=0.795 for a positive classification threshold of 0.2, and precision=0.973, recall=0.865 for a negative classification threshold of zero in experiment 6. Of course, those thresholds were swept for the evaluation set rather than being learned on held out data, so they should be interpreted as suggestive rather than definitive. But the spread between precision and recall and the asymmetry of the optimal classification thresholds do suggest that further gains may be available from alternative classifier designs.

Table 4.4 shows that, when the threshold ranges go higher (e.g., POS $\leq$ 0.4), classifier IIb still works better than classifier I on classifying positive words, but worse than Word Classifier I on classifying negative words. The table also shows that, as the threshold ranges go higher, the effectiveness of the classifiers trained on the expanded lexicon drops faster than that of the classifiers trained on the NTU lexicon. This indicates that the expanded lexicon has made the association between the characters and their lexicons less sharp when the character coverage of the lexicons has been increased. That is, when the lexicon was expanded, characters with positive sentiment tendency went into the negative lexicon, and vice versa.

In the next section, we report the experiments on classifying the attitude in Chinese sentences using the expanded lexicon and word classifier IIb trained on the expanded lexicon with varying threshold.

4.4 Chinese Sentence Attitude Classification Experiments

We treated words as our basic features for Chinese sentence attitude classification, although characters would also have been a reasonable choice. Since there are no word boundaries (other than those that are coincident with punctuation marks and sentence boundaries) in written Chinese, we segmented the sentences using a one-best partition by the Stanford Segmenter [176] to get words. Although segmentation of Chinese sentences is a research problem itself, one-best segmentation techniques offer an obvious starting point for our purpose.

For both subjectivity detection and polarity classification, two baseline systems

(B1 and B2) were designed: B1 is an intuitive lower baseline, and B2 is a higher baseline. Based on the two baseline systems, we added various components or revised various parameters to test whether these revisions make any difference.

4.4.1 Baseline 1 and its Variants

Here we describe Baseline 1 and system variants based on this basic framework. There are three purposes of these systems:

- to establish a basic framework for subjectivity detection and polarity classification;
- to explore the approaches of polarity classification by examining an aggregated polarity score and its two separate components—positivity and negativity.
- to test the effect of the thresholds of positivity and negativity on the performance of polarity classification.

B1: Baseline 1 Algorithm

The goal of Baseline 1 is to establish a simple framework for Chinese sentence subjectivity detection and polarity classification. Basically it uses the expanded lexicon to accumulate subjectivity density information of a sentence and to aggregate polarity with a negation mechanism, then explores a threshold for judging whether the sentence is subjective, positive, negative, neutral, not subjective. It has a straightforward algorithm which takes the following five steps:

(1) Accumulate a subjectivity score from the words of a sentence using the expanded lexicon. For each word in a sentence, accumulate an subjectivity score. That is, if a word is in the lexicon of operator, indicator, POS, NEG, NEU, POS + NEU, NEG + NEU, POS + NEU + NEG, and intensifier, it contributes 1 point to the subjectivity score.

(2) Compute a subjectivity density score by normalizing the accumulated subjectivity score with the length of the sentence (i.e., the number of words in the sentence).

(3) Accumulate a polarity score from the words in a sentence using the "1-word adjacency negation" approach. If a word is in the lexicon of pure POS, and it is not immediately preceded by a negation word, it gets 1 point, otherwise it gets -1 point. If the word is in the lexicon of NEG, and it is not immediately preceded by a negation word, it gets -1 point, otherwise it gets 1 point. If the word is in the POS + NEU lexicon, contextual handling for negation is applied. That is, if the word is immediately preceded by a negation word (or character) and its previous context (i.e., the accumulated polarity of its previous words) is positive, reverse its polarity, that is, it gets -1 point, otherwise it gets 1 point. If the word is in the NEG + NEU lexicon, contextual handling for negation is also applied. That is, if the word is immediately preceded by a negation word (or character) and its previous context is negative, reverse the polarity of this word, that is, it gets 1 point, otherwise it gets -1 point.

(4) Detect subjectivity. If the sentence has at least 1 *opinion operator* or its subjectivity density is ≥ 0.5, it is considered as having a strong subjectivity signal and

so is classified as subjective. We tried several other subjectivity density thresholds (e.g., 0.2, 0.3, 0.4, 0.6, 0.8), and 0.5 was a reasonably good threshold based on the classification performance.

(5) Compute polarity. If a sentence is computed as subjective, its polarity is computed as: POS if the accumulated polarity is ≥ 1, NEG if the accumulated polarity is ≤ -1, and NEU if the accumulated polarity is $= 0$.

Baseline 1 Variants

Social psychology studies on attitude have found that both positive and negative attitudes toward an attitude object can be simultaneously activated [16]. An aggregated polarity score does not reveal the degrees of positivity and negativity in the attitude composition; for instance, it does not distinguish a neutral attitude (i.e., low positivity and low negativity) from an ambivalent attitude (i.e., high positivity and high negativity). Therefore we were motivated to examine positivity and negativity when classifying polarity. The variants of Baseline 1 were designed to examine the separate contributions of positivity and negativity to the polarity of sentence attitude and the effect of the thresholds of positivity and negativity on the polarity of sentence attitude.

For simplicity of writing, we define a sentence with a *strong subjectivity signal* as having at least one opinion operator or its subjectivity density is ≥ 0.5, a sentence with a *weak subjectivity signal* as having no opinion operator and its subjectivity density is < 0.5; we also define a *strongly subjective* sentence as a sentence with a strong subjectivity signal, a *seemingly neutral* sentence as a sentence with a strong subjectivity signal but a zero aggregated polarity, a *weakly subjective* sentence as a sentence with a weak subjectivity signal. The variants and their main features are described below. Figure 4.5 briefly summarizes the basic framework and features of the systems.

- **A1B1** Treat *seemingly neutral* sentences as negative or neutral. When accumulating the polarity score, also accumulate positivity and negativity separately. If the aggregated polarity score is 0, it can be either neutral or ambivalent. Ambivalent sentences are worth further analysis in an attempt to guess whether it tends to be positive or negative. Here we guess it as negative. That is, if the aggregated polarity is 0, but either positivity $\geq$ 2 or negativity $\leq$ -2, it is negative. If the aggregated polarity is 0, and positivity $\leq$ 1 or negativity $\geq$ -1, it is neutral.

- **A2B1** Similar to A1B1, but treat *seemingly neutral* sentences as positive or neutral.

- **A3B1** Use positivity, negativity and aggregated polarity information to judge whether a *weakly subjective* sentence is really positive, negative, neutral, or not subjective at all. If a sentence has a strong subjectivity signal, consult B1; otherwise, if its positivity $\geq$ 2 or negativity $\leq$ -2, then if its aggregated polarity $\geq$ 1, it is positive; if its aggregated polarity $\leq$ -1, it is negative; if its aggregated polarity is = 0, it is neutral. Otherwise (i.e., positivity $<$ 2 and negativity $>$ -2), it is not subjective. Positive, negative, and neutral sentences are classified as subjective.

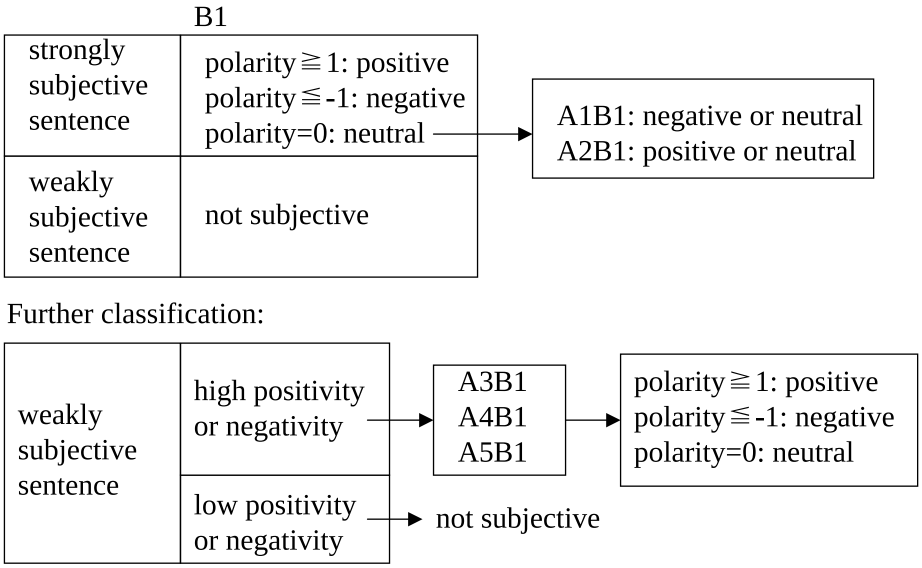
- **A4B1** Deal with *weakly subjective* sentences. Similar to A3B1, except that a higher positivity and negativity threshold is used, that is, positivity $\geq$ 3 or negativity $\leq$ -3.

- **A5B1** Deal with *weakly subjective* sentences. Similar to A3B1, except that an even higher positivity and negativity parameter is used, that is, positivity $\geq$ 4 or negativity $\leq$ -4.

Results for B1 and B1 Variants

Table 4.5 displays the results of B1 and its variants for detecting subjective sentences and their polarity. Since the data in the strict standard (Strict-M) is sparse, that is, 2/3 of sentences that appear in the lenient standard do not appear in the strict standard, we focus on the results under the lenient standard (Lenient-M). In order to examine the effect of using different gold standards on system rankings, the system performance evaluated against the lenient standards created by the three synthetic annotators is also provided.

Table 4.5 shows that the baseline B1 can detect subjective sentences reasonably well because an F score of 0.7683 is almost the same as our baseline for NTCIR-6 which is the highest among the 5 teams in NTCIR-6 [157, 199]. We notice that A3B1, A4B1, and A5B1 improve the F score over B1 for subjectivity detection, indicating that using positivity and negativity information separately helps detect subjective sentences with a weak subjectivity signal. We also notice that A1B1,


B1
strongly subjective sentence
polarity≧1: positive
polarity≦-1: negative
polarity=0: neutral
A1B1: negative or neutral
A2B1: positive or neutral
weakly subjective sentence
not subjective
Further classification:
weakly subjective sentence
high positivity or negativity
A3B1
A4B1
A5B1
polarity≧1: positive
polarity≦-1: negative
polarity=0: neutral
low positivity or negativity
not subjective

Run	Treat	Features
A1B1	seemingly neutral as negative	pos $\geq$ 2 or neg $\leq$ -2; polar = 0
	seemingly neutral as neutral	pos $\leq$ 1 or neg $\geq$ -1; polar = 0
A2B1	seemingly neutral as positive	pos $\geq$ 2 or neg $\leq$ -2; polar = 0
	seemingly neutral as neutral	pos $\leq$ 1 or neg $\geq$ -1; polar = 0
A3B1	weakly subjective as positive	pos $\geq$ 2 or neg $\leq$ -2; polar $\geq$ 1
	weakly subjective as negative	pos $\geq$ 2 or neg $\leq$ -2; polar $\leq$ -1
	weakly subjective as neutral	pos $\geq$ 2 or neg $\leq$ -2; polar = 0
A4B1	weakly subjective as positive	pos $\geq$ 3 or neg $\leq$ -3; polar $\geq$ 1
	weakly subjective as negative	pos $\geq$ 3 or neg $\leq$ -3; polar $\leq$ -1
	weakly subjective as neutral	pos $\geq$ 3 or neg $\leq$ -3; polar = 0
A5B1	weakly subjective as positive	pos $\geq$ 4 or neg $\leq$ -4; polar $\geq$ 1
	weakly subjective as negative	pos $\geq$ 4 or neg $\leq$ -4; polar $\leq$ -1
	weakly subjective as neutral	pos $\geq$ 4 or neg $\leq$ -4; polar = 0

Figure 4.5: Baseline B1 and its variants: framework and features (pos: positivity, neg: negativity, polar: polarity).

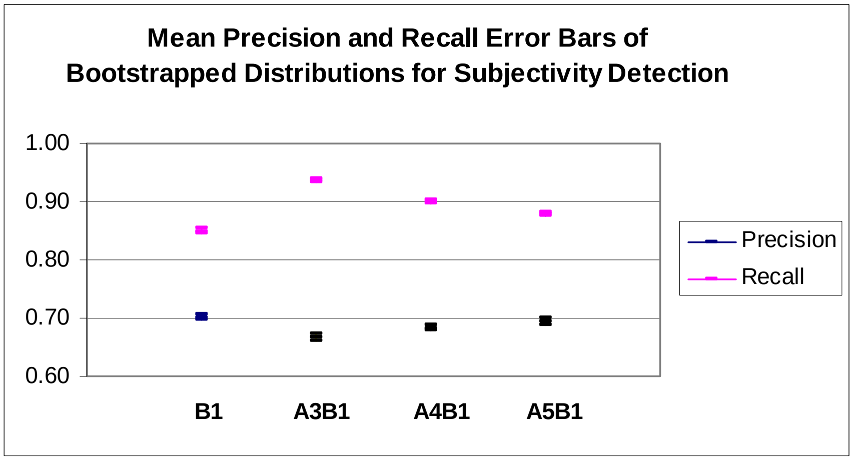
Runs	Lenient-M		Lenient-1		Lenient-2		Lenient-3	
B0	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	0.5902	0.2637						
R	1	0.4530						
F	0.7422	0.3362						
B1	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	0.7012	0.3328	0.6865	0.3572	0.6103	0.3029	0.6351	0.3247
R	0.8496	0.4033	0.7997	0.4161	0.8469	0.4203	0.8273	0.4229
F	0.7683	0.3647	0.7388	0.3844	0.7094	0.3520	0.7186	0.3673
A1B1	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	save	0.3557	0.2836	0.3755	same	0.3175	same	0.3290
R	as	0.4310	0.9150	0.4374	as	0.4407	as	0.4286
F	B1	**0.3897**	0.7388	0.4041	B1	0.3691	B1	0.3723
A2B1	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	same	0.3346	same	0.3752	same	0.3097	same	0.3191
R	as	0.4055	as	0.4371	as	0.4297	as	0.4156
F	B1	**0.3667**	B1	0.4038	B1	0.3600	B1	0.3610
A3B1	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	0.6663	0.3349	0.6731	0.3682	0.5724	0.2997	0.6108	0.3197
R	0.9356	0.4702	0.9087	0.4970	0.9206	0.4820	0.9221	0.4826
F	**0.7783**	**0.3912**	0.7734	0.4230	0.7059	0.3696	0.7349	0.3846
A4B1	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	0.6834	0.3465	0.6804	0.3734	0.5871	0.3070	0.6244	0.3276
R	0.9002	0.4564	0.8616	0.4728	0.8857	0.4631	0.8843	0.4640
F	**0.7769**	**0.3940**	0.7603	0.4173	0.7061	0.3692	0.7319	0.3840
A5B1	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	0.6941	0.3541	0.6841	0.3768	0.5974	0.3119	0.6322	0.3312
R	0.8790	0.4484	0.8329	0.4587	0.8664	0.4524	0.8606	0.4509
F	**0.7757**	**0.3957**	0.7512	0.4138	0.7072	0.3693	0.7289	0.3819

Table 4.5: Results for baseline B1 and its variants (P=Precision, R=Recall, F=F_1 measure. Opin: Opinion; Pola: Polarity. B0 is the random baseline that guesses the popular categories. Bold scores are those higher than baseline B1).

A2B1, A3B1, A4B1 and A5B1 improve the F score of polarity classification over B1, and A1B1 is statistically significantly better than A2B1 (statistical tests are introduced below), indicating that there are positive and negative sentences (more negative sentences than positive ones) among the *seemingly neutral* sentences, and using positivity and negativity information together with the aggregated polarity helps sentence polarity classification.

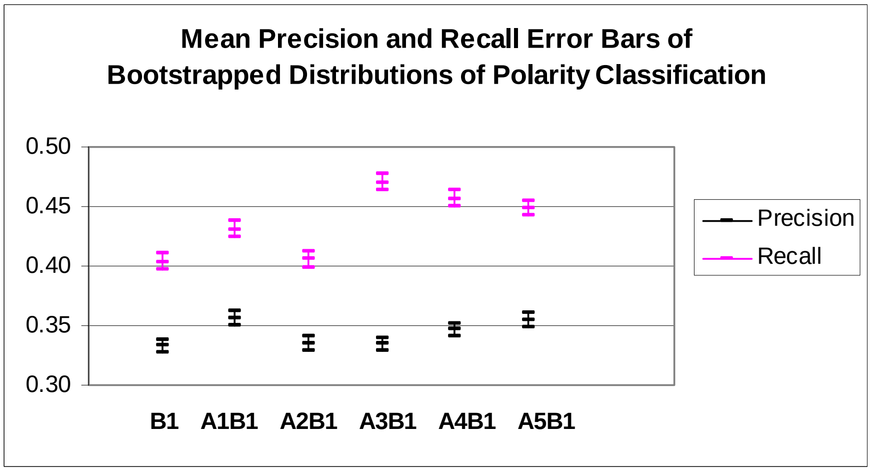
To test whether there is any statistical significance between the systems, we have performed bootstrap resampling on precision and recall data separately. Figure 4.6 displays the error bars of bootstrapped precision and recall statistics of B1 and its variants for sentence subjectivity detection. The precision scores of A3B1, A4B1 and A5B1 are statistically significantly higher than B1 (at $p < 0.05$) but their recall scores are statistically significantly lower than B1.

Figure 4.7 displays the error bars of bootstrapped precision and recall statistics of B1 and its variants for sentence polarity classification. Both precision and recall of A1B1, A3B1 and A5B1 are statistically significantly higher than those of B1, so their F scores are statistically significantly higher than that of B1. A3B1's precision is not statistically significantly higher that of B1, but its recall is statistically significantly higher than that of B1. A2B1 is almost the same as B1, but A1B1 is statistically significantly better than A2B1, indicating that there are significantly more negative sentences than positive sentences among the *seemingly neutral* sentences.



Runs	Precision			Recall		
	mean	std	error bar	mean	std	error bar
B1	0.7012	0.0057	[0.6955, 0.7069]	0.8496	0.0049	[0.8447, 0.8545]
A3B1	0.6663	0.0054	[0.6609, 0.6717]	0.9356	0.0033	[0.9323, 0.9389]
A4B1	0.6834	0.0055	[0.6779, 0.6889]	0.9002	0.0041	[0.8961, 0.9043]
A5B1	0.6941	0.0056	[0.6885, 0.6997]	0.8790	0.0043	[0.8747, 0.8833]

Figure 4.6: Error bars of bootstrapped precision and recall of B1 and its variants for subjectivity detection.



	PRECISION			RECALL		
Runs	mean	std	error bar	mean	std	error bar
B1	0.3328	0.0058	[0.3270, 0.3386]	0.4033	0.0067	[0.3966, 0.4100]
A1B1	0.3557	0.0058	[0.3499, 0.3615]	0.4310	0.0067	[0.4243, 0.4377]
A2B1	0.3346	0.0057	[0.3289, 0.3403]	0.4055	0.0066	[0.3989, 0.4121]
A3B1	0.3349	0.0053	[0.3296, 0.3402]	0.4702	0.0067	[0.4635, 0.4769]
A4B1	0.3465	0.0056	[0.3409, 0.3521]	0.4564	0.0065	[0.4499, 0.4629]
A5B1	0.3541	0.0057	[0.3484, 0.3598]	0.4484	0.0066	[0.4418, 0.4550]

Figure 4.7: Error bars of bootstrapped precision and recall of B1 and it variants for polarity classification.

4.4.2 Baseline 2 and its Variants

Since baseline B1 and its variants work reasonably well on subjectivity detection, but do not work so well on polarity classification—A5B1 has the highest F score on polarity classification which is still slightly lower than CUHK's 0.405, the highest score among the 5 teams in NTCIR-6. So we want to build better systems for polarity classification. Inspired by the previous conclusion that there are positive and negative sentences (more negative sentences than positive ones) among the *seemingly neutral* sentences, we hypothesize that among the sentences with high subjectivity density, the *seemingly neutral* sentences are mostly negative. In order to improve polarity classification, we have also tested more complex negation mechanisms. Here we describe Baseline 2 and its variants which incorporate these ideas.

B2: Baseline 2 Algorithm

We hypothesize that if a sentence has a strong subjectivity signal and a zero aggregated polarity, it is more likely to be negative than to be neutral. Therefore B2 is similar to B1 except the final step for computing polarity: if a sentence has a strong subjectivity signal, its polarity is computed as: positive if the aggregated polarity is ≥ 1, negative if the aggregated polarity is < 1, no neutral sentences are reported for sentences with strong subjectivity signals. That is, the sentences with strong opinion signals that were classified as neutral by B1 are classified as negative by B2.

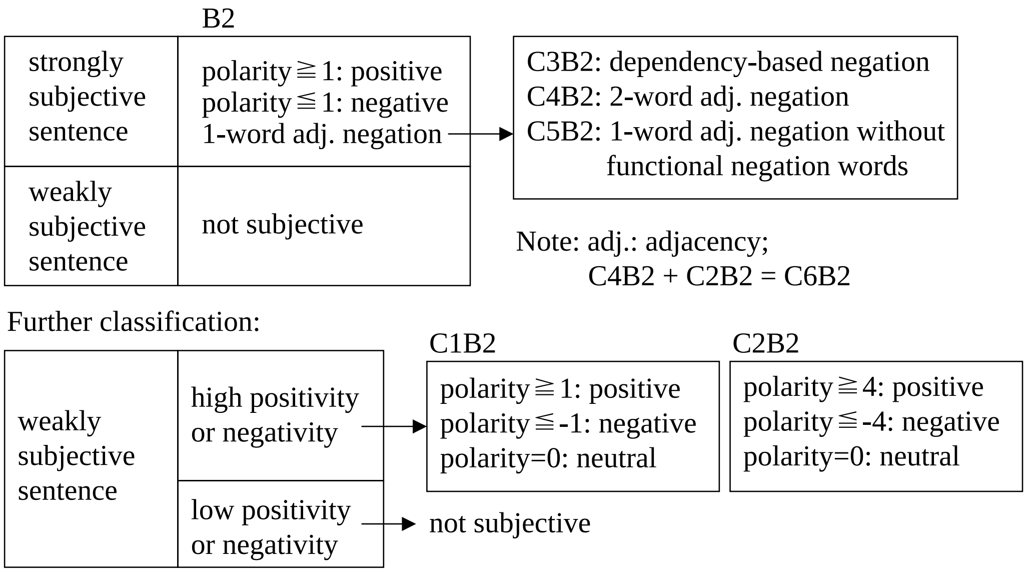
Baseline 2 Variants

Based on B2, we have experimented with more complex negation mechanisms, replacing the simple immediate negation (*1-word adjacency negation*) with longer distance negation (*2-word adjacency negation* and *dependency-based negation*), and repeating the experiments using separate positivity and negativity information together with aggregated polarity for polarity classification. The variants and their main features are described below. Figure 4.8 briefly summarizes the basic framework and features of these systems.

- **C1B2** This is the combination of B2 and A5B1. That is, using B2 for computing the polarity for the sentences with strong opinion signals, and using A5B1 for computing subjectivity and polarity for the sentences with weak subjectivity signals. That is, if a sentence has no opinion operators and has low subjectivity density, but has a high positivity or negativity score (positivity $\geq$4 or negativity $\leq$-4), if its aggregated polarity score is > 1, positive; < -1, negative; or =0, then neutral.

- **C2B2** Similar to C1B2, but with a higher aggregated polarity threshold for detecting subjective sentences from those with weak opinion signals. That is, replace the aggregated polarity thresholds (> 1 for POS, < -1 for NEG, or $= 0$ for NEU) with higher thresholds (> 4 for POS, < -4 for NEG, or $= 0$ for NEU).

- **C3B2** Based on B2, except that the negation approach is based on *dependency-based relationship* rather than the *1-word adjacency negation* approach.



Run	Treat	Features
C1B2	weakly subjective as positive	pos $\geq$ 4 or neg $\leq$ -4; polar $\geq$ 1
	weakly subjective as negative	pos $\geq$ 4 or neg $\leq$ -4; polar $\leq$ -1
	weakly subjective as neutral	pos $\geq$ 4 or neg $\leq$ -4; polar $=$ 0
C2B2	weakly subjective as positive	pos $\geq$ 4 or neg $\leq$ -4; polar $\geq$ 4
	weakly subjective as negative	pos $\geq$ 4 or neg $\leq$ -4; polar $\leq$ -4
	weakly subjective as neutral	pos $\geq$ 4 or neg $\leq$ -4; polar $=$ 0
C3B2	strongly subjective sentences	dependency-based negation
C4B2	strongly subjective sentences	2-word adjacency negation
C5B2	strongly subjective sentences	1-word adjacency negation
		without functional negation words
C6B2	strongly subjective sentences	using C4B2
	weakly subjective sentences	using C2B2

Figure 4.8: Baseline B2 and its variants: framework and features (pos: positivity, neg: negativity, polar: polarity).

- **C4B2** Based on B2, except that the negation approach is *2-word adjacency negation* rather than *1-word adjacency negation*.

- **C5B2** Based on B2, but the functional negation words (mainly verbs) are ignored in the *1-word adjacency negation* approach.

- **C6B2** This is the integration of C2B2 and C4B2, which is the higher polarity threshold plus the *2-word adjacency negation* approach.

Results for B2 and B2 Variants

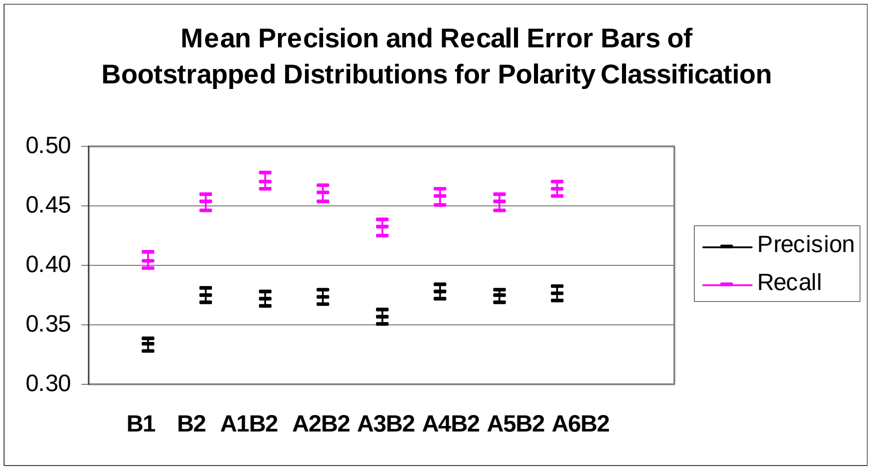
Table 4.6 displays the performance of the CUHK classifier (the best polarity classification system in NTCIR-6) and the results of baseline B2 and its variants. Figure 4.9 shows the bootstrapped precision and recall statistics of B2 and its variants for polarity classification.

Table 4.6 shows that the F scores of both subjectivity detection and polarity classification of B2 are higher than those of the CUHK system [157, 200] in terms of the Lenient-M standard. However, it is unfair to compare our current systems with CUHK because we have used the test collection to tune the parameters (such as 0.5 for the subjectivity density) and CUHK did not get the chance to tune their parameters. The purpose of the comparison here is to show that B2 is a high baseline.

B2's performance on polarity classification is statistically significantly higher than B1 as shown in Figure 4.9. This confirms our earlier hypothesis that if a sentence has a strong subjectivity signal and a zero aggregated polarity, it is more

Runs	Lenient-M		Lenient-1		Lenient-2		Lenient-3	
CUHK	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	0.818	0.522						
R	0.519	0.331						
F	0.635	0.405						
B2	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	0.7012	0.3737	0.6865	0.3864	0.6103	0.3186	0.6351	0.3116
R	0.8496	0.4528	0.7997	0.4501	0.8469	0.4421	0.8273	0.4060
F	0.7683	**0.4095**	0.7388	0.4158	0.7094	0.3703	0.7186	0.3526
C1B2	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	0.6941	0.3713	0.6841	0.3873	0.5974	0.3130	0.6322	0.3146
R	0.8790	0.4702	0.8329	0.4714	0.8664	0.4539	0.8606	0.4282
F	0.7757	**0.4150**	0.7512	0.4252	0.7072	0.3705	0.7289	0.3627
C2B2	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	0.6988	0.3728	0.6868	0.3875	0.6041	0.3153	0.6346	0.3138
R	0.8621	0.4599	0.8145	0.4596	0.8536	0.4455	0.8417	0.4162
F	0.7719	**0.4118**	0.7452	0.4205	0.7075	0.3693	0.7236	0.3578
C3B2	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	same	0.3563	0.6865	0.3666	0.6103	0.2947	0.6357	0.2960
R	as	0.4317	0.7997	0.4271	0.8469	0.4089	0.8273	0.3856
F	B2	0.3904	0.7388	0.3945	0.7094	0.3425	0.7186	0.3350
C4B2	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	same	0.3770	same	0.3876	same	0.3195	same	0.3132
R	as	0.4568	as	0.4515	as	0.4434	as	0.4079
F	B2	**0.4131**	A3B	0.4171	A3B2	0.3714	A3B2	0.3543
C5B2	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	same	0.3734	same	0.3853	same	0.3181	same	0.3113
R	as	0.4524	as	0.4489	as	0.4415	as	0.4056
F	B2	**0.4091**	A3B2	0.4147	A3B2	0.3698	A3B2	0.3523
C6B2	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	0.6985	0.3761	0.6863	0.3886	same	0.3163	0.6343	0.3153
R	0.8617	0.4640	0.8140	0.4609	as	0.4470	0.8413	0.4182
F	0.7716	**0.4154**	0.7447	0.4216	A2B2	0.3705	0.7233	0.3595

Table 4.6: CUHK system and results for baseline 2 and its variants (P=Precision, R=Recall, F=F-measure; Opin: Opinion; Pola: Polarity; Bold scores are those higher than CUHK's).



Runs	Precision			Recall		
	mean	std	error bar	mean	std	error bar
B1	0.3328	0.0058	[0.3270, 0.3386]	0.4033	0.0067	[0.3966, 0.4100]
B2	0.3737	0.0059	[0.3678, 0.3796]	0.4528	0.0069	[0.4459, 0.4597]
C1B2	0.3713	0.0057	[0.3656, 0.3770]	0.4702	0.0065	[0.4637, 0.4767]
C2B2	0.3728	0.0059	[0.3669, 0.3787]	0.4599	0.0067	[0.4532, 0.4666]
C3B2	0.3563	0.0060	[0.3503, 0.3623]	0.4317	0.0068	[0.4249, 0.4385]
C4B2	0.3770	0.0057	[0.3713, 0.3827]	0.4568	0.0067	[0.4501, 0.4635]
C5B2	0.3534	0.0059	[0.3675, 0.3793]	0.4524	0.0067	[0.4457, 0.4591]
C6B2	0.3761	0.0060	[0.3703, 0.3819]	0.4640	0.0065	[0.4575, 0.4705]

Figure 4.9: Error bars of bootstrapped precision and recall of B1, B2 and B2 variants for polarity classification.

likely to be negative than to be neutral in the NTCIR-6 collection.

Both C1B2 and C2B2 improve the polarity classification performance over B2, indicating that positivity and negativity information is useful for polarity classification of sentences with weak subjectivity signals.

Both C4B2 and C6B2 improve the polarity classification performance over B2, indicating that the *2-word adjacency negation* approach works better than the *1-word adjacency negation* approach. However, as shown in Figure 4.9, there is no statistically significant difference between the precision scores of B2, C4B2, and C6B2; there is no statistically significant difference between the recall of B2 and C4B2. But C6B2's recall is statistically significantly higher than B2's. This indicates that *2-word adjacency negation* together with higher polarity thresholds helps sentence polarity classification.

C5B2 is almost the same as B2, indicating that ignoring the functional negation words in the negation mechanism is probably fine.

C3B2's polarity classification performance is statistically significantly lower than B2, indicating that the *dependency-based negation* approach does not work as well as the *1-word adjacency negation* approach. There might be several reasons. One might be that the dependency relationships identified in my way are not all correct; the other might be that the words in the dependency relationship are not in any lexicon. This is worth further testing in the next section, which addresses applying word semantic orientation classification to sentence attitude classification.

4.4.3 Applying Word Semantic Orientation Classification

Now we test whether word semantic orientation classification can be useful for sentence-level subjectivity and polarity classification. Here we use the framework of Baseline B1 because it is a lower baseline; if we cannot make any improvement over a lower baseline by using word classification, that means we have not found a good way applying word semantic orientation classification.

Classifiers Based on B1 and Word Semantic Orientation Classification

When aggregating sentence polarity from word semantic orientation, baseline B1 does not take into account the words that are not in the lexicons. Here we add a word semantic orientation classification component to B1. That is, the polarity of a word is computed if a word does not appear in any of the following lexicon: operator, indicator, intensify, quantity, negation, functional negation, POS, NEG, NEU, POS+NEU, NEG + NEU, POS + NEU + NEG. Recall that the word classification algorithm cannot distinguish a neutral word from a non-opinionated word. For a word that is computed as not positive and not negative, it is safer to consider it as non-opinionated than to consider it as neutral because we want to distinguish between opinionated and non-opinionated sentences and we have already had a relatively big sentiment lexicon. Word semantic orientation classifier IIb trained on the expanded lexicon was used.

There are two variables that affect the aggregated polarity of a sentence — the polarity of the words in the sentence and the negation mechanism. Since the

Run	Negation Mechanism	Word SO Threshold
W0B1	1-word adjacency	>0 POS, <0 NEG
W1B1	1-word adjacency	$\geq$ 0.5 POS, $\leq$-0.5 NEG
W2B1	1-word adjacency	$\geq$0.6 POS, $\leq$-0.6 NEG
W3B1	1-word adjacency	$\geq$0.8 POS, $\leq$-0.8 NEG
W4B1	2-word adjacency	>0 POS, <0 NEG
W5B1	2-word adjacency	$\geq$0.5 POS, $\leq$-0.5 NEG
W6B1	2-word adjacency	$\geq$0.8 POS, $\leq$-0.8 NEG
W7B1	dependency-based	$\geq$0.8 POS, $\leq$-0.8 NEG

Table 4.7: Systems applying word semantic orientation (SO) classification to sentence polarity classification.

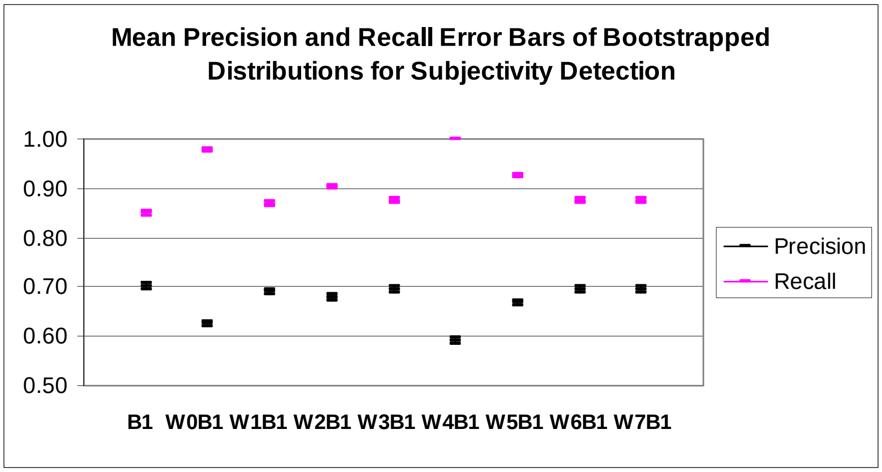
word semantic orientation classification algorithm can compute a polarity score for any word, a word with a computed prior polarity can be either a signal (if it is computed correctly) or a noise (if it is computed incorrectly), so we want to increase the confidence that the semantic orientation of the word is correctly computed by increasing the threshold of positivity and negativity. The negation mechanism has also a direct impact on the aggregated polarity of a sentence. Table 4.7 displays the main features of these two variables for each system.

Result

Table 4.8 displays the performance of classifiers that are based on Baseline B1 and word semantic orientation classification. Figure 4.10 displays error bars of the bootstrapped precision and recall statistics of these systems for sentence subjectivity detection. Figure 4.11 displays the bootstrapped precision and recall statistics of these systems for sentence polarity classification.

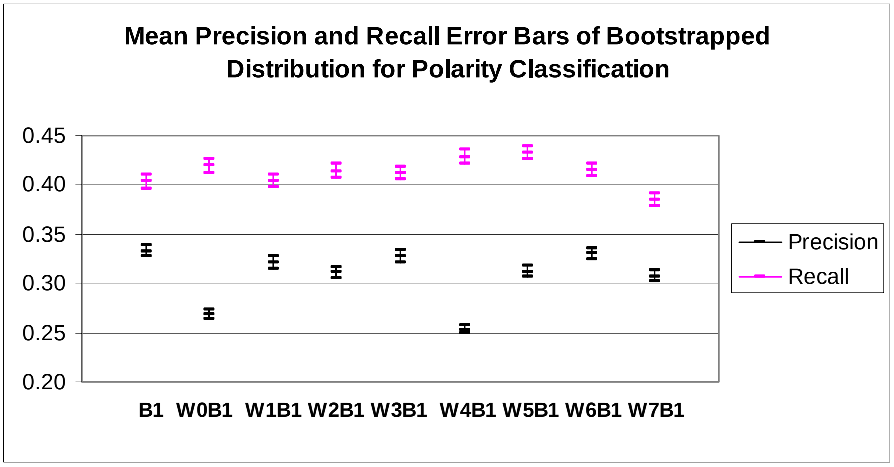
Runs	Lenient-M		Lenient-1		Lenient-2		Lenient-3	
B1	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	0.7012	0.3328	0.6865	0.3572	0.6103	0.3029	0.6351	0.3247
R	0.8496	0.4033	0.7997	0.4161	0.8469	0.4203	0.8273	0.4229
F	0.7683	0.3647	0.7388	0.3844	0.7094	0.3520	0.7186	0.3673
W0B1	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	0.6261	0.2684	0.6409	0.3125	0.5412	0.2482	0.5794	0.2495
R	0.9771	0.4189	0.9616	0.4688	0.9674	0.4436	0.9722	0.4186
F	0.7632	0.3275	0.7691	0.3750	0.6941	0.3183	0.7261	0.3126
W1B1	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	0.6904	0.3210	0.6809	0.3350	0.5995	0.2927	0.6262	0.2938
R	**0.8678**	0.5808	0.8228	0.4048	0.8631	0.4213	0.8462	0.3971
F	**0.7690**	0.1663	0.7452	0.3666	0.7075	0.3454	0.7198	0.3377
W2B1	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	0.6792	0.3113	0.6726	0.3347	0.5892	0.2847	0.6198	0.2953
R	0.9032	0.4139	0.8598	0.4279	0.8973	0.4335	0.8860	0.4221
F	0.7753	0.3553	0.7548	0.3756	0.7113	0.3437	0.7294	0.3475
W3B1	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	0.6953	0.3276	0.6820	0.3519	0.6019	0.2985	0.6308	0.3128
R	**0.8740**	**0.4119**	0.8242	0.4252	0.8666	0.4297	0.8525	0.4227
F	**0.7745**	**0.3650**	0.7464	0.3851	0.7104	0.3523	0.7251	0.3595
W4B1	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	0.5908	0.2532	0.6146	0.2964	0.5157	0.2376	0.5495	0.2359
R	0.9996	0.4284	0.9996	0.4822	0.9993	0.4604	0.9996	0.4290
F	0.7427	0.3183	0.7612	0.3672	0.6803	0.3134	0.7092	0.3044
W5B1	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	0.6673	0.3120	0.6651	0.3260	0.5788	0.2815	0.6121	0.2920
R	0.9257	0.4328	0.8870	0.4348	0.9196	0.4472	0.9129	0.4355
F	0.7756	0.3626	0.7602	0.3726	0.7104	0.3455	0.7328	0.3496
W6B1	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	same	0.3300	same	0.3523	same	0.2973	same	0.3148
R	as	**0.4148**	as	0.4258	as	0.4280	as	0.4254
F	W3B1	**0.3676**	W3B1	0.3856	W3B1	0.3509	W3B1	0.3619
W7B1	Opin	Pola	Opin	Pola	Opin	Pola	Opin	Pola
P	same	0.3066	same	0.3275	same	0.2750	same	0.2920
R	as	0.3855	as	0.3958	as	0.3959	as	0.3947
F	W3B1	0.3416	W3B1	0.3584	W3B1	0.3246	W3B1	0.3357

Table 4.8: Results for classifiers based on B1 and word classification (P=Precision, R=Recall, F=F_1 measure. Opin: Opinion; Pola: Polarity. Bold scores indicate statistically significant difference from B1).



	PRECISION			RECALL		
Runs	mean	std	error bar	mean	std	error bar
B1	0.7012	0.0057	[0.6955, 0.7069]	0.8496	0.0049	[0.8447, 0.8545]
W0B1	0.6261	0.0053	[0.6208, 0.6314]	0.9771	0.0021	[0.9750, 0.9792]
W1B1	0.6904	0.0055	[0.6849, 0.6959]	0.8678	0.0046	[0.8632, 0.8724]
W2B1	0.6792	0.0056	[0.6736, 0.6848]	0.9032	0.0040	[0.8992, 0.9072]
W3B1	0.6953	0.0056	[0.6897, 0.7009]	0.8740	0.0044	[0.8696, 0.8784]
W4B1	0.5908	0.0051	[0.5857, 0.5959]	0.9996	0.0003	[0.9993, 0.9999]
W5B1	0.6673	0.0055	[0.6618, 0.6728]	0.9257	0.0036	[0.9221, 0.9293]
W6B1	0.6953	0.0056	[0.6897, 0.7009]	0.8740	0.0044	[0.8696, 0.8784]
W7B1	0.6953	0.0056	[0.6897, 0.7009]	0.8740	0.0044	[0.8696, 0.8784]

Figure 4.10: Error bars of bootstrapped precision and recall of B1 and systems using word semantic orientation classification for subjectivity detection. W6B1 and W7B1 are the same as W3B1 for subjectivity detection.



Runs	PRECISION			RECALL		
	mean	std	error bar	mean	std	error bar
B1	0.3328	0.0058	[0.3269, 0.3385]	0.4033	0.0067	[0.3966, 0.4100]
W0B1	0.2684	0.0048	[0.2636, 0.2732]	0.4189	0.0066	[0.4123, 0.4255]
W1B1	0.3210	0.0057	[0.3153, 0.3267]	0.4034	0.0066	[0.3968, 0.4100]
W2B1	0.3113	0.0055	[0.3058, 0.3168]	0.4139	0.0066	[0.4073, 0.4205]
W3B1	0.3276	0.0058	[0.3218, 0.3334]	0.4119	0.0066	[0.4053, 0.4185]
W4B1	0.2532	0.0045	[0.2487, 0.2577]	0.4284	0.0065	[0.4219, 0.4349]
W5B1	0.3120	0.0053	[0.3066, 0.3172]	0.4328	0.0067	[0.4261, 0.4395]
W6B1	0.3300	0.0057	[0.3243, 0.3357]	0.4148	0.0065	[0.4083, 0.4213]
W7B1	0.3066	0.0055	[0.3011, 0.3121]	0.3855	0.0063	[0.3792, 0.3918]

Figure 4.11: Error bars of bootstrapped precision and recall of B1 and systems using word semantic orientation classification for polarity classification.

For subjectivity detection, Table 4.8 shows that, except for W0B1 and W4B1 which use a low threshold for word classification, all the other systems work better than B1, indicating that a low threshold for word classification (i.e., >0 for positive words and <0 for negative words) has classified many non-opinionated words as opinionated. Figure 4.10 shows that the precision scores of W1B1 and W3B1 (same as W6B1 and W7B1) are not statistically significantly lower than B1, but their recall scores are statistically significantly higher than B1. This indicates that words computed with higher positivity or negativity scores have more reliable semantic orientations and are useful for sentence subjectivity detection.

For polarity detection, Table 4.8 shows that, only W3B1 and W6B1 work better than B1. Figure 4.11 shows that the precision scores of W3B1 of W6B1 are not statistically significantly lower than B1, but their recall score are statistically significantly higher than B1. There is no statistically significant difference between W3B1 and W6B1 although both precision and recall of W6B1 are higher than those of W3B1. This indicates that words computed with higher positivity or negativity scores have more reliable semantic orientations and are also useful for sentence polarity classification, and the *2-word adjacency negation* approach seems to be slightly better than the *1-word adjacency negation* approach.

W7B1 is statistically significantly worse than B1 and W3B1, indicating that the dependency-based negation does not work well. There might be several reasons. One is that the dependency parser itself can make errors; spot-checks of several parsed sentences did find errors. A second reason is that the dependency relationships extracted for the negation words can be too long or not correct. A third reason is that the dependency-base negation mechanism simply may not work well with the sentence polarity aggregation approach; that is, the dependency relationship per se might be correct, but the polarity aggregation approach does not benefit from the dependency relationship in some cases. For instance, the sentence "the Iraq war policy is bad, although not failed or fatal" is negative based on the *1-word adjacency negation* approach because "not" negates "failed" only, the sentence polarity aggregation algorithm accumulates a polarity score of -1 (i.e., bad: -1, not failed: 1, fatal: -1); but it is positive based on the *dependency-based negation* approach because the dependency relationship to be extracted for "not" is "not failed or fatal", the sentence polarity aggregation algorithm accumulates a polarity score of 1 (i.e., bad: -1, not failed or fatal: 2).

4.5 Test Collection Revisited

In Section 4.2, we have addressed that the test collection has low inter-annotator agreement on sentence attitude annotations. We know that in information retrieval, inter-annotator agreement on relevance judgement is low, but a test collection with low agreement can still be useful as long as the low agreement does not affect system rankings. Similarly, the judgement of sentence subjectivity and polarity is itself a subjective task, and it is natural that the annotators may disagree on some of the judgements. Therefore what we care about is whether the low agreement on the attitude annotations has an effect on the rankings of the systems that perform these two tasks.

Standards	Lenient-1	Lenient-2	Lenient-3
Lenient-M	0.348	0.094	0.758
Lenient-1		-0.391	0.530
Lenient-2			0.033

Table 4.9: Kendall's τ correlation of 12 systems for sentence subjectivity detection.

We have evaluated all our systems for subjectivity detection and polarity classification against the gold standard (Lenient-M) and the three synthetic annotator's annotations (i.e., Lenient-1, Lenient-2, Lenient-3) as shown in Table 4.5, Table 4.6 and Table 4.8. We have computed Kendall's τ (pronounced as tau) correlations of the ranks of all the systems using these four standards. Kendall's τ ranges from -1 to 1, with 1 for two sets of exactly same ranks and -1 for two sets of exactly reverse ranks. For either subjectivity detection or polarity classification, exactly same systems (such as A1B1 and A2B1 for subjectivity detection) are removed from the rankings. Since Kendall's τ does not handle tied rankings, when two or more different systems with a same F score (i.e., difference too small to be detected by the systems) are encountered, a permutation of the tied ranks is performed, and an average τ is taken across the permutated ranks. Note that the τ score computed with permutation of tied ranks is lower than without permutation.

Table 4.9 shows that the Kendall's τ correlation for subjectivity detection ranges from -0.391 to 0.758, and Table 4.10 shows that the τ correlation for polarity classification ranges from 0.472 to 0.819. This indicates that the four standards do make a difference for system rankings.

For subjectivity detection, Table 4.9 shows that Lenient-2 has a very low or

Standards	Lenient-1	Lenient-2	Lenient-3
Lenient-M	0.781	0.819	0.424
Lenient-1		0.724	0.472
Lenient-2			0.472

Table 4.10: Kendall's τ correlation of 21 systems for sentence polarity classification.

even negative τ correlation with Lenient-M, Lenient-1 and Lenient-3, therefore it is the least useful standard. Lenient-3 has the highest correlation with Lenient-M. For polarity classification, Table 4.10 shows that Lenient-3 has a positive but low τ correlation with Lenient-M, Lenient-1 and Lenient-2, therefore it is the worst standard. Lenient-2 has the highest correlation with Lenient-M. Since a sentence's subjectivity and polarity are annotated by a same annotator, we cannot use Lenient-3 for subjectivity detection and Lenient-2 for polarity classification. Therefore, we can use either Lenient-M or Lenient-1 for the both tasks.

When systems are ranked by different standards, it is natural that the rankings will be different to some extent, but we expect the important rankings, such as those of a baseline and its alternative systems, should remain. A simple test was done to validate this hypothesis. We have three groups of systems: B1 series in Table 4.5, B2 series in Table 4.6, and W series which apply word semantic orientation classification in Table 4.8. From each group, we took the baseline (or the worst system in W series), the best system, and a system between them by the Lenient-M standard for polarity classification. Table 4.11 displays the rankings of the systems by the two standards in ascending order of subjectivity detection and polarity classification effectiveness. Same systems or systems with same effectiveness were removed from

Subjectivity Detection		Polarity Classification	
Lenient-M	Lenient-1	Lenient-M	Lenient-1
B1	B1	W7B1	W7B1
W1B1	C6B2	W1B1	W1B1
C6B2	W1B1	B1	B1
W7B1	W7B1	W6B1	W6B1
A5B1	A5B1	A3B1	A5B1
A3B1	A3B1	A5B1	B2
		B2	C2B2
		C2B2	C6B2
		C6B2	A3B1

Table 4.11: System rankings using two standards.

ranking for subjectivity detection. It shows that the important orderings are not changes, for instance, W7B1 are worse than W1B1 and W6B1 (W7B1 $<$ W1B1 $<$ W6B1), A3B1 and A5B1 are better than B1 (A3B1 and A5B1 $>$ B1), C2B2 and C6B2 are better than B2 (C6B2 $>$ C2B2 $>$ B2).

4.6 Discussion

We did reasonably well on sentence subjectivity detection; our baseline (B1) achieved an F_1 of 0.7683 (P=0.7012, R=0.8496). We made much effort trying to improve the effectiveness of sentence polarity classification. Although we created classifiers that worked only slightly better than CUHK's, our systems are still bad at sentence polarity classification; our best system (C6B2) still achieved a low F_1 of 0.4154 (P=0.3761, R=0.4640).

The reason we are still poor at sentence polarity classification is mainly that our simple approach of aggregating sentence polarity from word polarity only partially modeled reality. For instance, if a sentence has more positive words than negative words, our simple aggregation approach will predict the sentence as positive. However, in the real world, sentence polarity may not be generated that way. Furthermore, our approach does not model sarcasm, irony, and attitude that must be judged based on the context outside a single sentence.

We may ask the question: which is more important, subjectivity detection or polarity classification? Subjectivity and polarity may have different importance to the user, depending on the user's needs. If the user wants to automatically assemble opinionated sentences and then manually assess the attitude the individual sentences express, subjectivity detection alone may be adequate. In this case, subjectivity detection serves a first filter to enhance the user's productivity.

Subjectivity detection is also important for distinguishing ambivalence from neutrality. A single aggregated polarity score would not distinguish ambivalence from neutrality, so we need to examine the degrees of positivity and negativity in the attitude composition. By definition, ambivalence is both highly positive and highly negative, and neutrality is both weakly positive and weakly negative. Therefore, strong subjectivity and zero polarity implies ambivalence, whereas weak subjectivity and zero polarity implies neutrality.

For sentence subjectivity and polarity classification, we have evaluated and optimized our classifiers with F_1. Precision and recall might have different importance to the two tasks. To subjectivity detection, recall is probably more important than precision, because if missing a subjective sentence, we will not be detecting its polarity at all. Our subjectivity classifiers consistently achieve higher recall than precision. However, this does not mean we do not care about precision because a false-alarm subjective sentence may confuse an ambivalence classifier. For polarity classification, precision is probably more important than recall in statistical aggregation applications, because missing some sentences with a certain polarity can still lead to reasonable aggregate estimates. Our polarity classifiers consistently achieve lower precision than recall, however. Therefore, we are more satisfied with subjectivity classification than polarity classification.

4.7 Implications for Future Work

People express opinions and attitudes all the time, some in recorded discourses, some not. In specific domains and contexts, the people may be voters, customers, employers, collaborators, competitors, etc. When a large number of people express their attitudes and make them electronically available, the demand for automatic detection of the attitudes emerges. That is one reason why automatic attitude mining has been an active research topic in the past decade.

Sentiment lexicons and word semantic orientation classification are the starting points for automatic attitude analysis in sentences, passages and documents. This study has demonstrated useful approaches for performing these two tasks. Sentence-level attitude analysis can be useful for passage-level attitude analysis which is one of our ultimate goals.

This study has also found many problems with current techniques, and proposed new research questions and design issues, such as investigating the way annotators judge sentence attitude. Finding solutions to these problems and issues will advance the state of art of automatic attitude classification.

We have manually built decision trees for sentence attitude classification so that we can get insights into the problem. In the future we may apply machine learning techniques such as C4.5 to automatically build decision trees.

The current test collection does not distinguish ambivalence from neutrality. We will suggest to NTCIR that they include ambivalence in their opinion analysis test collection in the future.

Sentence-level attitude analysis may also have implications for other information processing and summarization tasks that require analysis of a person's attitude. For instance, argumentation or debate systems may need to summarize the attitudes behind a person's arguments. In other words, attitude classification can be a standalone application and can also be a component of other systems that summarize attitudes.

4.8 Summary

In this chapter, we have addressed the methods for exploring the research questions for Chinese sentence subjectivity detection and polarity classification, experiment design, experiment results, findings, and discussion.

We have created a relatively large Chinese sentiment lexicon by collecting an available Chinese sentiment lexicon (NTU's lexicon), rekeying two Chinese sentiment dictionary books, extracting annotated words from NTCIR-6 training documents, translating an English sentiment lexicon into Traditional Chinese, and expanding the combined Chinese sentiment lexicon with functional prefixes and suffixes. Ultimately the resulting Chinese sentiment lexicon is manually classified into nine sub-lexicons: opinion operators, opinion indicators, intensifiers (mainly adverbs), quantifiers (mainly subjective quantity words), negation words, functional negation words, positive words, negative words, and neutral words. We can share with the community most of the resources that are not constrained by copyright.

We have extended Ku et al.'s algorithm for character-based classification for the semantic orientation of individual Chinese words by (1) automatically enhancing the training lexicon by stripping negation characters and flipping the polarity of the resulting term, and (2) augmenting the classifier with polarity-flipping when negation characters are encountered. A statistically significant improvement on the accuracy has been achieved over Ku et al.'s approach (when a relatively tight threshold range of (-0.2, 0.2) is used to classify neutral words).

We have built Chinese sentence subjectivity detection and polarity classification systems better than CUHK's, which was the best system for NTICR-6. The design of our sentence attitude classifiers has been guided by the social psychology theory that both positive and negative attitudes can be activated toward an attitude object simultaneously. The classifiers use a lexicon, a negation approach, the word semantic orientation classification module, and the following linguistic and quantitative features: subjectivity density, aggregated polarity, positivity, and negativity.

We have investigated three negation handling approaches:

- *1-word adjacency negation*: the negation is applied to the word immediately following it;
- *2-word adjacency negation*: the negation is applied to the two words immediately following it;
- dependency-based negation: the dependency relationship between a negation word and its related words in a sentence is applied based on dependency parsing.

We have manually built decision trees for both sentence subjectivity detection and polarity classification. A sentence is first segmented into words, and its length (i.e., the number of words in the sentence) is calculated. The expanded and categorized lexicon is used to count the number of sentence words found in the lexicon; for the words not found in the lexicon, the word semantic orientation classifier is optionally applied to compute the semantic orientation. A subjectivity density score is computed using the sentence length, the number of words in the lexicon, and the number of words optionally computed (with the word semantic orientation classification algorithm) as subjective or having a certain semantic orientation. If a sentence has a high subjectivity density score or has at least one opinion operator, the sentence is classified as strongly subjective; otherwise, its positivity and negativity are further examined to classify its subjectivity.

The polarity of a sentence is aggregated from the polarity of words it contains. The expanded lexicon, a negation mechanism, and optionally the word semantic orientation classifier are used to accumulate the aggregated polarity of the sentence.

A subjective sentence is first detected as having a strong or weak subjectivity signal before polarity classification is performed. Heuristic rules based on subjectivity, aggregated polarity, positivity, and negativity are used to classify the polarity of the sentence.

Our experiments on sentence attitude classification have found that:

1. words with higher computed prior-polarity scores are more reliable indicators of sentence subjectivity and polarity. Words with low computed prior-polarity scores can be neutral or non-opinionated, and thus may be assigned a wrong polarity. Our current algorithm for word semantic orientation classification is useful but not perfect.

2. the *2-word adjacency negation* approach works better than *1-word adjacency negation* approach, which is consistent with our intuition for the Chinese language. *Dependency-based negation* does not work as well as *1-word adjacency negation* perhaps because our current approach to dependency relationship extraction is not sufficiently reliable.

3. positivity and negativity are useful features for sentence attitude classification, especially when the sentence has a weak subjectivity signal.

4. A *seemingly neutral* sentence which has a strong subjectivity signal but a zero aggregated polarity is more likely to be negative than to be actually neutral in the NTCIR-6 collection.

5. It is difficult to classify the polarity of a sentence with a weak subjectivity

signal. Presently we do not have a good means to deal with this.

Since the test collection has low inter-annotator agreement on sentence attitude annotations, we were concerned whether the low agreement had an effect on the rankings of the systems that perform these the tasks (i.e., subjectivity detection and polarity classification). Since each sentence was annotated by three annotators, we had three synthetic annotator's annotations (i.e., Lenient-1, Lenient-2, Lenient-3). We have evaluated all our systems for subjectivity detection and polarity classification against the gold standard (Lenient-M) and the additional three standards. We found that Lenient-M and Lenient-1 yielded sufficiently stable system rankings to be useful as a basis for comparing alternative classifier designs.

We have manually built decision trees for sentence subjectivity and polarity classification, which allow us get insights to the problems as described above. Understanding of these problems will help build classifiers using machine learning approaches.

We did reasonably well on sentence subjectivity classification but still poorly on sentence polarity classification. We wish to understand how the NTCIR annotators judge sentence subjectivity and polarity so that we can design alternative approaches. More generally we wish to understand how human beings judge the attitude of a sentence, a document segment, and a document. We also plan to encourage NTCIR to distinguish ambivalence from neutrality in the test collection in the future.

Chapter 5

Conclusion

Our ultimate goal is to build a system that helps users compare attitudes toward the same aspects of a topic across languages. Such a system will need to:

- search the blogosphere and newswire collections in Chinese and English,
- identify text segments on specific aspects of the topic,
- detect the polarity and ambivalence of attitudes toward those aspects in each language, and
- align the aspects across languages for the purpose of comparing the attitudes, which requires a graphical user interface.

That challenge problem can guide the development of specific technologies for aspect and attitude classification. We have studied two problems derived from that goal — bilingual segment aspect classification and Chinese sentence attitude classification. In this chapter, we first summarize major approaches to the two problems, our findings and contributions; then address the limitations of our study, and future work towards realizing our ultimate goal.

5.1 Findings and Discussion

5.1.1 Bilingual Aspect Classification

In monolingual text classification, both training and test data are in the same language. Cross-language text classification arises when training data are not in the same language, but rather in some other language. There have been a few studies on cross-language document classification since 1999. The typical finding of those studies is that the effectiveness of cross-language document classification is worse than that of monolingual document classification.

Our study has focused on using training data in two languages (i.e., English and Chinese) at the same time for aspect classification, and in cases where relatively little training data is available. Assuming document segments are appropriate as aspect instances, we have investigated the problem of whether supporting-language training examples can be useful for main-language aspect classification and found the answer to be yes. In other words, our bilingual aspect classification approach which applies both main-language and supporting-language training examples improves classification effectiveness over our monolingual aspect classification approach which applies main-language training examples only. Here when one language (e.g., English) is used as main-language, the other language (e.g., Chinese) is used as supporting-language.

Our general approach to the problem is as follows. For each topic that is of interest to a user, we first retrieved a set of relevant documents for each language. For each aspect of interest to the user when browsing the two sets of retrieved documents, a few document segments were selected from both languages by the user as training examples. The training examples in one language were translated into the other language and served as additional training examples. Then the system classified the remaining aspect instances into the aspect specified by the user.

In our experiments, we first automatically split a document into segments using a variant of TextTiling. These document segments served as aspect instances. Then for each topic, we retrieved a set of relevant document segments, which were used to perform local Latent Semantic Analysis (LSA) to reduce the dimensionality of the term space. Translation probabilities from statistical machine translation were used to translate the training example vectors from one language into the other. The translated supporting-language training example vectors were folded into the main-language local LSA space. Three variants of k-Nearest-Neighbor (kNN) classifier were used to classify remaining document segments to the aspects in the main-language LSA space.

To evaluate our classifiers, we designed an aspect classification test collection in which two annotators determined consecutive sentences in a document and annotated them as aspect instances for 176 aspects from 50 topics in English and Chinese using the TDT3 and TDT4 news collections. We processed a document into document segments as aspect instances using a variant of TextTiling, and the document segments were then mapped onto the annotated document segments in the raw gold standard. Two test collections were created for our experiments. Inter-annotator agreement was calculated using automatically generated document segments. The average Cohen's kappa was 0.22 for English aspects and 0.50 for Chinese aspects, indicating that the annotators did a better job annotating document segments for Chinese aspects than for English aspects, perhaps because their native language was Chinese.

Through our experiments, we were able to conclude that, when only a few main-language training examples were available, a few additional supporting-language training examples would also be useful. That is, when each language is, in turn, the main language, the user must denote some segments in both languages. In this case, a no-cost improvement can be achieved by using the training examples from the supporting-language. But when a fixed number of training examples was used, the classification performance was generally correlated with the percentage of main-language training examples. This implies that supporting-language training examples should generally be used as supplements to, rather than substitutes for, main-language training examples.

We also found that, for monolingual aspect classification, more training examples generally yielded better classification effectiveness; for bilingual aspect classification, a point of diminishing returns occurs after a few foreign-language training examples have been added. This suggests that the usefulness of supporting-language training examples is constrained by translation errors.

When projecting the translated supporting-language training examples into the main-language local LSA space, we also tested the idea of the idea of moving the translated training examples toward the main-language training examples in the LSA space until their centroids meet. It does not work well in our current experiment, probably because the centroid correction is overfit to the limited training data.

Our results suggest that supervised classification can benefit from hand-annotating a few same-language examples, and that when performing classification in bilingual collections it is useful to label some examples in each language.

5.1.2 Chinese Attitude Classification

There has been active research on English subjectivity analysis and polarity classification for more than a decade at the scale of words, sentences, and entire documents. Work on Chinese attitude classification, by contrast, is considerably more recent. So we elected to work on Chinese attitude classification in this study. Ideally we should have worked on Chinese attitude classification at the scale of document segments; however, the best presently available test collection for Chinese attitude classification is focused on sentence-level classification, and it seems reasonable to expect that sentence-level attitude can be aggregated into segment-level attitude as a system implementation approach, so our study is focused on Chinese sentence attitude classification.

Regardless of the scale of attitude classification, words typically have provided the base feature set that attitude classifiers exploited, so we started by creating a relatively large Chinese sentiment lexicon by leveraging existing Chinese and English lexical resources and expanding the lexicon using affixes. The lexicon was manually categorized into 9 categories. The resulted positive and negative lexicon enlarged the National Taiwan University's original positive and negative lexicon by roughly

4 and 9 times, respectively, thus substantially increasing the coverage of our lexical categories, and introducing some degree of robustness against inconsequential (and equally valid) differences in Chinese word segmentation.

Turney and Littman [178] and Yu and Hatzivassiloglou [207] have estimated the semantic orientation of an unknown English word by looking outside the unknown word and examining the strength of its association with a set of known positive and negative words. Ku et al. [93] at National Taiwan University adopted a complementary approach, looking inside, rather than outside, an unknown Chinese word. They first estimated the sentiment polarity of individual Chinese characters, using a set of positive and negative Chinese words as training data, and then inferred the polarity of an unknown word from those character polarities. We extended that model to account for characters that indicate negation, both during training and during classification. This results in small (7% and 3%, respectively) but statistically significant relative F_1 improvements when classifying positive words (at threshold 0.2) and negative words (at threshold 0) in cross-validation experiments, both with their 11,092-word Chinese lexicon and our expanded 80,789-word Chinese lexicon.

When estimating the word semantic orientation, a higher threshold indicates higher confidence of the semantic orientation being estimated. In an extrinsic evaluation, we applied this module to sentence subjectivity and polarity classification; we found that only the words with higher computed prior-polarity scores are useful. Words with low computed prior-polarity scores may be neutral or non-opinionated, or they may simply be assigned the wrong polarity. This suggests that the current algorithm is useful but not perfect. In particular, it cannot distinguish neutral (but subjective) words from non-subjective words.

We have built Chinese sentence subjectivity detection and polarity classification systems better than CUHK's, which was the best system for NTICR-6. The design of our sentence attitude classifiers was guided by the social psychology theory of attitude that both positive and negative attitudes can be activated toward an attitude object. We took a shallow linguistic analysis approach to sentence subjectivity and polarity classification. The classifiers used a lexicon, a negation approach, the word semantic orientation classification module, and the following linguistic and quantitative features: subjectivity density, aggregated polarity, positivity, and negativity.

We manually built decision trees for both sentence subjectivity detection and polarity classification. The expanded lexicon, and optionally the word semantic orientation classifier, the length of a sentence were used to calculate the subjectivity score for the sentence. If a sentence had a high subjectivity density score or had at least one opinion operator, the sentence was classified as strongly subjective; otherwise, its positivity score and negativity score were further examined to classify its subjectivity.

The polarity of a sentence was aggregated from the polarity of words it contains. The expanded lexicon, a negation mechanism, and optionally the word semantic orientation classifier were used to accumulate the aggregated polarity of the sentence. A subjective sentence was first detected as having a strong or weak subjectivity signal before polarity classification was performed. Heuristic rules based on subjectivity, aggregated polarity score, positivity score, and negativity score were used to classify the polarity of the sentence.

We have investigated three negation handling approaches:

- *1-word adjacency negation*: the negation is applied to the word immediately following it;
- *2-word adjacency negation*: the negation is applied to the two words immediately following it;
- dependency-based negation: the dependency relationship between a negation word and its related words in a sentence is applied based on dependency parsing.

Our sentence attitude classification experiments found that, the *2-word adjacency negation* approach works better than *1-word adjacency negation* approach, which is consistent with our intuition about the Chinese language. *Dependency-based negation* does not work as well as the *1-word adjacency negation* approach, probably because of limitations in dependency relationship extraction.

Since the test collection has relatively low inter-annotator agreement on sentence attitude annotations, we were concerned about whether the low agreement on the attitude annotations would have an effect on the preference order between systems that perform these the tasks (i.e., subjectivity detection and polarity classification). Since each sentence was annotated by three annotators, we formed three synthetic annotators (i.e., Lenient-1, Lenient-2, Lenient-3). We have evaluated all our systems for subjectivity detection and polarity classification against the gold standard (Lenient-M) and these additional three standards. We found that Lenient-M and

Lenient-1 yielded sufficiently stable system rankings to be useful as a basis for comparing alternative classifier designs.

We did reasonably well on sentence subjectivity classification, but still not as well as we would have liked on sentence polarity classification. The lower effectiveness of sentence polarity classification likely results from two factors:

1. the shallow linguistic analysis approach, which aggregates sentence polarity from words without analyzing the syntactic structure of the sentence, and
2. a failure to more faithfully replicate the ways in which humans determine the polarity of a sentence from linguistic units (such as words, phrases, and clauses) and from extra-linguistic cues (such as irony and sarcasm) that draw on world knowledge.

Subjectivity and polarity may have different importance to the user, depending on the user's needs. If the user intends to manually check the polarity of individual sentences, subjectivity detection may be more important. In this case, subjectivity detection serves a first filter to enhance the user's productivity. On the other hand, if the user wants the aggregated polarity of a set of sentences, reasonably good polarity classification is essential.

Subjectivity detection is also important for distinguishing ambivalence from neutrality. An aggregated polarity score does not distinguish ambivalence from neutrality, so we need to examine the degrees of positivity and negativity separately. By definition, ambivalence is both highly positive and highly negative, and neutrality is both weakly positive and weakly negative. Therefore, strong subjectivity and zero polarity implies ambivalence, whereas weak subjectivity and zero polarity implies neutrality.

Since we did reasonably well on subjectivity detection and still poorly on polarity classification, when designing an interactive system, we could present the subjectivity and polarity of each sentence, and the aggregated polarity of a set of sentences, but we would warn the user that the polarity may be not reliable.

5.2 Contributions

The contributions of the study of bilingual aspect classification are threefold.

- First, we have proposed bilingual aspect classification, which integrates monolingual and cross-language text classification techniques, as a new research challenge.
- Second, we have proposed and tested successful approaches for bilingual aspect classification, and so demonstrated that the task is possible.
- Third, we have designed a test collection for this task that can be shared with the research community.

English attitude classification for news, blogs, and online product reviews has been an active research area in natural language processing for the past decade. Chinese attitude classification is a relatively new research area. Earlier work on English attitude classification provides a rich array of techniques for possible application to Chinese, some of which have been investigated in this dissertation. We have made the following major contributions to Chinese sentence attitude classification:

- Verified that the Lenient-M and Lenient-1 test collection yield sufficiently stable system rankings to be useful as a basis for comparing alternative classifier designs.
- Extended an existing character-based approach for classifying the semantic orientation of a Chinese word, and verified that the results with high classification confidence are useful for sentence subjectivity detection and polarity classification.
- Created a relatively large Chinese sentiment lexicon by collecting existing Chinese and English sentiment lexicons, automatically translating the English sentiment lexicon, manually pruning the translated lexicon, combining these lexicons, and automatically further expanding the combined lexicon. This lexicon can be shared with other researchers.
- Built the best presently available systems for Chinese word semantic orientation classification, for Chinese sentence subjectivity detection and for Chinese sentence polarity classification.

5.3 Limitations

Any experimental investigation is necessarily limited; many compromises have been made in the design process in order to focus on the most important research questions. Some of the limitations are common to all text classification experiments that involve the use of test collections, while others are unique to this study. Our investigation of bilingual aspect classification has the following limitations that are unique to our study:

- We focused only on aspects for which no two segments shared any common sentence. In the real world, aspects may overlap to some degree, and some sentences might actually be associated with multiple aspects.
- We assumed that a fixed document segment was a reasonable surrogate for an aspect instance. Moreover, we did not optimize the size of the document segments.
- We selected 100 dimensions for LSA based on experience from a different setting, rather than optimizing that parameter for our task.

Our investigation of Chinese sentence attitude classification has the following major limitations that are unique to our study:

- Our Chinese sentiment lexicon was prepared and pruned in one month, and errors may have occurred when rekeying the Simplified Chinese sentiment dictionaries into Traditional Chinese, translating the English sentiment lexicon into Simplified Chinese, converting Simplified Chinese into Traditional Chinese, and manually pruning the combined lexicon. Most importantly, when categorizing the lexicon into nine sub-lexicons, the investigator personally made subjective categorization decisions and subjective polarity judgments for some of the words (especially translated words). Therefore the resulting lexicons may not be completely representative of the decisions an independent annotator would have made.

- The conclusions were drawn based on experiments with a relatively small range of system parameters and parameter thresholds. We did not test all the possible combinations of the parameters and parameter thresholds.

- The conclusions were drawn based on experiments using the NTCIR-6 test collection, for which the inter-annotator agreement is lower than we would wish to have. We do not presently have other test collections for news or other domains.

- Differences in the segmentation of Chinese can result in different classification results, but we tried only the one-best segmentation provided by the Stanford Segmenter. Different dependency parsers can also generate different results, but we tried only the Stanford Parser. Moreover, the dependency relationships extracted from the parse of a sentence are based on heuristic rules inferred by manually examining a very small sample of parsed sentences in the test collection; therefore the rules may be incomplete and sometimes incorrect.

- We treat the semantic orientation of the words in our lexicons as context-free, which is straightforward, but almost surely suboptimal.

- Our techniques infer attitude only from direct use of language with no access to external knowledge or social context.

- We have focused on sentence-level attitude classification, while what we ultimately need is document segment level attitude classification.

5.4 Future Work

Much work can be done in the future on bilingual aspect classification, Chinese attitude classification, and other tasks that support our ultimate goal.

5.4.1 Bilingual Aspect Classification

Bilingual aspect classification involves at least five basic issues:

- defining what we mean by aspect,
- choosing a language pair,
- choosing classification methods,
- selecting and using aspect training examples, and
- choosing text genre (e.g., news, blogs).

We made a series of technical decisions along the way:

- We have operationally defined non-overlapping document segments as aspect instances and aspects as mutually exclusive categories;
- extracted global DF vectors from the whole collection for segments;
- took 100 dimensions to build the local LSA space;

- tested the idea of moving the translated supporting-language training examples toward the main-language training examples all the way until their centroids met; and
- used kNN and its two variants as the classification methods.

Using document segments as aspect instances opens some interesting questions:

- Will different people focus on similar sets of aspects for the same topic?
- What is the relationship between the size of document segments and the characteristics (such as broadness or specificity) of the aspect and the topic?
- How can the size of document segments be optimized?
- What is the effect of text genre on this optimal partition into document segments?

Exploring these questions might lead us to better understand the best way to split a document into segments.

Natural aspect instances may or may not overlap. If we allow aspect instances to overlap (i.e., to share sentences), that naturally suggests creating overlapping segments (or to split documents in multiple ways).

We have used only news articles for our experiment. Different text genres, such as blogs, might have an effect on the size of document segment, the retrieved relevant documents for each topic, and translation technique, which is worth investigation.

When extracting the DF vectors from the index of document segments, we used global DF vectors extracted from the whole collection. We could explore using local

DF vectors extracted from the documents retrieved for a topic, or some smoothed combination of local and global DF statistics.

When building the local LSA space for each main-language topic, we selected 100 dimensions based on experience from a different setting; we could optimize that parameter for our task.

When using the supporting-language training examples, we translated them and projected them into main-language local LSA space. We suspect that translation error and projection error might have occurred. We could use only the translated supporting-language training examples to perform main-language aspect classification to investigate the translation error. We could also move the translated supporting-language training examples toward the main-language training examples a little bit, or half the way, to investigate the projection error.

We only tried kNN for bilingual aspect classification. Different classification methods could produce different results, even with the same features, so other classification methods would be worth trying.

We have investigated bilingual aspect classification with English and Chinese. Human languages are complex, and each language possesses unique features that might have some effect on translation and classification. For instance, translating from Chinese to English is difficult due to ambiguous word segmentation for Chinese, but monolingual Chinese information retrieval is usually quite effective (because any consistent segmentation will suffice). Until more language pairs are tried, our results can provide at best one glimpse into the effects of these types of issues on bilingual aspect classification.

Finally, our design ideas would benefit from experience with use of our techniques by real users. For example, with monolingual users, a machine translation module to help users translate queries and select document segments as aspect training examples will be needed.

5.4.2 Chinese Attitude Classification

Much work can be done in the future on Chinese semantic orientation polarity classification, Chinese sentence subjectivity detection and polarity classification, and test collection improvement.

For character-based Chinese word semantic orientation classification, if we can detect non-subjective words before applying the current word polarity classification approach, we can solve a part of the problem of the current approach, which cannot distinguish neutral words from non-subjective words. For example, named entities may not express a semantic orientation even if they contain characters with a positive or negative attitude tendency. If we can identify more features of a word, such as whether it is a named entity, its part of speech (POS), we may be able to perform better word semantic orientation classification.

We have manually built decision trees for sentence subjectivity and polarity classification, which allowed us get insights to the problems as described above. Based on the understanding of these problems, we could build classifiers using machine learning approaches, such as C4.5, in the future.

The current shallow linguistic analysis approach which aggregates a sentence's polarity from its words may not scale well to bigger linguistic units. A long sentence may have an argumentation logic with fluctuations of attitudes between clauses. The context of the sentence, that is, the sentences before and after the sentence of interest, may also have an influence on the interpretation of that sentence.

We assumed sentence attitude could be aggregated from the words it contained, but our approach of aggregating sentence polarity from word polarity resulted in poor classification effectiveness. This is probably due to a failure to more faithfully replicate the ways in which the annotators determine the polarity of a sentence from the linguistic units (such as words, phrases, and clauses) and from extra-linguistic cues (such as irony and sarcasm) that draw on world knowledge. We will do more failure analysis to understand better the problem of our approach, but more fundamentally, an investigation of how the attitude annotators determined the sentence subjectivity and polarity might help us design better approaches to sentence polarity classification. Furthermore, a human subject study of how attitude in text documents is conveyed and interpreted might help us to design better sentence attitude classification approaches.

Most attitude models have polarity as one dimension and polarity intensity as a different dimension. When intensity is high enough, all the attitude can become negative. For instance, extreme love may become pornography, and overly intensive appraisal may make people feel uncomfortable. So investigating the intensity of polarity may help us understand the polarity of attitude better.

We used the NTCIR-6 test collection to evaluate our systems. Each sentence was annotated by three persons, and there were a total of seven annotators. NTU has just released data on which person annotated which document. This will allow us to compose gold standards which minimize the number of annotators, and thus minimize the variations in annotation consistency.

Currently neither English attitude classification nor Chinese attitude classification have distinguished ambivalence from neutrality. The current Chinese Opinion Analysis Pilot Task test collection does not distinguish ambivalence from neutrality either. We will help NTCIR to get ambivalence into the test collection in the future.

Text genre apparently has an effect on attitude classification because different text genres have different linguistic features. We studied English blog attitude analysis in another study. We wish to investigate Chinese blog attitude analysis in the future.

Finally, our assumption that sentence-level attitude classification can be usefully aggregated to perform segment-level attitude classification needs to be tested by developing and evaluating alternative approaches to that task.

5.4.3 Building Interactive Systems

We proposed an ambitious goal, which aimed to help end users to compare attitudes about aspects of a topic in English and Chinese news and blogs. We have built aspect classifiers and Chinese sentence attitude classifiers for news. Due to genre differences, the components trained on news might not work as well on blogs; therefore most system components (document segment splitter, search engine, aspect classifier, and attitude classifier) will also need to be trained on blogs.

Of course, the system components ultimately also need to be integrated into an interactive system for end users, which is not a trivial task. Graphical user interfaces need to be created for searching topics, selecting aspects, and comparing attitudes towards aspects of those topics in the two languages. Simple interfaces can be easily built, but building the right ones is a research topic in its own right.

In our design, we assume the users are bilingual, able to read and write both English and Chinese. If the user is monolingual, we need to use state-of-the-art cross-language information retrieval (CLIR) and machine translation techniques to build additional modules for retrieving documents and selecting document segments in a foreign language.

In our ultimate goal, we started with separate collections of news and blogs. Real collections may include both news and blogs. A classifier for distinguishing between news and blogs might therefore be needed. Trying to automatically identify news based on newsworthiness of events and ideas is a difficult task. Newsworthiness of events or ideas is determined by seven factors:

1. impact: events that are likely to affect many people;
2. timeliness: events that are immediate, recent;
3. prominence: events involving well-known people or institutions;
4. proximity: events in the circulation or broadcast area;
5. conflict: events that reflect clashes between people or institution;
6. the bizarre: events that deviate sharply from the expected and the experiences

of everyday life; and

7. currency: events and situations that are being talked about [117].

We can probably simplify the task, however, by exploiting systematic genre differences between the two sources, such as formality [124].

To make our envisional system more useful, we may also compare the attitudes on aspects of a topic across demographic groups, rather than sorting only by languages and source type. A classifier for associating documents with a population might be built for this purpose. A population might, for example, be roughly characterized by geographic location, such as English speaking people in China or Singapore, or Chinese speaking people in the U.K. or the U.S. This task raises two related sub-tasks:

- distinguishing among languages. There are systematic language encoding differences among some languages such as between simplified Chinese and traditional Chinese. There are also differences within a same type of language across different geographic locations, such as between British English and American English, or between Chinese in mainland China and Chinese in Singapore. Some words or phrases (e.g., slang) are much more common in British English than American English. Similarly, Chinese writers in mainland China and Chinese writers in Singapore often spell transliterated foreign names differently.

- detecting geographic locations of news and blog sources. News articles are usually marked with news agency names, which can indicate the country of origin. In addition, document contents can also be used to infer the likely location of the writer, such as when the author mentions local restaurants and theaters.

Both language and geographic location cues could be used together to sort documents by demographic characteristics, and more sophisticated techniques (e.g., based on average word length) could also be tried.

5.5 Summary

The goal of this dissertation has been to explore the design of tools to help users make sense of subjective information in English and Chinese and to compare attitudes on aspects of a topic in English and Chinese document collections. This big challenge problem can guide the development of specific technologies for aspect and attitude classification. We have explored two sub-problems—bilingual aspect classification and Chinese sentence attitude classification, with news collections. We achieved encouraging results for both.

We applied monolingual and cross-language text classification techniques to build bilingual aspect classification systems using variants of the k-Nearest-Neighbor technique. Through our experiments with the test collection that we created for bilingual aspect classification, we concluded that when only a few main-language training examples were available, a few supporting-language training examples would also generally be useful.

We adopted a shallow linguistic analysis approach to classifying the subjectivity and polarity of a Chinese sentence. We created a relatively large Chinese sentiment lexicon by using existing Chinese and English lexical resources, and extended an existing character-based approach for automatically classifying the semantic orientation of Chinese words. The results of our sentence subjectivity and polarity classifiers proved to be more effective than the best previously reported results on the NTCIR-6 test collection. We did reasonably well on sentence subjectivity detection, but still poorly on sentence polarity classification.

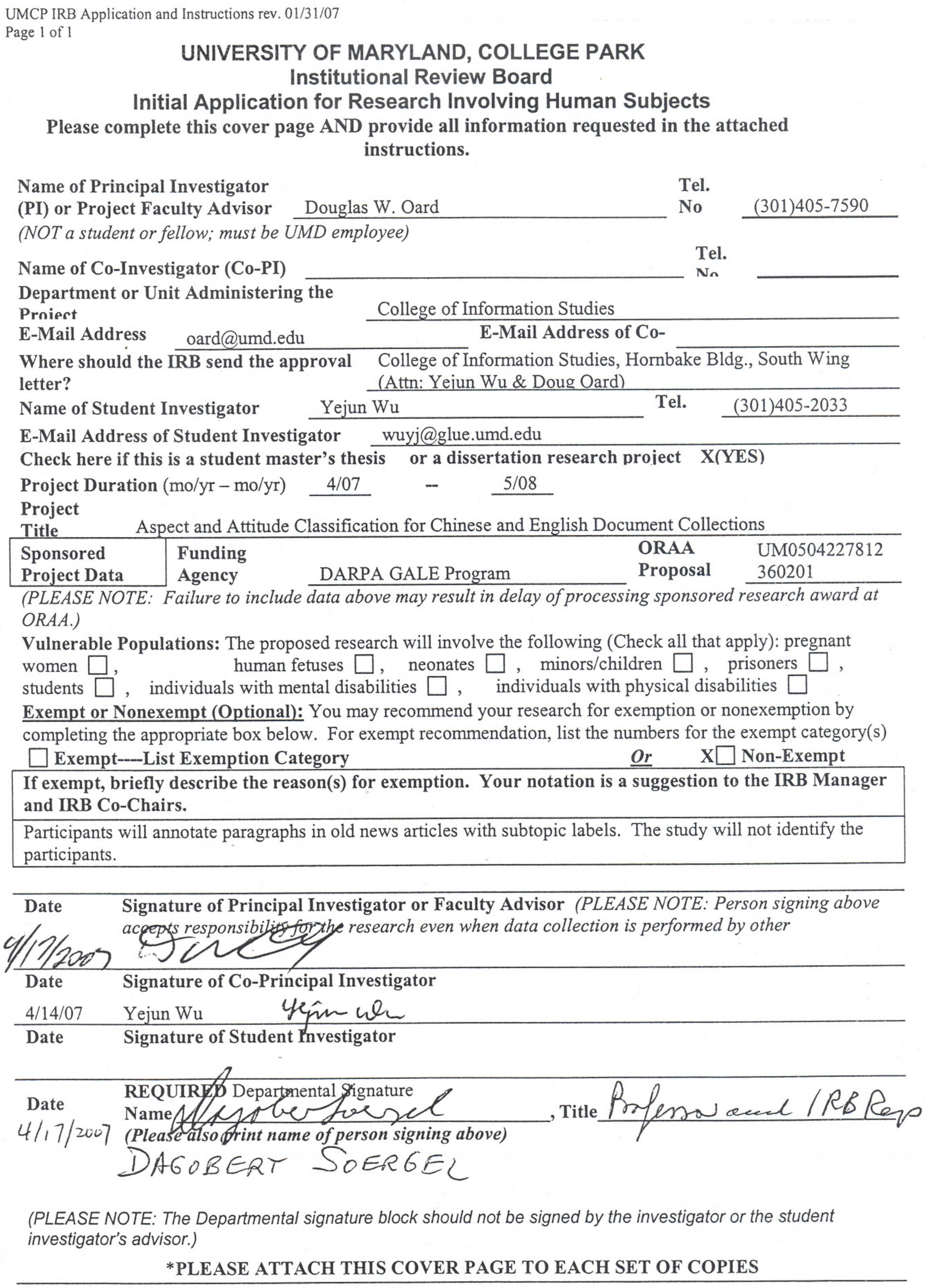
Much work can be done in the future on bilingual aspect classification, Chinese attitude classification, and other tasks that would help to realize our ultimate goal. But it has been said that such a journey of one thousand li[1] begins with a single step, and we have now taken that first step.

[1] Li is a traditional Chinese unit of distance, which now has a standardized length of half a kilometer. http://en.wikipedia.org/wiki/Li_(unit) (last visited on January 4, 2008).

Appendix A

Approved IRB Application

Attached here is our original Institutional Review Board (IRB) application. Before the annotators started annotation, a training session was provided to them (see Section 3.2.3). The granularity of “segment numbers” denoted in the instructions was not clearly defined when the application was prepared. They were actually sentence numbers, and the annotators were instructed to freely define segments. The third participant was not actually recruited because the two annotators worked as the “third participant” (i.e., as an independent annotator to support evaluation of inter-annotator agrement), each one for the other.



UMCP IRB Application and Instructions rev. 01/31/07
Page 1 of 1

UNIVERSITY OF MARYLAND, COLLEGE PARK
Institutional Review Board
Initial Application for Research Involving Human Subjects

Please complete this cover page AND provide all information requested in the attached instructions.

Name of Principal Investigator (PI) or Project Faculty Advisor Douglas W. Oard **Tel. No** (301)405-7590
(NOT a student or fellow; must be UMD employee)

Name of Co-Investigator (Co-PI) ______ **Tel. No** ______

Department or Unit Administering the Project College of Information Studies

E-Mail Address oard@umd.edu **E-Mail Address of Co-** ______

Where should the IRB send the approval letter? College of Information Studies, Hornbake Bldg., South Wing (Attn: Yejun Wu & Doug Oard)

Name of Student Investigator Yejun Wu **Tel.** (301)405-2033

E-Mail Address of Student Investigator wuyj@glue.umd.edu

Check here if this is a student master's thesis or a dissertation research project X(YES)

Project Duration (mo/yr – mo/yr) 4/07 -- 5/08

Project Title Aspect and Attitude Classification for Chinese and English Document Collections

Sponsored Project Data **Funding Agency** DARPA GALE Program **ORAA Proposal** UM0504227812 360201

(PLEASE NOTE: Failure to include data above may result in delay of processing sponsored research award at ORAA.)

Vulnerable Populations: The proposed research will involve the following (Check all that apply): pregnant women □, human fetuses □, neonates □, minors/children □, prisoners □, students □, individuals with mental disabilities □, individuals with physical disabilities □

Exempt or Nonexempt (Optional): You may recommend your research for exemption or nonexemption by completing the appropriate box below. For exempt recommendation, list the numbers for the exempt category(s)

□ **Exempt----List Exemption Category** ***Or*** X□ **Non-Exempt**

If exempt, briefly describe the reason(s) for exemption. Your notation is a suggestion to the IRB Manager and IRB Co-Chairs.

Participants will annotate paragraphs in old news articles with subtopic labels. The study will not identify the participants.

Date **Signature of Principal Investigator or Faculty Advisor** *(PLEASE NOTE: Person signing above accepts responsibility for the research even when data collection is performed by other*

4/17/2007 Doug

Date **Signature of Co-Principal Investigator**

4/14/07 Yejun Wu Yejun Wu

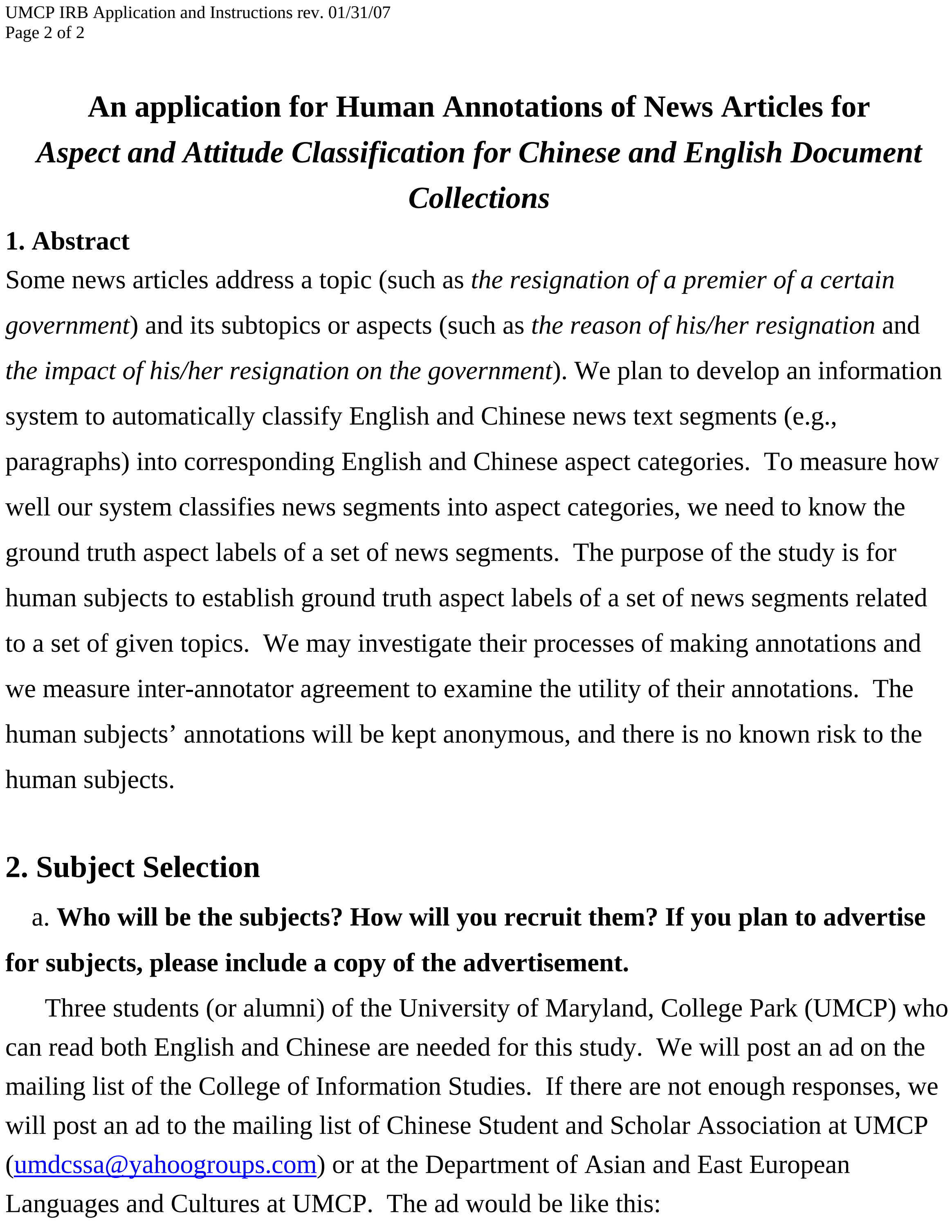
Date **Signature of Student Investigator**

Date 4/17/2007 **REQUIRED** Departmental Signature
Name Dagobert Soergel, **Title** Professor and IRB Rep
(Please also print name of person signing above)
DAGOBERT SOERGEL

(PLEASE NOTE: The Departmental signature block should not be signed by the investigator or the student investigator's advisor.)

***PLEASE ATTACH THIS COVER PAGE TO EACH SET OF COPIES**

Figure A.1: Approved IRB application (page 1)



An application for Human Annotations of News Articles for *Aspect and Attitude Classification for Chinese and English Document Collections*

1. Abstract

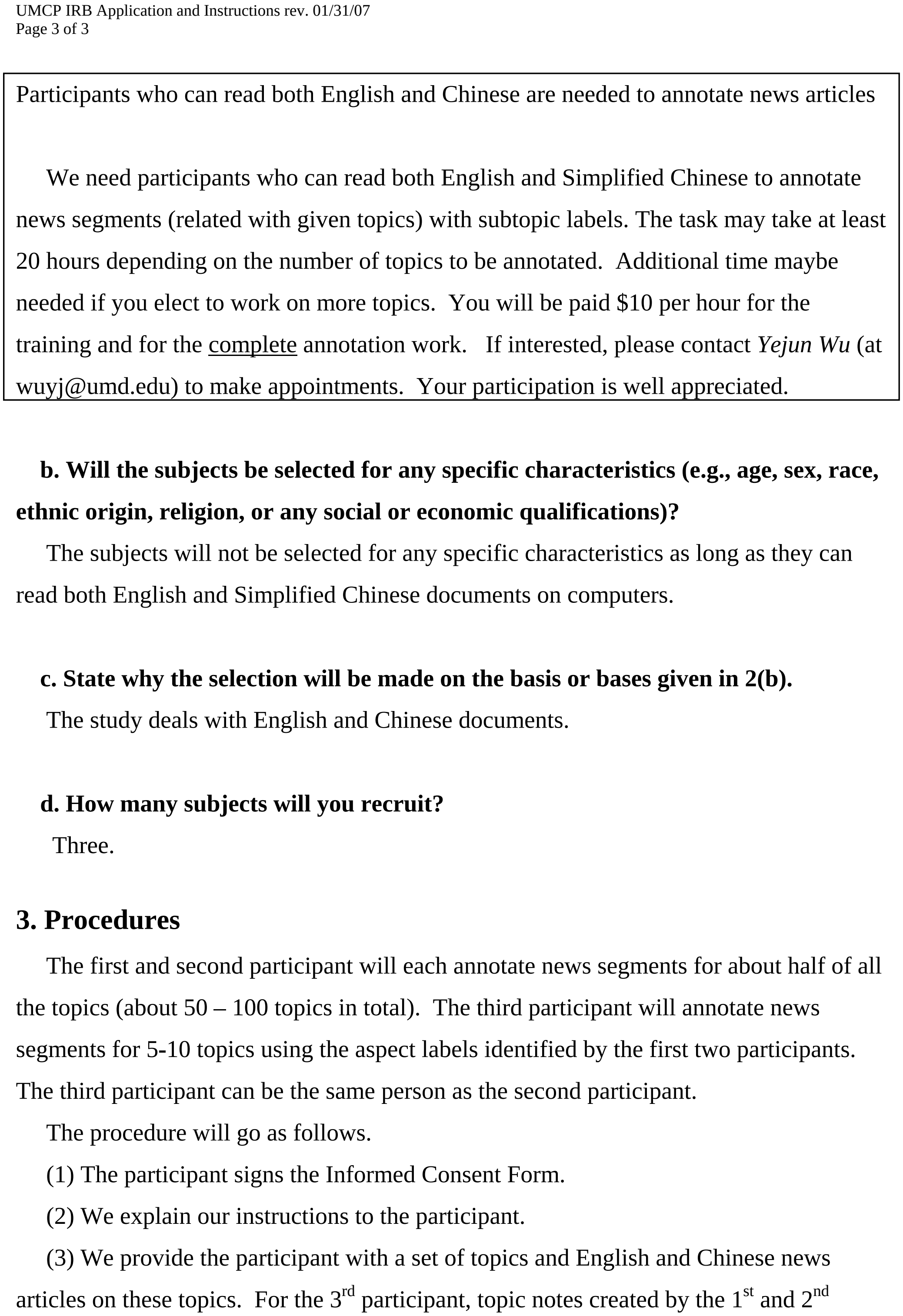
Some news articles address a topic (such as *the resignation of a premier of a certain government*) and its subtopics or aspects (such as *the reason of his/her resignation* and *the impact of his/her resignation on the government*). We plan to develop an information system to automatically classify English and Chinese news text segments (e.g., paragraphs) into corresponding English and Chinese aspect categories. To measure how well our system classifies news segments into aspect categories, we need to know the ground truth aspect labels of a set of news segments. The purpose of the study is for human subjects to establish ground truth aspect labels of a set of news segments related to a set of given topics. We may investigate their processes of making annotations and we measure inter-annotator agreement to examine the utility of their annotations. The human subjects' annotations will be kept anonymous, and there is no known risk to the human subjects.

2. Subject Selection

a. **Who will be the subjects? How will you recruit them? If you plan to advertise for subjects, please include a copy of the advertisement.**

Three students (or alumni) of the University of Maryland, College Park (UMCP) who can read both English and Chinese are needed for this study. We will post an ad on the mailing list of the College of Information Studies. If there are not enough responses, we will post an ad to the mailing list of Chinese Student and Scholar Association at UMCP (umdcssa@yahoogroups.com) or at the Department of Asian and East European Languages and Cultures at UMCP. The ad would be like this:

Figure A.2: Approved IRB application (page 2)



UMCP IRB Application and Instructions rev. 01/31/07
Page 3 of 3

Participants who can read both English and Chinese are needed to annotate news articles

We need participants who can read both English and Simplified Chinese to annotate news segments (related with given topics) with subtopic labels. The task may take at least 20 hours depending on the number of topics to be annotated. Additional time maybe needed if you elect to work on more topics. You will be paid $10 per hour for the training and for the complete annotation work. If interested, please contact *Yejun Wu* (at wuyj@umd.edu) to make appointments. Your participation is well appreciated.

b. Will the subjects be selected for any specific characteristics (e.g., age, sex, race, ethnic origin, religion, or any social or economic qualifications)?

The subjects will not be selected for any specific characteristics as long as they can read both English and Simplified Chinese documents on computers.

c. State why the selection will be made on the basis or bases given in 2(b).

The study deals with English and Chinese documents.

d. How many subjects will you recruit?

Three.

3. Procedures

The first and second participant will each annotate news segments for about half of all the topics (about 50 – 100 topics in total). The third participant will annotate news segments for 5-10 topics using the aspect labels identified by the first two participants. The third participant can be the same person as the second participant.

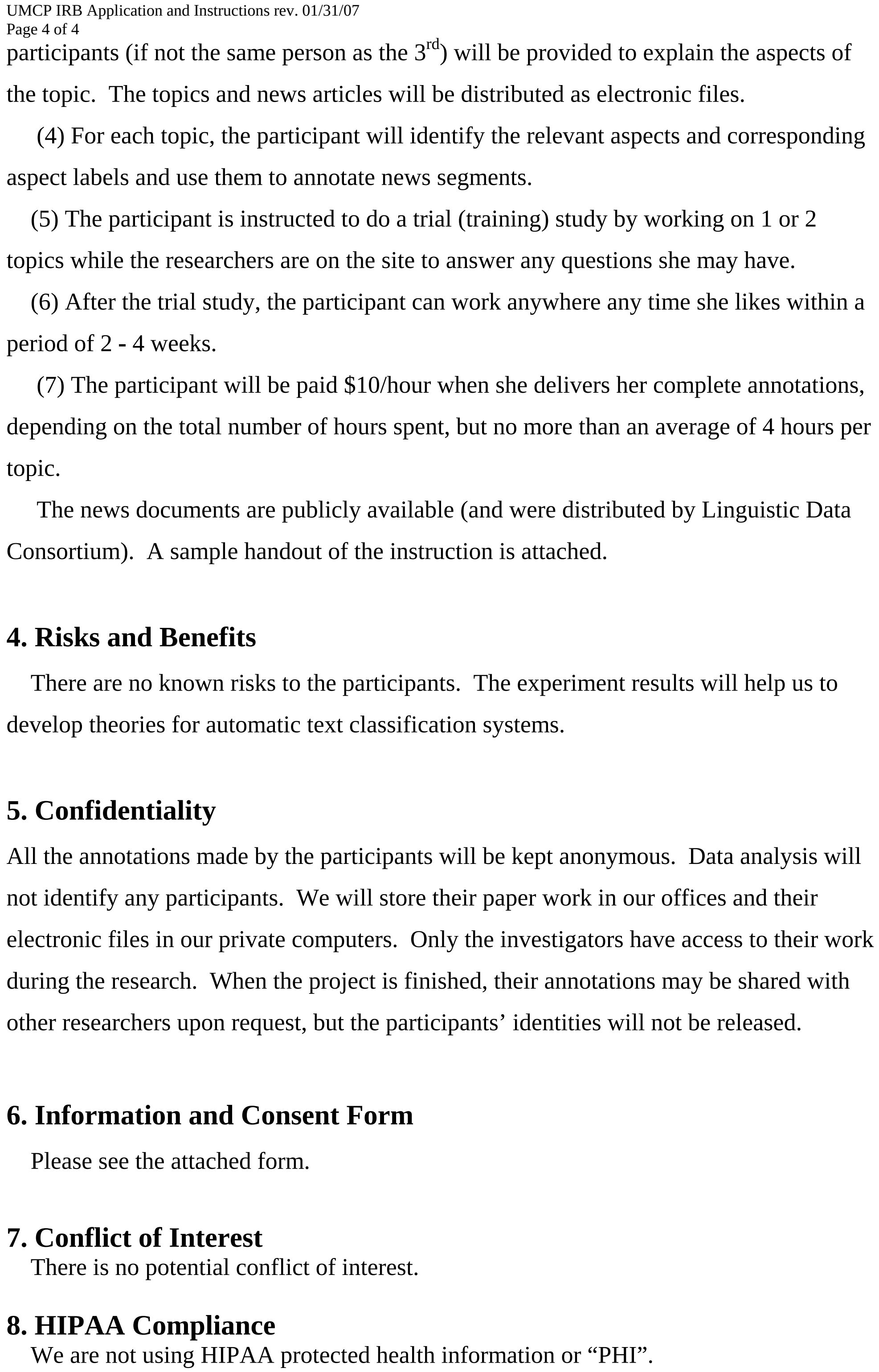
The procedure will go as follows.

(1) The participant signs the Informed Consent Form.

(2) We explain our instructions to the participant.

(3) We provide the participant with a set of topics and English and Chinese news articles on these topics. For the 3rd participant, topic notes created by the 1st and 2nd

Figure A.3: Approved IRB application (page 3)



UMCP IRB Application and Instructions rev. 01/31/07
Page 4 of 4

participants (if not the same person as the 3rd) will be provided to explain the aspects of the topic. The topics and news articles will be distributed as electronic files.

(4) For each topic, the participant will identify the relevant aspects and corresponding aspect labels and use them to annotate news segments.

(5) The participant is instructed to do a trial (training) study by working on 1 or 2 topics while the researchers are on the site to answer any questions she may have.

(6) After the trial study, the participant can work anywhere any time she likes within a period of 2 - 4 weeks.

(7) The participant will be paid $10/hour when she delivers her complete annotations, depending on the total number of hours spent, but no more than an average of 4 hours per topic.

The news documents are publicly available (and were distributed by Linguistic Data Consortium). A sample handout of the instruction is attached.

4. Risks and Benefits

There are no known risks to the participants. The experiment results will help us to develop theories for automatic text classification systems.

5. Confidentiality

All the annotations made by the participants will be kept anonymous. Data analysis will not identify any participants. We will store their paper work in our offices and their electronic files in our private computers. Only the investigators have access to their work during the research. When the project is finished, their annotations may be shared with other researchers upon request, but the participants' identities will not be released.

6. Information and Consent Form

Please see the attached form.

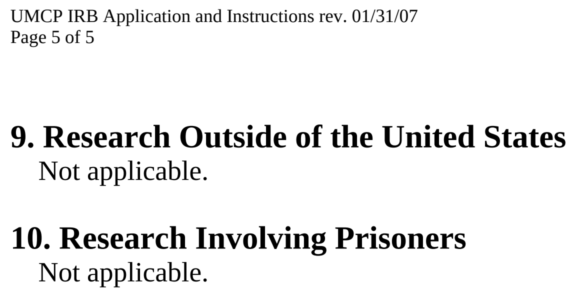
7. Conflict of Interest

There is no potential conflict of interest.

8. HIPAA Compliance

We are not using HIPAA protected health information or "PHI".

Figure A.4: Approved IRB application (page 4)



UMCP IRB Application and Instructions rev. 01/31/07
Page 5 of 5

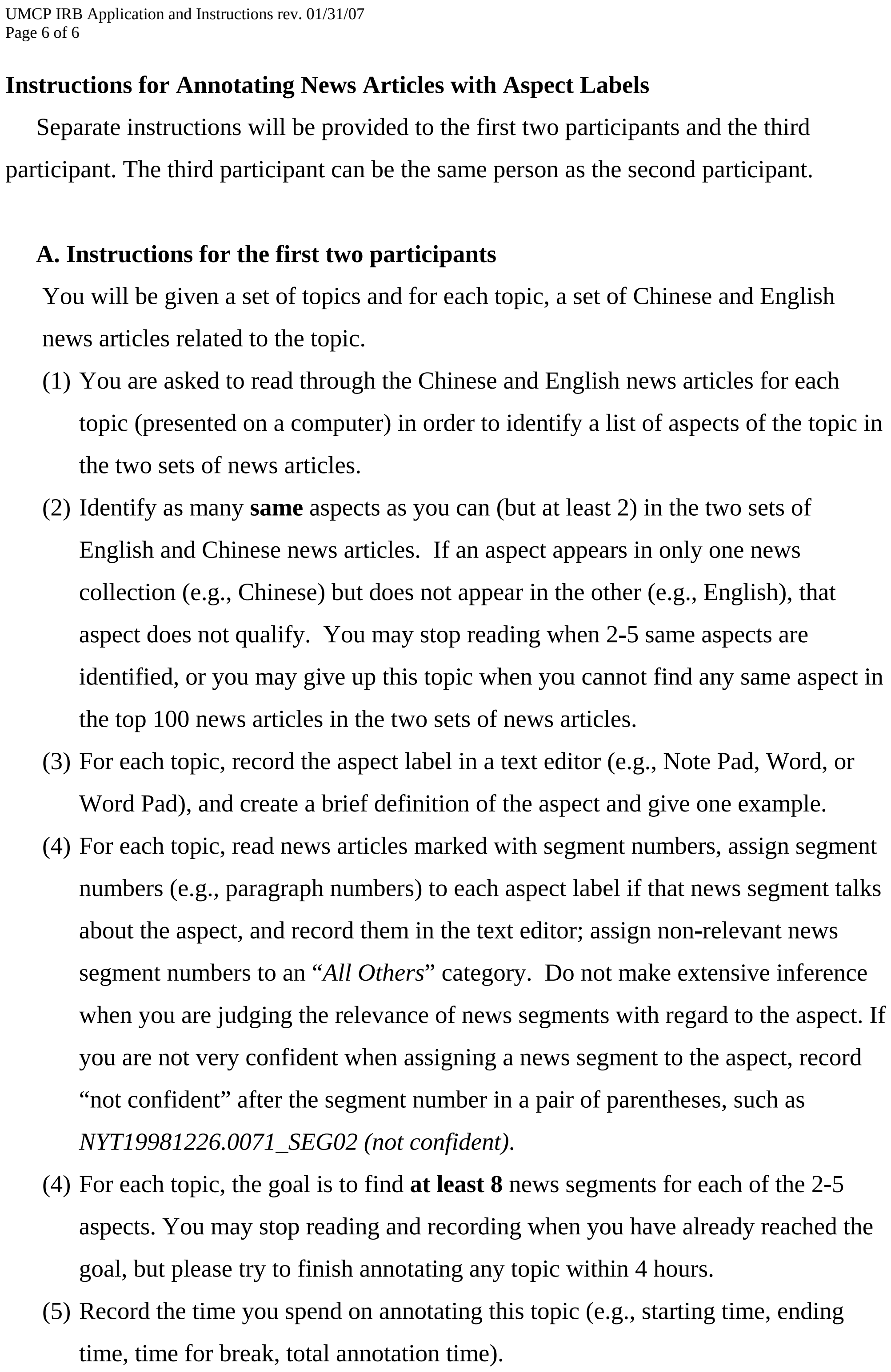
9. Research Outside of the United States

Not applicable.

10. Research Involving Prisoners

Not applicable.

Figure A.5: Approved IRB application (page 5)



UMCP IRB Application and Instructions rev. 01/31/07
Page 6 of 6

Instructions for Annotating News Articles with Aspect Labels

Separate instructions will be provided to the first two participants and the third participant. The third participant can be the same person as the second participant.

A. Instructions for the first two participants

You will be given a set of topics and for each topic, a set of Chinese and English news articles related to the topic.

(1) You are asked to read through the Chinese and English news articles for each topic (presented on a computer) in order to identify a list of aspects of the topic in the two sets of news articles.

(2) Identify as many **same** aspects as you can (but at least 2) in the two sets of English and Chinese news articles. If an aspect appears in only one news collection (e.g., Chinese) but does not appear in the other (e.g., English), that aspect does not qualify. You may stop reading when 2-5 same aspects are identified, or you may give up this topic when you cannot find any same aspect in the top 100 news articles in the two sets of news articles.

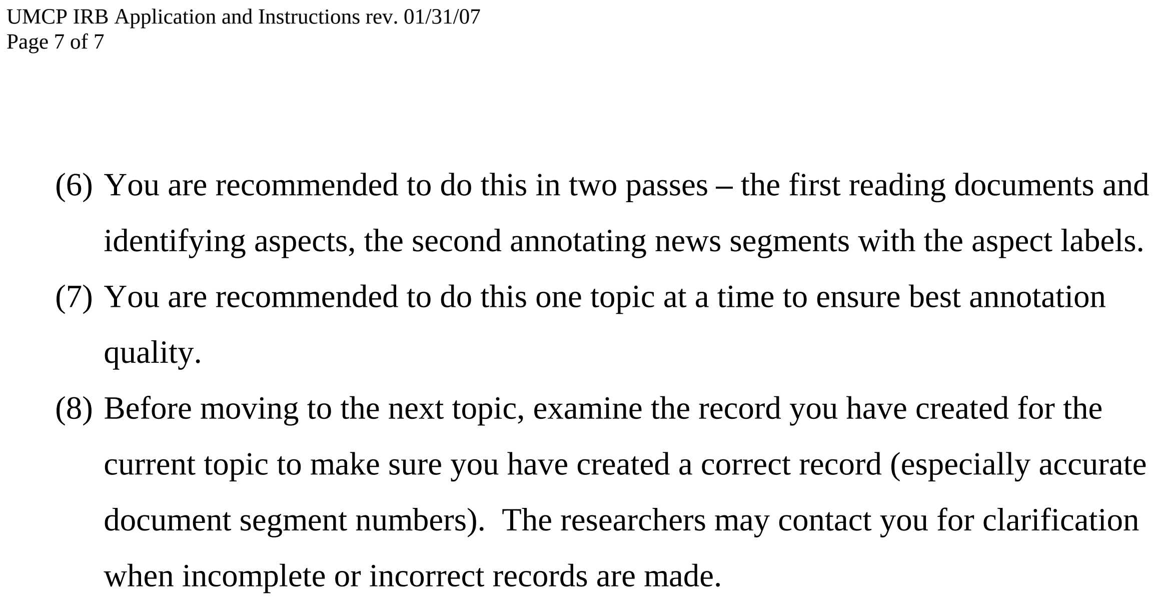
(3) For each topic, record the aspect label in a text editor (e.g., Note Pad, Word, or Word Pad), and create a brief definition of the aspect and give one example.

(4) For each topic, read news articles marked with segment numbers, assign segment numbers (e.g., paragraph numbers) to each aspect label if that news segment talks about the aspect, and record them in the text editor; assign non-relevant news segment numbers to an "*All Others*" category. Do not make extensive inference when you are judging the relevance of news segments with regard to the aspect. If you are not very confident when assigning a news segment to the aspect, record "not confident" after the segment number in a pair of parentheses, such as *NYT19981226.0071_SEG02 (not confident).*

(4) For each topic, the goal is to find **at least 8** news segments for each of the 2-5 aspects. You may stop reading and recording when you have already reached the goal, but please try to finish annotating any topic within 4 hours.

(5) Record the time you spend on annotating this topic (e.g., starting time, ending time, time for break, total annotation time).

Figure A.6: Approved IRB application (page 6)



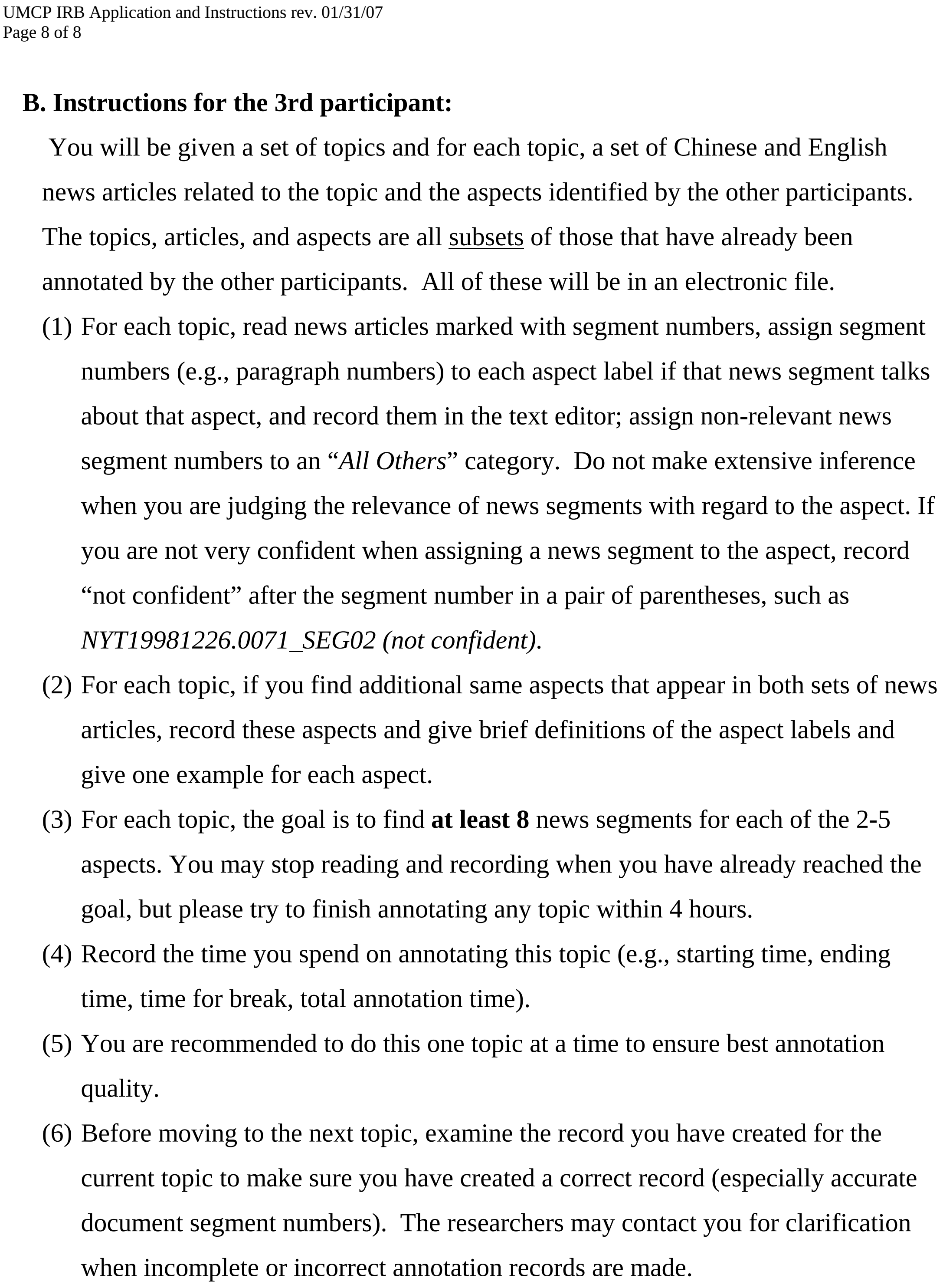
UMCP IRB Application and Instructions rev. 01/31/07
Page 7 of 7

(6) You are recommended to do this in two passes – the first reading documents and identifying aspects, the second annotating news segments with the aspect labels.

(7) You are recommended to do this one topic at a time to ensure best annotation quality.

(8) Before moving to the next topic, examine the record you have created for the current topic to make sure you have created a correct record (especially accurate document segment numbers). The researchers may contact you for clarification when incomplete or incorrect records are made.

Figure A.7: Approved IRB application (page 7)



B. Instructions for the 3rd participant:

You will be given a set of topics and for each topic, a set of Chinese and English news articles related to the topic and the aspects identified by the other participants. The topics, articles, and aspects are all subsets of those that have already been annotated by the other participants. All of these will be in an electronic file.

(1) For each topic, read news articles marked with segment numbers, assign segment numbers (e.g., paragraph numbers) to each aspect label if that news segment talks about that aspect, and record them in the text editor; assign non-relevant news segment numbers to an "*All Others*" category. Do not make extensive inference when you are judging the relevance of news segments with regard to the aspect. If you are not very confident when assigning a news segment to the aspect, record "not confident" after the segment number in a pair of parentheses, such as *NYT19981226.0071_SEG02 (not confident).*

(2) For each topic, if you find additional same aspects that appear in both sets of news articles, record these aspects and give brief definitions of the aspect labels and give one example for each aspect.

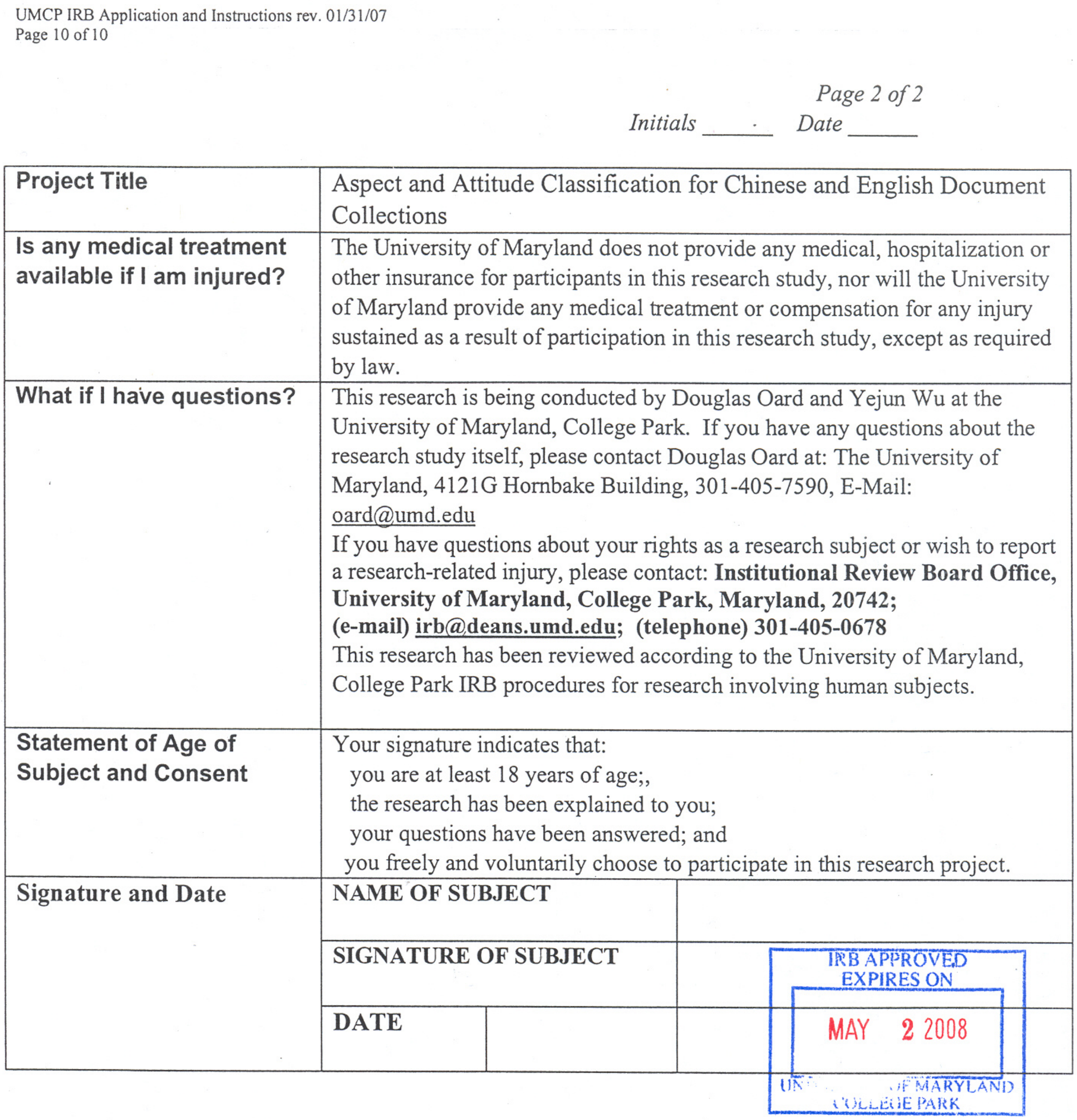
(3) For each topic, the goal is to find **at least 8** news segments for each of the 2-5 aspects. You may stop reading and recording when you have already reached the goal, but please try to finish annotating any topic within 4 hours.

(4) Record the time you spend on annotating this topic (e.g., starting time, ending time, time for break, total annotation time).

(5) You are recommended to do this one topic at a time to ensure best annotation quality.

(6) Before moving to the next topic, examine the record you have created for the current topic to make sure you have created a correct record (especially accurate document segment numbers). The researchers may contact you for clarification when incomplete or incorrect annotation records are made.

Figure A.8: Approved IRB application (page 8)



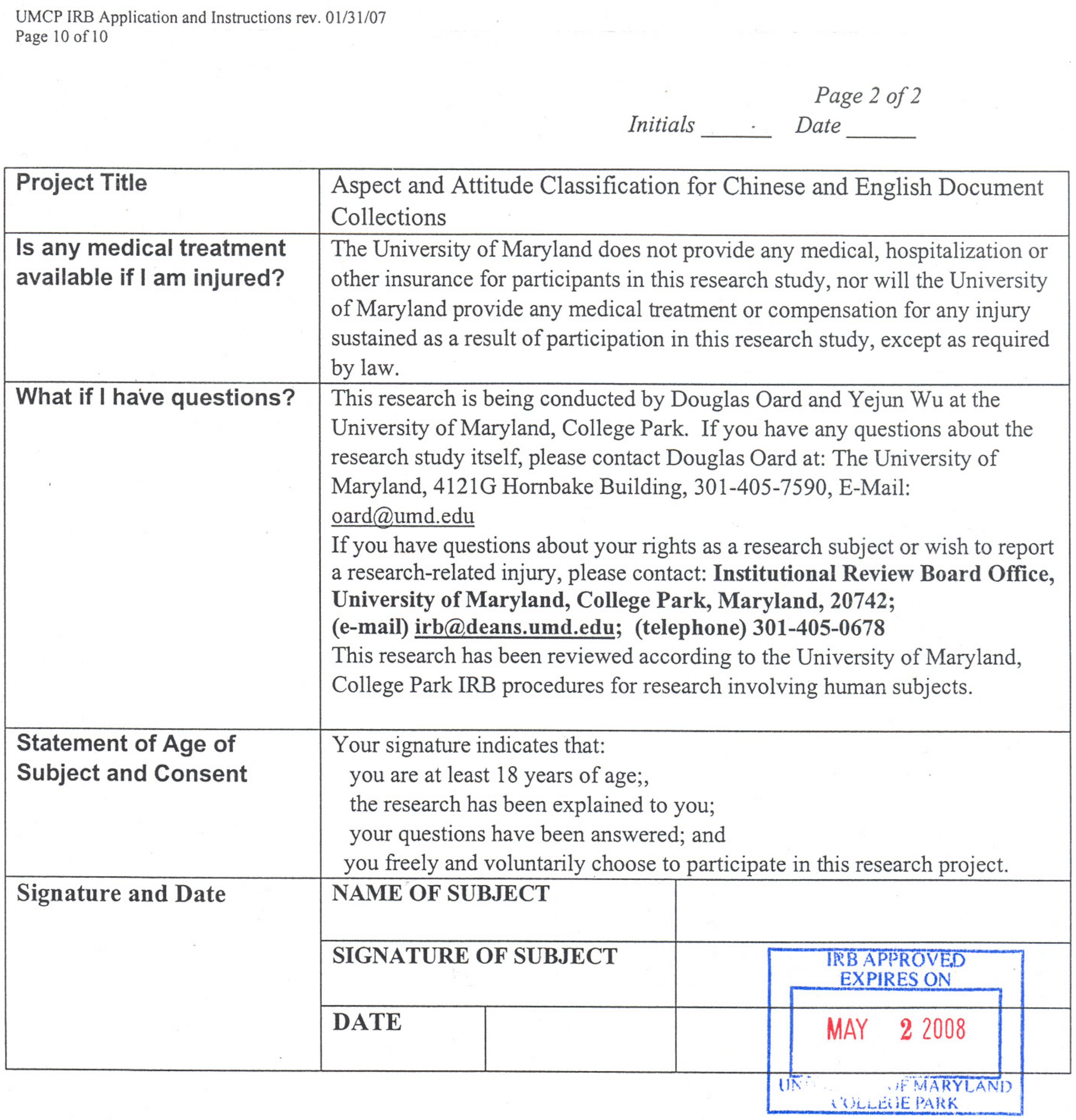
UMCP IRB Application and Instructions rev. 01/31/07
Page 10 of 10

Page 2 of 2
Initials ______ *Date* ______

Project Title	Aspect and Attitude Classification for Chinese and English Document Collections
Is any medical treatment available if I am injured?	The University of Maryland does not provide any medical, hospitalization or other insurance for participants in this research study, nor will the University of Maryland provide any medical treatment or compensation for any injury sustained as a result of participation in this research study, except as required by law.
What if I have questions?	This research is being conducted by Douglas Oard and Yejun Wu at the University of Maryland, College Park. If you have any questions about the research study itself, please contact Douglas Oard at: The University of Maryland, 4121G Hornbake Building, 301-405-7590, E-Mail: oard@umd.edu If you have questions about your rights as a research subject or wish to report a research-related injury, please contact: **Institutional Review Board Office, University of Maryland, College Park, Maryland, 20742; (e-mail) irb@deans.umd.edu; (telephone) 301-405-0678** This research has been reviewed according to the University of Maryland, College Park IRB procedures for research involving human subjects.
Statement of Age of Subject and Consent	Your signature indicates that: you are at least 18 years of age;, the research has been explained to you; your questions have been answered; and you freely and voluntarily choose to participate in this research project.
Signature and Date	**NAME OF SUBJECT** **SIGNATURE OF SUBJECT** **DATE**

IRB APPROVED
EXPIRES ON
MAY 2 2008
UNIVERSITY OF MARYLAND
COLLEGE PARK

Figure A.9: Approved IRB application (page 9)



UMCP IRB Application and Instructions rev. 01/31/07
Page 10 of 10

Page 2 of 2
Initials _____ *Date* ______

Project Title	Aspect and Attitude Classification for Chinese and English Document Collections		
Is any medical treatment available if I am injured?	The University of Maryland does not provide any medical, hospitalization or other insurance for participants in this research study, nor will the University of Maryland provide any medical treatment or compensation for any injury sustained as a result of participation in this research study, except as required by law.		
What if I have questions?	This research is being conducted by Douglas Oard and Yejun Wu at the University of Maryland, College Park. If you have any questions about the research study itself, please contact Douglas Oard at: The University of Maryland, 4121G Hornbake Building, 301-405-7590, E-Mail: oard@umd.edu If you have questions about your rights as a research subject or wish to report a research-related injury, please contact: **Institutional Review Board Office, University of Maryland, College Park, Maryland, 20742; (e-mail) irb@deans.umd.edu; (telephone) 301-405-0678** This research has been reviewed according to the University of Maryland, College Park IRB procedures for research involving human subjects.		
Statement of Age of Subject and Consent	Your signature indicates that: you are at least 18 years of age;, the research has been explained to you; your questions have been answered; and you freely and voluntarily choose to participate in this research project.		
Signature and Date	**NAME OF SUBJECT**		
	SIGNATURE OF SUBJECT		
	DATE		

IRB APPROVED
EXPIRES ON
MAY 2 2008
UNIVERSITY OF MARYLAND
COLLEGE PARK

Figure A.10: Approved IRB application (page 10)

Appendix B

TDT3 and TDT4 Topics Annotated

The following 50 topics were annotated by the two annotators. 36 of them were kept after validating and mapping. 33 topics were in the first operational test collection; 17 in the second operational test collection. The Chinese aspects in Topic 30027 were used as training data for classifying English aspects only.

Topic	Title	Kept	TC1	TC2
30001	Cambodian Government Coalition	√	√	
30002	Hurricane Mitch	√	√	
30003	Pinochet Trial	√	√	√
30005	Osama bin Laden Indictment			
30006	NBA Labor Disputes	√	√	√
30007	Congolese Rebels vs. Pres. Kabila	√	√	√
30008	November APEC Summit Meeting	√	√	
30010	Car Bomb in Jerusalem	√		
30011	Anwar Ibrahim Case	√	√	√
30015	Holbrooke-Milosevic Meeting	√	√	
30020	Asian Games in Thailand	√	√	
30022	Chinese Dissidents Sentenced	√	√	√
30024	Newt Gingrich Resigns from US House	√	√	√
30027	Russian Financial Crisis	√	√	
30028	Turkey - Syria Tension			
30030	Taipei Mayoral Elections			

Table B.1: The 50 topics annotated by the two annotators (Part 1).
(Kept: topics kept after validation and mapping; TC1: in the first operational test collection; TC2: in the second operational test collection.)

Topic	Title	Kept	TC1	TC2
30031	Shuttle Endeavour Mission for Space Station	√	√	
30033	Euro Introduced	√	√	√
30034	Indonesia - East Timor Conflict	√	√	
30035	Taiwanese Negotiator Visits China	√	√	
30036	Nobel Prizes Awarded			
30038	Olympic Bribery Scandal			
30041	Jiang's Historic Visit to Japan			
30042	PanAm Lockerbie Bombing Trial	√	√	√
30044	Kurd Separatist Abdullah Ocalan Arrested	√	√	
30045	Mobil - Exxon Merger	√	√	
30046	House Speaker-Designate Livingston Resigns			
30047	Space Station Module Zarya Launched	√	√	√
30048	IMF Bailout of Brazil	√	√	√
30049	North Korean Nuclear Facility?	√	√	
30050	U.S. Mid-Term Elections	√		
30053	Clinton's Gaza Trip	√	√	
30054	China Human Rights Treaty			
30059	Russian Politico Starovoitova Assassinated	√	√	√
31023	Kyoto Energy Protocol	√	√	√
31032	Yeltsin's Illness	√	√	√
31037	Japanese and Russian Leaders Meet			
31038	American Embassy Bombing Trial			
31050	China Renews Ban on Opposition Parties			
40004	Russian Nuclear Submarine Kursk Sinks	√	√	
40007	Presidential Power Struggle in Yugoslavia	√	√	√
40037	Australian Open Tennis Championship	√	√	
40038	Earthquake hits India's Gujarat State	√	√	√
40043	Madeline Albright Visits North Korea	√	√	√
40059	Attack on the USS Cole	√	√	
41012	Trouble in the Ivory Coast			
41018	Arab League Summit meeting in Cairo			
41024	Congolese President Laurent Kabila Feared Dead	√	√	√
41025	End of the Line for Peruvian President Alberto Fujimori			
41026	Crash of Singapore Airlines Flight SQ006	√		

Table B.2: The 50 topics annotated by the two annotators (Part 2).

Appendix C

An Example Topic

Here is an example topic from TDT3. The entire description of each topic was provided to the annotators.

- 30001. Cambodian Government Coalition[1]
- Seminal Event
    - WHAT: Cambodia's People's Party party beats the FUNCINPEC party in national elections; later the two parties agree to a coalition government.
    - WHO: Hun Sen, Leader of the People's Party and Prince Norodom Ranariddh, leader of FUNCINPEC
    - WHERE: Phnom Penh, Cambodia
    - WHEN: Elections take place in July; coalition formed in November, 1998.
- Topic Explication. Hun Sen's narrow victory in the elections led to protests by FUNCINPEC supporters, and violent government crackdowns against the protesters. After a three-month deadlock, the two parties agreed on a coalition government leaving Hun Sen as sole Prime Minister. On topic: stories about the election itself (campaigns, results of the election); citizens' responses to the election (protests); government efforts to stop the protests; negotiations

[1] http://projects.ldc.upenn.edu/TDT3/topics.html (last visited on November 26th, 2008.

between the two parties; details of the agreement reached between the parties; reactions of Cambodian citizens and world leaders to the agreement.

- Rule of Interpretation Rule 1: Elections
- Related Articles: APW19981113.0251, NYT19981125.0292
- More examples: Yes, Brief.

BIBLIOGRAPHY

[1] Allport, G. W., 1935. Attitudes. In C. Murchison (Ed.), *A handbook of Social Psychology*. Worcester, Mass: Clark University Press.

[2] Antle, Alissa, 2004. Supporting children's emotional expression and exploration in online environments. *IDC 2004*, June 1–3, 2004, College Park, Maryland.

[3] Armitage, Christopher J. and Conner, Mark, 2004. The effects of attitudinal ambivalence on attitude- intension-behavior relations. In Geoffrey Haddock and Gregory R. Maio (Eds.), *Contemporary Perspectives on the Psychology of Attitudes*. New York: Psychology Press.

[4] Baeza-Yates, Ricardo and Ribeiro-Neto, Berthier, 1999. Modern Information Retrieval. New York: ACM Press.

[5] Bell, David A., Guan, J.W. and Bi, Yaxin, 2005. On Combining Classifier Mass Functions for Text Categorization. *IEEE Transactions on Knowledge and Data Engineering October*, 17(10), 1307–1319.

[6] Bel, Nuria, Koster, Cornelis H. A., and Villegas, Marta, 2003. Cross-lingual text categorization. *European Conference on Digital Libraries (ECDL)*, 18(11), 613–620.

[7] Belkin, Nicholas J., 1980. Anomalous states of knowledge as a basis for information retrieval. *The Canadian Journal of Information Science*, 5, 133–143.

[8] Belo, Robert, 2004. Blogs take on the mainstream. *BBC News Online*. 31 December, 2004.

[9] Blood, Rebecca, 2002. *The weblog handbook: practical advice on creating and maintaining your blog.* Perseus Publishing, Cambridge, MA.

[10] Breckler, Steven J., 1984. Empirical validation of affect, behavior, and cognition as distinct components of attitude. *Journal of Personality and Social Psychology*, 47, 1191–1205.

[11] Breckler, Steven J., 2004. Hold still while I measure your attitude: Assessment in the throes of ambivalence. In Geoffrey Haddock and Gregory R. Maio (Eds.), *Contemporary Perspectives on the Psychology of Attitudes*. New York: Psychology Press. 77–92.

[12] Brill, Eric, 1995. Transformation-based error-driven learning and natural language processsing: A case study in part of speech tagging. *Computational Linguistics*, 21(4), 543–565.

[13] Brislin, R., 1993. Understanding Culture's Influence on Behavior. Harcourt Brace, Fort Worth.

[14] Brown, Ralf D., Pierce, Thomas, Yang, Yiming, Carbonell, Jaime G., 1999. Link detection - results and analysis. *TDT'99.* http://www.nist.gov/speech/tests/tdt/tdt99/papers/cmu_sld/ (last visited on October 16, 2008).

[15] Bruce, Rebecca, and Wiebe, Janyce, 1999. Recognizing subjectivity: a case study of manual tagging. *Natural Language Engieering*, 1(1): 1–16.

[16] Cacioppo, John T. and Bernston, Gary G., 1994. Relationship between attitudes and evaluative space: a critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin*, 115(3), 401–423.

[17] Cacioppo, John T., Gardner, W. L., and Berntson, G. G., 1997. Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review*, 1, 3–25.

[18] Cacioppo, John T. and Tassinary, Louis G., 1989. The concept of attitudes: a psychophysiological analysis. Huge Wagner and Antony Manstead (Ed.), *Handbook of Social Psychophysiology.* New York: John Wiley & Sons Ltd. Chapter 12. 309–346.

[19] Case, O. and Donald O., 1991. The collection and use of information by some American historians: a study of motives and methods. *Library Quarterly*, 61(1), 61–82.

[20] Chan, S., 1992. Families with Asian roots. In E. Lynch and M. Hanson (eds.), *Developing Cross-Cultural Competence: A Guide for Working with Young Children and Their Families*, Baltimore: Paul H. Brooks.

[21] Chen, Mao and Singh, Jaswinder P. 2001. Computing and using reputations for Internet ratings. *EC'01*, October 14–17, 2001, Tampa, Florida, USA.

[22] Crampton, Thomas. 2005. French police fear that blogs have helped incite rioting. Newyork Times, November 10, 2005. http://www.nytimes.com/2005/11/10/international/europe/10blogs.html (last visited June 6, 2007).

[23] Dahlgren, Peter. 1981. TV news and the suppression of reflexivity. In E. Katz & T. Szecsko (Eds.), *Mass Media and Social Change*. Beverley Hills, CA: Sage. 101–113.

[24] Das, Sanjiv R. and Chen, Mike Y. 2006. Yahoo! for Amazon: sentiment extraction from small talk on the web. *8th Asia Pacific Finance Association Annual Conference (2001).* 2006 version available at: http://scumis.scu.edu/˜srdas/chat.pdf

[25] Dasarathy, Belur V., 1991. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.* McGraw-Hill Computer Science Series. IEEE Computer Society Press, Las Alamitos, California.

[26] Dave, Kushal, Lawrence, Steve, and Pennock, David M. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *WWW2003*, May 20–24, 2003, Budapest, Hungary.

[27] Deerwester, Scott, Dumais, Susan T., Furnas, George W., Landauer, Thomas K., and Harshman Richard, 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.

[28] Dervin, Brenda, 1983. An overview of sense-making research: concepts, methods and results to date. *Paper presented at the International Communication Association Annual Meeting*, Dallas, May 1983.

[29] Dervin, Brenda, 1988. Measuring aspects of information seeking: A test of quantitative/qualitative methodology. In M. Burgoon (Ed.), *Communications Yearbook*, 419–444.

[30] Dervin, Brenda, 1992. From the mind's eye of the user: The sense-making qualitative-quantitative methodology. In Jack D. Glazier & Ronald R. Powell (eds). *Qualitative Research in Information Management*, 1992. Englewood, Colorado: Libraries Unlimited, Inc., 61–84.

[31] Dervin, Brenda, 1998. Sense-making theory and practice: an overview of user interests in knowledge making and use. *Journal of Knowledge Management*, 2(2), 36–46.

[32] Devitt, Ann and Ahmad, Khurshid, 2007. Sentiment polarity identification in financial news: a cohesion-based approach. *Proceedings of the Association of Computational Linguistics (ACL2007)*, Prague, June 28–30, 2007. 984–991.

[33] Donsbach, Wolfgang. 2004. Psychology of news decisions: Factors behinds journalists' professional behavior. *Journalism*, 5(2), 131–157.

[34] Durbin, Stephen D., Neal Richter, J. Neal, and Warner, Doug, 2003. A system for affective rating of texts. In *Proceedings of the KDD Workshop on Operational Text Classification Systems (OTC-3)*, 2003.

[35] Eagly A., and Chaiken, S., 1993. The Psychology of Attitudes. Fort Worth, TX: Harcourt Brace Jovanovich.

[36] Eagly A. and Chaiken, S., 1998. Attitude structure and function. In D. T. Gilbert, S. T. Fiske, and G. Lindzey (Eds.), *The Handbook of Social Psychology, 4th ed., Vol. 1.* New York: McGraw-Hill. 269-322.

[37] Edelman, Russ, 2003. Making sense of search. *AIIM E-Doc Magazine, 17(3)*, 64.

[38] Eiser, J. Richard, 1975. Attitudes and the use of evaluative language: a two way process. *Journal for the Theory of Social Behavior*, 5, 235-248.

[39] Eiser, J. Richard, 1980. Cognitive Social Psychology: A Guidebook to Theory and Research. London: McGraw-Hill.

[40] Eiser, J. Richard, 1986. *Social Psychology: Attitudes, Cognition and Social Behavior*. Cambridge: Cambridge University Press.

[41] Eiser, J. Richard, 1987. The Expression of Attitude. Springer-Verlag.

[42] Eiser, J. Richard and Stroebe, W., 1972. *Categorization and Social Judgment.* London: Academic Press.

[43] Ekman, Paul, 1994. Moods, emotions, and traits. In Paul Ekman and Richard J. Davidson (ed.), *The Nature of Emotion*, Oxford: Oxford Univeristy Press. 56–58.

[44] Ekman, Paul, 1994. All emotions are basic. In Paul Ekman and Richard J. Davidson (ed.), *The Nature of Emotion*, Oxford: Oxford Univeristy Press. 15–19.

[45] Ekman, Paul and Davidson, Richard J., 1994. Afterword: What is the function of emotions? In Paul Ekman and Richard J. Davidson (ed.), *The Nature of Emotion*, Oxford: Oxford Univeristy Press. 137–139.

[46] Ellsworth, Phoebe C., 1994. Some reasons to expect universal antecedents of emotion. In Paul Ekman and Richard J. Davidson (ed.), *The Nature of Emotion*, Oxford: Oxford Univeristy Press. 150–162.

[47] Engholm, C., 1991. *When Business East Meets Business West: The Guide to Practice and Protocol in the Pacific Rim.* John Wiley & Sons, New York.

[48] Esuli, Andrea and Sebastiani, Fabrizio, 2006. Sentiwordnet: a publicly available lexical resource for opinion mining. *Proceedings of LREC 2006.* http://citeseer.ist.psu.edu/esuli06sentiwordnet.html (last visited on April 20, 2008)

[49] Fellbaum, Christiane, 1998. WordNet: An Electronic Lexical Database. Cambridge, Massachusetts: MIT Press.

[50] Festa, Paul, 2003. Newsmaker: blogging comes to Harvard. CNET News.com, available at: http://news.com.com/2008-1082-985714.html?tag=fd_nc_1 (last visited on October 14, 2008).

[51] Fishbein, Arjen, 1980. *Understanding Attitudes and Predicting social Behavior.* Englewood Cliffs, N.J.: Prentice-Hall, 1980.

[52] Fishbein, M. and Ajzen, K., 1975. *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research.* Reading, MA: Addison-Wesley.

[53] Forman, George, 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3. 1289–1305.

[54] Franz, Martin, McCarley, J. Scott, Ward, Todd and Zhu, Wei-Jing, 2001. Unsupervised and supervised clustering for topic tracking. *SIGIR'01*, September 9-12, New Orleans, Louisiana.

[55] Franz, Martin. 2007. unpublished correpondence between J. Scott Olsson and Martin Franz; message passed by J. Scott Olsson. In J. Scott Olsson and Douglas W. Oard, 2007. Improving text classification for oral history archives with temporal domain knowledge. *SIGIR'07*, July 23–27, 2007, Amsterdam, The Netherlands.

[56] Friedl, Herwig and Stampfer, Erwin, 2001. Cross-validation. http://citeseer.ist.psu.edu/435313.html (last visited on November 20, 2007)

[57] Frijda, Nico H., 1994. Varieties of affect: emotions and episodes, moods, and sentiments. In Paul Ekman and Richard J. Davidson (ed.), *The Nature of Emotion*, Oxford: Oxford Univeristy Press. 59–67.

[58] Fu, Hong, 2006. On the definition and scope of modern chinese affixes — taking five kinds of modern Chinese textbooks as examples. *Journal of Guizhou Educational Institute*, 5, 2006 (in Chinese).

[59] Gliozzo, Alfio and Strapparava, Carlo, 2006. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, July 2006, Sydney. 553–560.

[60] Gordon, Andrew, Kazemzadeh, Abe, Nair, Anish and Petrova, Milena, 2003. Recognizing expressions of common psychology in English text. *Proceedings of the 41st Annual Meeting of the Association of Computational Linguisitcs (ACL-03)*, 208–215.

[61] Grefenstette, Gregory, Qu, Yan, Shanahan, James G., Evans, David A., 2004. Coupling niche browsers and affect analysis for an opinion mining application. *RIAO-2004*.

[62] Haddock, Geoffrey and Huskinson, Thomas L. H., 2004. Individual differences in attitude structure. In Geoffrey Haddock and Gregory R. Maio (Eds.), *Contemporary Perspectives on the Psychology of Attitudes.* New York: Psychology Press, 36-56.

[63] Halavais, A., 2002. Blogs and the "social weather". *Internet Research 3.0*, Maastricht, The Netherlands, October, 2002.

[64] Hatzivassiloglou, Vasileios, Klavans, Judith, Holcombe, Melissa, Barzilay, Regina, Kan, Min-Yan, and McKeown, Kathleen, 2001. SIMFINDER: A flexible clustering tool for summarization. *Proceedings of the Workshop on Summarization in NAACL-01.*

[65] Hatzivassiloglou, Vasileois, and McKeown, Katheleen, 1997. Predicting the semantic orientation of adjectives. *ACL-97*, 1997, 174–181.

[66] Hatzivassiloglou, Vasileios, and Wiebe, Janyce, 2000. Effects of adjective orientation and gradability on sentence subjectivity. *COLING-2000.*

[67] Hearst, Marti A., 1997. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics*, 23(1), 33–64. http://www.sims.berkeley.edu/˜hearst/tiling-about.html (last visited on November 20, 2007)

[68] Henley, Nancy M., Miller, Michelle D., Beazley, Jo Anne, et al., 2002. Frequency and specificity of referents to violence in news reports of anti-gay attacks. Discourse & Society, 13(1), 75–104. http://das.sagepub.com/cgi/reprint/13/1/75 (last visited on November 11, 2007)

[69] Herring, Susan C., Scheidt, Lois Ann, Wright, Elijah, and Bonus, Sabrina, 2004. Bridging the Gap: A Genre Analysis of Weblogs. *Proceedings of the 37th Hawaii International Conference on System Sciences*, 2004.

[70] Herring, Susan C., Scheidt, Lois Ann, Bonus, Sabrina, and Wright, Elijah, 2005. Weblogs as a bridging genre. *Information, Technology & People*, 18(2), 142–171.

[71] Hofstede, G., 1980. *Culture's Consequences. Sage*, Newbury Park, California.

[72] Hofstede, G., 1991. *Cultures and Organizations: Software of the Mind.* McGraw Hill, London.

[73] Hofstede, G. and Bond, M., 1984. 'Hofstede's culture dimensions: an independent validation using Rokeach's value survey. *Journal of Cross-Cultural Psychology*, 15, 417–413.

[74] Hoshino-Browne, Etsuko, Zanna, Adam S. et al., 2004. Investigating attitudes cross-culturally. In Geoffrey Haddock and Gregory R. Maio (Eds.), *Contemporary Perspectives on the Psychology of Attitudes.* New York: Psychology Press.

[75] Howell, David C., 2002. *Statistical Methods for Psychology*, 5th Ed. Pacific Grove, CA: Duxbury of Thomason Learning. Chapter 8.

[76] Hsu, Chih-Wei et al., 2006. A practical guide to support vector classification. http://www.csie.ntu.edu.tw/∼cjlin/papers/guide/guide.pdf (last visited on April 19, 2008).

[77] Hu, Minqing and Liu, Bing, 2004. Mining and summarizing customer reviews. *KDD'04*, August 22–25, 2004, Seattle, Washington, USA.

[78] Hull, David, 1994. *Information Retrieval Using Statistical Classification.* Doctoral dissertation, Stanford University.

[79] Irwin, Harry, 1996. Communicating with Asia - Understanding People and Customs. Allen & Unwin, Sydney.

[80] Irwin, Harry and More, E., 1994. *Managing Corporate Communication*, Allen & Unwin, Sydney.

[81] Johnson, C., 1994. *The Empowerement of Asia.* Australian Centre for American Studies, Sydney.

[82] Judd, C. M, and Kulik, J. A., 1980. Schematic effects of social attitudes on information processing and recall. *Journal of Personality and Social Psychology, 38*, 369–578.

[83] Kamps, Jaap and Marx, Maarten, 2002. Words with attitude. *Proceedings of the First International Conference on Global Wordnet*, CIIL, Mysore, India, 332–341.

[84] Karlgren, Jussi and Cutting, Douglass, 1994. Recognizing text genres with simple metrics usign discriminant analysis. *Proceedings of the Fifteenth International Conference on Computational Linguistics (COLING-94)*, 1071–1075.

[85] Kaplan, K. L., 1972. On the ambivalence-indifference problem in attitude theory and measurement: A suggested modification of the semantic differential technique. *Psychological Bulletin*, 77, 361–372.

[86] Kennedy, Alistair and Inkpen, Diana, 2005. Sentiment classification of movie and product reviews using contextual valence shifters. In *Proceedings of FINEXIN 2005, Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations*, Ottawa, May 2005.

[87] Kessler, Brett, Numberg, Geoffrey, and Schutze, Hinrich, 1997. Automatic detection of text genre. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, 32–38.

[88] Kim, Soo-Min and Hovy, Eduard, 2004. Determining the sentiment of opinions. *Coling 2004*.

[89] Kluver, Randy, 2000. Globalization, Informatization, and Intercultural Communication, *American Communication Journal*, 3(3).

[90] Kohavi, Ron, 1995. A study of cross-validation and bootstrap for accuracy estimatino and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence.* San Mateo, CA: Morgan Kaufmann, 1995. 1137–1143. http://citeseer.ist.psu.edu/kohavi95study.html (last visited on April 20, 2008).

[91] Krech, D. and Crutchfield, R. S., 1948. *Theory and Problems of Social Psychology.* New York: McGraw-Hill.

[92] Krishnamurthy, Sandeep, 2002. The multidimensionality of blog conversations: the virtual enactment of September 11 *Internet Research 3.0*, Maastricht, The Netherlands, October, 2002.

[93] Ku, Lun-Wei, Liang, Yu-Ting, and Chen, Hsin-Hsi, 2006. Opinion extraction, summarization and tracking in news and blog corpora. *AAAI Spring Symposium Technical Report SS-06-03*, Palo Alto, California, 2006.

[94] Ku, Lun-Wei, Wu, Tung-Ho, Lee, Li-Ying and Chen, Hsin-Hsi, 2005. Construction of an evaluation corpus for opinion extraction. *Proceedings of NTCIR-5 Workshop*, Tokyo, Japan, 2005. 513–520.

[95] Larsen, J. T., McGraw, A. P., and Cacioppo, J. T., 2001. Can people feel happy and sad at the same time? *Journal of Personaligy and Social Psychology*, 81, 684–696.

[96] Lasica, J. D., 2001. Blogging as a form of journalism. *USC Annenberg Online Journalism Review.*

[97] Lasica, J. D., 2001. Weblogs: A new source of news. *USC Annenberg Online Journalism Review.*

[98] Lavine, Howard, 2004. Attitude ambivalence in the realm of politics. In Geoffrey Haddock and Gregory R. Maio (Eds.), *Contemporary Perspectives on the Psychology of Attitudes.* New York: Psychology Press. 94–119.

[99] Levow, G., Oard, D, and Resnik, P., 2005. Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management: Special Issue on Cross-language Information Retrieval*, 41(4), 523–547, 2005.

[100] Levy, Roger and Manning, Christopher D., 2003. Is it harder to parse Chinese, or the Chinese treebank? *ACL 2003.*

[101] Lewis, D., 1991. Evaluating text categorization. *HLT Workshop on Speech and Natural Language*, 312–318. http://acl.ldc.upenn.edu/H/H91/H91-1061.pdf (last visited on April 20, 2008).

[102] Likert, R., 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 140, 5-53.

[103] Lin, Dekang, 1998. Automatic retrieval and clustering of similarwords. *Proceedings of COLING-ACL'98*, 768-773.

[104] Lin, Dekang, 1998. Dependency-based evaluation of MINIPAR. *Workshop on the Evaluation of Parsing Systems*, Granada, Spain.

[105] Liu, Bing, Hu, Minqing, and Cheng, Junsheng, 2005. Opinion observer: Analyzing and comparing opinions on the web. *WWW 2005*. May 10-14, 2005, Chiba, Japan.

[106] Lustig, M. and Koester, J., 1993. *Intercultural Competence: Interpersonal Commuinication Across Cultures*. Harper Collins, New York.

[107] Lyman, Peter, Varian, Hal R., et al., 2003. How much information? 2003. Regents of the University of California. October 27, 2003.

[108] Lyons, John, 1977. *Semantics*. Volumn 1. Cambridge, England: Cambridge University Press.

[109] Macleod, Catherine, Grishman, Ralph and Meyers, Adam, 1998. *COMLEX Syntax Reference Manual*, Proteus Project, NYU.

[110] Maio, Gregory R., Bell, D. W., and Esses, Victoria M., 1996. Ambivalence and persuasion: The processing of messages about immigrant groups. *Journal of Experimental Social Psychology*, 32, 513-536.

[111] Maio, Gregory R. and Esses, Victoria M., et al., 2004. The function-struture model of attitudes - incorporating the need for affect. In Geoffrey Haddock and Gregory R. Maio (Eds.), *Contemporary Perspectives on the Psychology of Attitudes*. New York: Psychology Press, 9-33.

[112] Maio, Gregory R. and Haddock, Geoffrey, 2004. Theories of attitude. In Geoffrey Haddock and Gregory R. Maio (Eds.), *Contemporary Perspectives on the Psychology of Attitudes*. New York: Psychology Press.

[113] Manning, Christopher D. and Schutze, Hinrich, 2000. *Foundations of Statistical Natural Language Processing*. Chapter 16, Text Categorization. Cambridge, Massachusetts: The MIT Press.

[114] Marneffe, Marie-Catherine de, MacCartney, Bill, and Manning, Christopher D., 2006. Generating typed dependency parsers from phrase structure parses. *LREC 2006*.

[115] Moore, David S. and McCabe, George D., 1989. Introduction to the Practice of Statistics (Chapter 18), 5th Edition. New York: W. H. Freeman. available at: http://www.insightful.com/Hesterberg/bootstrap/ (last visited on April 20, 2008).

[116] McNair, Brian, 1998. *The Sociology of Journalism*. New York: Arnold.

[117] Mencher, Melvin 1991. *News Reporting and Writing*, 5th edition. Dubuque, IA: Wm. C. Brown Publishers.

[118] Metzler, D. and Croft, W.B., 2004. Combining the language model and inference network approaches to retrieval. *Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval*, 40(5), 735–750.

[119] Mihalcea, Rada, Banea, Carmen, and Wiebe, Jance, 2007. Learning multilingual subjective language via cross-lingual projections. *Proceedings of the Association for Computational Linguistics (ACL 2007)*, Prague, June 23–30, 2007. 976–983. http://www.cs.unt.edu/˜rada/papers.html (last visited April 22, 2008).

[120] Miller, George A., 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4). Speicial Issue. 235–312.

[121] Morinaga, Satoshi, Yamanishi, Kenji, et al, 2002. Mining product reputations on the web. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002).*

[122] Mullen, Tony and Collier, Nigel, 2004. Sentiment Analysis using Support Vector Machines with Diverse Information Sources. *Proceedings of EMNLP 2004.*

[123] Nasukawa, Tetsuya and Yi, Jeonghee, 2003. Sentiment analysis: capturing favorability using natural language processing. *K-CAP'03*, October 23-25, 2003, Sanibel Island, Florida, USA.

[124] Nowson, Scott, Oberlander, Jon, and Gill, Alastair J., 2005. Weblogs, Genres, and Individual Differences. *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, Stresa, Italy, 2005.

[125] Oard, Douglas W., 1996. Adaptive vector space text filtering for monolingual and cross-language applications. Ph.D. Dissertation, University of Maryland, College Park.

[126] Oard, Douglas W., 2007. Unpublished correspondence on statistical significance tests for bootstrapped precision and recall.

[127] Oard, Douglas W., Elsayed, Tamer, Wang, Jianqiang, Wu, Yejun, et al., 2006. TREC-2006 at Maryland: Blog, Enterprise, Legal and QA Tracks. *TREC 2006*. Gaithersburg, Maryland, November 14–17, 2006.

[128] Och, F. J. and Ney, H., 2000. Improved statistical alighment models. *Proceedings of the 38th Annual Conference of the Association for Computational Linguistics*, Hongkong, China, October 2000. 440–447.

[129] Olsson, J. Scott, 2006. An analysis of the coupling between training set and neighborhood sizes for the kNN classifier. SIGIR 2006, SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.

[130] Olsson, J. Scott, Oard, Douglas W., and Hajic, Jan, 2005. Cross-language text classification. *Proceedings of SIGIR 2005*, August 15–19, Salvador, Brazil. 645–646.

[131] Osgood, C. E. and Tannenbaum, P. H., 1955. The principle of congruity in the prediction of attitude change. *Psychological Review*, 1955, 62, 42–55.

[132] Oskamp, Stuart, 1991. *Attitudes and opinions*, 2nd Ed. Englewood Cliffs, New Jersey: Prentice Hall.

[133] Pang, Bo, Lee, Lillian, and Vaithyanathan, Shivakumar, 2002. Thumbs up? sentiment classification using machine learning techniques. *Proceedings of EMNLP 2002*, 79–86.

[134] Pang, Bo and Lee, Lillian, 2005. See stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *ACL 2005.*

[135] Petty, Richard E. and Cacioppo, John T., 1981. *Attitudes and Persuasion: Classic and Contemporary Approaches*, Dubuque, Iowa: Wm. C. Brown Company Publishers.

[136] Petty, R. E., Wegener, D. T., and Fabrigar, L. R., 1997. Attitudes and attitude change. *Annual Review of Psychology*, 48, 609–647.

[137] Pratkanis, A. R. , 1989. The cognitive representation of attitudes. In A. R. Pratkanis, S. J. Breckler, and A. G. Greenwald (Eds.), *Attitude Structure and Function*, Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 71–98.

[138] Qu, Yan and Furnas, George W. , 2004 *CHI 2005.* April 2-7, 2004. Portland, Oregon, USA.

[139] Quirk, Randolph, Greenbaum, Sidney, Leech, Geofrey, and Svartvik, Jan, 1985. *A Comprehensive Grammar of the English Language*. New York: Longman.

[140] Rasmussen, Edie, 1992. Clustering algorithms. In William B. Frakes and Ricardo Baeza-Yates (Eds.), *Information Retrieval Structures & Algorithms*. 1992. Upper Saddle River, New Jersey: Prentice Hall, Inc.

[141] Ray, Tiernan, 2003. Why blogs haven't stormed the business world. E-Commerce Times, available at: http://www.ecommercetimes.com/perl/story/21389.html, April 29, 2003 (last visited on November 2, 2008).

[142] Riggs, Tracy and Wilensky, Robert , 2001. An algorithm for automated rating of reviewers. *JCDL'01*, June 24-28, 2001, Roanoke, Virginia.

[143] Rigutini, Leonardo, Maggini, Marco, and Liu, Bing, 2005. An EM based training algorithm for cross-language text classification. *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*. September 2005, Compiegne, France. 7 pages.

[144] Riloff, Ellen, 1996. Automatically Generating Extraction Patterns from Untagged Text. *AAAI-96*, 1044–1049.

[145] Riloff, Ellen and Wiebe, Janyce, 2003. Learning extraction patterns for subjective expressions. *EMNLP-2003*.

[146] Robertson, S., Walker, S, Sparck-Jones, K., Hancock-Beaulieu, M. M., 1996. Okapi at TREC-3. *Text Retrieval Conference.* http://citeseer.ist.psu.edu/robertson96okapi.html (last visited on April 20, 2008).

[147] Ropp, P. (Ed.), 1990. *Heritage of China: Contemporary Perspectives on Chinese Civilization.* University of California Press, Berkeley, California.

[148] Russell, Daniel M., Stefik, Mark J., Pirolli, Peter, Card, Stuart K., 1993 The cost structure of sensemaking. *INTERCHI'93.* April 1993. 269–276.

[149] Salvetti, Franco, Lewis, Stephen, and Reichenbach, Christoph, 2004. Automatic opinion polarity classification of movie reviews. *Colorado Research in Linguistics*, 17(1), June 2004.

[150] Savolainen, Reijo, 1993. The sense-making theory: reviewing the interests of a user-centered approach to information seeking and use. *Information Processing & Management*, 29(1), 13–28.

[151] Savolainen, Reijo , 1999. Information use, gap-bridging and sense-making. Paper present as a non-divisional workshop held at the meeting of the International Commuication Association, San Francisco, May 1999. http://communication.sbs.ohio-state.edu/sense-making/meet/1999/meet99savolainen.html (last visited on April 20, 2008)

[152] Schapire, R. E. and Singer, Y. , 2000. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39 (2/3), 135–168.

[153] Schwartz, B., 1985. China's Cultural Values. Occassional Paper No. 18, Center for Asian Studies, Arizona State University, Tempe, Arizona.

[154] Schwartz, S. H., 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology*, 25. New York: Academic Press. 1–65.

[155] Schwarz, Norbert, Groves, Robert M., and Schuman, Howard, 1996. Survey Methods. In The Handbook of Social Psychology, McGraw Hill. 1–37.

[156] Sebastiani, Fabrizio, 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), March 2002. 1–47.

[157] Seki, Yohei, Evans, David Kirk, and Ku, Lun-Wei, et al., 2007. Overview of opinion analysis pilot task at NTCIR-6. *Proceedings of NTCIR-6 Workshop Meeting*, May 15–18, 2007, Tokyo, Japan. 265–278.

[158] Shavitt, Sharon and Brock, Timothy C. (Eds.), 1994. *Persuasion - Psychological Insights and Perspectives.* Boston: Allyn and Bacon.

[159] Shi, Jilin and Zhu, Yinggui (Ed.). *Bao Yi Ci Ci Dian (Positive Dictionary).* Sichuan Dictionary Press, Chengdu, Sichuan, China. 2005.

[160] Slavens, Thomas P., 2003. *Reference interviews, questions, and materials*, 3rd Ed., Lanham, MD: Scarecrow Press, Inc.

[161] Soboroff, Ian, 2004. On evaluating Web Ssearch with very few relevant documents. *SIGIR'04*, July 25-29, 2004, Sheffield, South Yorkshire, UK.

[162] Sparck-Jones, K., Walker, S. and Robertson, S. E., 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 1 and Part 2. *Information Processing & Management*, 36(6), 779–808, 809–840.

[163] Spertus, Ellen, 1997. Smokey: automatic recognition of hostile messages. In *Innovative Applications of Artificial Intelligence (IAAAI'97).*

[164] Steinbach, Michael, Karypis, George, and Kumar, Vipin, 2000. A comparison of document clustering techniques. *KDD Workshop on Text Mining, 2000.* http://www-users.cs.umn.edu/ karypis/publications/Papers/PDF/doccluster.pdf

[165] Stewart, J., 1991. A postmoden look at traditional communication postulates. *Western Journal of Speech Communication*, 55, Fall, 354–79.

[166] Stone, Philip J., Dunphy, Dexter C.,Smith, Marshall S. , and Ogilvie, Daniel M., 1966. The General Inquirer: A Computer Approach to Content Analysis. Cambridge, Massachusetts: The MIT Press. The General Inquirer lexicon is available for reserach at http://www.wjh.harvard.edu/˜inquirer/ (last visited on April 20, 2008).

[167] Stoyanov, Ves, Cardie, Claire and Wiebe, Janyce, 2005. Multi-perspective question answering using the OpQA corpus. *HLT/EMNLP 2005*, October 6–8, 2005.

[168] Tajfel, H., 1959. Quantitative judgment in social perception. *British Journal of Psychology*, 50, 16–29.

[169] Tao, J., 1990. The Chinese moral ethos and the concept of individual rights. *Journal of Applied Philosophy*, 7(2), pp, 119–127.

[170] Taylor, Robert S., 1968. Question negotiation and information seeking in libraries. *College and Research Libraries, 29*, 178–189.

[171] Thompson, D., 1996. *The Pocket Oxford Dictionary of Current English.* Oxford: Oxford University Press.

[172] Thompson, M. M., Zanna, M. P., and Griffin, D. W., 1995. Let's not be indifferent about (attitudinal) ambivalence. In R. E. Petty and J. A. Krosnick (Eds.), *Attitude Strength: Antecedents and Consequences.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc. pp. 361–385.

[173] Thurstone, L. L., 1928. Attitudes can be measured. *American Journal of Sociology*, 33, 529–544.

[174] Thurstone, L. L., 1931. The measurement of social attitudes. *Journal of Abnormal and Social Psychology*, 26, 249–269.

[175] Tourangeau, R., Rips, L., and Rasinksi, K., 2000. An introduction and a point of view. In *The Psychology of Survey Response.* New York: Cambridge University Press.

[176] Tseng, Huihsin, Chang, Pichuan, Andrew, Galen, Jurafsky, Daniel and Manning, Christopher. A Conditional Random Field Word Segmenter. In *Fourth SIGHAN Workshop on Chinese Language Processing.* 2005.

[177] Turney, Peter, 2002. Thumbs up or thumb down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL-2002)*, 2002, 417–424.

[178] Turney, Peter and Littman, Michael, 2002. Unsupervised learning of semantic orientation from a hundred-billion-word courpus. Technical Report ERB-1094, National Research Council of Canada, May 15, 2002.

[179] Turney, Peter and Littman, Michael, 2003. Measuring praise and criticism: inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21. 315–346.

[180] Upshaw, H. S. and Ostrom, T. M., 1984. Psychological perspective in attitude research. In *J. R. Eiser (Ed.), Attitudinal Judgment.* New York: Springer-Verlag.

[181] van der Pligt, J. and Eiser, J. Richard, 1984. Dimensional salience, judgment, and attitudes. In J. Richard Eiser (Ed.), *Attitudinal Judgment.* New York: Springer-Verlag.

[182] Walther, Joseph B., 1996. Computer-mediated communication: Impersonal, interpersonal, adn hyperpersonal interaction. *Communication Research*, 23(1), 3–34.

[183] Wang, Jianqiang, 2005. *Matching Meaning for Cross-Language Information Retrieval*. Ph.D. thesis, University of Maryland, College Park.

[184] Wang, Jianqiang and Oard, Douglas W., 2006. Combining bidirectional translation and synonym for cross-language information retrieval. SIGIR'06, August 6–11, 2006, Seattle, Washington, USA., 202–209.

[185] Wayne, Charles L., 2000. Topic detection and tracking in English and Chinese *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, 2000, Hong Kong, China, 165–172.

[186] Weaver, David H. & Wu, Mei, 1998. *The American Journalists in the 1990s: U.S. News People at the End of an Era*. Mahawah, NJ: Lawrence Erlbaum.

[187] Merriam-Webster's Words of the Year 2004. Mirriam-Webster Online. Available at http://www.m-w.com/info/04words.htm (last visited on October 5, 2008).

[188] Weiss, S. M. and Kulikowski, C. A., 1990. *Computer Systems That Learn*. Morgan Kaufmann.

[189] Wiebe, Janyce, 2000. Learning subjective adjectives from corpora. *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)*, 2000, 735–740.

[190] Wiebe, Janyce and Mihalcea, Rada, 2006. Word sense and subjectivity. *Proceedings of COLING-ACL 2006*.

[191] Wiebe, Janyce and Riloff, Ellen, 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-05)*, 2005.

[192] Wiebe, Jance, Wilson, Theresa, and Bell, Matthew, 2001. Identifying collocations for recognizing opinions. *Proceedings of ACL/EACL'01 Workshop on Collocation*, Toulouse, France, July 2001.

[193] Wiebe, Janyce, Wilson, Theresa, Bruce, Rebecca, et al, 2002. Learning Subjective Language. Tech Report TR-02-100, Dept. of Comp. Science, Univ. of Pittsburg.

[194] Wilks, Yorick and Stevenson, Mark. 1998. The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Journal of Natural Language Engineering*, 4(2), 135–144.

[195] Wilson, Theresa, Wiebe, Janyce, and Hoffmann, Paul, 2005. Recognizing contextual polarity in phrase-level sentiment analysis. *HLT-EMNLP*, 2005.

[196] Wilson, Theresa, Wiebe, Janyce and Hwa, Rebecca, 2004. Just how mad are you? Finding strong and weak opinion clauses. *AAAI-2004*.

[197] Wilson, Theresa, Wiebe, Janyce, and Hoffmann, Paul, 2005. Recognizing contextual polarity in phrase-level sentiment analysis. *HLT-EMNLP 2005.*

[198] Wu, Mingfang, Fuller, Michael, and Wilkinson, Ross, 2000. Teh role of a judge in a user based retrieval experiment. *SIGIR 2000*, Athens, Grece. 331-333.

[199] Wu, Yejun and Oard, Douglas W., 2007. NTCIR-6 at Maryland: Chinese opinion analysis pilot task. *Proceedings of the 6th NTCIR Workshop on Evaluation of Information Access Technologies.* May 2007, Tokyo, Japan. 344-349

[200] Xu, Ruifeng, Wong Kam-Fai, and Xia, Yunqing, 2007. Opinmine - opinon analysis system by CUHK for NTCIR-6 Pilot Task. *Proceedings of NTCIR-6 Workshop Meeting*, May 15–18, Tokyo, Japan. 350–357.

[201] Yang, Xipeng, 2003. Thoughts on roots and affixes. *Chinese Language Learning*, 2, 2003 (in Chinese).

[202] Yang, Yiming, Ault, Tom, et al., 2000. Improving text categorization methods for event tracking. In *SIGIR 2000*, Athens, Greece.

[203] Yang, Yiming and Liu, Xin, 1999. A re-examination of text categorization methods. *The 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, 42–49.

[204] Yang, Yiming and Pedersen, Jan O., 1997. A comparison study on feature selection in text categorization. *ICML'97*

[205] Yang, Ling and Zhu, Yinggui (Ed.). *Bian Yi Ci Ci Dian (Negative Dictionary).* Sichuan Dictionary Press, Chengdu, Sichuan, China. 2005.

[206] Yi, Jeonghee, Nasukawa, Tetsuya, Bunescu, Razvan, and Niblack, Wayne, 2003. Sentiment Analyzer: Extracting of Sentiments towards a Given Topic using NLP Techniques. In *The Third IEEE International Conference on Data Mining (ICDM'03)*, Nov. 2003.

[207] Yu, Hong and Hatzivassiloglou, Vasileios, 2003. Toward answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. *EMNLP-2003*, 129–136.

[208] Zajonc, R. B. 1980. Feeling and thinking: Preferences need no inferences. American Psychologist, 35, 151–175.

[209] Zajonc, R. B. 1984. On primacy of affect. In Klaus R. Scherer and Paul Ekman (Ed.), *Approaches to Emotion*, Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

[210] Zannar, M. P. and Rempel, J. K. 1988. Attitudes: a new look at an old concept. In D. Bar-Tal and A. W. Kruglanski (Eds.), *The Social Psychology of Knowledge*, Cambridge, UK: Cambridge University Press, 315–334.

[211] Zhong, Bu. 2006. *Searching for Meaning: Multi-level Coginitive Processing of News Decision Making among U.S. and Chinese Journalists.* Doctoral Dissertation, University of Maryland, College Park, 2006.

[212] Zimmerman, D. P. 1987. Effects of computer conferencing on the language use of emotionally distributed adolescents. *Behavior Research Methods, Instruments, & Computers*, 19, 224–230.