ABSTRACT

Title of thesis: An Information Retrieval Test Collection for English SMS Conversations

Rashmi Sankepally
Master of Information Management, 2015

Thesis directed by: Dr. Douglas W. Oard
College of Information Studies

Information retrieval research for informal conversational settings differs in important ways from the more traditional goal of document retrieval. The goal of this research is to build an information retrieval test collection from informal conversational messages and to demonstrate the use of that collection to compare the retrieval effectiveness of some information retrieval systems. The test collection is based on the Linguistic Data Consortium's collection of more than 8,000 English SMS (Short Message Service) conversations, which contain more than 120,000 individual messages. The collection is described, followed by a description of the processes for creating and collecting topics, performing relevance judgments, and establishing baseline results. The findings indicate that traditional approaches for building information retrieval test collections can reasonably be applied to pre-clustered SMS conversations, but that the process of creating relevance judgments is somewhat more challenging and thus the reliable detection of differences in system effectiveness is somewhat more complex.

AN INFORMATION RETRIEVAL TEST COLLECTION FOR ENGLISH SMS CONVERSATIONS

by

RASHMI SANKEPALLY

Masters thesis 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
Master of Information Management
2015

Advisory Committee:
Professor Douglas W. Oard, Chair/Advisor
Professor Brian Butler
Associate Professor Jimmy Lin

Acknowledgments

First and foremost I'd like to thank my advisor, Dr. Doug Oard for giving me an invaluable opportunity to work on very interesting projects over the past two years. He has always been available for help and advice and has pulled me through the most difficult odds at times. It has been a pleasure to work with and learn from such an exceptional individual and I look forward to work with him for my doctoral studies.

I would also like to thank Professor Brian Butler and Professor Jimmy Lin for agreeing to serve on my thesis committee, for their suggestions during the proposal and for sparing their invaluable time reviewing this thesis.

I would also like to acknowledge help and support from the CLIP lab members. Especially, Jiaul Paik's help in system development and contribution to different ideas is gratefully acknowledged. Thanks to Mossaab Bagdouri for his help in setting up the relevance judgment platform.

I owe my deepest thanks to my family - my mother, father and sister and friends who have always stood by me and guided me. Special thanks to Padma Malini Devi Dasi and Deva Prastha Dasa for their affection and blessings.

I would like to acknowledge financial support from the National Science Foundation (NSF) and from the Defense Advanced Research Projects Agency (DARPA) for funding my masters education and research.

It is impossible to mention all that I am grateful to, and I apologize to those I have left out. Thank you all and thank God!

Table of Contents

List of Abbreviations

IR	Information Retrieval
LDC	Linguistic Data Consortium
TREC	Terabyte REtrieval Conference
CLEF	Conference and Labs of the Evaluation Forum
NII	National Institute of Informatics
NTCIR	NII Testbeds and Community for Information access Research
SMS	Short Message Service
AP	Average Precision
MAP	Mean Average Precision
nDCG	normalized Discounted Cumulative Gain

Chapter 1: Introduction

Text processing and research on short text has caught on over the past decade with the increasing popularity of microblogging sites and people's willingness to express their feelings, thoughts, opinions and ideas in the form of short text. With the ubiquity of smartphone usage, many people have turned to using short messaging services like Whatsapp, Viber, Telegram, etc., to drop in quick messages that are archived and do not require immediate response. Such services are also being extensively used for real-time chatting due to their simplicity and convenience.

It is interesting to capture information from this kind of content for the following main reasons among others:

- It is reflective of the real intentions of people as it is an intimate exchange of messages between two parties who are often well-known to each other, as against being a mere portrayal of oneself in front of the general public as is the case with Facebook posts, tweets, etc.

- Information retrieval research has to move from formal settings to more informal settings. SMS could be a good genre to explore more in this area.

- There is much more information generated in conversations than through any

other means. Pew Research says that an average text messaging user in the USA sends or receives about 40 messages per day [1].

- Many people have recently become interested in lifelogging, the hobby of recording large portions of their lives using various means. Creating an effective search system for their historical SMS (Short Message Service) conversations could prove useful in this.

- No such research has been done in the past due to lack of availability of data and the myriad privacy concerns involved.

There has not been an information retrieval shared task using SMS conversations as the document collection. Hence, there is a requirement for a test collection to begin exploring such options. SMS text is informal and characterized by a number of abbreviations, acronyms, spelling errors, extra punctuations, emoticons and lack of grammar.

Example of a typical SMS text:
C U @7 in the longue then. :) HAND!!!
(Normal English translation: See you at 7 in the lounge then..:Smiley: Have a nice day!)

The Linguistic Data Consortium (LDC) at the University of Pennsylvania has made collections of various genres of data like chat, SMS, discussion forums and blogs in three different languages (English, Chinese and Egyptian Arabic) as part of the

BOLT (Broad Operational Language Technology) program. The initiative is funded by DARPA (Defense Advanced Research Projects Agency), USA. The LDC's collection of SMS/chat genre in English is used in the current work.

1.1 Research Questions

The main research goal of this work is to explore the possibility of building an Information Retrieval (IR) test collection from the collection of SMS conversations in such a way that it is useful for evaluating retrieval systems. With that goal, current work addresses the following research questions:

1. How to build an information retrieval test collection of informal conversational messages?

2. Is such a test collection useful for evaluating future retrieval systems designed for such informal conversations?

1.2 Contributions of the Thesis

The contributions of this thesis are that we have:

- Formalized a set of topics for the collection,

- Obtained relevance judgments for a sample of SMS conversations for each topic, and

- Shown that reliable comparisons can be done using the test collection

1.3 Organization of the Thesis

The next chapter provides background on the process used to create test collections in the past and also describes findings from prior work that are relevant to the current work. Chapter 3 discusses exploratory work done on the SMS collection. Chapter 4 details the process of building an IR test collection for the SMS conversations. Chapter 5 summarizes our findings and conclusions from this research.

Chapter 2: Building IR Test Collections

2.1 Overview

Reviewing the standard methodologies followed for building various information retrieval test collections offers an understanding that is crucial for extending and refining different stages of that process for informal conversational text.

IR test collections have been made for various purposes with various kinds of document collections in the past. Particularly, TREC (the Text REtrieval Conference) was instituted in 1992 with one of its goals being to develop large and reusable test collections for producing appropriate evaluation resources and to encourage IR research on large collection retrieval experiments. In addition NTCIR (NII Testbeds and Community for Information access Research) and CLEF (formerly the Cross-Language Evaluation Forum and now known as Conference and Labs of the Evaluation Forum) was instituted subsequently in the years 1997 and 2000 respectively. They had similar goals as TREC, to create reusable test collections for evaluation of different systems in the domain of Information Retrieval. As is discussed in the following section, work done in various tracks of these forums serve as good background for the current work.

2.2 Experimental and Computational Background

Below we discuss some previous IR research work done in informal conversational contexts.

Oard et al [2] have made a reusable test collection from audio recordings of interviews collected from Holocaust survivors. They used search-guided relevance assessments for making relevance judgments in about 10,000 thematic segments from 625 hours of interviews with 246 individuals. More than 100 topics were developed from actual user requests.

A line of work focuses on retrieval experiments on tweets. The TREC Microblog Track [3] uses a test collection consisting of 16 million tweets and 60 queries. In addition, it uses time stamp information for the user's search queries and the tweets to get more recent (for the adhoc retrieval task) and subsequent (filtering task) tweets. Traditional IR techniques of pooling and collecting relevance assessments were employed to create the reusable test collection.

More recently a pilot research study has been done in retrieving opinions from discussion forum threads by Dietz et al [4]. They tested a range of forum retrieval techniques to differentiate between opinionated and factual forum posts. They performed their experiments on a subset of forum data consisting of 262,000 threads and 5.5 million posts provided by the LDC. Their approach uses the highest among all passage level judgments of a document as the measure of document relevance. Further, they also collected 10 million Wikipedia articles and 134 million news documents for experimenting with different query expansion techniques. They ob-

serve that many successful adhoc retrieval techniques are only as good as baseline techniques for this task. Different query expansion techniques in combination with pseudo relevance feedback models (RM3) have yielded good results.

The TREC Web track [5] focuses on effectiveness and robustness in its risk sensitive task. It has a test collection of 1 billion pages and around 100 topics. One important consideration of this track is that they have topics that reflect aspects of authentic Web usage. They developed topics from the logs and data resources of commercial search engines. The Question Answering (QA) TREC track [6] attempted to create a test collection from a document collection of 528,000 articles from popular news agencies and 200 fact-based, short-answer questions. They found that their test collection was stable to the extent that it yielded good correlations between the system rankings obtained from using different qrels (query relevance files, which consist of judgments for a sample of topic-document pairs) namely, mutiple-judge qrels (some function of assessments from multiple assessors) and 1-judge qrels (assessments from a single assessor), thereby establishing the validity of their evaluation. But at the same time, the test collection was not reusable. This was as a result of assessors having different opinions about whether a given answer string correctly answered a question.

In addition, Borlund's work on the concept of relevance [7] acknowledges and formalizes the multi-dimensional, dynamic nature of relevance. It concludes that relevance should be judged in relation to information need rather than according to the query (words used to describe the information need). It suggests the use of simulated work tasks and graded relevance assessments for robust IR evaluation.

2.3 Summary

The main takeaways from the related prior work are:

- It is important for the test collection to have topics that reflect real-user intentions.
- Performing relevance assessment with clear guidelines using graded relevance scores has been shown to give more stable results.
- Using traditional IR techniques for creating test collections have shown to yield reasonable results in various informal settings.

Applicability of these approaches for query formulation, pooling, relevance judgments, and evaluation of systems on the SMS conversation collection is explored in the current work.

Chapter 3: SMS Document Collection

3.1 Overview

This chapter discusses the initial exploratory work done to understand the nature of the document collection. It describes the collection and packaging efforts by Song et al [8] and goes on to detail the cleaning and analysis work performed on it in order enable its use for further experiments.

The SMS/chat document collection includes a combination of donated and collected messages from recruited participants. The English collection was released in three phases (R1, R2 and R3). These conversations were collected from two sources: either from the LDC's collection platform or from donations made by participants of their SMS or chat archives on the LDC online portal [8].

3.2 The Document Collection

For donating messages, participants were asked to make their contribution of messages on an online service where they had the option of redacting specific content from their messages before submitting. In addition, conversations were collected using a collection platform. The collection platform initiates the conversation by

sending a message to both the participants, who may or may not be known to each other before. It then records and stores all the messages that they exchange thereafter.

The collection includes SMS as well as chat conversations. The SMS conversations contain mobile messages from participants while the chats include instant messaging messages. Both SMS and chat conversations include conversations about various topics, which were not suggested by the collectors. An extract from an SMS conversation is given below as an example:

⟨?xml version="1.0" encoding="UTF-8"? ⟩

⟨conversation id="SMS_ ENG_ 20110925.0000" medium="sms" ⟩

⟨messages ⟩

...

⟨message id="m47" medium="sms" subj_id="130015" date="2012-02-09 17:59:21 -0500" ⟩

⟨body ⟩How are things? You should come home in April got a killer deal for cirque de soliel and we need a single to round out the buy one get one! Miss ya! ⟨/body ⟩

⟨/message ⟩

⟨message id="m48" medium="sms" subj_id="130014" date="2012-02-09 18:38:36 -0500" ⟩

⟨body ⟩Ooh, tempting. But I think i'm saving my vacation days (and my travel budget) for summer.⟨/body ⟩

⟨/message ⟩

⟨message id=“m49” medium=“sms” subj_id=“130014” date=“2012-02-09 18:38:52 -0500” ⟩

⟨body ⟩Miss you too!⟨/body ⟩

⟨/message ⟩

...

⟨/messages ⟩

⟨/conversation ⟩

Of the 9,106 SMS conversations, 824 were collected and 8,282 were donated. Chat conversations are usually done at a single point of time, in one go, whereas SMS conversations may have large time breaks in between. Only the donated SMS conversations were used for all further analysis for homogeneity.

3.3 Data Formatting and Processing by LDC

A conversation is always between a pair of participants. Group messages involving more than 2 participants are broken into smaller conversations between the person who donated and each other participant in the group conversation. For example, a single archive with 10 participants would result in 9 conversations with one person in common (the participant who donated the archive).

Participant IDs and message IDs were sequentially assigned to all conversations. Participant IDs are assigned consistently within each donated archive, but the

participant IDs were not normalized across donated archives, as such information was not consistently available in the donations.

3.4 Preliminary Data Analysis

As is described in more detail in the next chapter, relevance assessments required for the test collection creation were performed in two phases - phase 1 and phase 2. The document collection differed between the two phases. For phase 1, all the 8282 donated conversations were used as the content to be searched. Among the donated SMS conversations, a total of 55 conversations have only a single participant. These conversations contain no replies from the recipient. These have been removed from the collection for phase 2 experiments. Further, the 40 longest conversations (by number of messages) are removed from the collection for phase 2. This ensured the number of messages in each conversation to lie between 2 and 303. The subset of 8,187 donated conversations that remain after these removals were used for phase 2. More analyses for the phase 2 document collection are detailed in the following:

3.4.1 Number of Conversations, Number of Messages and Number of Characters

For the 8,187 SMS conversations thus obtained, Figure 3.1 shows the sorted bar plot of number of messages for all conversations. Figure 3.2 shows the sorted bar plot of number of characters for all conversations. The collection contains a

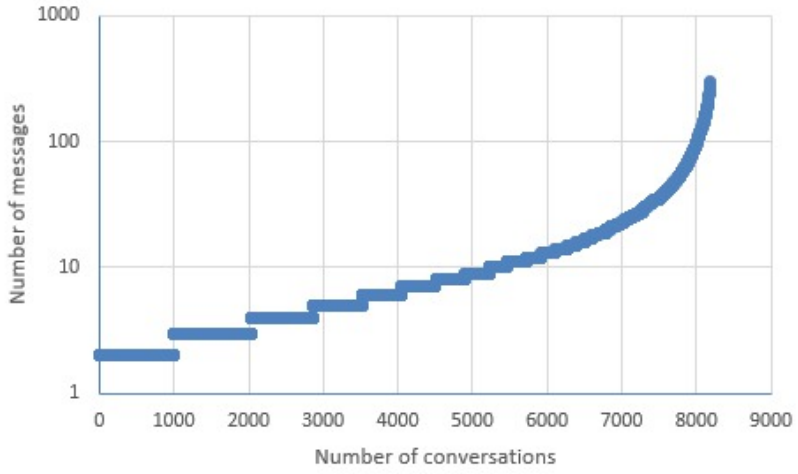

Figure 3.1: Plot of number of messages per conversation, sorted

total of 121,114 messages. 8,003 conversations have fewer than 100 messages and only 184 conversations have between 100 and 303 messages.

3.4.2 Redactions

Participants had the opportunity to redact a portion or all of their messages before donating them. These redactions appear as #'s in the messages. There are a total of 202 conversations having at least one redaction. 151 of these conversations have only one message redacted. These conversations are retained and redactions are ignored for the purpose of this collection, as we would not expect that they would seriously affect the content of the conversation.

In addition the LDC staff audited the donated SMS conversations and flagged all conversations that had personal identifying information(PII) or sensitive content. Such conversations were not included in the final release of data.

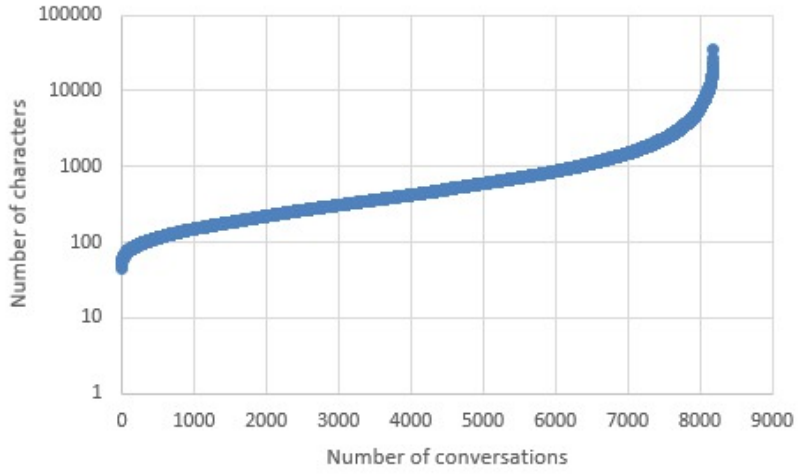

Figure 3.2: Plot of number of characters per conversation, sorted

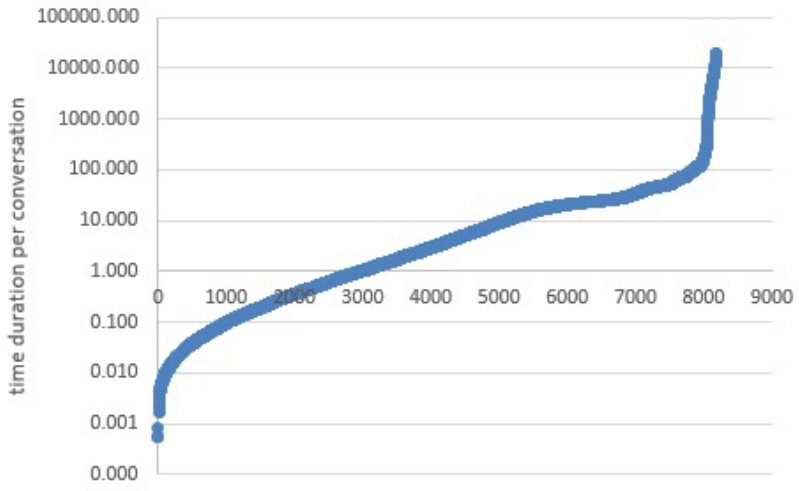

Figure 3.3: plot of conversation durations(in hours), sorted

3.4.3 Conversation Durations

The time durations for all the donated SMS are sorted and plotted in Figure 3.3. While 6,439 conversations last no more than 24 hours, 1,748 conversations last longer than 24 hours, to a maximum of 2 years.

3.5 Summary

While phase 1 experiments use 8282 set of donated messages, phase 2 experiments use 8,187 SMS conversations, each having between 2 and 303 messages. Participants had the option to redact specific messages, but they were not considered separately for this test collection because the extent of redaction was in general observed to be small. Most of the conversations lasted less than 24 hours, but about 20% of them lasted more than 24 hours.

Chapter 4: SMS IR Test Collection

4.1 Overview

An information retrieval test collection for SMS conversations has been built. Usually an IR test collection consists of the following four main components:

- Document collection
- Topics
- Results from a diversity of ranked retrieval systems
- Relevance judgments for a sample of documents, designed to support specific evaluation measures

The current chapter describes each of these components for the SMS IR test collection in detail.

4.2 Document Collection and Indexing

In our case, each document is a conversation containing a sequence of time-stamped text messages exchanged between two participants. Preliminary data analysis on the collection has been discussed in Chapter 3. We obtained relevance as-

sessments in two phases: phase 1 and phase 2. The document collection, topics and systems used for pooling differed between these phases.

For our experiments, we decided to use the 8,282 donated SMS conversations for phase 1, and 8,187 conversations (after further removing single-participant conversations and very long conversations) for phase 2. The document collection was indexed using Indri. The index enables the unit of retrieval to be either an entire conversation or a message from a conversation. We later decided that the unit of retrieval for this thesis is to be the entire conversation.

4.3 Topic Development

Topics have been developed manually based on ideas from reading about 100 of the SMS conversations and by gathering ideas from other external sources discussed in more detail below. These topics are intended to be reflective of the real-user needs from this kind of informal conversational context.

We read about 100 conversations manually. This helped us in understanding different facets of the conversations like the language usage, general type of topics discussed, some indication about the people who are conversing, and similar things. A seed set of broad topics was created based on this knowledge to aid in the topic development process.

Some topics have been developed by taking ideas from the topic listings for TREC's Robust track and TREC's Microblog. The Microblog track's topics were considered for the years 2011, 2012 and 2013 as the SMS collection has messages

whose time-stamps fall in these years, and hence might have conversations pertaining similar topics. Topics that are used for general conversations for modelling opinions, experiences and behavior of people have been used.

Each of the topics is re-written in TREC format with title, description and narrative parts. The title part is the real query and contains what a user might first type when searching on a certain topic. The description part expresses the information need in the form of a single English sentence. To adhere to the BOLT guidelines on queries [9], the description part is usually written in the form of a single question. The narrative part, which expands on the description and clarifies definitions of terms, is intended principally to aid assessors in their judgments. In addition, topics also have an extra field at the beginning called "type". The type can be 'opinion', 'experience', 'behavior' or 'knowledge'. These types are based on Oard's classification of types of questions of social media [10].

Writing queries in TREC format has two advantages: (i) it provides a useful degree of specificity to guide the relevance judgment process, and (ii) any combination of the fields can be used as queries for generating system runs.

All the topics of the collection are shown in Appendix A. The following is an example topic from the collection:

⟨top lang='en' type='experience' ⟩

⟨num ⟩009 ⟨/num ⟩

⟨title ⟩dealing with stress ⟨/title ⟩

⟨desc ⟩What are some helpful things that people do when they are stressed?

Type	Phase 1	Phase 2	Total
Opinion	3	5	8
Behavior	5	5	10
Experience	7	10	17
Knowledge	0	1	1

Table 4.1: Number of topics by type from Phase 1 and Phase 2

⟨/desc ⟩

⟨narr ⟩Relevant conversations should include information about dealing with stress that is based on real situations in which people have gone through some stressful experience due to work, courses, family pressures, health, financial difficulties or similar situations. ⟨/narr ⟩

⟨/top ⟩

In this manner 62 topics have been developed, each with its title, description and narrative parts. Indri's interactive retrieval was used to perform a preliminary manual triage for each topic to see if the collection has any relevant content. Based on this informal check, 36 topics have been selected for use in the test collection. 15 of these topics were used for phase 1 experiments, and the remaining 22 were used for phase 2 experiments. The number of topics used in each phase are given in Table 4.1.

4.4 Ranked Retrieval Systems

A diversity of ranked retrieval systems have been used to retrieve ranked lists of conversations for each topic. The following were the four types of IR models that were used:

1. Language Model (lm) :This is Indri's default retrieval model and is based on language modelling [11] and inference networks [12]. Dirichlet smoothing with $\mu = 1000$ was used.

2. Query Expansion (qe): The pseudo-relevance feedback model in Indri was used for this system. It uses a relevance model (RM3) in a language model framework. 20 documents and 30 terms were used for query expansion [13]. lm was used for retrieval using the expanded query.

3. BM25 (bm25): BM25 is a ranking function based on a probabilistic retrieval framework [14] that is widely used in information retrieval tasks. Indri has an implementation of BM25.

4. word2vec (word2vec): word2vec, introduced by Mikolov et al [15], is a tool from computational linguistics that represents the semantics of words based on vector representations from their distributions in large text collections. To enrich the pools, a system was developed that expands queries based on word2vec representations. Google's word2vec code [16] was used to train on the latest Wikipedia dump [17]. The word2vec CBAG (clustered bag of words) model with context set to 10 and number of iterations set to 5 was used. Each

query word was represented as a 100-dimensional vector. Other words whose vectors are close to the query word vectors (based on cosine similarity) are used to expand the query. lm was used for retrieval using the expanded query.

For phase 1, only the first three systems were used, without stemming. Retrieval runs were obtained for T, TD and TDN queries separately (T-title, D-description, N-narrative). This results in 9 system runs. A pooling depth of 50 was used because in past work that depth has been shown to be sufficient to highly correlate with exhaustive assessments while making obtaining human assessments affordable [18]. For phase 2, all the systems except word2vec additionally used the Porter stemmer.

4.5 Relevance Judgments - Human assessment

This section describes the process of obtaining relevance assessments for the sampled topic-conversation pairs from independent assessors. The assessment was performed in two phases (phase 1 and phase 2) in order to perform topic development, human assessment and system development in parallel.

4.5.1 Assessor Recruitment and Training

Three graduate students from the University of Maryland were recruited to perform the assessments. One of them was from China and the other two were from India. They were non-native English speakers. Two of them have been in the USA for less than a year and one of them was in USA for a year and a half. The top 10

documents for each of the 10 topics from three ranked lists obtained from simple Indri systems (using qe, bm25 and lm) were used for training. These were not used for the final qrels (query relevance file), which has final judgments for the pooled topic-conversation pairs.

4.5.2 Judgment Process

Conversations were given to the assessors in a readable format. The conversations originally in XML format (as explained in Section 3.2) were formatted to look like actual conversations between people, so they do not contain the markup tags. The topics were shown to assessors in their entirety including their title, description and narrative parts. The principal unit of judgment is the entire conversation. Assessors were asked to assign a relevance category to each topic-conversation pair that they were presented with, based on whether the information need is addressed by any part of the conversation. They had the following four options for assessment:

HREL : Conversation has highly relevant content and is worthy of being a top result for the topic.

REL : Conversation has somewhat relevant content, which may be minimal. Relevant information must be present.

NON : Conversation does not provide useful information about the topic, but may be useful for some other topic (i.e., it has some intelligible information about some topic, which may or may not appear in the set of topics provided to the assessors.)

JUNK : Conversation has no useful information for any purpose. It is either spam

or too short to convey anything useful, although it might have terms from the topic.

In addition, assessors identified all messages that they felt were relevant to the topic. These serve two purposes. First, they are useful in tracing back to see what the assessors think is relevant content. Such back-tracing was used extensively for examining cases of disagreement while characterizing inter-assessor agreement. Second, they may be used for a later exploratory study on finding 'where in the document the answer lies'.

4.5.3 Inter-assessor Agreement

Pair-wise inter-assessor agreement scores have been computed by including some duplicate topic-conversation pairs for all the assessors. In phase 1, queries 015, 017 and 020, amounting to a total of 487 topic-conversation pairs, were judged by all three assessors. In phase 2, queries 023, 024 and 032, amounting to a total of 421 topic-conversation pairs were judged by all three assessors. Cohen's kappa and positive overlap between pairs of assessors were computed on binarized judgments after treating scores of NON and JUNK as non-relevant and scores of REL and HREL as relevant.

For the first phase, inter-assessor agreements were very low. We employed two measures to compute inter-assessor agreement levels. Cohen's Kappa gives the chance corrected agreement and positive overlap is the size of intersection of relevant document sets divided by the size of union of relevant document sets [19]. Table 4.2

Pairs of assessors	Kappa measure	Positive overlap
A-B	0.109	0.103
B-C	0.187	0.119
C-A	0.175	0.113

Table 4.2: Inter-assessor agreement measurements for phase 1 relevance assessments

Pairs of assessors	Kappa measure	Positive overlap
A-B	0.328	0.208
B-C	0.203	0.125
C-A	0.498	0.355

Table 4.3: Inter assessor agreement measurements for phase 2 relevance assessments

shows the kappa and positive overlap values for the assessments. Table 4.3 contains agreement measures for phase 2. It can be seen that agreement levels have improved in phase 2. Assessors met after phase 1 to discuss cases of disagreement. This might have helped in improving their agreement scores. The agreement between assessors A and C in phase 2 has a kappa value of 0.498 which is moderate [20].

4.6 Results

To build an information retrieval test collection, the systems that contribute to the pool (the documents given to assessors to judge) must be good in several respects. This section details different ways that we adopted to measure goodness of systems.

4.6.1 Retrieval Effectiveness

Most importantly, the systems that contribute to the pools must find relevant documents. Different metrics can used to measure the retrieval effectiveness of a system. We used Mean Average Precision (MAP) and normalized Discounted Cumulative Gain (nDCG) to measure retrieval effectiveness.

In order to obtain these metrics, a qrels (query relevance) file has to be created by combining judgments from all the assessors. Despite the low agreement levels between any pair of assessors (Tables 4.2 and 4.3), experimental results from prior work show that comparative evaluation of systems may nevertheless be possible in spite of variations in multiple relevance judgments [19]. To explore this, systems from both the phases are ranked based on the following different qrels files, for the three queries for which all three assessors' judgments are available separately, for each phase. Binarization was performed by considering the assessments of JUNK and NON as non relevant and assessments REL and HREL as relevant.

- origA : Binarized judgments only from assessor A
- origB : Binarized judgments only from assessor B
- origC : Binarized judgments only from assessor C
- union : Binarized judgments obtained by the union of origA, origB and origC (i.e., relevant if any are relevant)
- intersect : Binarized judgments obtained by the intersection of origA, origB and origC (i.e., relevant only if all are relevant)

rank	origA	origB	origC	union	intersect
1	td-bm25	tdn-bm25	tdn-bm25	td-bm25	tdn-bm25
2	tdn-bm25	td-bm25	td-qe	tdn-bm25	td-qe
3	td-qe	td-qe	td-bm25	t-bm25	td-bm25
4	t-bm25	td-lm	t-qe	td-qe	td-lm
5	td-lm	t-bm25	t-lm	td-lm	t-qe
6	t-qe	t-lm	td-lm	t-qe	t-lm
7	t-lm	t-qe	t-bm25	t-lm	t-bm25
8	tdn-qe	tdn-lm	tdn-qe	tdn-lm	tdn-qe
9	tdn-lm	tdn-qe	tdn-lm	tdn-qe	tdn-lm

Table 4.4: System rankings obtained from different qrels from phase 1

Tables 4.4 and 4.6 show the system rankings (systems sorted in the decreasing order of their Mean Average Precision value) obtained from the different qrels from phase 1 and phase 2 respectively. Systems are represented using a simple naming convention - the fields (t,d,n) of topics that are used for querying followed by the name of the retrieval model. For example, td-bm25 is a system that used both title and description as the query and bm25 as the retrieval model.

System rankings are compared using two measures - (i) Kendall's τ [21], a commonly used measure for finding correlation between ranked lists, which is based on counts of concordant and discordant pairs, and (ii) Tau Average Precision or τ_{AP} [22], which is an improvement over Kendall's τ that gives more weight to concordant pairs among the best systems. Table 4.5 gives these values for system rankings obtained from different pairs of qrels.

Based on the values of τ_{AP} from phase 1, B agrees with A and C more often than any other possible pair. So, a qrels file has been constructed for all topics

qrel_pair	Kendall's τ	Tau AP (τ_{AP})
origA-origB	0.78	0.61
origB-origC	0.61	0.62
origC-origA	0.61	0.52
union-origA	0.89	0.89
union-origB	0.78	0.58
union-origC	0.50	0.45
union-intersect	0.61	0.40
intersect-origA	0.72	0.62
intersect-origB	0.72	0.71
intersect-origC	0.89	0.89

Table 4.5: Correlations between system rankings from phase 1 from different qrels

rank	origA	origB	origC	union	intersect
1	t-qe	t-qe	tdn-qe	td-qe	t-qe
2	t-lm	t-lm	tdn-lm	t-qe	t-lm
3	td-word2vec	t-word2vec	tdn-bm25	t-lm	tdn-qe
4	td-bm25	t-bm25	t-lm	tdn-qe	t-bm25
5	td-qe	tdn-qe	t-qe	td-lm	tdn-lm
6	t-bm25	tdn-lm	td-qe	t-bm25	t-word2vec
7	t-word2vec	td-word2vec	td-lm	td-bm25	td-qe
8	td-lm	td-qe	t-bm25	tdn-lm	tdn-bm25
9	tdn-qe	td-lm	td-bm25	tdn-bm25	td-lm
10	tdn-bm25	tdn-bm25	td-word2vec	td-word2vec	td-bm25
11	tdn-lm	td-bm25	t-word2vec	t-word2vec	td-word2vec
12	tdn-word2vec	tdn-word2vec	tdn-word2vec	tdn-word2vec	tdn-word2vec

Table 4.6: System rankings obtained from different qrels from phase 2

qrel_pair	Kendall's τ	Tau AP (τ_{AP})
origA-origB	0.42	0.50
origB-origC	0.29	0.11
origC-origA	-0.06	-0.11
union-origA	0.33	0.25
union-origB	0.32	0.26
union-origC	0.45	0.30
union-intersect	0.55	0.38
intersect-origA	0.21	0.36
intersect-origB	0.73	0.73
intersect-origC	0.48	0.23

Table 4.7: Correlations between system rankings of phase 2 from different qrels

in phase 1, by combining individual judgments from A, B and C and using B's judgments for cases in which all three judgments were available.

Similarly in phase 2, we see that B seems to be more consistent with A and C according to τ_{AP}. This contradicts the interassessor agreement scores (Table 4.3) obtained for B with other assessors. These fluctuating results are likely due to the small sample of 3 topics for which all three assessors' judgments were available. Note that inter-assessor agreement scores were computed on a larger sample of data (421 cases for phase 2) compared to the correlation measurements between ranked lists (12 cases for phase 2) and hence seem more reliable. So we decided to use A's jdgments for cases in which all three judgments were available, for constructing final qrels for phase 2.

With the final qrels decided, we could now calculate the effectiveness measures. For computing MAP (Mean Average Precision), judgments were binarized by considering the assessments of JUNK and NON as non-relevant and the assess-

ments of REL and HREL as relevant. Table 4.8 shows the total number of relevant conversations for each of the 15 topics from phase 1 and each of the 21 topics from phase 2. Topics 020, 026, 042, 045 have no relevant conversations in their pools and hence were not used in measuring retrieval effectiveness. For calculating nDCG (normalized Discounted Cumulative Gain), JUNK and NON assessments were taken as 0, a REL assessment was taken as 1 and a HREL assessment was taken as 2.

Table 4.9 shows MAP and nDCG values for the 14 queries from phase 1. It can seen that MAP values range between 0.31 and 0.42, indicating that on average every third document is relevant. nDCG values range between 0.49 and 0.62, indicating that systems perform as good as 50% of the best ranking , on average. These scores indicate that the systems exhibit reasonably good retrieval performance. Table 4.10 shows results from phase 2. The title (T) query systems performed better in phase 2.

4.6.2 Contribution to Pooling

It is important that systems used to build test collections enrich the assessment pools with different sets of conversations that are potentially relevant. A conversation is deemed potentially relevant if it is judged relevant by at least one assessor. There are 190 such potentially relevant conversations in phase 1 and 121 in phase 2. Tables 4.11 and 4.12 show the number of potentially relevant conversations contributed uniquely by each system respectively from phase 1 and phase 2. It can be observed that bm25 systems are good contributors in phase 1, while qe systems

topic number	**# relevant**
002	3
003	1
004	19
005	7
008	1
009	13
010	2
011	29
012	4
013	5
015	34
017	3
019	21
020	0
021	5
total	147

topic number	**# relevant**
023	2
024	12
026	0
032	8
034	2
036	6
037	1
039	9
041	2
042	0
043	27
045	0
047	3
050	1
051	16
054	1
055	3
056	1
057	6
061	1
062	9
total	110

Table 4.8: Total number of relevant conversations for each topic from phase 1 (left) and phase 2 (right)

System	MAP
td-bm25	0.418
td-qe	0.412
tdn-bm25	0.412
td-lm	0.390
t-bm25	0.360
tdn-qe	0.351
t-qe	0.329
t-lm	0.317
tdn-lm	0.313

System	nDCG
tdn-bm25	0.616
td-bm25	0.615
td-qe	0.586
t-bm25	0.571
td-lm	0.571
tdn-qe	0.547
t-qe	0.523
t-lm	0.523
tdn-lm	0.485

Table 4.9: MAP and nDCG scores of systems from phase 1 relevance assessments in decreasing order

System	MAP
t-qe	0.362
t-bm25	0.361
t-lm	0.357
td-qe	0.350
t-word2vec	0.349
td-lm	0.339
tdn-lm	0.325
td-word2vec	0.323
tdn-qe	0.322
td-bm25	0.292
tdn-bm25	0.266
tdn-word2vec	0.193

System	nDCG
t-bm25	0.523
t-qe	0.522
t-lm	0.520
td-qe	0.517
t-word2vec	0.510
td-lm	0.506
td-bm25	0.489
tdn-qe	0.475
td-word2vec	0.472
tdn-lm	0.465
tdn-bm25	0.458
tdn-word2vec	0.309

Table 4.10: MAP and nDCG scores of systems from phase 2 relevance assessments in decreasing order

system	T	TD	TDN	≥ 2
bm25	2	6	29	23
qe	0	1	1	0
lm	1	3	10	1
≥ 2	7	10	5	91

Table 4.11: Number of unique relevant conversations contributed by each system in phase 1

system	T	TD	TDN	≥ 2
bm25	0	0	6	0
qe	0	0	0	0
lm	0	0	0	0
word2vec	7	3	3	7
≥ 2	5	1	11	78

Table 4.12: Number of unique relevant conversations contributed by each system in phase 2

are poorer contributors. In phase 2, there are not many unique contributions from individual systems. Many relevant conversations have been retrieved by multiple systems. Also note that the number of potentially relevant conversations dropped from 190 for 15 topics in phase 1 to 121 for 21 topics in phase 2, indicating that assessors have been more strict in phase 2 on an average. It is encouraging to note that 20 relevant conversations were uniquely contributed by word2vec systems.

4.6.3 Reliable Detection of Differences

Ultimately, the goal of an IR test collection is to distinguish between alternative system designs. As a way to measure how different two systems are from

Better	Worse	p for AP
td-qe	tdn-lm	0.016
td-lm	tdn-lm	0.026
td-bm25	tdn-lm	0.039
tdn-bm25	tdn-lm	0.039
td-qe	td-lm	0.048

Better	Worse	p for nDCG
tdn-qe	tdn-lm	0.021
td-bm25	tdn-lm	0.026
td-qe	tdn-lm	0.033

Table 4.13: Pairs of systems that have statistically significant differences between topic-wise Average Precision and nDCG scores from phase 1 relevance assessments

each other, topic-wise Average Precision (AP) values and topic-wise normalized Discounted Cumulative gain (nDCG) values obtained from every pair of systems are compared using two tailed paired t-tests for statistical significance, for both phase 1 and phase 2 systems. The p-values for phase 1 and phase 2 systems that have statistically significant differences between their retrieved ranked lists (considering $p < 0.05$ as statistical significance) are given in Tables 4.13 and 4.14, respectively. In phase 1, It can be seen that 4 systems are significantly better than tdn-lm in terms of AP scores and 3 systems are significantly better than tdn-lm with respect to nDCG scores. In phase 2, 8 systems are statistically significantly better than tdn-word2vec in terms of Average Precision and 10 systems are better in terms of nDCG.

4.6.4 Evaluating Future Retrieval Systems

Below is the process to evaluate future retrieval systems using this collection:

- Obtain topic-wise AP and nDCG values using the final relevance judgment

Better	Worse	p for AP
t-bm25	tdn-word2vec	0.016
tdn-qe	tdn-word2vec	0.019
t-bm25	tdn-bm25	0.030
tdn-lm	tdn-word2vec	0.033
t-qe	tdn-word2vec	0.041
td-qe	tdn-word2vec	0.042
td-bm25	tdn-word2vec	0.043
t-bm25	td-bm25	0.046
t-lm	tdn-word2vec	0.046

Better	Worse	p for nDCG
tdn-qe	tdn-word2vec	0.002
td-bm25	tdn-word2vec	0.002
t-bm25	tdn-word2vec	0.004
tdn-bm25	tdn-word2vec	0.004
td-qe	tdn-word2vec	0.005
tdn-lm	tdn-word2vec	0.007
td-lm	tdn-word2vec	0.007
t-qe	tdn-word2vec	0.008
t-lm	tdn-word2vec	0.008
t-word2vec	tdn-word2vec	0.040

Table 4.14: Pairs of systems that have statistically significant differences between topic-wise Average Precision and nDCG scores from phase 2 relevance assessments

file.

- Determine whether the system is better than the baselines by performing statistical significance tests.

4.7 Discussion

It can seen from Table 4.2 that the inter-assessor agreement is very low for phase 1. After phase 1, assessors met to discuss cases of disagreement and identify reasons for this. Some of the reasons that came to light were as follows:

- Conversations may have abrupt topic shifts, sometimes making them not very comprehensible.

- Some conversations were too long and thus confused the assessors.

- The scope of some queries was not sufficiently clear, leaving room for different interpretations in some instances [7].

- The assessors are foreign nationals who have stayed in the United States for less than 2 years and hence lack the background and context to entirely understand social conventions in these conversations.

Agreement levels improved slightly in phase 2 (Table 4.3), but the agreement between assessors is not our ultimate goal. Hence, we compared system rankings from different relevance judgments to measure the effect of agreement on ranking systems. This was done only for 3 topics, however and yielded contradictory results in phase 2.

4.8 Summary

The process adopted for building a test collection for SMS conversations has been outlined. We were able to make reliable comparisons between systems despite variations in relevance judgments from multiple assessors. We have probed into the reasons for low agreements. We have given a procedure to evaluate future systems using the collection.

Chapter 5: Conclusions

5.1 Overview

We built an information retrieval test collection for a set of SMS conversations. Our findings from this research indicate that traditional approaches for building information retrieval collections can be extended to SMS conversations. We found from our process of obtaining human assessments that the conversations pose some challenges. We have outlined a procedure to evaluate systems despite these challenges.

5.2 Limitations

- The number of topics was 14 in phase 1 and 17 in phase 2. Although a total of 31 topics is larger than what Sanderson and Zobel [23] consider as small topic sets($\leq$ 25), this is still smaller than the 50 topics they recommend for the reliability of statistical significance tests.

- We only experimented with entire conversations as the units of retrieval and have not experimented with individual messages or smaller sets of messages as retrieval units.

- Pools for obtaining relevance judgments were created with limited range of systems, so future retrieval systems may retrieve results that are not represented in the present pools.

5.3 Future Work

It would be interesting to develop message retrieval systems that return smaller sets of SMS messages in place of entire conversations. We asked our assessors to mark the relevant messages. We may be able to evaluate such systems without any additional human assessment.

We now have a set of topics with relevance judgments for future experiments. Future IR systems designed for retrieval of informal conversational text can use our collection for evaluation. The queries and relevance judgments are published at: `https://www.terpconnect.umd.edu/~rashmi/SMSTestCollection.zip`

Appendix A: List of Topics

Listed below are the 36 topics that were used for the collection. The first 15 in order (numbered between 002 to 021) are phase 1 topics and the next 21 topics (numbered between 023 to 062) are phase 2 topics:

```
<top lang='en' type='opinion'>

<num> 002 </num>

<title> new Xbox release </title>

<desc> How is the new Xbox different from the older versions? </desc>

<narr> Xbox is a video gaming brand owned by Microsoft. It represents
a series of video game consoles. Xbox One has been released recently
succeeding its former versions - Xbox and Xbox 360. Relevant
conversations should contain discussions of Xbox One, including
discussions of what new features it has compared to older
models. </narr>

</top>

<top lang='en' type='opinion'>

<num> 003 </num>

<title> Bruins </title>

<desc> Do people like the Boston Bruins?  </desc>
```

<narr> To be relevant, conversations should be about Boston Bruins, a sports team. The discussion could be how the team performs, their standing, recent matches played by the team, scores and other such things. </narr>

</top>

<top lang='en' type='experience'>

<num> 004 </num>

<title> living with parents </title>

<desc> What are the pros and cons of living with parents? </desc>

<narr> People need to choose whether they should continue living with their parents after a certain age or when they have sufficient earnings to become financially independent. Relevant conversations should include discussions of what it is like to live with parents compared to living away from home. </narr>

</top>

<top lang='en' type='experience'>

<num> 005 </num>

<title> fun on weekends </title>

<desc> What do people do for fun on weekends? </desc>

<narr> Relevant conversations should describe real experiences of people spending their weekends doing exciting activities, or discussing their weekend plans for having fun. They should contain an element of recreation in them, that is more than just doing away with routine chores. Weekends could also include long weekends with a holiday following or preceding the usual weekend days (Saturday and Sunday in the USA). </narr>

</top>

<top lang='en' type='experience'>

<num> 008 </num>

<title> disc golf </title>

<desc> Where and how do people play disc golf? </desc>

<narr> Disc golf has gained popularity recently. It is played with a frisbee, but with rules similar to those of golf. Relevant conversations should contain information about how the game is played, indications of places where the game is played, or information about groups who play it. </narr>

</top>

<top lang='en' type='experience'>

<num> 009 </num>

<title> dealing with stress </title>

<desc> What are some helpful things that people do when they are stressed? </desc>

<narr> Relevant conversations should include information about dealing with stress that is based on real situations in which people have gone through some stressful experience due to work, courses, family pressures, health, financial difficulties or similar situations. </narr>

</top>

<top lang='en' type='experience'>

<num> 010 </num>

<title> surprise birthday party </title>

<desc> How do people plan surprise parties for their friends or family? </desc>

<narr> To be relevant, conversations should contain instances of people conversing about planning a surprise birthday party. Descriptions of actual surprise parties are relevant only if there is mention of the planning. </narr>

</top>

<top lang='en' type='behavior'>

<num> 011 </num>

<title> birthday ideas </title>

<desc> How do people spend their birthday? </desc>

<narr> Relevant conversations should contain indications of what a person did on his or her birthday. </narr>

</top>

<top lang='en' type='behavior'>

<num> 012 </num>

<title> disobeying rules </title>

<desc> When and why do people break rules? </desc>

<narr> There are lots of instances in which people break rules of one kind or another. For example, lots of people jaywalk even though it is illegal to do so. Relevant conversations should describe situations in which people broke a legal rule. </narr>

</top>

<top lang='en' type='experience'>

<num> 013 </num>

<title> motivation for working out </title>

<desc> What motivates people to work out? </desc>

<narr> Relevant conversations should describe reasons why people were motivated to exercise. This workout could be going to a gym, playing a sport, going out for a jog, or any similar activity. Relevant conversations may address the benefits of exercise generally, or they may indicate beneficial results from specific workout sessions.
</narr>

</top>

<top lang='en' type='opinion'>

<num> 015 </num>

<title> taking time off </title>

<desc> In what situations do people take time off from work? </desc>

<narr> Relevant conversations should include discussions about taking off from work. This may include planned holidays, sick days, so-called "personal days", or emergencies that require arriving late or departing early from work. </narr>

</top>

<top lang='en' type='experience'>

<num> 017 </num>

<title> quit smoking </title>

<desc> How do people go about quitting smoking? </desc>

<narr> What have people done to quit smoking? How do they feel after quitting? Conversations about such things are relevant. </narr>

</top>

<top lang='en' type='behavior'>

<num> 019 </num>

<title> password sharing </title>

<desc> In what situations do people share passwords? </desc>

<narr> Relevant conversations should contain instances in which people share passwords such as for wifi, netflix, any proprietary software or any other such service. </narr>

</top>

<top lang='en' type='behavior'>

<num> 020 </num>

<title> plagiarism at school </title>

<desc> In what situations do people deliberately commit academic plagiarism? </desc>

<narr> Copying homework, using online resources without proper citation and such things constitute academic plagiarism. Find instances in which people express an intention to commit plagiarism or admit they have committed plagiarism. </narr>

</top>

<top lang='en' type='behavior'>

<num> 021 </num>

<title> recycling wastes </title>

<desc> Do people want to recycle? </desc>

<narr> In US, recycling bins are found in many places. But do people actually care enough to recycle the stuff that can be recycled? Relevant conversations contain instances and discussions about recycling waste and its importance. </narr>

</top>

<top lang="en" type="opinion">

<num> 023 </num>

<title> cherry blossoms DC </title>

<desc> How do people like the National Cherry Blossom festival in Washington, DC? </desc>

<narr> The National Cherry Blossom festival is held in March or April every year at Washington, DC. To be relevant, a conversation would contain discussions of the Cherry Blossom festival in Washington, DC. </narr>

</top>

<top lang="en" type="experience">

<num> 024 </num>

<title> pet care </title>

<desc> How do people care for their pets? </desc>

<narr> Relevant conversations would have information about caring for pet animals. They might deal with cleaning, diet, exercise, veterinary care, or similar things </narr>

</top>

<top lang="en" type="experience">

<num> 026 </num>

<title> buying clothes online </title>

<desc> What are people's experiences buying clothes online? </desc>

<narr> Buying many things online is becoming increasingly common, but some people are reluctant to buy clothes online. What experience have people had with buying clothes online? Conversations that describe actual experiences would be relevant, even if they are the experiences of other people. Conversations in which people only express opinions without describing any real experiences would not be relevant. </narr>

</top>

<top lang="en" type="behavior">

<num> 032 </num>

<title> vegan diet balance </title>

<desc> How do vegans balance their diet? </desc>

<narr> What food items do vegans include in their diet in order to make it balanced and healthy? Do they take additional supplements (such as vitamin pills)? </narr>

</top>

<top lang="en" type="opinion">

<num> 034 </num>

<title> farmers markets </title>

<desc> What do people think about farmers' markets? </desc>

<narr> Farmers' markets feature a retail market where food items are sold directly by farmers to consumers. To be relevant, conversations would contain people expressing their opinions on farmers' markets. </narr>

</top>

<top lang="en" type="experience">

<num> 036 </num>

<title> selling on eBay </title>

<desc> What experience do people have with selling on eBay? </desc>

<narr> Relevant conversations would contain instances in which people discuss different details selling things on eBay. This might, for example, include what was made available for sale, how those things were marketed, or opinions grounded in experience about whether selling on eBay is generally successful and profitable. </narr>

</top>

<top lang="en" type="opinion">

<num> 037 </num>

<title> U.S. gas prices </title>

<desc> How do people explain changes in U.S. automobile gasoline prices? </desc>

<narr> Conversations that suggest reasons why U.S. gasoline prices fluctuate would be relevant, regardless of whether those suggested reasons have a factual basis. Conversations about gas prices in other countries would not be relevant. </narr>

</top>

<top lang="en" type="experience">

<num>039</num>

<title> college tuition planning </title>

<desc> How do students plan to pay their college tuition? </desc>

<narr> Tuition is the money that is paid to an educational institution for enrollment and registration in courses. It is typically paid at the beginning of each semester. Conversations that describe financial planning by parents or students for paying college tuition fees would be relevant. These discussions might address saving money by earning income and managing of current spending, the use of tax advantaged college savings plans, obtaining help from family members, or similar things. </narr>

</top>

<top lang="en" type="experience">

<num> 041 </num>

<title> airport security <title>

<desc> What is it like to go through airport security in the United States? </desc>

<narr> A relevant conversation would contain discussions of actual experiences with airport security. Discussion topics might include how long it took to pass through the security in a specific airport, what procedures were involved, what alternatives were available, or what problems were encountered. The experience need not be from the person who is reporting it; second-hand reports of actual experiences would also be relevant. Comparisons to airport security in other countries would be relevant only if specific information is provided about airport security in the USA as a part of the conversation. Information from the public sources that does not involve reports of actual experiences would not be relevant. </narr>

</top>

<top lang="en" type="opinion">

<num> 042 </num>

<title> public schools </title>

<desc> What do people think about the quality of public schools for elementary, middle and secondary school education in the United States? </desc>

<narr> To be relevant, conversations should contain opinions about the quality of education in the public school system in USA from kindergarten though the twelfth grade. To be relevant, the conversation must include at least one instance of a mention or implication regarding education quality. Conversations that solely address schedules, procedures, safety, or other issues not clearly related to education quality or that solely address child care, pre-school programs, colleges, or universities would not be relevant. </narr>

</top>

<top lang="en" type="experience">

<num> 043 </num>

<title> public transport </title>

<desc> When do people prefer to use public transportation? </desc>

<narr> Public surface transportation services such as buses, subways, and trains in the United States ranges from good in some places to nonexistent in others. Many residents of United States own a car and rarely use public transportation; others rely almost exclusively on public transportation. Conversations in which people's preferences regarding the use of public surface transportation are mentioned, or in which those preferences can be inferred from reports of their actual use of public transportation, would be relevant. Public transportation is intended to mean scheduled overland transportation services that are available to all members of the public, whether operated by private companies or by government. Neither on-demand road transportation services such as taxis or Uber nor common-carrier water or air transportation such as ferries or airlines are considered public transportation for this purpose. </narr>

</top>

<top lang="en" type="behavior">

<num> 045 </num>

<title> heroic acts </title>

<desc> Find accounts of selfless heroic acts by individuals or groups for the benefit of others or for a cause. </desc>

<narr> Relevant conversations will contain a description of specific real life acts in which individuals or groups of people displayed courage or exceptional selflessness for a greater cause. General statements concerning heroic acts would not be relevant. </narr>

</top>

<top lang="en" type="experience">

<num> 047 </num>

<title> parenting regrets </title>

<desc> What are some instances in which parents later regret their behavior or actions towards their pre-teen children? </desc>

<narr> Relevant conversations would mention real experiences in which one or more participants express regret regarding something that they now regard as mistaken behavior when raising their children. These regrets might be over specific acts or over general patterns of behavior. To be relevant, the children must have been 12 years of age or younger at the time of the regretted actions. Regrets by the children regarding their parents' actions and actions described by others as mistakes would not be relevant unless described and agreed with by the parent involved. </narr>

</top>

<top lang="en" type="knowledge">

<num> 050 </num>

<title> comic books </title>

<desc> What are some comic books that people talk about? </desc>

<narr> Relevant conversations will include people mentioning one or more specific comic books. To be relevant, the identity of at least one comic book must be discernible, but the name need not be explicitly stated. In the case of a series, it is not necessary that a specific issue in that series be identified. It does not matter whether the comic book has actually been read, but generic references to popular comic book characters such as Superman without specific reference to their appearance in a comic book would not be relevant. </narr>

</top>

<top lang="en" type="behavior">

<num> 051 </num>

<title> texting and driving </title>

<desc> Find real instances in which people texted while they were driving. </desc>

<narr> Conversations that contain evidence of people texting or having texted while they were driving would be relevant. This could include mentions of previously having texted while driving, or it could include messages that were clearly sent while driving. For this purpose, driving is defined as being the operator of a motor vehicle (car, truck, bus, motorcycle, airplane; but not bicycle) with the motor operating. Reports of texting while stopped at a traffic light would be relevant, even if the motor is temporarily stopped while at the light for fuel conversation purposes. </narr>

</top>

<top lang="en" type="opinion">

<num> 054 </num>

<title> Walmart low prices </title>

<desc> What do people believe are the reasons for the low prices offered by Walmart stores? </desc>

<narr> Walmart is one of the largest retailer chains in the world. Relevant conversations will include indications of about what people believe about how it is that Walmart manages to offer its goods

at low prices. Reports of actual prices, without any indication of how those prices were kept low, would not be relevant. </narr>

</top>

<top lang="en" type="behavior">

<num> 055 </num>

<title> hobbies </title>

<desc> What are some examples of people's hobbies? </desc>

<narr> Relevant conversations would include some mention of some hobby engaged in by some person. The person with the mentioned hobby need not be a participant in the conversation. To be a hobby, the activity must be unpaid and unreimbursed and engaged in with enjoyment as a substantial objective. For example, exercise might be a hobby if enjoyment is indicated in some way, but it would not be a hobby if engaged in solely for health reasons. Note that the same activity that is a hobby for one person (e.g., building homes as a charitable activity) might be paid employment for another. </narr>

</top>

<top lang="en" type="experience">

<num> 056 </num>

<title> Amtrak train service </title>

<desc> What do people think about Amtrak train service? </desc>

<narr> Amtrak is the principal operator of long-distance passenger trains in the United States. Relevant conversations would include mention of actual experiences with the Amtrak train service, opinions about the utility of the service to specific people based on its route structure and schedules and similar details. All types of experiences with the Amtrak would be relevant, including such aspects of Amtrak operators as call centers, ticket offices, on board service, announcements on board and in stations, cleanliness, and service disruptions. </narr>

</top>

<top lang="en" type="experience">

<num> 057 </num>

<title> US unemployment </title>

<desc> What are the actual experiences of people who are unemployed in the United States? </desc>

<narr> Relevant conversations will contain accounts of experiences related to unemployment by at least one person who was a resident of the United States at the time. Experiences may be specific or general, and they may involve physical activities or feelings. To be relevant, the experience must be clearly related to unemployment in some way; mere mention of some activity of daily living (e.g., sleep) by a person who is unemployed would not be relevant unless something about that experience was clearly affected by unemployment status. All people who are receiving unemployment compensation are considered to be unemployed for the purpose of this question. In addition, people who are not currently employed but who wish to be employed are considered to be unemployed. Retired people are not considered to be unemployed unless they are now actively seeking reemployment. </narr>

</top>

<top lang="en" type="behavior">

<num> 061 </num>

<title> junk food </title>

<desc> Which food items are thought of as junk food? </desc>

<narr> Food that is unhealthy, highly processed or containing high levels of calories might be described as junk food. Relevant conversations would identify one or more types of food as "junk food," either using that term or some other equally clear indication (e.g., by mention of "empty calories"). The food might be described generically (e.g., candy) or specifically (e.g., Snickers). </narr>

</top>

<top lang="en" type="experience">

<num> 062 </num>

<title> online TV </title>

<desc> When do people prefer watching programs on their computer rather than on their TV? </desc>

<narr> With the advent of streaming video services such as Netflix, some people are "cutting the cord" and moving to Internet-only connectivity for all of their media services, including television. Others sometimes watch television and sometimes watch programs on streaming video. Relevant conversations would reflect people's preferences for when or where they would choose one means of video delivery over another, or other information about how they would make that choice. Both actual experiences and expectations would be relevant, and both first-hand reports of personal experiences or expectations and second-hand reports of personal experiences or expectations of others would be relevant. </narr>

</top>

Bibliography

[1] Pew Research Center. How Americans Use Text Messaging. `http://www.pewinternet.org/2011/09/19/how-americans-use-text-messaging/`, 2011. visited on 2015-04-18.

[2] Douglas W. Oard, Dagobert Soergel, David Doermann, Xiaoli Huang, G. Craig Murray, Jianqiang Wang, Bhuvana Ramabhadran, Martin Franz, Samuel Gustman, James Mayfield, Liliya Kharevych, and Stephanie Strassel. Building an information retrieval test collection for spontaneous conversational speech. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 41–48. ACM, 2004.

[3] Ian Soboroff, Iadh Ounis, J Lin, and I Soboroff. Overview of the trec 2012 microblog track. In *21st Text REtrieval Conference, Gaithersburg, Maryland*, 2012.

[4] Laura Dietz, Ziqi Wang, Samuel Huston, and W. Bruce Croft. Retrieving opinions from discussion forums. In *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13, pages 1225–1228. ACM, 2013.

[5] Kevyn Collins-Thompson, Paul Bennett, Fernando Diaz, Charles LA Clarke, and Ellen M Voorhees. Trec 2013 Web Track Overview. In *22nd Text REtrieval Conference, Gaithersburg, Maryland*, 2014.

[6] Ellen M Voorhees and Dawn M Tice. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 200–207. ACM, 2000.

[7] Pia Borlund. The concept of relevance in IR. *Journal of the Association for Information Science and Technology*, 54(10):913–925, Aug 2003.

[8] Zhiyi Song, Stephanie Strassel, Haejoong Lee, Kevin Walker, Jonathan Wright, Jennifer Garland, Dana Fore, Brian Gainor, Preston Cabe, Thomas Thomas, Brendan Callahan, and Ann Sawyer. Collecting natural SMS and chat conversations in multiple languages: The BOLT phase 2 corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014*, pages 1699–1704, 2014.

[9] Linguistic Data Consortium. BOLT IR Query Development Guidelines, May 2013.

[10] Doug Oard. Answering Questions from Conversations, Dec 2012. Key Note talk at Workshop on Question Answering for Complex Domains, 24th International Conference on Computational Linguistics, Mumbai, India.

[11] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281. ACM, 1998.

[12] Howard Turtle and W. Bruce Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, Jul 1991.

[13] Yuanhua Lv and ChengXiang Zhai. A comparative study of methods for estimating query language models with pseudo feedback. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1895–1898. ACM, 2009.

[14] Stephen E. Robertson and Karen Sparck Jones. Document retrieval systems. chapter Relevance Weighting of Search Terms, pages 143–160. Taylor Graham Publishing, London, UK, 1988.

[15] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[16] Google. Google word2vec code. `http://code.google.com/p/word2vec/`, 2013. visited on 2015-04-18.

[17] Wikipedia. Wikipedia data dump. `https://dumps.wikimedia.org/enwiki/latest/`, 2015. visited on 2015-04-18.

[18] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 307–314. ACM, 1998.

[19] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, 2000.

[20] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.

[21] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, pages 81–93, 1938.

[22] Emine Yilmaz, Javed A Aslam, and Stephen Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 587–594. ACM, 2008.

[23] Mark Sanderson and Justin Zobel. Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 162–169. ACM, 2005.