ABSTRACT

Title of dissertation:	Fine-Grained Linguistic Soft Constraints on Statistical Natural Language Processing Models
	Yuval Marton, Doctor of Philosophy, 2009
Dissertation directed by:	Professor Philip Resnik, Department of Linguistics and Institute for Advanced Computer Studies

This dissertation focuses on effective combination of data-driven natural language processing (NLP) approaches with linguistic knowledge sources that are based on manual text annotation or word grouping according to semantic commonalities. I gainfully apply fine-grained linguistic soft constraints – of syntactic or semantic nature – on statistical NLP models, evaluated in end-to-end state-of-the-art statistical machine translation (SMT) systems. The introduction of semantic soft constraints involves intrinsic evaluation on word-pair similarity ranking tasks, extension from words to phrases, application in a novel distributional paraphrase generation technique, and an introduction of a generalized framework of which these soft semantic and syntactic constraints can be viewed as instances, and in which they can be potentially combined.

Fine granularity is key in the successful combination of these soft constraints, in many cases. I show how to softly constrain SMT models by adding fine-grained weighted features, each preferring translation of only a specific syntactic constituent. Previous attempts using coarse-grained features yielded negative results. I also show how to softly constrain corpus-based semantic models of words ("distributional

profiles") to effectively create word-sense-aware models, by using semantic word grouping information found in a manually compiled thesaurus. Previous attempts, using hard constraints and resulting in aggregated, coarse-grained models, yielded lower gains.

A *novel paraphrase generation technique* incorporating these soft semantic constraints is then also evaluated in a SMT system. This paraphrasing technique is based on the Distributional Hypothesis. The main advantage of this novel technique over current "pivoting" techniques for paraphrasing is the independence from parallel texts, which are a limited resource. The evaluation is done by augmenting translation models with paraphrase-based translation rules, where fine-grained scoring of paraphrase-based rules yields significantly higher gains.

The model augmentation includes a novel *semantic reinforcement component:* In many cases there are alternative paths of generating a paraphrase-based translation rule. Each of these paths reinforces a dedicated score for the "goodness" of the new translation rule. This augmented score is then used as a soft constraint, in a weighted log-linear feature, letting the translation model learn how much to "trust" the paraphrase-based translation rules.

The work reported here is the first to use distributional semantic similarity measures to improve performance of an end-to-end phrase-based SMT system. The unified framework for statistical NLP models with soft linguistic constraints enables, in principle, the combination of both semantic and syntactic constraints – and potentially other constraints, too – in a single SMT model.

Fine-Grained Linguistic Soft Constraints on Statistical Natural Language Processing Models

by

Yuval Yehezkel Marton

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
2009

Advisory Committee:
Professor Philip Resnik, Chair/Advisor
Professor Amy Weinberg, Advisor
Professor William Idsardi, Member
Professor Chris Callison-Burch, Special Member (JHU)
Professor Bonnie Dorr, Dean's Representative

Acknowledgments

My deep gratitude to my advisor and dissertation committee chair, Philip Resnik, for seeing this through, and his patience along this rocky road. His ideas were always excellent additions, together with his inspiring presentation skills and advice. Many thanks go also to my other advisor, Amy Weinberg, for her sharp grasp of any matter, and her challenging questions that taught me to think more critically and precisely. I am grateful to Bill Idsardi for his very useful questions and comments, his kind meta-academic help, and his engaging phonology and cognitive science classes. I also had the pleasure and privilege of collaborating with, and learning from, Chris Callison-Burch of John Hopkins University. He generously shared data, technology, and ideas with me. An ordinary post-talk discussion with CCB (as he is better known in some circles) turned into the beginning of a research direction that not only contributed a lot to my dissertation, but I am sure will also continue to influence my future research. Last, but not least in my dissertation committee, I would like to extend many thanks to Bonnie Dorr, who agreed to serve as the Dean's representative. Her numerous knowledgeable comments throughout the entire dissertation contributed immensely. The amount of her work and attention far exceeded what a Dean's rep role usually entails, and I am most grateful for that.

My thanks also go to Norbert Hornstein, the Department of Linguistics chair, not only for being an inspiring and fun teacher of syntax, but mainly for endowing me with some important life lessons in departmental dynamics – and no less importantly – for the daily cookies, without which, who knows if I would have made it this far. No lesser thanks go to Lisa Hellerstein, my former advisor at Poly/NYU, for encouraging me to pursue this direction, and for her valuable and supportive post-graduation advice. Additional thanks go to Carol Whitney, for very engaging discussions, for sharing her academic experience, and for encouraging and supporting my efforts in research, although visual word recognition research eventually ended outside the dissertation.

I would like to thank Mona Diab for her help with the verb test set, and Raluca Budiu for her help and clarifications regarding the GLSA method and its implementation details. Many thanks to David Chiang for his Hiero code and for collaborating with me; to Saif Mohammad, for finding the time (typically between 2-5 AM) to collaborate with me while working full time on other projects, and not having given up until our work was accepted; to Chris Dyer for discussions and his most appreciated help with code and data; and to Adam Lopez for early illuminating discussions and his implementation of pattern matching with Suffix Array. I would also like to thank Mary Harper and her students Denis Filimonov and Zhongqiang Huang for letting me use their computing resources in times of need. For useful discussions and their good collegiality, thanks also go to the rest of my CLIP Lab PIs and colleagues, including Doug Oard, Louiqa Raschid, Smara Muresan, Hendra Setiawan, Matt Snover, Asad Sayeed, Michael Subotin, and Vlad Eidelman.

Thanks are also due to the present and past administrative team, Kathi Faulkingham, Robert Magee, and Kim Kwok, for helping me find my way through the troubled sea of forms, reimbursements, waivers and applications.

I am deeply indebted to my friends Irit Dekel and Michael Weinman, who showed me the light when I felt I was lost, Donny Inbar, who never tired of cyber-whipping me to complete the dissertation, and all my other good friends, especially Raz, Tami, Serge, Stephanie, Shuki, and Steven, for listening empathically, and nourishing my body and soul. I would like to conclude with lots of thanks to my family and mainly my mother, Ruchama, my father, Michael, and his wife, Orit, for supporting and encouraging me all along the way.

The work in Chapter 2 was supported in part by DARPA prime agreement HR0011-06-2-0001. The MIRA part of this research was supported in part by DARPA contract HR0011-06-C-0022 under subcontract to BBN Technologies and HR0011-06-02-001 under subcontract to IBM.

The work in Chapter 3 was supported, in part, by the National Science Foundation under Grant No. IIS-0705832, and in part, by the Human Language Technology Center of Excellence.

The work in Chapter 4 was partially supported by the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-2-001 and NSF award 0838801, by the EuroMatrixPlus project funded by the European Commission, and by the US National Science Foundation under grant IIS-0713448.

I thank the anonymous reviewers for their valuable feedback on the parts that extended peer-reviewed previous publications. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the sponsors.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The potential of computer applications focusing in natural language processing (NLP) is quite appealing: almost-instantaneous automatic translation, automatic document or news summarization, human–computer interface in natural language (speech recognition, natural language understanding, natural language generation, etc.), and semantic analysis (topic classification, sentiment analysis, information retrieval and clustering, etc.) are just a few examples. However, the current output quality of NLP applications still falls far behind that of humans, in spite of vast research efforts. Can linguistic information (and particularly, constraints) reliably improve the output quality of statistical machine translation and other statistical NLP tasks? Current research trends concentrate on hybrid approaches, combining detailed linguistic analysis – manually crafted rule-based or linguistic annotation-based or linguistic resource-based models – with automatically learned statistical text corpus-based models. Yet several recent hybrid research attempts yielded negative results (compared with "pure statistical" or "pure linguistic" approaches), or were limited in applicability, granularity and gains.

Why concentrate on hybrid approaches? The general assumption is that (a) current statistical tools can easily use "brute force" to calculate relations such as

word co-occurrence statistics over large corpora of electronic text, but are too weak or lack sufficient information to do well in NLP tasks such as inferring meaning of words and phrases; and that (b) current linguistic resources encapsulate more relevant knowledge, but are too coarse for certain tasks – due to too coarse linguistic theories or infeasible amount of human labor required for detailed analysis – resulting in low coverage and inexact analysis of linguistic phenomena. Hybrid approaches attempt to benefit from the best of all worlds: augment statistical tools with linguistic information, while increasing coverage and accuracy, compared with using statistical tools alone, or linguistic analysis alone. Then why is it hard to gainfully apply hybrid approaches?

> *This thesis tests the hypothesis that if one loosens the overly restrictive application of linguistic knowledge in standard natural language applications and/or if one uses linguistic knowledge in a finer-grained manner than is currently used in natural language applications, significant gains may be achieved, as measured by widely accepted evaluation methods, such as the* BLEU *score for statistical machine translation.*

Specifically, this dissertation explores effective combination of (a) statistical data-driven NLP approaches, which use minimally processed large corpora of text, with (b) linguistic analysis or knowledge approaches, which use linguistic resources that are based on manual annotation, such as syntactic parses, or word groupings according to semantic commonalities, such as thesaurus-based "concept" listings.

Soft constraints (explained below) are key combinatory element in this work. I explore here ways of gainfully applying fine-grained soft constraints – of syntactic or semantic nature – on statistical NLP models, focusing on evaluation of such hybrid knowledge/corpus-based models in end-to-end state-of-the-art statistical machine translation systems. Evaluation tasks or sub-tasks include word-pair similarity ranking and paraphrase generation. I introduce a unified NLP corpus-based model with soft constraints, and show how two seemingly different linguistic constraints and two seemingly different NLP tasks can be viewed as instances of the generalized model.

Soft constraints are a mathematical means to bias a model towards certain directions or areas – e.g., to search more intensively in certain parts of the search space – without totally precluding the rest of the model's universe. In contrast, *hard constraints* totally preclude parts of the model's universe. Constraints are often theory-driven. For example, the belief that translation should be done progressively on syntactic constituents such as a noun phrase (NP), can be realized as a soft syntactic constraint, leading a translation model to prefer translating such phrases over word sequences that do not constitute a syntactic phrase. In the previous sentence, syntactic phrases such as "for example" (a preposition phrase, PP) would be preferred in such a model, while the non-syntactic phrases "example, the" and "phrases over" would be dispreferred, perhaps rightfully so – although other non-syntactic phrases such as "there is" might have enough support in the data to be rightfully translated as a unit, corresponding to, say, the German "es gibt" or

the Hebrew "yesh" (transcribed here in Latin letters). Alternatively, such a belief can be realized as a hard syntactic constraint, banning the translation model from considering translation of any word sequence that does not form a syntactic constituent in the translated sentence, even a potentially well-supported sequence such as "there is".

The importance of using *fine-grained* soft constraints is demonstrated in several settings and aspects:

For syntactic constraints, previous attempts to constrain statistical machine translation (SMT) models yielded negative results. The approach there was to constrain the models by adding a single weighted feature, preferring translation units ("spans") that are syntactic constituents in the source sentence over other word sequences. In Chapter 2 I show positive results with constraining SMT models by adding *finer-grained* weighted features, each preferring translation of only a specific syntactic constituent. These translation models remain data-driven (corpus-based), but are constrained, or biased, by syntactic parsing information – an automatic technique for syntactic structure tagging that is based on manual annotations. I show that using parsing tags denoting conventional syntactic constituents (such as NP or VP) is more useful than including "non-classical" tags (denoting parentheses, unparsed fragments, and so on). Detailed parsing information, which is not available via mere "flat NP chunking", is shown to be useful, too, in several language pairs and test sets. In order to avoid feature selection problems and better evaluate the advantage of using fine constraint granularity, feature weights are optimized not

only with the current *de facto* standard Minimum Error Rate Training (MERT), but also with the newer Margin-Infused Relaxed Algorithm (MIRA), one of whose advantages is handling a large number of features well.

For semantic constraints, previous related work created distributional corpus-based semantic models that were aggregated models of thesaurus-based "concepts" – models of groups of related words, and not models of individual words. Word sense modeling was done by mapping the target word to each aggregated model that was based on a concept to which the target word belonged. Each such concept-based model served as a coarse sense-specific model of the target word. In Chapter 3 I introduce hybrid semantic models that are also corpus-based, and are only *biased* toward each concept-based model, effectively creating finer-grained sense-specific models of the *individual* target word. These models achieve better scores than either the corresponding "pure" word-based or concept-based models in word-pair semantic similarity ranking tasks.

I extend these hybrid semantic models from modeling words to modeling word sequences (phrases), and their semantic similarity capability from verification (given words or phrases x and y, return their semantic similarity score) to active semantic problem solving, i.e., paraphrase generation (given a word or phrase x return another word or phrase y that is most similar to x semantically). In Chapter 4 I present a novel paraphrasing technique, which assumes the Distributional Hypothesis, using a large monolingual text corpus. I show how this technique can be used to augment

a translation model with translations of phrases unknown to the model, but whose paraphrases' translation *are* known to the model.

A noteworthy novelty in the translation model augmentation is the use of semantic reinforcement, by unifying alternative paths for generating a particular paraphrastic translation rule: The different paths serve as reinforcing evidence for the goodness of that rule, in proportion to the semantic distance score of each path. For example, if some unknown phrase f is a paraphrase of known phrases f_1 and f_2, each translating to some phrase e in the target language according to the model, then there are two paths for creating a new translation rule from the unknown f to e. A default approach might create a separate new rule for each path, making these new rules compete with one another in order to enter the final sentence translation derivation during "decoding" time; or it might use only the "best path" – the path with the highest paraphrase similarity score. However, here *all paths* reinforce the model's confidence in using a single new translation rule from f to e, by increasing the new rule's associated semantic score in proportion to the paraphrase scores of f to f_1, and f to f_2, respectively. This associated semantic score is implemented in a weighted log-linear feature, enabling the system to tune the weight as it learns how much to "trust" the new translation rules. Performance of fine-grained and coarse-grained associated scoring is compared, too.

So far there have been only few research attempts to connect SMT to distributional semantic similarity methods, and none that involve an end-to-end SMT system. The usage of explicit or implicit semantic knowledge in SMT has gained

momentum, but for paraphrase generation, most current work uses "pivoting" techniques. "Pivoting" here refers to techniques of generating paraphrases by translating to another language (or languages) and back. These techniques have a weakness of relying on relatively limited resources: bi-directional translation phrase tables, typically derived from sentence-aligned bilingual parallel texts that are standardly used in SMT. In contrast, distributional paraphrasing techniques have the advantage of using monolingual corpora, which are relatively abundant. However, pivoting techniques benefit from using human linguistic knowledge implicit in the bilingual sentence alignments, whereas distributional techniques do not. I explore how these competing advantages play out. The work reported here is the first to use distributional similarity measures to improve performance of end-to-end phrase-based SMT systems, simulated for "low-density" languages.

In addition to evaluating soft syntactic and semantic constraints in end-to-end state-of-the-art SMT settings, I also argue that these linguistic soft constraints can be viewed as instances of a generalized statistical NLP model (Chapter 5). Each soft constraint can simply be added to the model linearly as a weighted term. I take this analogy even further, and extend the *de facto* standard model to explicitly include the target sense of the translated or paraphrased word or phrase: Given a word, or generally a phrase u, potentially in context, return the semantically closest phrase v, under certain restrictions, taking potentially different senses of u and v into account. Sense-aware shortest semantic distance means that for the target sense s of the target phrase u, return a phrase v that has sense r, such that v in sense r

is semantically closest to u in sense s.[1] The difference between tasks lay in the restrictions, which are task-specific: In a translation task, v must be in the target language; in a paraphrasing task, v must be in the same language, and formally non-identical to u.

To recap, the way this dissertation handles the question of why it is hard to gainfully apply soft linguistic constraints to data-driven (corpus-based) models, especially in SMT, is by breaking it up to the following questions:

Chapter 2: Can the use of fine-grained syntactic information in soft constraints improve SMT quality, in spite of previous negative results with coarser information?

Chapter 3: Can the use of soft constraints, resulting in fine-grained semantic models, improve semantic distance measure quality, over previous positive results with hard constraints and coarser models?

Chapter 4: Can the use of soft constraints with fine-grained semantic models, when extended from modeling words to phrases and used in paraphrase generation, improve SMT quality, too? Also,

- Can distributional techniques for paraphrase generation for SMT do as well as, or better than "pivoting" techniques, in spite of the fact that the latter benefit from implicit linguistic knowledge in sentence-aligned parallel texts?

[1] If context cannot be used to determine the current sense of u, then v must have a sense that is closest to one of the senses of u, closer than any sense of any other phrase v' to any sense of u.

- Can semantic reinforcement (evidence from similar paths or rules) for scoring paraphrase-based translation rules improve SMT quality?
- Can fine-grained semantic scoring for paraphrase-based translation rules improve SMT quality?

Chapter 5: Is it possible to unify the frameworks of soft syntactic constraints and soft semantic constraints, and propose a tunable (task-specific optimization) unified linear statistical NLP model, with linguistic resource-based soft constraints, of which the syntactic and semantic constraints models can be viewed as instances? What possible benefits this might have?

A few stylistic remarks:

1. Throughout the introduction I mainly use the term "word sequence" when referring to any sequence of words, regardless of syntactic constituency, and the term "phrase" mainly in the linguistic sense of a syntactic constituent (e.g., a noun phrase). However, in the SMT literature, the term "phrase" is commonly used in the non-linguistic sense. I follow this SMT terminology in Chapter 2. In order to help the reader to disambiguate this term, when referring to a syntactic phrase, it is mentioned with a part of speech, as in "noun phrase", or an equivalent acronym such as "NP".

2. Due to the fact that the main topics covered by Chapters 2 through 4 are usually categorized as different sub-areas, background and related work are covered in each of these chapters, instead of one centralized location.

In pursuing this doctoral research direction, I was inspired by issues of linguistic representation in the brain, although I make no cognitive or neuroscientific claims in this dissertation. Two "classic" views on linguistic representations in the brain are abstraction (or generative) approaches and exemplar-based approaches. Abstraction approaches assume that linguistic input is generalized to possibly pre-defined abstract symbols, after which the individual instances of the input become inaccessible, and that linguistic representation and processing only use and operate on the abstract symbols. Exemplar-based approaches assume that there are no pre-defined abstraction categories, and generalizations are made *ad-hoc* over the existing body of the currently known exemplars. However, there is a growing body of literature arguing that in their pure, extreme form, none of these classic views can serve as a good model of linguistic representation in the brain. I invite the reader to consider whether, similarly perhaps consequently, none of these extreme approaches can best serve in NLP applications either. That is, if one regards exemplar-based approaches analogous to data-driven corpus-based statistical NLP models, and abstraction approaches analogous to hard constraints such as syntax-directed machine translation (following the example above, a syntax-directed system would not consider translation of word sequences that are not syntactic constituents). Rather, a data-driven approach that generalizes over *linguistically-biased* patterns, yet without forcing all data into a small set of rules, word groupings, or symbols, is likely to fare better. I leave this as food for thought for the reader, and do not attempt to support this view in the dissertation.

Chapter 2

Soft Syntactic Constraints for Hierarchical Phrased-Based Translation

2.1 Introduction

This chapter focuses solely on one type of soft constraints: *soft syntactic constraints*, evaluated in statistical machine translation (SMT).[1] Next chapters focus on another type: soft *semantic* constraints, evaluated in several tasks. I show in Chapter 5 that models containing any of these soft linguistic constraints can be viewed as instances of a unified model.

The statistical revolution in machine translation, beginning with Brown et al. (1990) and Brown et al. (1993) in the early 1990s, replaced an earlier era of detailed language analysis with automatic learning of shallow source-target mappings from large parallel corpora. Over the last several years, however, the pendulum has begun to swing back in the other direction, with researchers exploring a variety of statistical models that take advantage of source- and particularly target-language syntactic analysis (e.g., Cowan et al., 2006; Zollmann and Venugopal, 2006; Marcu et al., 2006; Galley et al., 2006 and numerous others).

[1] Much of this chapter draws on Marton and Resnik (2008).

Chiang (2005) distinguishes statistical machine translation approaches that are "syntactic" in a *formal* sense, from those that are syntactic in a *linguistic* sense: Formally syntactic approaches go beyond the finite-state underpinnings of phrase-based models, using hierarchical grammars such as synchronous context-free grammar (SCFG). Linguistically syntactic approaches take advantage of *a priori* language knowledge in the form of annotations derived from human linguistic analysis or treebanking. The two forms of syntactic modeling are doubly dissociable: current research frameworks include systems that are finite state but informed by linguistic annotation prior to training (e.g., Koehn and Hoang, 2007; Birch et al., 2007; Hassan et al., 2007), and also include systems employing context-free models trained on parallel text without benefit of any prior linguistic analysis (e.g. Chiang, 2005; Chiang, 2007; Wu, 1997). Over time, however, there has been increasing movement in the direction of systems that are syntactic in both the formal and linguistic senses. See Table 2.1.

	data-driven	**linguistically syntactic**
word-based or "flat" phrase-based	IBM models (Brown et al., 1993), Pharaoh (Koehn, 2004b), Moses (Koehn et al., 2007)	Koehn and Hoang, 2007; Birch et al., 2007; Hassan et al., 2007; Cherry, 2008, ...
hierarchical, formally syntactic	ITG (Wu, 1997), SCFG: Hiero (Chiang, 2005; Chiang, 2007), ...	Cowan et al., 2006; Zollmann and Venugopal, 2006; Marcu et al., 2006; Galley et al., 2006; Marton and Resnik, 2008; Chiang et al., 2008; Xiong et al., 2009; DeNeefe and Knight, 2009, ...

Table 2.1: Formally syntactic and linguistically syntactic SMT approaches are doubly dissociable.

In any such system, there is a natural tension between taking advantage of the linguistic analysis, versus allowing the model to use linguistically unmotivated mappings learned from parallel training data. The tradeoff often involves starting with a system that exploits rich linguistic representations and relaxing some part of it. For example, DeNeefe et al. (2007) begin with a tree-to-string model, using treebank-based target language analysis, and find it useful to modify it in order to accommodate useful "phrasal" chunks that are present in parallel training data but not licensed by linguistically motivated parses of the target language. Similarly, Cowan et al. (2006) focus on using syntactically rich representations of source and target parse trees, but they resort to phrase-based translation for modifiers within clauses. Finding the right way to balance linguistic analysis with unconstrained data-driven modeling is clearly a key challenge.

Here I address this challenge from a less explored direction. Rather than starting with a system based on linguistically motivated parse trees, I begin with a model that is syntactic only in the formal sense. I then introduce soft constraints that take source-language parses into account to a limited extent. Introducing syntactic constraints in this restricted way allows us to take maximal advantage of what can be learned from parallel training data, while effectively factoring in key aspects of linguistically motivated analysis. As a result, I obtain substantial improvements in performance for both Chinese-English and Arabic-English translation.

In Section 2.2 I review related work. Then, in Section 2.3, I briefly review the Hiero statistical MT framework (Chiang, 2005, 2007), upon which this chap-

ter builds, and I discuss Chiang's initial effort to incorporate soft source-language constituency constraints for Chinese-English translation. In Section 2.4, I suggest that an insufficiently fine-grained view of constituency constraints was responsible for Chiang's lack of strong results, and introduce finer grained constraints into the model. I also introduce a novel type of syntactic constraints, penalizing source-side translation units that cross the boundaries of syntactic constituents. Section 2.5 demonstrates the value of these constraints via substantial improvements in Chinese-English translation performance, and extends the approach to Arabic-English. I show improvements when optimizing the model using the practically standard Minimum Error Rate Training (MERT) weight optimization algorithm, and also when using the newer Margin Infused Relaxed Algorithm (MIRA), one of which advantages is handling a large amount of features. Section 2.6 discusses the results, and I conclude in Section 2.7 with a summary and potential directions for future work.

2.2 Related Work

The amount of work involving syntactic knowledge with statistical machine translation (SMT) is vast. There are now yearly workshops dedicated to this very topic.[2] See Lopez (2008b) for a recent comprehensive survey. I will concentrate here on approaches that attempt to relax, or "soften", syntactic constraints in SMT decoding, especially those pertaining to the source language. Other related work, such as work involving the use of syntactic constraints for word alignment (e.g.,

[2] `http://www.cs.ust.hk/~dekai/ssst` - Workshop on Syntax and Structure in Statistical Translation (SSST)

Gildea, 2003; Smith and Eisner, 2006; Cherry and Lin, 2006), or syntactic language modeling (Charniak et al., 2003; Birch et al., 2007; Hassan et al., 2007 and many others), will not be covered here. Since this topic has attracted more interest with and following the publication of the core of this work (Marton and Resnik, 2008), I will start with reviewing prior work in the next sub-section, and continue with work published at the same conference or following my work, in the following sub-section.

2.2.1 Related Prior Work

For ease of exposition, it is useful to map the relevant literature along two axes: (1) use of syntactic parsing information of the source language vs. the target language, and (2) starting from a syntactic commitment (parser-based, syntax-directed approach) and relaxing it vs. starting from a data-driven approach and adding syntactic constraints. This mapping is illustrated in Figure 2.1, where the top chart represents the state of relevant literature before the publication of this work (Marton and Resnik, 2008), and the bottom chart situates this work together with past work and other work published at the same time. It is hard to directly compare the related work because of the diversity in training sets, language models, syntactic information, translation "decoder" used, and so on. However, many of these research efforts found it useful to relax hard syntactic constraints in some way, as detailed below. Adding soft syntactic constraints, instead of using – or relaxing – hard syntactic constraints was less explored. The charts illustrate the relative "vacuum" in the upper left *adding source-side syntactic constraints* quadrant before

the publication of the core work in this chapter at the ACL 2008 conference, and the growing interest of the research community in this quadrant, with and following that publication. I show later in this chapter how to gainfully add soft syntactic constraints to a hierarchical phrase-based SMT.

Prior work concentrates in the lower right *relaxing target-side syntax-directed models* quadrant, with some cases of using source-side syntactic parses as well. Among approaches using parser-based syntactic models, several researchers have attempted to reduce the strictness of syntactic constraints in order to better exploit shallow correspondences in parallel training data. Section 2.1 has already briefly noted Cowan et al. (2006), who relax parse-tree-based alignment to permit alignment of non-constituent sub-phrases on the source side, and translate modifiers using a separate phrase-based model, and DeNeefe et al. (2007), who modify syntax-based extraction and binarize trees (following Wang et al., 2007b) to improve phrasal coverage. Similarly, Marcu et al. (2006) relax their syntax-based system by rewriting target-side parse trees on the fly, adding an intermediate, fictive, "non-syntactic" tree node (non-terminal symbol spanning only part of a syntactic constituent), in order to avoid the loss of "non-syntactifiable" phrase pairs such as *the mutual* in both source and target languages.

Zollmann and Venugopal (2006), lower right quadrant, start with a target language parser and use it to provide constraints on the extraction of hierarchical phrase pairs. Unlike Hiero (see Section 2.3), which uses one "unnamed" non-terminal symbol (X), their translation model uses a full range of "named" nonterminal symbols

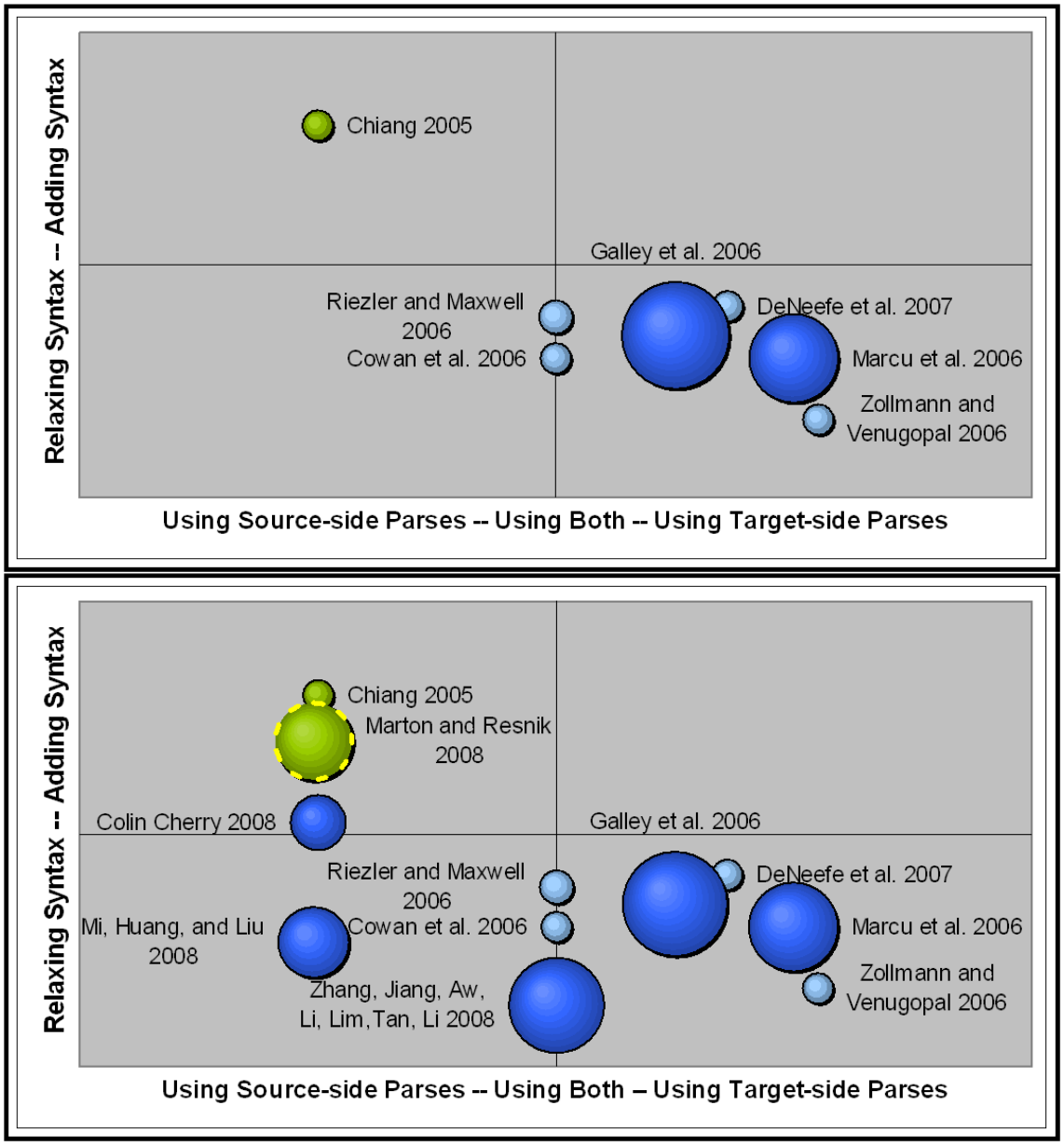

Figure 2.1: Representative syntax-aware literature before June 2008 (top) and after (bottom). Circle size denotes relative gain in BLEU score when using syntactic information, with larger circles denoting larger gains, while the smallest (and light) circles denoting no significant gains (negative results). Note that BLEU gains cannot be compared due to differences in training sets, language pairs, language models, syntactic information, etc. Top *adding syntax* quadrants are relatively empty, with Chiang (2005) showing negative result.

in the synchronous grammar, corresponding to syntactic parsing tags. As an alternative way to relax strict parser-based constituency requirements, they explore the use of phrases spanning generalized, categorial-style constituents in the parse tree, e.g. type NP/NN denotes a phrase like *the great* that lacks only a head noun (say, *wall*) in order to comprise an NP.

A soft-constraint approach that can also be viewed as coming from the data-driven side, adding syntax, is taken by Riezler and Maxwell (2006). They use LFG dependency trees on both source and target sides, and relax syntactic constraints by adding a "fragment grammar" for unparsable chunks. Their work is located accordingly on the border between the two lower quadrants. They decode using Pharaoh, augmented with their own log-linear features (such as $p(e_{snippet}|f_{snippet})$ and its converse), side by side to "traditional" lexical weights. Riezler and Maxwell (2006) do not achieve higher BLEU scores, but do score better according to human grammaticality judgments for in-coverage cases.

Setiawan et al. (2007) employ a "function-word centered syntax-based approach", with synchronous CFG and extended ITG models for reordering phrases, and relax syntactic constraints by only using a small number function words (approximated by high-frequency words) to guide the phrase-order inversion. This line is further developed in Setiawan et al. (2009), see next sub-section.

In addition, various researchers have explored the use of hard linguistic constraints on the source side, e.g. via "chunking" noun phrases and translating them separately (Owczarzak et al., 2006), or by performing hard reorderings of source

parse trees in order to more closely approximate target-language word order (Wang et al., 2007a; Collins et al., 2005).

Quirk et al. (2005) and Quirk and Menezes (2006) use phrasal SMT with example-based (EBMT) elements. They use source-side syntactic dependency "treelets" that are projected onto "flat" target-side phrases via unsupervised word alignments. They relax the sub-tree ordering by using an ordering model on freely ordered sub-treelets. Several such syntax-aware features are combined in a log-linear framework. Their work can be mapped to the lower left *relaxing source-side syntax* quadrant, near the border of the lower right quadrant.

Eisner (2003) learns probabilistic synchronous tree substitution grammar (STSG) from unaligned trees in sentence-aligned parallel parsed text. STSG is similar to synchronous tree adjoining grammar (STAG; Shieber and Schabes, 1990), excluding adjunction adjoining, and is weakly equivalent to SCFG. Bilingual syntactic alignment is relaxed by allowing null treelets on both sides. DeNeefe and Knight (2009) apply a less restricted STAG variant, tree insertion grammar (TIG), generalizing Nesson et al. (2006), using LDC treebank-style target-side (English) trees, and non-named non-terminals (X) projected on the source-side. They further relax the syntactic constraints with "fail-safe monotone translation rules in case of parse failures and extremely long sentences" and associated features, weighted in a log-linear framework.

2.2.2 More Recent Related Work

The research direction of the work described in this chapter has attracted considerable attention, both directly (follow-up work such as that of Xiong et al., 2009, detailed below), or indirectly (general interest of the research community in this direction, potentially independently of this work). Therefore, to emphasize this traction and shift of interest, the more recent work, which was published with or following this work, is discussed in this sub-section, separately.

In addition to Setiawan et al. (2009) and DeNeefe and Knight (2009) which were mentioned above, several other publications concerning source side soft syntactic constraints were published at the same time as, or after, Marton and Resnik (2008).

Cherry (2008) published at the same time as Marton and Resnik (2008). He incorporates source-side syntactic dependency trees as soft syntactic constraints in a weighted "syntactic cohesion" feature in a log-linear framework in a phrased-based system, Moses (Koehn et al., 2007). The use of source side soft syntactic constraints in a log-linear model is similar to my work, however my work uses syntactic constituency parses (as opposed to dependency trees), in a hierarchical phrase-based SMT system, Hiero (as opposed to the "flat" phrase-based Moses). Moses translates monotonously in the target language, occasionally breaking up the order of source side phrases used for the translation; the cohesiveness constraint discourages the decoder from shuffling their order in a way inconsistent with the

dependency tree. The syntactic constraints presented later in this chapter also encourage the use of source-side phrases that is consistent with the source side parse tree. However, these syntactic constraints do not directly affect phrase order; rather, a derivation with a hierarchical translation rule, i.e., a rule with a gap (X), which is higher in the synchronous CFG tree, is rewarded if this gap, which connects the current translation rule in the derivation, is consistent with a syntactic constituent on the source side (see Sections 2.3.2 and 2.4). An overall small improvement in BLEU is shown, with about 1 BLEU point improvement for a small subset that has "uncohesive baseline translations". They also showed preference of human raters for the cohesive output over baseline, for sentences in the uncohesive subset. The notion of cohesive constraint is extended in Bach et al. (2009), where violation of source-side cohesiveness is penalized recursively and "softly", in proportion to number of words in violation in the applied translation rule.

Mi et al. (2008) use a source-side parse packed forest in decoding. There, alternative parses compete but also reinforce repeating sub-trees. The authors "soften" the syntactic forest constraint by adding a "default translation hyper-edge" for monotone translation in order to increase coverage, using the flat phrase-based Pharaoh (Koehn, 2004b). They show 1.7 BLEU points gain over 1-best parses in Chinese-English translation task. The SMT system described in this chapter uses only 1-best source-side parse tree, not a forest. Here, too, monotone translation is used as a last resort (the so-called "glue rules" in Hiero).

Zhang et al. (2008) use parses of both source and target languages, and relax the syntactic constraint by translating synchronous "tree sequences". The largest parse sub-trees that exactly "cover" the source phrase in sequence, are used for translation, but the source phrase does not have to exactly match a single syntactic constituent. In the extreme case this reduces to monotone decoding. They, too, test their system in a Chinese-English translation task, and report gains of 1.4, 2.2, and 3.4 BLEU points over STSG, Moses, and SCFG baselines, respectively. The work in this chapter uses only source-side parses, and does not involve sequences of syntactic constituents in a single rule.

Xiong et al. (2009) re-implemented the Marton and Resnik (2008) XP+ feature (see Section 2.4) in a bracketing transduction grammar system (Wu, 1997), and obtained over 1 BLEU point gain over their syntax-unaware baseline, in Chinese-English translation task. They compared using this feature with using two variants of their syntax-derived bracketing (SDB) features, which estimate probabilities of source-side phrase cohesion (in other words, probability that in the target side, the words that are translation of the words in that source phrase will not enclose translation of source-side words outside the phrase). These probabilities are estimated using syntactic features such as the subsuming source-side tree or sub-trees, and whether the trees exactly span the phrase, contain it, or that the phrase crosses the boundaries of the sub-trees. They achieve even larger gains of up to 1.7 BLEU points over their baseline with these features.

Setiawan et al. (2009) follow Setiawan et al. (2007) with a model that uses two neighboring function words as a soft constraint guiding a Hiero decoder in deciding whether to invert (re-order) the corresponding target phrases. The syntactic knowledge is still only approximated via the function words, without using parsing information, as in the previous work of the first author. The usage of soft syntactic constraints via log-linear features is similar to the work in this chapter. However, the work here does use parsing information. They achieve gains of up to 1.5 BLEU points over their baseline.

Venugopal et al. (2009) use soft syntactic constraints to make syntactic similarities between different derivations reinforce the similar parts, rather than have the entire derivations compete, as is standardly done, including the work described here. This technique alleviates the "spurious ambiguity" problem, and results in improvements of about 1 BLEU point in a small data size Chinese-English translation task (model using 0.6M words in a limited domain, IWSLT06[3]), and somewhat less in a medium data size task (model using a subset of 67M words of the NIST broadcast news MT05 set).

Hanneman and Lavie (2009) relax a syntax-directed manually written tree-to-tree translation rule system by adding a non-syntactic constituent parsing tag for any "phrase". They use it to incorporate non-syntactic "flat" phrase-based translations to increase coverage. They introduce a "syntax-prioritized technique" to increase coverage efficiently and without loss of translation quality in a French-

[3]International Workshop on Spoken Language Translation 2006

English translation task. These authors relax theory-driven manually constructed hard syntactic constraints (the syntactic rules), while the work here uses wider coverage, data-driven, automatically extracted syntax-unaware rules, which are biased towards syntactic translation units via parsing information-based soft constraints.

2.3 Hierarchical Phrase-based Translation

2.3.1 Hiero

Hiero (Chiang, 2005; Chiang, 2007), which is used in the experiments reported in this chapter, is a hierarchical phrase-based statistical MT framework that generalizes phrase-based models by permitting phrases with gaps. Formally, Hiero's translation model is a weighted synchronous context-free grammar (SCFG). Hiero employs a generalization of the standard non-hierarchical phrase extraction approach in order to acquire the synchronous rules of the grammar directly from word-aligned parallel text. Rules have the form

$$X \rightarrow \langle \bar{e}, \bar{f} \rangle$$

where $\bar{e}$ and $\bar{f}$ are phrases containing terminal symbols (words) and possibly co-indexed instances of the nonterminal symbol X.[4] For example, the translation rule

$$X \rightarrow \langle \textit{the green } X_1 \textit{ sleeps } X_2 \ , \ \textit{la } X_1 \textit{ verte dort } X_2 \rangle$$

could translate the English *the green caterpillar sleeps under a leaf* to the French *la chenille verte dort sous une feuille*, or the English *the green idea sleeps furiously* to the French *la idée (l'idée) verte dort furieusement.* All co-indexed occurrences of X would have to be translated with another such rule, e.g., $X \rightarrow \langle \textit{idea}, \textit{idée} \rangle$ or $X \rightarrow \langle \textit{furiously}, \textit{furieusement} \rangle$. The English (source) side of the nested rule will substitute a source side occurrence of X in the containing rule, while the target side of the nested rule will synchronously substitute the occurrence of X in the containing rule which was co-indexed with the substituted source side X. Since Hiero is SCFG-based, the choice of what nested rule to use is independent of the containing rule.

Associated with each rule is a set of translation model features, $\phi_i(\bar{f}, \bar{e})$; for example, one intuitively natural feature of a rule is the phrase translation probability or log-probability $\phi(\bar{f}, \bar{e}) = \log p(\bar{e}|\bar{f})$, directly analogous to the corresponding feature in non-hierarchical phrase-based models like Pharaoh (Koehn et al., 2003). In addition to this phrase translation probability feature, Hiero's feature set includes the inverse phrase translation probability $\log p(\bar{f}|\bar{e})$, lexical weights $\text{lexwt}(\bar{f}|\bar{e})$ and

[4]This is slightly simplified: Chiang's original formulation of Hiero has two nonterminal symbols, X and S. The latter is used only in two special "glue" rules that permit complete trees to be constructed via concatenation of subtrees when there is no better way to combine them.

lexwt($\bar{e}|\bar{f}$), which are estimates of translation quality based on word-level correspondences (Koehn et al., 2003), and a rule penalty allowing the model to learn a preference for longer or shorter derivations; see Chiang (2007) for details.

These features are combined using a log-linear model, with each synchronous rule contributing

$$\sum_i \lambda_i \phi_i(\bar{f}, \bar{e}) \tag{2.1}$$

to the total log-probability of a derived hypothesis. Each λ_i is a weight associated with feature ϕ_i, and these weights are typically optimized using minimum error rate training (Och, 2003).

As noted in Section 2.1, Hiero is only formally syntactic, and is not linguistically aware beyond the capability to handle rules with gaps (synchronous CFG). Next, I discuss past and present attempts to make Hiero syntactically aware also in the linguistic sense.

2.3.2 Soft Syntactic Constraints

When looking at Hiero rules, which are acquired automatically by the model from parallel text, it is easy to find many cases that seem to respect linguistically motivated boundaries. For example,

$$X \rightarrow \langle \text{jingtian } \mathrm{X}_1, \mathrm{X}_1 \text{ this year} \rangle,$$

seems to capture the use of *jingtian/this year* as a temporal modifier when building linguistic constituents such as noun phrases (*the election this year*) or verb phrases (*voted in the primary this year*). However, it is important to observe that nothing in the Hiero framework actually *requires* nonterminal symbols to cover linguistically sensible constituents, and in practice they frequently do not. This rule could just as well be applied with X_1 covering the "phrase" *submitted and* to produce non-constituent substring *submitted and this year* in a hypothesis like *The budget was submitted and this year cuts are likely.*

Chiang (2005) conjectured that there might be value in allowing the Hiero model to favor hypotheses for which the synchronous derivation respects linguistically motivated source-language constituency boundaries, as identified using a parser. He tested this conjecture by adding a soft constraint in the form of a "constituency feature": if a synchronous rule $X \rightarrow \langle \bar{e}, \bar{f} \rangle$ is used in a derivation, and the span of $\bar{f}$ is a constituent in the source-language parse, then a term λ_c is added to the model score in expression (2.1).[5] A hard constraint would prevent the application of any rules violating syntactic boundaries; however, using the *soft* constraint weighted feature allows the model to boost the "goodness" for a rule if it is consistent with the source language constituency analysis, and to leave its score unchanged otherwise. The weight λ_c, like all other λ_i, is set during a tuning step, originally done via Minimum Error Rate Training (MERT; Och, 2003), and recently alternatively also via Margin Infused Relaxed Algorithm (MIRA; Cram-

[5]Formally, $\phi_c(\bar{f}, \bar{e})$ is defined as a binary feature, with value 1 if $\bar{f}$ spans a source constituent and 0 otherwise. In the latter case $\lambda_c \phi_c(\bar{f}, \bar{e}) = 0$ and the score in expression (2.1) is unaffected.

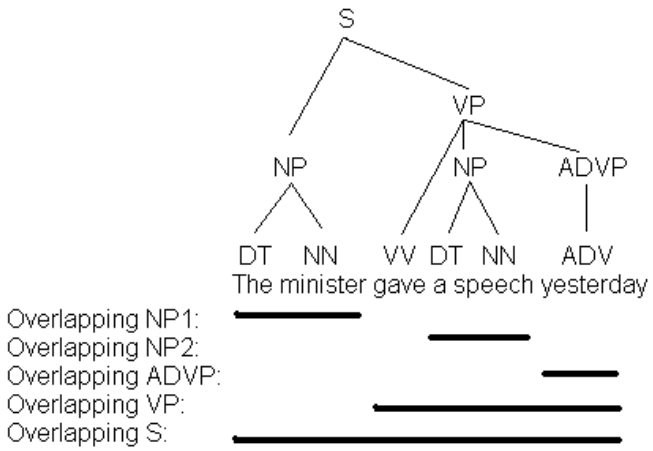

Figure 2.2: Illustration of Chiang's (2005) syntactic constituency feature, which does not distinguish among constituent types. Translation rules whose source side exactly spans any of the horizontal lines would be equally rewarded. A rule translating, say, *minister gave a* as a unit would not be rewarded. In this example English is used as the source language, for ease of readability.

mer and Singer, 2003; Crammer et al., 2006; Watanabe et al., 2007; Chiang et al., 2008). Either optimization process determines empirically the extent to which the constituency feature should be trusted.

Figure 2.2 illustrates the way the constituency feature worked, treating English as the source language for the sake of readability. In this example, λ_c would be added to the hypothesis score for any rule used in the hypothesis whose source side spanned *the minister*, *a speech*, *yesterday*, *gave a speech yesterday*, or *the minister gave a speech yesterday.* A rule translating, say, *minister gave a* as a unit would receive no such boost.

Chiang tested the constituency feature for Chinese-English translation, and obtained no significant improvement on the test set. The idea then seems essentially to have been abandoned; it does not appear in later discussions (Chiang, 2007).

2.4 Soft Syntactic Constraints, Revisited

On the face of it, there are any number of possible reasons Chiang's (2005) soft constraint did not work – including, for example, practical issues like the quality of the Chinese parses.[6] However, I focus here on two conceptual issues underlying his use of source language syntactic constituents.

First, the constituency feature treats all syntactic constituent types equally, making no distinction among them. For any given language pair, however, there might be some source constituents that tend to map naturally to the target language as units, and therefore more valuable in translation – and others that do not (Fox, 2002; Eisner, 2003; Koehn, 2003). Moreover, a parser may tend to be more accurate for some constituents than for others. Assigning a high weight also to noisy parsing tags or inconsistent tag pairing might have caused more damage than benefit to the overall translation quality.

Second, the Chiang (2005) constituency feature gives a rule additional credit when the rule's source side overlaps exactly with a source-side syntactic constituent. Logically, however, it might make sense not just to give a rule $X \rightarrow \langle \bar{e}, \bar{f} \rangle$ extra credit when $\bar{f}$ matches a constituent, but to incur a *cost* when $\bar{f}$ *violates* a constituent boundary. Using the example in Figure 2.2, one might want to penalize hypotheses containing rules where $\bar{f}$ is *the minister gave a* (and other cases, such as *minister gave*, *minister gave a*, and so forth).

[6] In fact, this turns out not to be the issue; see Section 2.5.

This accomplishes coverage of the logically complete set of possibilities, which include not only $\bar{f}$ matching a constituent exactly or crossing its boundaries, but also $\bar{f}$ being properly contained within the constituent span, properly containing it, or being outside it entirely. Whenever these latter possibilities occur, $\bar{f}$ will exactly match or cross the boundaries of some other constituent. Still in Figure 2.2, the second thick horizontal line (spanning *a speech*) is properly contained in the span of the fourth thick line, which is the verb phrase *gave a speech yesterday*; therefore, although using a rule whose source side has the span of the second line would not be rewarded by a VP-matching feature, and would not be penalized by a cross-VP-boundary feature, it *would* be rewarded by a NP-matching feature. Conversely, the fourth thick line (spanning the VP *gave a speech yesterday*) properly contains the span of the second one, which is the noun phrase *a speech*); a rule whose source side has the span of the fourth line would not be rewarded by a NP-matching feature, nor would it be penalized by a cross-NP-boundary feature, but it *would* be rewarded by a VP-matching feature. The first thick line (the NP *the minister* is entirely outside the span of the fourth line (the VP); it would not be affected by VP-sensitive features, but it would be rewarded by a NP-matching feature. A rule whose source side spans *minister gave a* would be penalized by both a cross-NP-boundary feature and a cross-VP-boundary feature.[7]

These observations suggest a finer-grained approach to the constituency feature idea, retaining the idea of soft constraints, but applying them using *various*

[7]To be precise, a binary branching parsing tree would achieve a logically complete set of possibilities; a tree with a larger maximal fan-out can include other possibilities such as *gave a speech* and *a speech yesterday*, which are neither rewarded not penalized by the proposed features.

soft-constraint constituency features. My first observation argues for distinguishing among constituent types (NP, VP, etc.). My second observation argues for distinguishing the benefit of matching constituents from the cost of crossing constituent boundaries. I therefore define a space of new features as the cross product

$$\{\text{CP}, \text{IP}, \text{NP}, \text{VP}, \ldots\} \times \{=, +\}.$$

where $=$ and $+$ signify matching and crossing boundaries, respectively. For example, $\phi_{\text{NP}=}$ would denote a binary feature that matches whenever the span of $\bar{f}$ exactly covers an NP in the source-side parse tree, resulting in $\lambda_{\text{NP}=}$ being added to the hypothesis score (expression (2.1)). Similarly, $\phi_{\text{VP}+}$ would denote a binary feature that matches whenever the span of $\bar{f}$ crosses a VP boundary in the parse tree, resulting in $\lambda_{\text{VP}+}$ being *subtracted from* the hypothesis score.[8] For readability from this point forward, I will omit ϕ from the notation and refer to features such as NP= (which one could read as "NP match"), VP+ (which one could read as "VP crossing"), etc.

In addition to these individual features, I define three more variants:

- For each constituent type, e.g. NP, I define a feature NP_ that ties the weights of NP= and NP+. If NP= matches a rule, the model score is incremented by $\lambda_{NP\_}$, and if NP+ matches, the model score is decremented by the same quantity.

[8]Formally, $\lambda_{\text{VP}+}$ simply contributes to the sum in expression (2.1), as with all features in the model, but weight optimization using minimum error rate training should, and does, automatically assign this feature a negative weight.

- For each constituent type, e.g. NP, I define a version of the model, NP2, in which NP= and NP+ are *both* included as features, with separate weights $\lambda_{NP=}$ and λ_{NP+}.

- I define a set of "standard" linguistic labels containing {CP, IP, NP, VP, PP, ADJP, ADVP, QP, LCP, DNP} and excluding other labels such as PRN (parentheses), FRAG (fragment), etc.[9] I define feature XP= as the disjunction of {CP=, IP=, ..., DNP=}; i.e. its value equals 1 for a rule if the span of $\bar{f}$ exactly covers a constituent having any of the standard labels. The definitions of XP+, XP_, and XP2 are analogous.

- Similarly, since Chiang's original constituency feature can be viewed as a disjunctive "all-labels=" feature, I also defined "all-labels+", "all-labels2", and "all-labels_" analogously.

2.5 Experiments

In the next section I describe experiments with soft syntactic constraints, implemented in weighted log-linear features. Section 2.5.1 describes experiments optimizing the feature weights with the *de facto* standard minimum error rate training (MERT), and Section 2.5.2 addresses the feature selection problem that arises in Section 2.5.1, using another weight optimization method.

[9] I map SBAR and S labels in Arabic parses to CP and IP, respectively, consistent with the Chinese parses. I map Chinese DP labels to NP. DNP and LCP appear only in Chinese. I ran no ADJP experiment in Chinese, because this label virtually aways spans only one token in the Chinese parses.

2.5.1 MERT Experiments

I carried out MT experiments for translation from Chinese to English and from Arabic to English, using a descendent of Chiang's Hiero system (Chiang et al., 2005), with binary disk grammar for Arabic-English translation, and a suffix array-based decoder implementation of Hiero, which became available later, for Chinese-English translation (Lopez, 2007; Lopez, 2008a). Language models were built using the SRI Language Modeling Toolkit (Stolcke, 2002) with modified Kneser-Ney smoothing (Chen and Goodman, 1998). Word-level alignments were obtained using GIZA++ (Och and Ney, 2000). The baseline model in both languages used the feature set described in Section 2.3; for the Chinese baseline I also included a rule-based number translation feature (Chiang, 2007).

In order to compute syntactic features, I analyzed source sentences using state of the art, tree-bank trained constituency parsers (Huang et al. (2008) for Chinese, and the Stanford parser v.2007-08-19 for Arabic (Klein and Manning, 2003a; Klein and Manning, 2003b)). In addition to the baseline condition, and baseline plus Chiang's (2005) original constituency feature, experimental conditions augmented the baseline with additional features as described in Section 2.4.

All models were optimized and tested using the BLEU metric (Papineni et al., 2002) with the NIST-implemented ("shortest") effective reference length, on lower-cased, tokenized outputs/references. Statistical significance of difference from the baseline BLEU score was measured by using paired bootstrap re-sampling (Koehn,

2004b), with a sample size of 2000 pairs. Statistical significance was determined in case the 95% confidence interval (CI) of the systems' BLEU score difference did not include zero. For conciseness, this is denoted as $p < .05$ below. Similarly, a 99% CI is denoted as $p < .01$, and so on for other CIs. The word "significant" is used below as a shorthand for "statistically significant" (at $p < .05$ unless specified otherwise). The associated t-test p-value for the significant cases was always $p < 0.0001$.

2.5.1.1 Chinese-English

For the Chinese-English translation experiments, I trained the translation model on the corpora in Table 2.2, totalling approximately 2.1 million sentence pairs after GIZA++ filtering for length ratio. Chinese text was segmented using the Stanford segmenter (Tseng et al., 2005).

LDC ID	**Description**
LDC2002E18	Xinhua Ch/Eng Par News V1 beta
LDC2003E07	Ch/En Treebank Par Corpus
LDC2005T10	Ch/En News Mag Par Txt (Sinorama)
LDC2003E14	FBIS Multilanguage Txts
LDC2005T06	Ch News Translation Txt Pt 1
LDC2004T08	HK Par Text (only HKNews)

Table 2.2: Training corpora for Chinese-English translation. LDC = The Linguistic Data Consortium at the University of Pennsylvania (`http://www.ldc.upenn.edu`)

Use	**Set**	**Size (sentences)**
Training	Table 2.2	2,100,000
Development	NIST MT03	919
Test	NIST MT06 (NIST part)	1,099
Test	NIST MT08	1,357

Table 2.3: Training development and test set sizes for Chinese-English translation

I trained a 5-gram language model using the English (target) side of the training set, pruning 4-gram and 5-gram singletons. For minimum error rate training and development I used the NIST MTeval MT03 set. Details are given in Table 2.3.

Table 2.4 presents the results. I first evaluated translation performance using the NIST MT06 (nist-text) set. Like Chiang (2005), I found that the original, undifferentiated constituency feature (Chiang-05) introduced a negligible, statistically insignificant improvement over the baseline. However, I found that several of the finer-grained constraints (IP=, VP=, VP+, QP+, and NP=) had achieved statistically significant improvements over baseline (up to .74 BLEU), and the latter three also improved significantly on the undifferentiated constituency feature. By combining multiple finer-grained syntactic features, I obtained significant improvements of up to 1.65 BLEU points (NP_, VP2, IP2, all-labels_, and XP+).

I also obtained further gains using combinations of features that had performed well; e.g., condition IP2.VP2.NP_ augments the baseline features with IP2 and VP2 (i.e. IP=, IP+, VP= and VP+), and NP_ (tying weights of NP= and NP+; see Section 2.4). Since component features in those combinations were informed by individual-feature performance on the test set, I tested the best performing conditions from MT06 on a new test set, NIST MT08. NP= and VP+ yielded significant improvements of up to 1.53 BLEU. Combination conditions replicated the pattern of results from MT06, including the same increasing order of gains, with improvements up to 1.11 BLEU.

Chinese	**MT06**	**MT08**
Baseline	.2624	.2064
Chiang-05	.2634	.2065
PP=	.2607	
DNP+	.2621	
CP+	.2622	
AP+	.2633	
AP=	.2634	
DNP=	.2640	
IP+	.2643	
PP+	.2644	
LCP=	.2649	
LCP+	.2654	
CP=	.2657	
NP+	.2662	
QP=	.2674^+	.2071
IP=	**.2680***+	.2061
VP=	.2683*	.2072
VP+	**.2693****++	**.2109***+
QP+	**.2694****++	.2091
NP=	**.2698****++	**.2217****++
Multiple / conflated features:		
QP2	.2614	
NP2	.2621	
XP=	.2630	
XP2	.2633	
all-labels+	.2633	
VP_	.2637	
QP_	.2641	
NP.VP.IP=.QP.VP+	.2646	
IP_	.2647	
IP2+VP2	.2649	
all-labels2	.2673*-	.2070
NP_	**.2690****++	.2101^+
IP2.VP2.NP_	**.2699****++	**.2105***+
VP2	**.2722****++	**.2123****++
all-labels_	**.2731****++	**.2125***++
IP2	**.2750****++	**.2132****+
XP+	**.2789****++	**.2175****++

Table 2.4: Chinese-English results. *,**: Significantly better than baseline ($p < .05, .01$, respectively). 36 ^: Almost significantly better than baseline ($p < .075$). +,++: Significantly better than Chiang-05 ($p < .05, .01$, respectively). -: Almost significantly better than Chiang-05 ($p < .075$).

2.5.1.2 Arabic-English

For Arabic-English translation, I used the training corpora in Table 2.5, approximately 100,000 sentence pairs after GIZA++ length-ratio filtering. I trained a trigram language model using the English side of this training set, plus the English Gigaword v2 AFP and Gigaword v1 Xinhua corpora. Development and minimum error rate training were done using the NIST MT02 set. Details are given in Table 2.6.

Table 2.7 presents the results. I first tested on on the NIST MT03 and MT06 (nist-text) sets. On MT03, the original, undifferentiated constituency feature did not improve over baseline. Two individual finer-grained features (PP+ and AdvP=) yielded statistically significant gains up to .42 BLEU points, and feature combinations AP2, XP2 and all-labels2 yielded significant gains up to 1.03 BLEU points.

LDC ID	**Description**
LDC2004T17	Ar News Trans Txt Pt 1
LDC2004T18	Ar/En Par News Pt 1
LDC2005E46	Ar/En Treebank En Translation
LDC2004E72	eTIRR Ar/En News Txt

Table 2.5: Training corpora for Arabic-English translation

Use	**Set**	**Size (sentences)**
Training	Table 2.5	100,000
Development	NIST MT02	663
Test	NIST MT03	1,357
Test	NIST MT06 (NIST part)	1,797
Test	NIST MT08	1,357

Table 2.6: Training development and test set sizes for Arabic-English translation

XP2 and all-labels2 also improved significantly on the undifferentiated constituency feature, by .72 and 1.11 BLEU points, respectively.

For MT06, Chiang's original feature improved the baseline significantly — this is a new result using his feature, since he did not experiment with Arabic. Improvements were also achieved by my IP=, PP=, and VP= conditions. Adding individual features PP+ and AdvP= yielded significant improvements up to 1.4 BLEU points over baseline, and in fact the improvement for individual feature AdvP= over Chiang's undifferentiated constituency feature approaches significance ($p < .075$).

More important, several conditions combining features achieved statistically significant improvements over baseline of up 1.94 BLEU points: XP2, IP2, IP, VP=.PP+.AdvP=, AP2, PP+.AdvP=, and AdvP2. Of these, AdvP2 is also a significant improvement over the undifferentiated constituency feature (Chiang-05), with $p < .01$. As I did for Chinese, I tested the best-performing models on a new test set, NIST MT08. Consistent patterns reappeared: improvements over the baseline up to 1.69 BLEU ($p < .01$), with AdvP2 again in the lead (also outperforming the undifferentiated constituency feature, $p < .05$). A translation example is brought in Section 2.6.

Arabic	**MT03**	**MT06**	**MT08**
Baseline	.4795	.3571	.3571
Chiang-05	.4787	**.3679****	**.3678****
VP+	.4802	.3481	
AP+	.4856	.3495	
IP+	.4818	.3516	
CP=	.4815	.3523	
NP=	.4847	.3537	
NP+	.4800	.3548	
AP=	.4797	.3569	
AdvP+	.4852	.3572	
CP+	.4758	.3578	
IP=	.4811	**.3636****	**.3647****
PP=	.4801	**.3651****	**.3662****
VP=	.4803	**.3655****	**.3694****
PP+	**.4837****	**.3707****	**.3700****
AdvP=	**.4823****	**.3711**-**	**.3717****
Multiple / conflated features:			
XP+	.4771	.3522	
all-labels2	**.4898**+**	.3536	.3572
all-labels_	.4828	.3548	
VP2	.4826	.3552	
NP2	.4832	.3561	
AdvP.VP.PP.IP=	.4826	.3571	
VP_	.4825	.3604	
all-labels+	.4825	.3600	
XP2	**.4859**+**	.3605^	**.3613****
IP2	.4793	**.3611***	.3593
IP_	.4791	**.3635***	**.3648****
XP=	.4808	**.3659****	**.3704**+**
VP=.PP+.AdvP=	**.4833****	**.3677****	**.3718****
AP2	**.4840****	**.3692****	**.3719****
PP+.AdvP=	.4777	**.3708****	**.3680****
AdvP2	.4803	**.3765**++**	**.3740**+**

Table 2.7: Arabic-English resutls. Results are sorted by MT06 BLEU score. *: Better than baseline ($p < .05$). **: Better than baseline ($p < .01$). +: Better than Chiang-05 ($p < .05$). ++: Better than Chiang-05 ($p < .01$). -: Almost significantly better than Chiang-05 ($p < .075$)

2.5.2 MIRA Experiments

One major weakness of the experiments described in Section 2.5.1 is the need for feature selection: no single constituent-sensitive feature, single constraint type (matching or crossing syntactic constituent boundaries), or single combination performed the best in all language pairs and test sets. Moreover, many a time feature combination resulted in performance drop. Feature selection was imposed by the limitations of the commonly used MERT algorithm (Och, 2003), whose runtime tends to soar, and performance to drop, when attempting to optimize weights of more than 20-25 features; this is a rule-of-thumb only, but it comes from many researchers' experience, including my own. This section addresses the feature selection problem by using MIRA (Crammer and Singer, 2003; Crammer et al., 2006; Watanabe et al., 2007; Chiang et al., 2008) instead of MERT.[10] Soft syntactic constraint features, similar to those described in Section 2.5.1, are tested in an Arabic-English translation task, with and without additional features: David Chiang's structural distortion features (Chiang et al., 2008). Unlike Section 2.5.1, here it is possible to tune all syntactic features in a single model. It is also worth noting that this experimentation is on a considerably larger scale than what is described in Section 2.5.1 and Marton and Resnik (2008).

Margin-Infused Relaxed Algorithm (MIRA) is a large margin classifier, as is support vector machine (SVM), attempting to best separate (classify) data points

[10]This section mainly draws on Chiang et al. (2008), and on personal communication with David Chiang, who, for the experiments described in this section, re-implemented the features described in Marton and Resnik (2008), and introduced the "structural distortion" features briefly mentioned in this section.

as they come, in an online fashion. MERT, by contrast, is a batch (offline) gradient decent algorithm. MERT updates feature weights iteratively, in an attempt to "climb" and maximize an objective function, typically BLEU ; MIRA also cares about the values of the feature weights, attempting iteratively to be close to as many hypotheses as possible, each hypothesis being a set of feature weights – one value for each feature.

The baseline model was Hiero with the following baseline features (Chiang, 2005; Chiang, 2007):

- two language models
- phrase translation probabilities $p(f \mid e)$ and $p(e \mid f)$
- lexical weighting in both directions (Koehn et al., 2003)
- word penalty
- penalties for:
  - automatically extracted rules
  - identity rules (translating a word into itself)
  - two classes of number/name translation rules
  - glue rules

The probability features were base-100 log-probabilities. Base-100 was chosen instead of the commonly used base-10 because in preliminary experimentation features

with large values tended to destabilize the MIRA training, and the larger base makes the probability features smaller in value.

The rules were extracted from all the allowable parallel text from the NIST 2008 evaluation (152+175 million words of Arabic+English, in 6,561,091 parallel sentence), aligned by IBM Model 4 using GIZA++ (union of both directions). Hierarchical rules were extracted from the most in-domain corpora[11] (4.2+5.4 million words in 170,863 parallel sentences) and phrases were extracted from the remainder. The coarse-grained distortion model was trained on the first 10,000 sentences of the training data.[12]

Two language models were trained, with the only difference being that one was trained on data similar to the English side of the parallel text, and the other on 2 billion words of English, mainly from the LDC English Gigaword 2. Both were 5-gram models with modified Kneser-Ney smoothing, lossily compressed using a perfect-hashing scheme similar to that of Talbot and Brants (2008) but using minimal perfect hashing (Botelho et al., 2005).

The documents of the NIST 2004 (newswire) and 2005 Arabic-English evaluation data were randomly partitioned into a tuning set (1178 sentences) and a development set (1298 sentences). The test data was the NIST 2006 Arabic-English evaluation data (NIST part, newswire and newsgroups, 1529 sentences).

[11] LDC2004T17, LDC2005E46, LDC2006E24, LDC2006E25, LDC2006E34, LDC2006E85, LDC2006E86, LDC2006E92, and LDC2006E93.

[12] From personal communication with David Chiang, these sentences were most likely taken from LDC2006E86 and LDC2006E93, which were used for extracting the hierarchical rules.

To obtain syntactic parses for this data, it was tokenized according to the Arabic Treebank standard using AMIRA (Diab et al., 2004), and parsed with the Stanford parser (Klein and Manning, 2003b). Then, the parsing trees were forced back into the MT system's tokenization.[13]

Both MERT and MIRA were run on the tuning set using 20 parallel processors. MERT was stopped when the score on the tuning set stopped increasing, as is common practice; MIRA was stopped when the score on the development set stopped

[13]The only notable consequence is that proclitic Arabic prepositions were fused onto the first word of their NP object, so that the PP and NP brackets were co-extensive.

LDC ID	**Description**
LDC2004T17	Arabic News Translation Text Part 1
LDC2004T18	Arabic English Parallel News Part 1
LDC2005E46	Arabic Treebank English Translation
LDC2004E13	UN Arabic English Parallel Text
LDC2006E24	GALE Y1 - Interim Release: Translations
LDC2006E25	GALE Y1 - Arabic English Parallel News Text
LDC2006E34	GALE Y1 Q2 Release - Translations V2.0
LDC2006E85	GALE Y1 Q3 Release - Translations
LDC2006E86	GALE Y1 Q3 Release - Word Alignment
LDC2006E92	GALE Y1 Q4 Release - Translations
LDC2006E93	GALE Y1 Q4 Release - Word Alignment
LDC2007E07	ISI Arabic-English Automatically Extracted Parallel Text

Table 2.8: Training corpora for Arabic-English translation (MIRA). The permissible parallel texts from the NIST MT 2008 evaluation (`http://www.itl.nist.gov/iad/mig/tests/mt/2008/doc/mt08_constrained.html`)

Use	**Set**	**Size (sentences)**
Training	Table 2.8	6,561,091
Tuning	NIST MT04 (newswire)	1,178
Development	NIST MT05	1,298
Test	NIST MT06 (NIST part, newswire and newsgroups)	1,529

Table 2.9: Training development and test set sizes for Arabic-English translation (MIRA)

increasing, and after no more than 20 iterations.[14] In these runs, MERT took an average of 9 passes through the tuning set and MIRA took an average of 8 passes. For comparison, Watanabe et al. (2007) report decoding their tuning data of 663 sentences 80 times.

2.5.2.1 Syntactic Features (MIRA)

For the MIRA experiments (including the MERT counterparts), the syntactic features were organized into coarse-grained and fine-grained sets, with minor differences in implementation from the features that were used in the experiments described in Sections 2.4 and 2.5.1.

Coarse-grained features As the basis for coarse-grained syntactic features, the following nonterminal labels were selected based on their frequency in the tuning data, whether they frequently cover a span of more than one word, and whether they represent linguistically relevant constituents: NP, PP, S, VP, SBAR, ADJP, ADVP, and QP. In addition to the twelve features in the baseline model, two features were defined: one which fires when a rule's source side span in the input sentence matches any of the above-mentioned labels in the input parse, and another which fires when a rule's source side span crosses a boundary of one of these labels (e.g., its source side span only partially covers the words in a VP subtree, and it also covers some

[14] This MIRA training policy was chosen to avoid overfitting. However, it was possible to use the tuning set for this purpose, just as with MERT: in none of these runs would this change have made more than a 0.2 BLEU difference on the development set.

or all or the words outside the VP subtree). These two features are equivalent to the previously defined $XP^{=}$ and XP^{+} feature combinations, respectively.

Fine-grained features The following nonterminal labels that appear more than 100 times in the tuning data were selected: NP, PP, S, VP, SBAR, ADJP, WHNP, PRT, ADVP, PRN, and QP. The labels that were excluded were mostly parts of speech, and non-constituent labels like FRAG. For each of these labels X, the following separate features were added: one that fires when a rule's source side span in the input sentence matches X, and a second feature that fires when a span crosses a boundary of X. These features are similar to the previously defined $X^{=}$ and X^{+}, except that the set here includes features for WHNP, PRT, and PRN.

2.5.2.2 Arabic-English (MIRA)

Table 2.10 shows the results of the experiments with the training methods and features described above. All significance testing was performed against the first line (MERT baseline) using paired bootstrap resampling (Koehn, 2004b).

MIRA is shown to be competitive with MERT when both use the baseline feature set. Indeed, the MIRA system scores significantly higher on the test set; but when the test set is broken down by genre, one can see that the MIRA system does slightly worse on newswire and better on newsgroups. (This is largely attributable to the fact that the MIRA translations tend to be longer than the MERT translations, and the newsgroup references are also relatively longer than the newswire references.)

When more features are added to the model, the two training methods diverge more sharply. When training with MERT, the coarse-grained pair of syntax features yields a small improvement, but the fine-grained syntax features do not yield any further improvement. By contrast, when the fine-grained features are trained using MIRA, they yield substantial improvements. Similar behavior for the structural distortion features can be observed: MERT is not able to take advantage of the finer-grained features, but MIRA is. Finally, using MIRA to combine both classes of features, 56 in all, produces the largest improvement, 2.6 BLEU points over the MERT baseline on the full test set. More details on MIRA implementation and the distortion features are out of the scope of this chapter, and can be found in Chiang et al. (2008).

			Dev	NIST 06 (NIST part)		
Train	Features	#	nw	nw	ng	nw+ng
MERT	baseline	12	52.0	50.5	32.4	44.6
	syntax (coarse)	14	52.2	50.9	33.0^{+}	45.0^{+}
	syntax (fine)	34	52.1	50.4	33.5^{++}	44.8
	distortion (coarse)	13	52.3	51.3^{+}	34.3^{++}	45.8^{++}
	distortion (fine)	34	52.0	50.9	34.5^{++}	45.5^{++}
MIRA	baseline	12	52.0	49.8^{-}	34.2^{++}	45.3^{++}
	syntax (coarse)	14	NA	$51.1^{?}$	NA	$46.3^{?}$
	syntax (fine)	34	53.1^{++}	51.3^{+}	34.5^{++}	46.4^{++}
	distortion (coarse)	13	NA	$51.6^{?}$	NA	$47.0^{?}$
	distortion (fine)	34	53.3^{++}	51.5^{++}	34.7^{++}	46.7^{++}
	distortion+syntax (fine)	56	$\mathbf{53.6}^{++}$	$\mathbf{52.0}^{++}$	$\mathbf{35.0}^{++}$	$\mathbf{47.2}^{++}$

Table 2.10: Comparison of MERT and MIRA on various feature sets. Key: # = number of features; nw = newswire, ng = newsgroups; + or ++ = significantly better than MERT baseline ($p < 0.05$ or $p < 0.01$, respectively), − = significantly worse than MERT baseline ($p < 0.05$). NA = value currently not available. $^{?}$ = significance currently not available.

2.6 Discussion

The results in Section 2.5 demonstrated, to my knowledge for the first time, that significant and sometimes substantial gains over baseline can be obtained by incorporating soft syntactic constraints into Hiero's translation model – and generally, incorporating source-side soft syntactic constraints into the decoding of a state-of-the-art SCFG SMT system.

Within each language pair tested, one can also see considerable consistency across multiple test sets, in terms of which constraints tend to help most. In the Chinese-English task, the top seven features combinations on the MT06 test set maintain the same rank order on MT08; the top 6 single features on MT06 maintain the same ranking on MT08, with minor permutations between neighboring features in Table tab:Chinese. In the Arabic-English task, the top eight feature combinations show some minor rank permutations between MT06 and MT08, although bigger permutations compared to MT03 (PP+.AdvP= and all-labels2 being the notable "offenders"); and the top five single features on MT06 maintain the same ranking on MT08, with only minor permutations on MT03;.

Furthermore, these results provide some insight into why the original approach may have failed to yield a positive outcome. For Chinese, I found that when I defined finer-grained versions of the exact-match features, there was value for some constituency types in biasing the model to favor matching the source language parse. Moreover, I found that there was significant value in allowing the model to be

sensitive to violations (crossing boundaries) of source parse sub-trees, as opposed to only matching of these syntactic constituent boundaries. These results confirm that parser quality was not the limitation in the original work (or at least not the only limitation), since in these experiments the parser was held constant.

Looking at combinations of new features, some "double-feature" combinations (VP2, IP2) achieved large gains, although note that more is not necessarily better: many combinations of more features did not yield better scores, and some did not yield any gain at all. No conflated feature reached significance, but it is not the case that all conflated features are worse than their same-constituent "double-feature" counterparts. For example, IP2 and IP_ achieve similar scores on the Arabic MT03, and MT06 test sets – but IP_ is about half a BLEU point higher than IP2 on MT08. However, on the Chinese MT06 test set, IP2 is about one point higher than IP_. On same test set, NP_ is about .7 BLEU higher than NP2.

I found no simple correlation between finer-grained feature scores (and/or boundary condition type) and combination or conflation scores. Since some combinations seem to cancel individual contributions, at least when optimized with MERT, I can conclude that the higher the number of participant features (of the kinds described here, optimized with MERT), the more likely a cancellation effect is; therefore, a "double-feature" combination is more likely to yield higher gains than a combination containing more features.

I also investigated whether non-canonical linguistic constituency labels such as PRN, FRAG, UCP and VSB introduce "noise", by means of the XP features

— the XP= feature is, in fact, simply the undifferentiated constituency feature, but sensitive only to "standard" XPs. Although performance of XP=, XP2 and all-labels+ were similar to that of the undifferentiated constituency feature, XP+ achieved the highest gain. Intuitively, this seems plausible: the feature says, at least for Chinese, that a translation hypothesis should incur a penalty if it is translating a substring as a unit when that substring is not a canonical source constituent.

Having obtained positive results with Chinese, I explored the extent to which the approach might improve translation using a very different source language. The approach on Arabic-English translation yielded large BLEU gains over baseline, as well as significant improvements over the undifferentiated constituency feature. A translation example is brought in Figure 2.3, where the noun phrase (NP) for the Syrian representative is broken in the baseline translation, but is correctly cohesively translated in the PP^+ model. Interestingly, this model is only sensitive to PPs, and yet the soft syntactic constraints seemed to have contributed to the SMT output quality nevertheless – perhaps due to a PP that contained the NP for the Syrian representative. A more in-depth future analysis is required to better understand this effect.

Comparing the two sets of experiments, one can see that there are definitely language-specific variations in the value of syntactic constraints; for example, AdvP, the top performer in Arabic, could not have possibly yielded gains directly in Chinese, since in these parses the AdvP constituents rarely spanned more than a single word. At the same time, some IP and VP variants seemed to do generally well in

Source	→... (PP (IN ب) (NP (NP (NN تعسن) (NP (NN مندوب) (NP (NNP سوريا) (NNP لدى)))) (DT ال) (NP (NN امم) (NP (NN ال) (JJ متحدة))))))) ... →
Gloss	...(PP (IN in) (NP (NP (NN appointment) (NP (NN representative) (NP (NNP syria) (NNP to)))) (DT the) (NP (NN nations) (NP (NN the) (JJ united))))))) ...
Reference	[the third decree ordered] the appointment of the syrian representative to the united nations ...
Baseline	... to appoint syria to the united nations representative ...
PP+	... to appoint a representative of syria to the united nations ...

Figure 2.3: Arabic-English translation example (MERT) for the PP$^+$ model. The noun phrase for the Syrian representative (underlined in each model) is broken in the baseline translation, but is correctly cohesively translated in the PP$^+$ model, even though this model is only sensitive to PPs, and the parsing information is sometimes noisy . Arabic source is presented word by word from left to right, to make it easier to read the parsing tags, and compare with the gloss (word by word literal translation) and the models' translation word order.

both languages. This makes sense, since — at least for these language pairs and perhaps more generally — clauses and verb phrases seem to correspond often on the source and target side. I found it more surprising that no NP variant yielded much gain in Arabic; this question will be taken up in future work.

Interestingly, in some cases gains were observed even in the presence of few or none of the tags that the feature was sensitive to, or that these tags did not span more than a single token in the test set. It might be possible that these features helped avoiding the pruning and deletion of important words; indeed the word penalty feature weight was affected, but further research is required to determine the cause. It is also worth noting that this source side soft syntactic constraints approach repeatedly

yielded gains in at least three independent implementations: Marton and Resnik, 2008; Chiang et al., 2008; Xiong et al., 2009 – using SCFG/MERT, SCFG/MIRA, and BTG/MERT (with inner sub-features set with MaxEnt), respectively.

The source side soft syntactic constraints approach presented here is particularly appealing because it can be used unobtrusively with any hierarchically-structured translation model. In principle, it can also be used in "flat" phrase-based SMT systems as well, with some modifications, as in the syntactic cohesion constraints applied by Cherry (2008) and others. It is also appealing in requiring to parse only the development and test sets, which are relatively short, and not the training set, which would result in a considerably longer training time (or the use of a larger computing cluster). The original approach's main drawback was the problem of feature selection, which was removed using MIRA (Chiang et al., 2008).

2.7 Conclusion

When hierarchical phrase-based translation was introduced by Chiang (2005), it represented a new and successful way to incorporate syntax into statistical MT, allowing the model to exploit non-local dependencies and lexically sensitive reordering without requiring linguistically motivated parsing of either the source or target language. An approach to incorporating parser-based constituents in the model was explored briefly, treating syntactic constituency as a soft constraint, with negative results.

In the work presented in this chapter, I returned to the idea of linguistically motivated soft constraints, and I demonstrated that they can, in fact, lead to substantial improvements in translation performance when integrated into the Hiero framework. I accomplished this using constraints that not only distinguish among constituent types, but which also distinguish between the benefit of matching the source parse bracketing, versus the cost of using phrases that cross relevant bracketing boundaries. I demonstrated improvements for Chinese-English translation, and succeeded in obtaining substantial gains for Arabic-English translation, as well. This approach repeatedly yielded positive results, not only when using Hiero with MERT, but also when using Hiero with MIRA, and in subsequent research by Xiong et al. (2009) using BTG with MERT.

These results contribute to a growing body of work on combining monolingually based, linguistically motivated syntactic analysis with translation models that are closely tied to observable parallel training data. Consistent with other researchers, I find that "syntactic constituency" may be too coarse a notion by itself; rather, there is value in taking a finer-grained approach, and in allowing the model to decide how far to trust each element of the syntactic analysis as part of the system's optimization process.

Chapter 3

Soft Semantic Constraints for Word-Pair Similarity Ranking

3.1 Introduction

This chapter introduces the notion of *soft semantic constraints*, in contrast to Chapter 2, which focuses on soft *syntactic* constraints. While the use of parsing information is relatively wide-spread, particularly in SMT, the soft semantic constraints and the hybrid semantic distance measures that employ them are new, and therefore would benefit from investigation of their properties and performance on a basic level and a more intrinsic evaluation first. This chapter investigates soft semantic constraints in semantic models of single words, evaluated in standard word-pair similarity ranking tasks.[1] The next chapter extends these models from single words to word sequences (phrases), and incorporates these soft semantic constraints in phrasal paraphrase generation, tested in SMT, similarly to Chapter 2.

Semantic distance is a measure of the closeness in meaning of two concepts. People are consistent judges of semantic distance. For example, one can easily tell that the concepts of "exercise" and "jog" are closer in meaning than "exercise" and "theater". Studies asking native speakers of a language to rank word pairs

[1] Much of this chapter draws on Marton et al. (2009b).

in order of semantic distance confirm this—average inter-annotator correlation on ranking word pairs in order of semantic distance has been repeatedly shown to be around 0.9 (Rubenstein and Goodenough, 1965; Resnik, 1999). Although the terms *semantic distance*, *semantic similarity*, and *semantic relatedness* are sometimes used inter-changeably in a loose manner, I will mostly follow a distinction detailed in Section 3.2. However, the title of this chapter is one such loose exception, aimed to avoid cumbersome phrasing.

A number of natural language tasks can be framed as semantic distance problems. For example: in word sense disambiguation (Banerjee and Pedersen, 2003; McCarthy, 2006), the target word or phrase's sense, which is closest in meaning to the target's context (if present), must be chosen; in machine translation (Lopez, 2008b), the target language translation hypothesis, which is closest in meaning to the source language phrase or sentence, must be chosen; in spelling correction, a substitute word, that is closer in both meaning to the neighboring words, and edit distance from the (mis-)spelled word, must be chosen; similarly in paraphrase generation, named entity resolution, determining textual entailment (Schilder and Thomson McInnnes, 2006), document summarization (Gurevych and Strube, 2004), (cross-language) information retrieval (Varelas et al., 2005), and so on. Thus, developing automatic measures that are in-line with human notions of semantic distance has received much attention. These automatic approaches to semantic distance rely on manually created lexical resources such as WordNet (Fellbaum, 1998), large text corpora, or both.

WordNet-based information content measures have been successful (Hirst and Budanitsky, 2005), but there are significant limitations on their applicability. They can be applied only if a sufficiently comprehensive WordNet exists for the language of interest (which is not the case for the "low-density" languages); even if there is a WordNet, a number of domain-specific terms may not be encoded in it; or, the WordNet may have too shallow a hierarchy for some word types (e.g, verbs). On the other hand, corpus-based distributional measures of semantic distance, such as cosine and α-skew divergence (Dagan et al., 1999), rely on raw text alone (Weeds et al., 2004; Mohammad, 2008). However, when used to rank word pairs in order of semantic distance or correct real-word spelling errors, they have been shown to perform poorly (Weeds et al., 2004; Mohammad and Hirst, 2006).

Mohammad and Hirst (2006) and Patwardhan and Pedersen (2006) argued that word sense ambiguity is a key reason for the poor performance of traditional distributional semantic distance measures, and they proposed hybrid approaches that are distributional in nature, but also make use of information in lexical resources such as published thesauri and WordNet. However, both these approaches can be applied to estimate the semantic distance between two terms *only* if both terms exist in the lexical resource they rely on. Lexical resources tend to have limited vocabulary and a large number of domain-specific terms are usually not included.

It should also be noted that values from different distance measures are not comparable (even after normalization to the same scale). That is, a similarity score of .75 as per one distance measure does not correspond to the same semantic distance

as a similarity score of .75 from another distance measure. All that can be inferred is that if w_1 and w_2 have a similarity score of .75 and w_3 and w_4 have a score of .5 by the *same* distance measure, then w_1–w_2 are closer in meaning than w_3–w_4. However, if in another distance measure, w_5 and w_6 have a score of .85, and w_7 and w_8 have a score of .4, one *cannot* infer that w_3–w_4 are closer in meaning than w_7–w_8. Moreover, one *cannot* infer that the semantic distance difference between the pairs w_1–w_2 and w_3–w_4 (.25), is smaller than between the pairs w_5–w_6 and w_7–w_8 (.45). Thus if one wishes to use two independent distance measures – in this case: one resource-reliant and one only corpus-dependent – then these two measures are not comparable (and hence cannot be used in tandem, e.g., in a linear combination), even if both rely—partially or entirely—on distributional corpus statistics.

In order to overcome this incomparability challenge, I propose a hybrid semantic distance method that combines the elements of a resource-reliant measure and a strictly corpus-dependent measure by imposing resource-reliant soft constraints on the corpus-dependent model – already at the co-occurrence counts stage, upon which the final value of the measure is based on. I choose the Mohammad and Hirst (2006) method as the resource-reliant method and not one of the WordNet-based measures because, unlike the WordNet-based measures, the Mohammad and Hirst method is distributional in nature and so lends itself immediately for combination with traditional distributional similarity measures. While WordNet-based measures rely mainly on "classical" relations such as *is-a*, and hence are mainly suitable for tasks of semantic similarity in its narrow sense (Morris and Hirst, 2004),

the approach taken here is more general in nature, and naturally applies also to any *semantic relatedness* task (see Section 3.2 for the distinction details).

Briefly, the proposed new hybrid method combines concept–word co-occurrence information (the Mohammad and Hirst distributional profiles of thesaurus concepts (DPC)) with word–word co-occurrence information, to generate word-sense-biased distributional profiles. The "pure" corpus-based distributional profile (a.k.a. *co-occurrence vector*, or *word association vector*), for some target word u, is biased with soft constraints towards each of the concepts c under which u is listed in the thesaurus, in order to create a distributional profile (DP) that is specific to u in the sense that is most related to the other words listed under c. For example, when measuring semantic distance between *water* and *bank*, if the latter is listed under two thesaurus concepts, say, Financial Institution and River, (meaning, it has two senses, each strongly related to one of these concepts), then its DP would be biased first towards the DPC of Financial Institution and then its distance from the DP of *water* would be measured; similarly, it would also be biased towards the DPC of River, and its distance from the DP of *water* would be measured again; assuming the distance of *water* to River (or more precisely, the River-biased *bank*) is shorter, the hybrid method would report it as the distance between *water* and *bank*.

Thus, this method can make more fine-grained distinctions than the Mohammad and Hirst method, and yet uses word sense information.[2] This proposed method

[2]Even though Mohammad and Hirst (2006) use thesaurus categories as coarse concepts, their algorithm can be applied using more finer-grained thesaurus word groupings as well. For example,

falls back gracefully to rely only on word-word co-occurrence information if any of the target terms is not listed in the lexical resource. Experiments on the word-pair ranking task[3] on three different datasets show that the this proposed hybrid measure outperforms all other comparable distance measures.

3.2 Background and Related Work

Strictly speaking, semantic distance/closeness is a property of lexical units—a combination of the surface form and word sense (Cruse, 1986).[4] Two terms are considered to be semantically close if there is a lexical semantic relation between them. Such a relation may be a classical relation such as hypernymy, troponymy, meronymy, and antonymy, or it may be what have been called an ad-hoc non-classical relation, such as cause-and-effect (Morris and Hirst, 2004). If the closeness in meaning is due to certain specific classical relations such as hypernymy and troponymy, then the terms are said to be semantically *similar*. Semantic *relatedness* is the term used to describe the more general form of semantic closeness caused by any semantic relation (Hirst and Budanitsky, 2005). So the nouns *liquid* and *water* are both semantically similar and semantically related, whereas the nouns *boat* and *rudder* are semantically related, but not similar. The challenge of measur-

in a Roget-style thesaurus, each such category is divided to paragraphs and even finer groupings divided by semicolon.

[3]This task involves producing similarity scores for each word-pair, and not only ranking of the pairs; but then the pairs are sorted by score, which produces a ranked list that is compared against a human-rated gold standard.

[4]The notion of semantic distance can be generalized, of course, to larger units such as phrases, sentences, passages, and so on (Landauer et al., 1998).

ing non-classical relations has also been coined "the tennis problem" (attributed to Roger Chaffin by Fellbaum, 1998): In a classical relation-based taxonomy such as WordNet, tennis equipment is under the category *artifact*, tennis players are under *person*, tennis court is under *location*, etc. They don't appear related, and hence a semantic measure based on such a taxonomy will fail to show their relatedness. Distributional semantic distance measures, which are non-classical, and better fit to cope with "the tennis problem", have been surveyed by Curran (2004), Weeds et al. (2004), and Mohammad (2008). Additional relevant research is discussed in the sub-sections below.

The next three sub-sections describe three kinds of automatic distance measures: (1) lexical-resource-based measures that rely on a manually created resource such as WordNet; (2) corpus-based measures that rely only on co-occurrence statistics from large corpora; and (3) hybrid measures that are distributional in nature, and that also exploit the information in a lexical resource.

3.2.1 Lexical-resource-based measures

WordNet is a manually-created hierarchical network of nodes (taxonomy[5]), where each node in the network represents a concept or word sense. An edge between two nodes represents a lexical semantic relation such as hypernymy (*is-a*) and troponymy (*has-part*). WordNet-based measures consider two terms to be close if

[5]I use the term "taxonomy" here in its wider sense, allowing also non-tree structure, that is, multiple inheritance relations.

they occur close to each other in the network (connected by only a few arcs; Lee et al., 1993; Rada et al., 1989), if their definitions share many terms (Banerjee and Pedersen, 2003; Patwardhan and Pedersen, 2006), or if they share a lot of information (Lin, 1998; Resnik, 1999 – which are in fact hybrid methods, described in Section 3.2.3). The length of each arc/link (distance between nodes) can be assumed a unit length, or can be computed from corpus statistics. Within WordNet, the *is-a* hierarchy is much more well-developed than that of other lexical semantic relations. So, not surprisingly, the best WordNet-based measures are those that rely only on the *is-a* hierarchy. Therefore, they are good at measuring semantic similarity (e.g., *doctor–physician*), but not semantic relatedness (e.g., *doctor–scalpel*). Further, the measures can only be used in languages that have a (sufficiently developed) WordNet. WordNet sense information has been criticized to be too fine grained or inadequate for certain NLP tasks (Agirre and Lopez de Lacalle Lekuona, 2003; Navigli, 2006). See Hirst and Budanitsky (2005) for a comprehensive survey of WordNet-based measures.

Lesk (1986) introduced a WSD method which relies on word glosses (definitions) in a dictionary. If a word has several senses listed in the dictionary, the gloss of each sense is compared with the glosses of the surrounding words, and the sense whose gloss has the most overlap in number of words, is chosen. Banerjee and Pedersen (2003), mentioned above, generalized this approach to a semantic relatedness measure that is based on the amount of word overlap in the glosses of two target words of interest.

3.2.2 Corpus-based measures

3.2.2.1 The distributional hypothesis and distributional profiles

Strictly corpus-based measures of distributional similarity rely on the distributional hypothesis. The distributional hypothesis, going back to Firth (1957) and even back to Harris (1940; 1954), assumes that words tend to have a typical distributional profile: They repeatedly appear next to specific other words in a typical rate of co-occurrence. Moreover, words close in meaning tend to appear in similar contexts (where context is taken to be the surrounding words in some proximity). Natural language processing (NLP) applications that assume the distributional hypothesis typically keep track of word co-occurrences in *distributional profiles* (DPs, a.k.a. *collocation vectors*, or *context vectors*). When specifically discussing traditional word-based DPs, as opposed to concept-based or hybrid DPs (see below), I denote them DPW. Each distributional profile DPW_u (for some word u) keeps counts of co-occurrence of u with all words within a usually fixed distance from each of its occurrences (a *sliding window*) in some training corpus. See examples in Table 3.1 and Figure 3.1.[6]

More advanced profiles keep "strength of association" (SoA) information between u and each of the co-occurring words, which is calculated from the counts of u, the counts of the other word, their co-occurrence count, and the count of all

[6]The dimensions of the DP co-occurrence vector can be defined arbitrarily, and do not have to correspond to the words in the vocabulary. The most notable alternative representation is the Latent Semantic Analysis and its variants (Landauer et al., 1998; Finkelstein et al., 2002; Budiu et al., 2006).

Collocate	Co-occurrence Count	Strength-of-Association (SoA)
'hanging'	8	12.20
'ventral'	6	18.44
'trousers'	14	62.44
...	...	...

Table 3.1: Numerical example of a distributional profile (DP) for word *cord*

Figure 3.1: Visual example of a distributional profile for word *bank*. Collocates' strength of association is proportional to their font size.

words in the corpus (corpus size). The information on the other words with respect to u is typically kept in a vector whose dimensions correspond to all words in the training corpus. This is described in Equation (3.1), where V is the training corpus vocabulary:

$$DP_u = \{< w_i, SoA(u, w_i) > | u, w_i \in V\} \text{ for all } i \text{ s.t. } 1 \leq i \leq |V| \tag{3.1}$$

Semantic similarity between words u and v can be estimated by calculating the similarity (vector distance) between their profiles. Slightly more formally, the distributional hypothesis assumes that if we had access to the hypothetical true (psycho-linguistic) semantic similarity function over word pairs, $semsim(u, v)$, then

$$\begin{aligned} &\forall u, v, w \in V, \ \ [semsim(u, v) > semsim(u, w)] \\ &\implies [psim(DPW_u, DPW_v) > psim(DPW_u, DPW_w)], \end{aligned} \tag{3.2}$$

where V is the language vocabulary, DPW_{word} is the distributional profile of $word$, and $psim()$ is a 2-place vector similarity function (all further described below). Paraphrasing and other NLP applications that are based on the distributional hypothesis assume entailment in the reverse direction: the right-hand-side of Formula (3.2) (profile/vector similarity) entails the left-hand-side (semantic similarity).

3.2.2.2 The sliding window and word association (SoA) measures.

Some researchers count *positional* collocations in a sliding window, i.e., the co-counts and SoA measures are calculated per relative position (e.g., for some word/token u, position 1 is the token immediately after u; position -2 is the token preceding the token that precedes u) (Rapp, 1999); other researchers use *non-positional* (which I dub here *flat*) collocations, meaning, they count all token occurrences within the sliding window, regardless of their positions in it relative to u (McDonald, 2000; Mohammad and Hirst, 2006).

Beside simple co-occurrence counts within sliding windows, other SoA measures include functions based on TF/IDF (Fung and Yee, 1998), mutual information (PMI) (Lin, 1998), conditional probabilities (Schuetze and Pedersen, 1997), chi-square test, and the log-likelihood ratio (Dunning, 1993). The formula for calculating log-likelihood ratios of words or phrases u and v is given in Equation (3.3):

$$
\begin{aligned}
LLR(u,v) = -2\log\lambda = \sum_{\substack{i,j\in\{1,2\} \\ k_{ij}>0}} k_{ij}\log\frac{k_{ij}N}{C_iR_j} = \\
= k_{11}\log\frac{k_{11}N}{Count(u)Count(v)} \\
+ k_{12}\log\frac{k_{12}N}{Count(u)[N-Count(v)]} \\
+ k_{21}\log\frac{k_{21}N}{Count(v)[N-Count(u)]} \\
+ k_{22}\log\frac{k_{22}N}{[N-Count(u)][N-Count(v)]}
\end{aligned}
\tag{3.3}
$$

where

$$C_1 = k_{11} + k_{12}$$

$$C_2 = k_{21} + k_{22}$$

$$R_1 = k_{11} + k_{21}$$

$$R_1 = k_{12} + k_{22}$$

$$k_{11} = Count(u, v) = \text{co-occurrence count of } u \text{ and } v$$

$$k_{12} = Count(u) - Count(u, v)$$

$$k_{21} = Count(v) - Count(u, v)$$

$$k_{22} = N - k_{11} - k_{12} - k_{21} = N - Count(u) - Count(v) + Count(u, v)$$

$$N = \sum_{i,j \in \{1,2\}} k_{ij} = \text{total number of tokens in the training corpus}$$

and $Count(.)$ = the number of times the token occurs in the training corpus, but note that the $-2\log$ can be ignored for our purposes.[7]

The formula for calculating point-wise mutual information (PMI) of words or phrases u and v is given in Equation (3.4):

$$PMI(u, v) = \log \frac{Count(u, v)N}{Count(u)Count(v)} = \log \frac{k_{11}N}{Count(u)Count(v)} \tag{3.4}$$

but note that N and the log can be stripped for our purposes.

[7]This formula resembles in form to the one in Rapp (1999), but there the value of k_{22} differs in one term, which makes $N \neq \sum_{i,j \in \{1,2\}} k_{ij}$.

For comparison, the maximum likelihood estimation (MLE) conditional probability strength-of-association measure $p(v|u)$, of words or phrases u and v as above, does not take into account $Count(v)$, or any k_{ij} directly besides k_{11}, or N (although one can argue that N was taken into account, but canceled out in the fraction). It is also asymmetric in u and v:

$$CP(u,v) = p_{MLE}(v|u) = \frac{Count(u,v)}{Count(u)} = \frac{k_{11}}{Count(u)} \tag{3.5}$$

3.2.2.3 Profile similarity measures.

A profile similarity function $psim(DPW_u, DPW_v)$, or generally: $psim(DP_u, DP_v)$, is typically defined as a two-place function, taking vectors as arguments, each vector representing a distributional profile of some word u and v, respectively, and whose cells contain the SoA of u (or v) with each word ("collocate") w_i in the known vocabulary. The vector representation allows for using well studied similarity measures, and also to intuitively think about the distance in geometric analogues, as illustrated in Figure 3.2.

Similarity can be estimated in several ways, e.g., the cosine coefficient, the Jaccard coefficient, the Dice coefficient (all proposed by Salton and McGill, 1983), α-skew divergence (Dagan et al., 1999), and the City-Block measure (Rapp, 1999). The formula for the cosine function for similarity measure is given in Eq. (3.6):

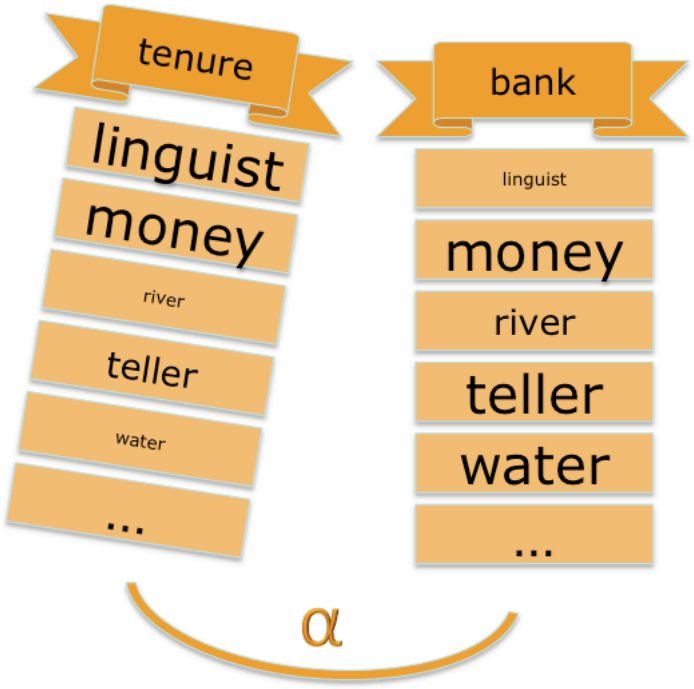

Figure 3.2: Visual example of distributional profile similarity between words *bank* and *tenure*. If each DP is represented as a vector, and each vector, in turn, is represented geometrically as a line (or hyper-plane), the similarity (or distance) between two vectors is represented as the angle α between them. Collocates' strength of association is proportional to their font size, as in Figure reffig:dp. and the.

$$psim(DP_u, DP_v) = cos(DP_u, DP_v) = \frac{\sum_{w_i \in V} SoA(u, w_i) SoA(v, w_i)}{\sqrt{\sum_{w_i \in V} SoA(u, w_i)^2} \sqrt{\sum_{w_i \in V} SoA(v, w_i)^2}} \quad (3.6)$$

The cosine is especially appealing, not only due to its successful track record in NLP similarity tasks. It is easy to compute, requires simple data structures (vectors) as input, and can be intuitively visualized: cosine of two two-dimensional vectors is inversely proportional to their angle α.[8] So two vectors that are identical or very similar (having a similarity score of, say, 1 or close to it on a scale of [0..1]), would make graphically a very small angle (zero or close to it); $\cos \alpha$ approaches 1 when α approaches 0. Conversely, vectors that are very dissimilar (having a similarity score of, say, 0 or close to it on a scale of [0..1]), would be perpendicular or close to it; and $\cos \alpha$ approaches 0 when α approaches a right angle. Any intermediate angle would result in intermediate cosine value, since this function is monotone in this scope. Each dimension of these vectors corresponds to one word in the known vocabulary. Although the graphic analogy is only intuitive in two dimensions, the formula – and similarity principle – can take any finite number of dimensions, i.e., any vocabulary size. Although cosine is not a probability, it uses the same convenient range [0..1], which makes it easy to combine or interpolate with other measures, if so desired.

In principle, any SoA can be used with any profile similarity measure. However, in practice, only some SoA/similarity measure combinations do well, and finding the

[8] To be precise, their smallest angle, $0 - 90°$, ignoring vector directionality.

best combination is still more art than science. Some successful combinations are cos_{CP} (Schuetze and Pedersen, 1997), Lin_{PMI} (Lin, 1998), $City_{LL}$ (Rapp, 1999), and Jensen–Shannon divergence of conditional probabilities (JSD_{CP}; a.k.a. Information Radius; Manning and Schütze, 1999).

These corpus-based measures are very appealing because they rely simply on raw text, but, as described earlier, when used to rank word pairs in order of semantic distance, or to correct real-word spelling errors, they perform poorly, compared to the WordNet-based measures. See Weeds et al. (2004), Mohammad (2008), and Curran (2004) for detailed surveys of distributional measures.

As Mohammad and Hirst (2006) point out, the DP of a word u conflates information about the potentially many senses of u. For example, consider the following. The noun *bank* has two senses River and Financial Institution. Assume that *bank*, when used in the Financial Institution sense, co-occurred with the noun *money* 100 times in a corpus. Similarly, assume that *bank*, when used in the River sense, co-occurred with the noun *boat* 80 times. So the DP of *bank* will have co-occurrence information with *money* as well as *boat*:

DPW(*bank*):

money,100; *boat*,80; *bond*,70; *fish*,77; ...

Assume that the DP of the word *ATM* is:

DPW(*ATM*):

money,120; *boat*,0; *bond*,90; *fish*,0; ...

Thus the distributional distance between the words *bank* and *ATM* will be some sort of an average of the semantic distance between the senses of *bank* and whatever senses "ATM" might have. However, for various natural language tasks, what is needed is the semantic distance between the intended senses of *bank* and *ATM*, which often also tends to be the semantic distance between their closest senses – in this case, most likely the financial senses.

3.2.3 Hybrid measures

Both Mohammad and Hirst (2006) and Patwardhan and Pedersen (2006) proposed measures that are not only distributional in nature but also rely on a lexical resource to exploit the manually encoded information therein as well as to overcome the sense-conflation problem (described in section 3.2.2). Since I essentially combine the Mohammad and Hirst method with a "pure" word-based distributional measure to create the proposed hybrid approach, I briefly describe their method here.

Mohammad and Hirst (2006) generate separate distributional profiles for the different senses of a word, without using any sense-annotated data. They use the categories in a Roget-style thesaurus (*Macquaries* (Bernard, 1986)) as coarse senses or concepts. There are about 1000 categories in a thesaurus, and each category has on average 120 closely related words. A word may be found in more than one category if it has multiple meaning. They use a simple unsupervised algorithm to determine the vector of words that tend to co-occur with each concept and the corresponding strength of association (a measure of how strong the tendency to

co-occur is). The target word u will be assigned one concept DP for each of the concepts that list u. These "distributional profiles of concepts" will be denoted DPCs. DPC(c) gives the number of times the concept (thesaurus category) c co-occurs with each of the words in a corpus. That is, the number of times any word associated with c co-occurs each of the words in the corpus.

Figure 3.3 shows a visual representation of example DPCs of the two concepts pertaining to *bank*, illustrating that the word *bank* is mapped to each of its senses (i.e., each of the concepts listing it in the thesaurus: FINANCIAL INSTITUTION and RIVER). It also illustrates that some collocates are more strongly associated with one sense (DPC) of *bank*, while others are more strongly associated with its other sense. For example, *money* is strongly associated with the FINANCIAL INSTITUTION sense (larger font), but not with the RIVER sense (smaller font). Conversely, *water* is strongly associated with RIVER. Below is also a numerical representation of such DPCs, partly with different collocates:[9]

DPC(FINANCIAL INSTITUTION):

money,1000; *boat*,32; *bond*,705; *fish*,0; ...

DPC(RIVER):

money,5; *boat*,863; *bond*,0; *fish*,948; ...

[9]The relatively large co-occurrence frequency values for DPCs as compared to DPWs is because a concept can be refered to by many words (on average 100).

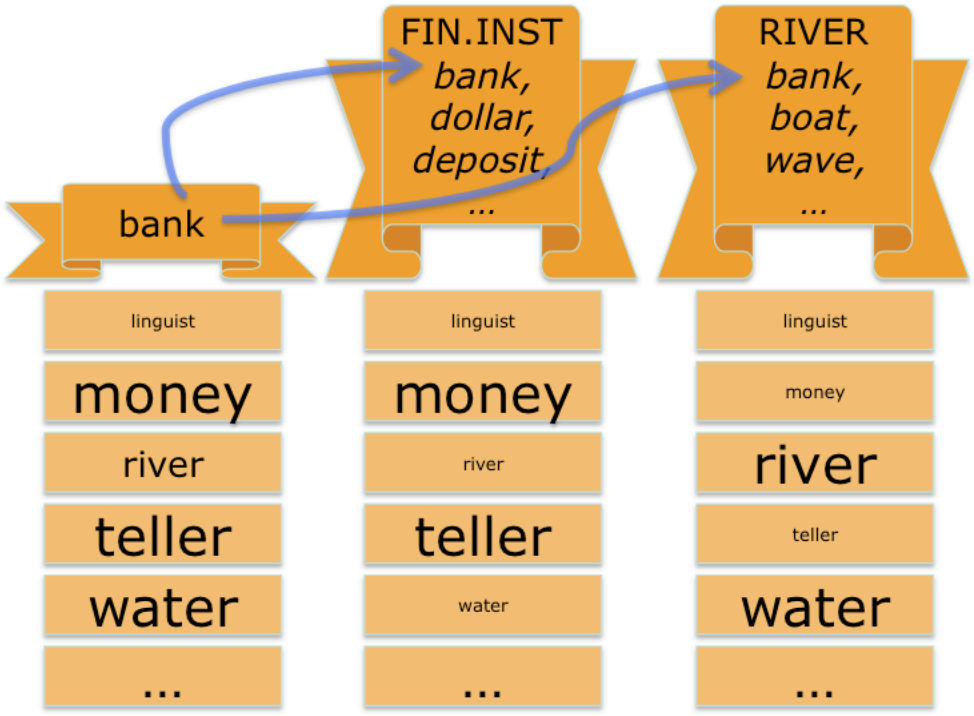

Figure 3.3: Visual example of concept-based distributional profiles serving as coarse word senses for word *bank*, illustrating that the word *bank* is mapped to each of its senses, here FINANCIAL INSTITUTION and RIVER. Some collocates are more strongly associated with one sense (DPC) of *bank*, while others are more strongly associated with its other sense. For example, *money* is strongly associated with the FINANCIAL INSTITUTION sense (larger font), but not with the RIVER sense (smaller font). Conversely, *water* is strongly associated with RIVER.

Here, too, one can see that *money* is strongly associated with the FINANCIAL INSTITUTION sense, but not with the RIVER sense. And conversely, *boat* is more strongly associated with RIVER.

The distance between two words u, v is determined by calculating the closeness of each of the DPCs of u to each of DPCs of v, and the closest DPC-pair distance is chosen. The strategy of choosing closest distance, or maximal similarity, has been taken before, e.g., Rada et al. (1989) and Resnik (1999).

Mohammad and Hirst (2006) show that their approach performs better than other strictly corpus-based approaches that they experimented with. However, all those experiments were on word-pairs that were listed in the thesaurus. Their approach is not applicable otherwise. Note also that if target words u and v appear under the same concept c, the semantic distance between u and v would be indistinguishable, since the concept-based similarity measure returns the semantic distance of the closest sense pair. For example, if the word *bank* has the two above-mentioned senses Financial Institution and River, and the word *wave* has the senses Physics and River, there are $2 \times 2 = 4$ DPC pairs to compare:

Financial Institution, Physics

Financial Institution, River

River, Physics

River, River

The last, identical pair would be returned, falsely representing synonymity between *bank* and *wave*. This is illustrated in Figure 3.4, and addressed in Sections 3.3 and 3.4 below. I show in these sections how cosine–log-likelihood-ratio (or any comparable distributional measure) can be combined with the Mohammad and Hirst DPCs to form a hybrid approach that is not limited to the vocabulary of a lexical resource, and uses a more fine-grained representation that alleviates the false synonymity problem.

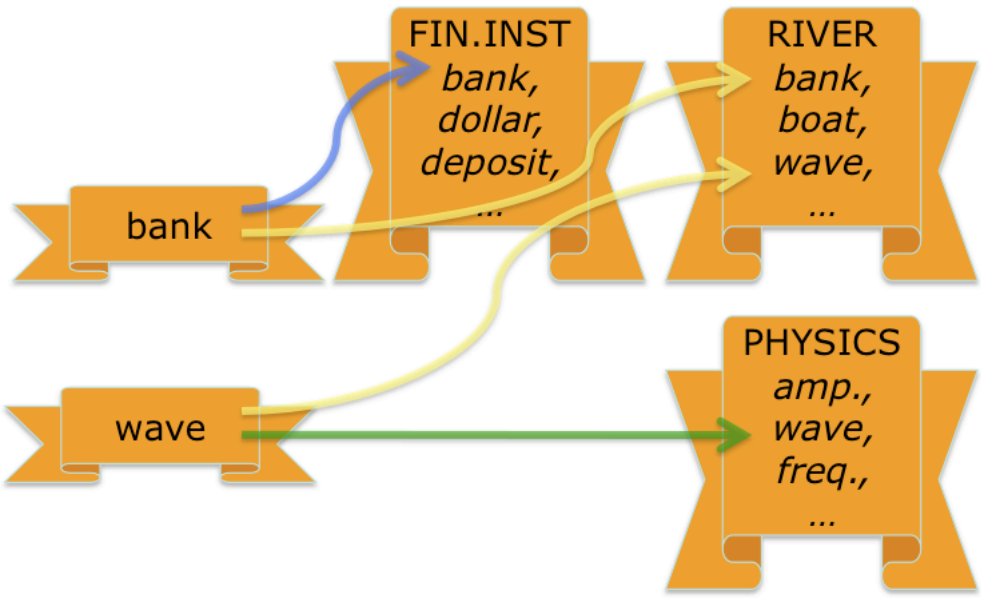

Figure 3.4: Problem of false synonymity representation with coarse DPCs due to mapping one sense of each target word to the same DPC: the concept-based similarity measure returns the semantic distance of the closest sense pair, which in this case is the identical sense pair (River,River).

Erk and Padó (2008) proposed a way of representing a word sense in context by biasing the target word's DP according to the context surrounding a target (specific) occurrence of the target word. They use dependency relations and selectional preferences of the target word and combine multiple DPs of words appearing in the context of the target occurrence, in a manner so as to give more weight to words co-occurring with both the target word and the target occurrence's context words. The advantage of their approach is that it does not rely on a thesaurus or WordNet. Its disadvantage is that it relies on dependency relations and selectional preferences information, which might not be available, or be of low quality for the language of interest. Also, the context information it uses in order to determine the word sense is quite limited (only the words surrounding a single occurrence – the target

occurrence of the target word) and hence the representation of that sense might not be sufficiently accurate. Since they treat each occurrence of the target word separately, their approach effectively assumes that each occurrence of a word has a unique sense.

Resnik (1999) introduced a hybrid model for calculating "information content" (Ross, 1976). In order to calculate it for a certain concept in the WordNet hierarchy, one traverses the concept's subtree, and sums the corpus-based word frequencies of all words under that concept, and all concept nodes in that subtree, recursively. A maximum likelihood log-probability estimation is then calculated by dividing that sum by the total number of word occurrences in the corpus, and taking the negative log. The semantic distance of two words is defined as the information content of the most informative common subsumer (the subsumer with the highest information content) of the two words, in the WordNet hierarchy. In case a word appears in more than one concept (i.e., it has more than one sense), the minimal distance between the cross product of its senses and the other word's senses is chosen. This measure is hybrid in the sense that it uses both a linguistic knowledge source and a large corpus of text, although it doesn't use the distributional contexts of the words in the corpus. Lin (1997) and Jiang and Conrath (1997) improved on this idea by incorporating the distance of each word from the lowest common subsumer, following the intuition that words that are closer to that subsumer are likely to be more similar than those that are far below it in the WordNet hierarchy.

3.3 New Distributional Measures with Soft Semantic Constraints

To recap previous sections about different types of distributional profiles, traditional distributional profiles of words (DPW) give word–word co-occurrence frequencies. For example, DPW(u) gives the number of times the target word u co-occurs with with all other words:

DPW(u):

w_1,f(u,w_1); w_2,f(u,w_2); w_3,f(u,w_3); ...

where f stands for co-occurrence frequency (and can be generalized to stand for any strength of association (SoA) measure such as the log-likelihood ratio, see third column in Table 3.1). Mohammad and Hirst create concept–word co-occurrence vectors, "distributional profiles of concepts" (DPCs), from non-annotated corpus. DPC(c) gives the number of times the concept (thesaurus category) c co-occurs with all the words in a corpus.

DPC(c):

w_1,f(c,w_1); w_2,f(c,w_2); w_3,f(c,w_3); ...

A target word u that appears under thesaurus concepts c_1, ..., c_n would be assigned to each of DPC(c_1), ..., DPC(c_n), respectively. Therefore, if a target word v also appears under some same concept c, the DPCs of u and v would be indistinguishable; also, if the target word does not appear in the thesaurus, this measure is inapplicable.

3.3.1 The hybrid-sense-proportional-counts method

In order to address the above-mentioned limitations (indistinguishable DPCs of u and v and vocabulary-limited applicability), one can use hybrid DPs that would benefit from both the word sense awareness of concept-based DPCs, and the large applicability of word-based DPWs. This can be achieved by using distributional profiles of word senses (DPWS(u_c)) that represent the strength-of-association (SoA) of the target word u, when used in sense c, with each of the words in the corpus:

DPWS(u_c):

w_1,f(u_c,w_1); w_2,f(u_c,w_2); w_3,f(u_c,w_3); ...

In order to get exact counts, one needs sense-annotated data. However, such data is expensive to create, and is scarce. Instead, one could estimate these counts from the DPW and DPC counts. One could use the concept-based DPCs as soft semantic constraints over the word-based DPWs (elaborated also in Section 5.5). The intuition here is to distribute each DPW co-occurrence count among the target's senses, in proportion to the relative co-occurrence with each sense, as estimated in the DPCs. This is expressed more formally in Equation 3.7:

$$f(u_c, w_i) = p(c|w_i) \times f(u, w_i) \tag{3.7}$$

where the conditional probability $p(c|w_i)$ is calculated from the co-occurrence frequencies in DPCs; and the co-occurrence count $f(u, w_i)$ is calculated from DPWs.

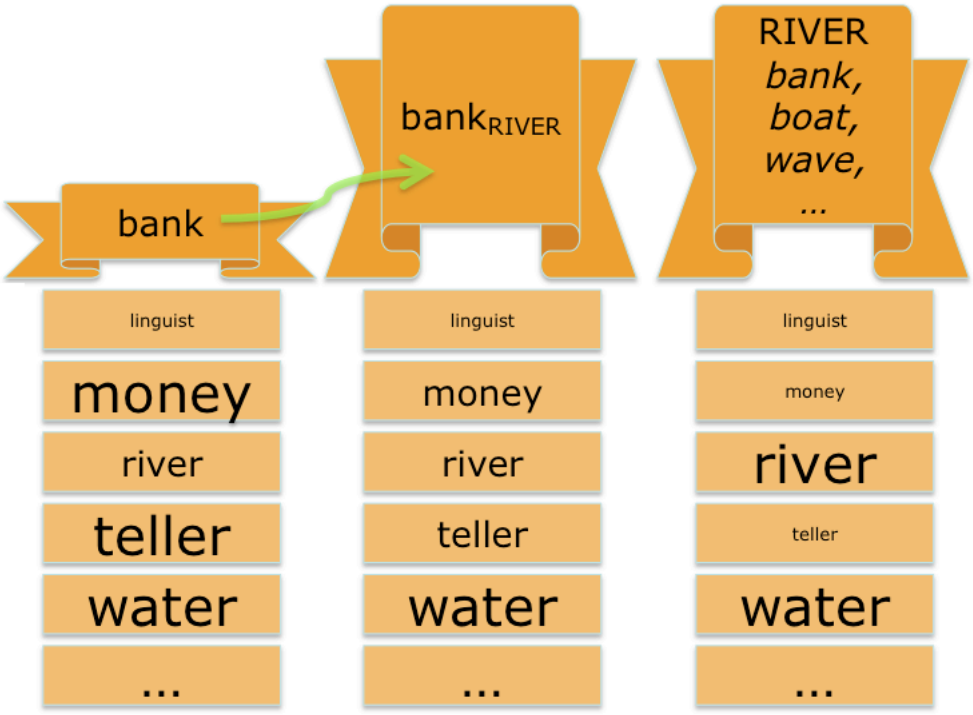

Figure 3.5: Visual example of a sense-aware distributional profile: the DPWS for the word *bank* in sense River. The *bank*'s strength of association with *money* in the DPWS is decreased relative to the DPW, since it is discounted in proportion to its value in the DPC of River, relative to its value in all the DPCs of *bank*.

If the target word is not in the thesaurus's vocabulary, then I assume uniform distribution over all concepts, and in practice I treat it as having a single sense, and take the conditional probability to be 1. Since the method takes sense-proportional co-occurrence counts, I will refer to this method as the **hybrid-sense-proportional-counts method** (or, **hybrid-proportional** for short). For example, Figure 3.5 visualizes an example DPWS of *bank*, created from the DPW of *bank* biased towards the River sense. In this example, the *bank*'s strength of association with *money* in the DPWS is decreased relative to the DPW, since it is discounted in proportion to its value in the DPC of that sense, relative to its value in all the DPCs of *bank*.

Below is also a numerical example of the DPWS of *bank*, here in the FINANCIAL INSTITUTION sense, calculated from its DPW and DPCs:

1. DPW(*bank*):

   money,100; *boat*,80; *bond*,70; *fish*,77; ...

2. (a) DPC(FINANCIAL INSTITUTION):

   money,1000; *boat*,32; *bond*,705; *fish*,0; ...

   (b) DPC(RIVER):

   money,5; *boat*,863; *bond*,0; *fish*,948; ...

3. DPWS($bank_{\text{FINANCIAL INSTITUTION}}$):

   money,$(\frac{1000}{1000+5} \times 100)$; *boat*,$(\frac{32}{32+863} \times 80)$; *bond*,$(\frac{705}{705+0} \times 70)$; *fish*,$(\frac{0}{0+948} \times 77)$; ...

Once the DPWS co-occurrence counts are calculated, any counts-based SoA and distance measures can be applied. For example, in this work I use log-likelihood ratio (Dunning, 1993) to determine the SoA between a word sense and co-occurring words, and cosine to determine the distance between two DPWS's log likelihood vectors (McDonald, 2000). I also contrast this measure with cosine of conditional probabilities vectors (Schuetze and Pedersen, 1997). Given two target words, the distance between each of their DPWS pairings is determined, and the closest DPWS-pair distance is chosen.

3.3.2 The hybrid-sense-filtered-counts method

Since the DPCs are created in an unsupervised manner, they are expected to be somewhat noisy. Therefore, I also experimented with a variant of the method proposed above, that simply makes use of whether the conditional probability $p(c|w_i)$ is greater than 0 or not:

$$f(u_c, w_i) = \begin{cases} f(u, w_i) & \text{If } p(c|w_i) > 0 \\ 0 & \text{Otherwise} \end{cases} \tag{3.8}$$

Since this method essentially filters out collocates that are likely not relevant to the target sense c of the target word u, I will refer to this method as the **hybrid-sense-filtered-counts method** (or, just **hybrid-filtered** for short). Below is an example hybrid-filtered DPWS of *bank* in the FINANCIAL INSTITUTION sense:

4. DPWS($bank_{\text{FINANCIAL INSTITUTION}}$):

   money,100; *boat*,80; *bond*,70; ...

Note that the collocate *fish* is now filtered (zeroed) out, compared with the hybrid-proportional DPWS example 3 above, whereas *bank's* co-occurrence counts with *money, boat,* and *bond* are left unchanged from the DPW example 1 (and are not sense-proportionally attenuated).

3.4 Experiments

I evaluated various semantic distance measures on the task of ranking word pairs in order of semantic distance. These included my new hybrid sense-biased methods as well as several baselines: the Mohammad and Hirst (2006) DPC-based methods, the traditional word-based distributional similarity methods, and several Latent Semantic Analysis (LSA)-based methods. I used three testsets and their corresponding human judgment gold standards: (1) the Rubenstein and Goodenough (1965) set of 65 noun pairs—denoted **RG-65**; (2) the WordSimilarity-353 (Finkelstein et al., 2002) set of 353 noun pairs (which include the RG-65 pairs) of which I discarded of one repeating pair—denoted **WS-353**; and (3) the Resnik and Diab (2000) set of 27 verb pairs—denoted **RD-00**.

3.4.1 Corpora and Pre-processing

I generated distributional profiles (DPWs and DPCs) from the *British National Corpus (BNC) (Burnard, 2000)*, which is a balanced corpus. I lowercased the characters, and stripped numbers, punctuation marks, and any SGML-like syntactic tags, but kept sentence boundary markers. The BNC contained 102,100,114 tokens of 546,299 types (vocabulary size) after tokenization. For the verb set, I also lemmatized this corpus.

I considered two words as co-occurring if they occurred in a window of ±5 words from each other. I stoplisted words that co-occurred with more than 2000 word types.

I use here cosine of the following SoA vectors: conditional probabilities (Schuetze and Pedersen, 1997), log-likelihood ratios (McDonald, 2000), and PMI (Lin, 1998).

3.4.2 Results

The Spearman rank correlations of the automatic rankings of the RG-65, WS353, and RD-00 testsets with the corresponding gold-standard human rankings are listed in Table 3.2.[10] The correlations were calculated using Richard Lowry's VassarStats statistical computation web site.[11] The higher the Spearman rank correlation, the more accurate is the distance measure.

3.4.2.1 Results on the RG-65 testset

Baselines. I replicated the traditional word-based distributional distance measure using cosine of vectors (DPs) containing conditional probabilities (**word-cos-cp**). Its rank correlation of .53 is close to the correlation of .54 reported in Mohammad and Hirst (2006), hereafter MH06. I replicated the MH06 concept-based approach (**concept-cos-cp**), and its bootstrapped variant that uses a smaller concept–word co-occurrence matrix (**concept*-cos-cp**). The latter yielded a correlation score .65,

[10] Certain experiments were not pursued as they were redundant in supporting my claims.

[11] `http://faculty.vassar.edu/lowry/corr_rank.html`

Method	RG-65	WS-353	RD-00
Baselines (replicated):			
Traditional distributional measures			
word-cos-cp	.53	.31	$.46^{+}$
word-cos-ll	.73	**.54**	$.50^{++}$
word-cos-pmi	.62	.43	.57
Mohammad and Hirst methods and variants			
concept-cos-cp	.62	.38	$.41^{+}$
concept*-cos-cp	.65	.33	$.43^{+}$
concept-cos-ll	.60	.37	$.43^{+}$
concept*-cos-ll	.64	.25	$.27^{-}$
concept*-cos-pmi	$.40^{++}$	.19	$.28^{-}$
Other (LSA and variants)			
LSA	.56	.47	$.55^{++}$
GLSA-cos-pmi	$.18^{-}$	n.p.	n.p.
GLSA-cos-ll	.47	n.p.	$.29^{-}$
New methods:			
hybrid-proportional-cos-ll	.72	.49	$.38^{+}$
hybrid-proportional*-cos-ll	.69	.46	$.39^{+}$
hybrid-filtered-cos-ll	.73	**.54**	$.53^{++}$
hybrid-filtered*-cos-ll	**.77**	**.54**	$.45^{+}$
hybrid-proportional*-cos-pmi	.58	.43	**.71**
hybrid-filtered*-cos-pmi	.61	.42	.64

Table 3.2: Spearman rank correlation on the noun-noun RG-65 (Rubenstein and Goodenough, 1965), the noun-noun WS-353 (Finkelstein et al., 2002), and the verb-verb RD-00 (Resnik and Diab, 2000) testsets, trained on BNC. Best correlations were always achieved by a hybrid (fine-grained) variant, with the strongest corresponding word-based method baseline on that test. Log-likelihood ratio (-ll) methods did best on noun pair test sets, while -pmi methods did best on the verb test set. '*' indicates the use of a smaller bootstrapped concept–word co-occurrence matrix. 'n.p.' indicates that the experiment was not pursued. All correlation scores are significant, $p < .001$, unless noted $^{+}$, $^{++}$for $p < .05, .01$, respectively, or insignificant: $^{-}$, $p > .1$.

close to the .69 reported in MH06. I also experimented with cosine of log-likelihood ratios (**word-cos-ll**), which obtained a correlation of .70 – best among the baseline methods, and cosine of PMI vectors (**word-cos-pmi**), which obtained a correlation of .62.

I conducted experiments with Latent Semantic Analysis (LSA; Landauer et al., 1998) and its GLSA variants (Budiu et al., 2006) as additional baselines. A limited vocabulary of the 33,000 most frequent words in the BNC and all test words was used in these experiments. (A larger vocabulary was computationally expensive and 33,000 is also the vocabulary size used by Budiu et al. (2006) in their LSA experiments.)

New Methods: Since word-cos-ll gave best noun-pair results among the baseline methods, and word-cos-pmi gave best verb-pair results among the baseline methods, I chose to concentrate on using them in the implementations of the hybrid method. The hybrid method variants presented in this chapter (**hybrid-proportional-cos-ll** and **hybrid-filtered-cos-ll**) were the best performers on the RG-65 test set. Particularly, they performed better than both the traditional word-distance measures (**word-cos-ll**), and the concept-based methods—variants of the MH06 method that are used with likelihood ratios (**concept-cos-ll, concept*-cos-ll**). The -pmi methods were all poorer performers than their -ll counterparts on the noun test sets. The -pmi hybrid variants obtained higher scores than the concept-based ones, but about the same scores as the word-based ones.

3.4.2.2 Results on WS-353 and RD-00 testsets

On WS-353, all the proposed hybrid methods out-performed their concept counterparts, and were on par with their word-based counterparts. On RD-00, **word-cos-pmi** out-performed all other word-based methods, and the hybrid -pmi methods were best performers with scores of .64 and .71.

The word-cos-ll, hybrid-proportional-cos-ll, and the two hybrid pmi results on RD-00 are better than any non-WordNet results reported by Resnik and Diab (2000), including their syntax-informed methods—the variants of Lin ("distrib", .43) and Dorr ("LCS", .39). In fact, the hybrid*-prop-cos-pmi and hybrid*-filt-cos-pmi results reach correlation levels of the WordNet-based methods reported there (.66–.68). Also, on WS-353, the hybrid sense-filtered variants and word-cos-ll obtained a correlation score higher than published results using WordNet-based measures (Jarmasz and Szpakowicz, 2003) (.33 to .35) and Wikipedia-based methods (Ponzetto and Strube, 2006) (.19 to .48); and very close to the results obtained by thesaurus-based (Jarmasz and Szpakowicz, 2003) (.55) and LSA-based methods (Finkelstein et al., 2002) (.56).

The lower correlation scores of all measures on the WS-353 test set are possibly due to it having politically biased word pairs (examples include: *Arafat–peace, Arafat–terror, Jerusalem–Palestinian*) for which BNC texts are likely to induce low correlation with the human raters of WS-353. This testset also has disproportionately many terms from the news domain.

The concept methods performed poorly on WS-353 partly because many of the target words do not exist in the thesaurus. For instance, there were 17 such word types that occurred in 20 WS-353 testset word pairs. When excluding these pairs, concept-cos-cp goes up from .38 to .45, and concept*-cos-pmi from .19 to .24. Interestingly, results of the hybrid methods show that they were largely unaffected by the out-of-vocabulary problem on the WS-353 dataset.

On the verbs dataset RD-00, while hybrid-proportional-cos-ll fared slightly better than word-cos-ll, using the smaller matrix seemed to hurt performance of hybrid*-prop-cos-ll compared to word-cos-ll. But results suggest that the -pmi methods might serve as a better measure than -ll for verbs, although this should be tested more rigorously.

Human judgments of semantic distance are less consistent on verb-pairs than on noun-pairs, as reflected in inter-rater agreement measures in Resnik and Diab (2000) and others. Thus, not surprisingly, the scores of almost all measures are lower for the verb data than the RG-65 noun data.

3.5 Discussion

The hybrid methods presented in this chapter obtained higher accuracies than all other methods on the RG-65 testset (all of whose words were in the published thesaurus), and on the RD-00 testset, and their performance was at least respectable on the WS-353 testset (many of whose words were not in the published thesaurus).

This is in contrast to the concept-distance methods which suffer greatly when the target words are not in the lexical resource (here, the thesaurus) they rely on, even though these methods can make use of co-occurrence information of words not in the thesaurus with concepts from the thesaurus.

Amongst the two hybrid methods proposed, the **sense-filtered-counts** method performed better using the smaller bootstrapped concept–word co-occurrence matrix whereas the sense-proportional method performed better using the larger concept–word co-occurrence matrix. I believe this is because the bootstrapping method proposed in Mohammad and Hirst (2006) has the effect of resetting to 0 the small co-occurrence counts. The noise from these small co-occurrence counts affects the sense-filtered-counts method more adversely (since any non-zero value will cause the inclusion of the corresponding collocate's full co-occurrence count) and so the bootstrapped matrix is more suitable for this method.

The results also show that the cosine of log-likelihood ratios method mostly performs better than cosine of conditional probabilities and the pmi methods on the noun sets. This further supports the claim by Dunning (1993) that log-likelihood ratio is much less sensitive than pmi to low counts. Interestingly, on the verb set, the pmi methods, and especially hybrid*-prop-cos-pmi, did extremely well. The differences between Equations (3.3) and (3.4) suggests that the last three terms in Equation (3.3) are helpful for computing semantic similarity of noun target words, but hurt that of verb targets. Further investigation is needed in order to determine if pmi is indeed more suitable for verb semantic similarity, and why.

3.6 Conclusion

Traditional distributional similarity conflates co-occurrence information pertaining to the many senses of the target words. Mohammad and Hirst (2006) showed how to use distributional measures in order to compute distance between coarse word senses (concepts, thesaurus categories). They obtained better results than traditional distributional similarity. However, their method required that the target words be listed in the thesaurus, which is often not the case for domain-specific terms and named entities. In this chapter, I presented hybrid methods (**hybrid-sense-filtered-counts** and **hybrid-sense-proportional-counts**) combining word–word co-occurrence information (traditional distributional similarity) with word–concept co-occurrence information (Mohammad and Hirst, 2006). This was done using soft constraints in such a manner that the method makes use of information encoded in the thesaurus when available, and degrades gracefully if the target word is not listed in the thesaurus. The presented method generates distributional profiles (DPs), which are word-sense-biased (denoted DPWS), from non-annotated corpus-based word-based DPs (DPW) and coarser-grained aggregated thesaurus-based "concept DPs" (DPC). I showed that the hybrid method, employing finer-grained soft semantic constraints than Mohammad and Hirst (2006), correlated with human judgments of semantic distance in most cases better than any of the other methods I replicated – word-based and concept-based alike.

Mohammad et al. (2007) showed that their method could be used to compute semantic distance in a resource poor language L_1 by combining its text with a thesaurus in a resource-rich language L_2 using an L_1–L_2 bilingual lexicon to create cross-lingual distributional profiles of concepts, that is, L_2 word co-occurrence profiles of L_1 thesaurus concepts. Since the method in this chapter makes use of the Mohammad and Hirst DPCs, it can just as well make use of their cross-lingual DPCs, to compute semantic distance in a resource-poor language, just as they did. I leave that for future work.

For future research I would also be interested in improving semantic distance measures for verb–verb, adjective–adjective, and cross-part-of-speech pairs, by exploiting specific information pertaining to these parts of speech in lexical resources in addition to purely co-occurrence information.

Chapter 4

Monolingually-Derived Phrasal Paraphrase Generation for Statistical Machine Translation

4.1 Introduction

This chapter extends the distributional profiles (DPs) and the semantic distance measures described in Chapter 3, from modeling single words (unigrams) to arbitrary word sequences. In addition, the semantic measures' power is extended, from verification (given words or phrases x and y, return their semantic similarity score) to active semantic problem solving, i.e., paraphrase generation (given a word or phrase x return another word or phrase y that is most similar to x semantically). These extensions are implemented within a new phrasal paraphrase generation technique, and are evaluated within a statistical machine translation (SMT) framework, with weighted log-linear features, similarly to the evaluation of soft syntactic constraints in Chapter 2. The paraphrase engine itself is general, and can incorporate any semantic distance measure.[1] As in Chapter 3, the "pure" corpus-based distributional semantic distance measure is compared with the hybrid knowledge / corpus-based measure, applied here in the service of paraphrase generation for SMT.

[1]Much of this chapter draws on Marton et al. (2009a).

Paraphrase generation is a task that serves various natural language processing (NLP) applications, such as natural language generation (NLG), summarization, information retrieval (IR), question answering (QA), and – as mentioned above – statistical machine translation (SMT). It is useful for SMT because it helps increasing translation coverage. Phrase-based SMT systems, flat and hierarchical alike (Koehn et al., 2003; Koehn, 2004a; Koehn et al., 2007; Chiang, 2005; Chiang, 2007), have achieved a much better translation quality than word-based ones (Brown et al., 1993), mainly by learning correct local dependency reordering, since phrases, spanning several words, inherently capture local word order; but untranslated words and phrases (including reordering of known words in unseen sequences) remain a major problem in SMT. According to Callison-Burch *et al.* (2006), a SMT system with a training corpus of 10,000 words learned only 10% of the vocabulary (i.e., 10% of the types, not of the tokens); the same system learned about 30% of the types with a training corpus of 100,000 words; and even with a large training corpus of nearly 10,000,000 words it only reached about 90% coverage of the source vocabulary. Coverage of higher order n-grams is even harder. This out-of-vocabulary (OOV) problem plays a major part in reducing machine translation quality, as reflected by both automatic measures such as BLEU (Papineni et al., 2002) and human judgment tests. Improving translation coverage accurately is therefore important for SMT systems.

The first solution that might come to mind is to use larger parallel training corpora. However, current state-of-the-art SMT systems cannot learn from non-aligned

corpora, while sentence-aligned parallel corpora (bitexts) are a limited resource (See Section 4.2 for discussion of automatically-compiled bitexts). Another direction might be to make use of non-parallel corpora for training. However, this requires developing techniques to extract alignments or translations from them, and in a sufficiently fast, memory-efficient, and scalable manner. One approach that can, in principle, better exploit both alignments from bitexts and make use of non-parallel corpora is the distributional collocational approach, e.g., as used by Fung and Yee (1998) and Rapp (1999). However, the systems described there are not easily scalable, and require pre-computation of a very large collocation counts matrix. Related attempts propose generating bitexts from comparable and "quasi-comparable" bilingual texts by iteratively bootstrapping documents, sentences, and words (Fung and Cheung, 2004), or by using a maximum entropy classifier (Munteanu and Marcu, 2005). Alignment accuracy remains a challenge for them.

Recent work has proposed augmenting the training data with paraphrases generated by pivoting through other languages (Bannard and Callison-Burch, 2005; Callison-Burch et al., 2006; Madnani et al., 2007). This indeed alleviates the vocabulary coverage problem, especially for the resource-poor, so-called "low density" languages. However, these approaches still require bitexts where one side contains the original source language.

The paradigm described in this chapter involves constructing monolingual distributional profiles (see Section 3.2.2.1) of out-of-vocabulary words and phrases in the source language; then, generating paraphrase candidates from phrases that co-

occur in similar contexts, and assigning them similarity scores. The highest ranking paraphrases are used to augment the translation phrase table. The table augmentation idea is similar to that of Callison-Burch et al. (2006), but the paradigm presented here does not require using a limited resource such as parallel texts in order to generate paraphrases. Moreover, this paradigm can, in principle, achieve large-scale acquisition of paraphrases with high semantic similarity.[2] However, using parallel training texts in pivoting techniques offers the potential advantage of implicit translational knowledge, in the form of sentence alignments, while the new approach is unguided in this respect. Therefore, I conducted experiments to find out how these relative advantages play out. I present here, to my knowledge for the first time, positive results of integrating distributional monolingually-derived paraphrases in an end-to-end state-of-the-art SMT system.

In the rest of this chapter I discuss related work in Section 4.2, describe distributional profiles of phrases in Section 4.3, and present the monolingually-derived paraphrase generation system in Section 4.4. I report experiments and results using "pure" corpus-based semantic distance measures and hybrid knowledge / corpus-based measures for paraphrasing in Section 4.5. I conclude by discussing the implications and future research directions in Section 4.6.

[2]The term "similarity" is used loosely here; see Section 3.2.

4.2 Related Work

This is not the first to attempt to ameliorate the out-of-vocabulary (OOV) words problem in statistical machine translation, and other natural language processing tasks. These attempts can be roughly divided into the following categories:

- augmenting current resources (typically parallel texts) with paraphrases of their elements,
- creating additional resources of same type (additional parallel texts), and
- using alternative resources (lesser or no reliance on parallel texts).

This work belongs to the first category, and therefore I mainly focus here on paraphrasing work. Paraphrase generation techniques can be described along various axes:

Number of languages: monolingual or multilingual textual resources.

Resource type: parallel text (bitext), comparable text, or non-related text / "mono-text" (one monolithic corpus).

Paraphrasing method: SMT (translating from and to the same language), pivoting (translating to another language and back), distributional (relying on similar contexts in which the paraphrases tend to occur), morphological and character-based analysis (compounds, edit distance), or other (e.g., time-locked bursts of terms such as *earthquake* in one or more languages).

Paraphrasing unit: word, phrase (any word sequence), syntactic constituent, paragraph, sentence, document...

Use of linguistic knowledge: syntactic information (parses), semantic information (WordNet hierarchy, thesaurus concepts, ...), none.

Paraphrasing object: paraphrasing source language elements in SMT, paraphrasing translation references (target language) elements in SMT, other (non-SMT-related, e.g., for document summarization).

This work uses monolingual, non-related text / mono-text in order to generate phrasal paraphrases with distributional techniques, optionally using semantic information, and extensible to using syntactic information as well. OOV phrases in the source language are paraphrased and then used to augment a SMT translation model (details in Sections 4.4 and 4.5).

This work is most closely related to that of Bannard and Callison-Burch (2005) and Callison-Burch et al. (2006), who also augment translation models with source-side paraphrases of the OOV phrases. Therefore I begin with describing their approach first. There, paraphrases are generated from bitexts of various language pairs, by "pivoting": translating the OOV phrases to an additional language (or languages) and back to the source language. This is illustrated in Figure 4.1. The quality of these paraphrases is estimated by marginalizing translation probabilities to and from the additional language side(s) e, as follows: $p(f_2|f_1) \approx \sum_e p(e|f_1)p(f_2|e)$. A major disadvantage of their approach is that it relies on the availability of parallel corpora

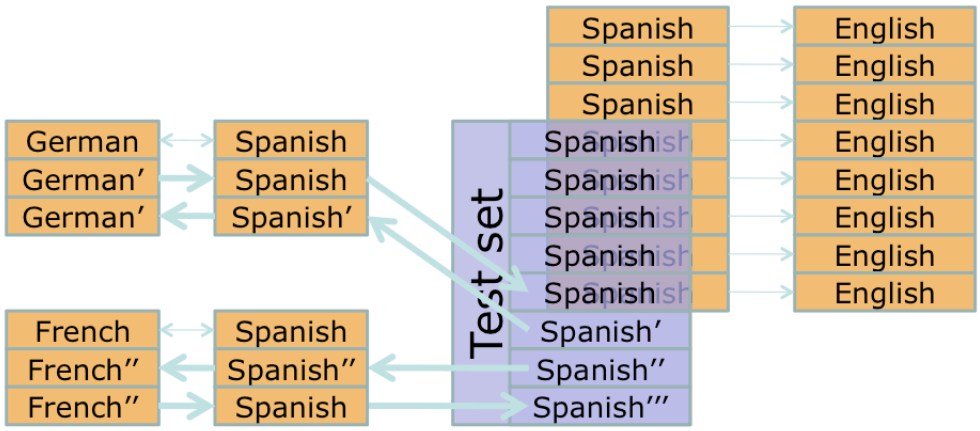

Figure 4.1: Pivoting technique for paraphrase generation. For a Spanish-to-English translation model, which encounters unknown source language (Spanish) phrases, augment the model by pivoting through other languages such as French or German. This requires translation models to and from these pivot languages, which are typically generated from sentence-aligned parallel texts. However, parallel texts are a limited resource.

in other languages. While this works for English and many European languages, it is far less likely to help when translating from other source languages, for which bitexts are scarce or non-existent. Also, the pivoting approach is inherently noisy in both the paraphrase candidates' correct sense, and their translational likelihood, because of the double translation step. The problem of incorrect sense translation is likely to be exacerbated with out-of-domain translation, i.e., when the test set is of a different genre than the bitexts. One advantage of the bitext-dependent pivoting approach is the use of the additional human knowledge that is encapsulated in the parallel sentence alignment. However, I argue that the ability to use much larger resources for paraphrasing should trump the human knowledge advantage.

More recently, Callison-Burch (2008) has improved performance of this pivoting technique by imposing syntactic constraints on the paraphrases. In one variant

the target phrase and its paraphrase are constrained to have the same parsing tag (e.g., NP), and in another variant, this constraint has been relaxed so that the phrase and its paraphrase must have the same Combinatory Categorial Grammar (CCG) super-tag sequence, but no longer need to have the same single constituent tag. The limitation of such an approach, in either variant, is the reliance on a good parser (in addition to reliance on bitexts), since a good parser is not available in all languages, especially not in resource-poor languages.

Habash and Hu (2009) show, using a similar pivoting technique to Callison-Burch et al. (2006) and a trilingual parallel text, that using English as a pivot language between Chinese and Arabic can actually outperform translation using a direct Chinese-Arabic bilingual parallel text. The authors suggest that this might be due to the fact that English is "half-way" between the other two languages in terms of word order properties. Wu and Wang (2008) show that it is possible to use pivoting technique for translation of a language pair even if there is little or no parallel text for this pair. They construct a "pivot" translation model, and in the case of having direct parallel text for that language pair, they build a standard translation model, and interpolate it with the "pivot" translation model. Max (2009) improves on the basic pivoting technique by taking the surrounding context of the target phrase, pivot phrase, and paraphrase candidates into account. Another approach using a pivoting technique augments the human reference translation with paraphrases, creating additional translation "references" (Madnani et al., 2007; Madnani, to appear). All approaches have shown gains in BLEU score.

Bond et al. (2008) also translate and back-translate in order to generate paraphrases, but they do not use another language. They improve SMT coverage by using a manually crafted monolingual HPSG grammar for generating meaning and grammar preserving paraphrases by parsing the English side and then converting it to an abstract semantic representation and back to English. This grammar allows for certain word reordering, lexical substitutions, contractions, and "typo" corrections. The paraphrases are then used to augment the training set. They test this method on both Japanese to English, and English to Japanese translation tasks, and achieve modest BLEU score gains in most cases.

Moving along the paraphrasing method axis, Barzilay and McKeown (2001) use direct translation in order to generate paraphrases, in contrast to this work and the above-mentioned pivoting approaches. They extract paraphrases from a monolingual parallel corpus, containing multiple translations of the same source. In addition to the parallel corpus usage limitations described above, this technique is further limited by the small size of such materials, which are even scarcer than the resources in the pivoting case. Barzilay and Lee (2003) focus on domain-specific sentential paraphrases, obtained from unannotated comparable corpora (and no longer dependent on parallel text). Paraphrasing patterns are learned by using multiple-sequence alignment and are represented by word lattice pairs. They demonstrate that sentential paraphrases are not always composed from word or phrase level paraphrases, and that the sentential paraphrase or its sub-part paraphrases, if any, might only be good in a specific domain.

Still on the paraphrasing method axis, much of the pre-pivoting research largely focused on morphological analysis in order to reduce type sparseness: Nissen and Ney (2004) explore morphological analysis of English and German tokens; Goldwater and McClosky (2005) employ stemming and lemmatizing for Czech-English alignments; Koehn and Knight (2003) propose a method for correctly splitting German compound words; Olteanu et al. (2006) translate German compound words if their parts are in the model's vocabulary. Mermer et al. (2007) use what they call "lexical approximation": they replace untranslated words with the closest known word, sharing certain features such as part-of-speech (POS) tag. Correct word segmentation (mainly in Chinese) in order to reduce OOV word rate has also produced a lot of research recently, e.g., Asahara et al. (2007), who use machine learning techniques, Demberg (2007) employing a universal, unsupervised model, Huang et al. (2007) who use a character-based word boundary classification, and Dyer et al. (2008) representing the input as a word lattice, with different word segmentation paths, optionally coming from different automatic word segmenters.

A non-morphologically-based representational approach suggested using back-off to character-based SMT for untranslated phrases (Vilar et al., 2007). Dolan et al. (2004) explore generating paraphrases by using edit-distance and by aligning headlines of time- and topic-clustered news articles; they do not address the OOV problem directly, as their focus is sentence-level paraphrases;. They use a standard SMT measure, alignment error rate (AER), and only report results of the alignment quality, and not of an end-to-end SMT system.

Next on the paraphrasing method axis are distributional methods: Work that relies on the distributional hypothesis using bilingual comparable corpora (without the need for bitexts), typically uses a seed lexicon for "bridging" source language phrases with their target languages paraphrases (Fung and Yee, 1998; Rapp, 1999; Diab and Finch, 2000). To date, reported implementations suffer from scalability issues, as they pre-compute and hold in memory a huge collocation matrix; I know of no report of using this approach in an end-to-end SMT system. Fung and Yee (1998) suggest an IR approach for translating OOV words from (non-parallel) comparable corpora. They compute profiles of word collocation counts (DPs) for source and target side words, with strength-of-association measures normalized by a TF/IDF-based measure, and then apply a "home-grown" similarity function between the OOV word collocation profile and target-side candidates' profile, via weighted "bridge words". They focus on 118 OOV Chinese words, and report that almost all unambiguous words find their translation within the first 100 candidates. However, only 6 words had the correct translation ranked first. Rapp (1999) shows collocation distributional measures can be helpful even in mining unrelated (non-parallel, non-comparable) texts. Diab and Finch (2000) also use collocation distributional measures to find translations from comparable corpora. They explore automatically acquiring the seed lexicon, and so do Haghighi et al. (2008).

Another IR approach is described in López-Ostenero et al. (2005), who focus on translating noun phrases by gathering candidate translations that have each content word aligned within the source phrase. In case no such candidate is found,

they back-off to translating word-by-word. They do not mention any further back-off when a single word's translation is unknown to the model.

Unsupervised learning of paraphrases has been studied in non SMT related previous work. One notable example is that of Lin and Pantel (2001)), who use syntactic dependency relations instead of simple co-occurrence counts, and a semantic measure that is based on similarity between paths in dependency trees. They are able to learn also paraphrases with gaps or variables (e.g., *X did Y* $\longleftrightarrow$ *Y was done by X*). Wu and Zhou (2003) also use dependency relations, and paraphrase the words in these relations by using their WordNet synsets.

Bilingual distributional paraphrasing work (Fung and Yee, 1998; Rapp, 1999; Diab and Finch, 2000; Haghighi et al., 2008) can also be viewed as belonging to the third category, freeing from or alleviating the reliance on parallel texts. As for the second category, aiming to reduce OOV rate by increasing parallel training set size without using more dedicated human translation: related work here concentrates on "harvesting" the World-Wide Web (Resnik and Smith, 2003; Oard et al., 2003; Munteanu and Marcu, 2005; Abdul-Rauf and Schwenk, 2009).

Paraphrasing research, in the sense of generating a different linguistic form with a similar meaning, is quite diverse, and can be characterized by even more axes: Paraphrases may be lexical (different words with similar meaning), or structural (e.g., switching between active and passive voice), or both. The paraphrasing target may or may not contain variables (gaps), as in *X gave Y to Z* or *threw X to the wolves*. And it may be lossy to some extent, in number of words and/or content,

with the extreme cases of summarization and translation (and some cases of textual entailment), if one views them as forms of paraphrasing. When using distributional methods, semantic distance may be a function of simple co-occurrence between two terms, or a function of other relations, such as syntactic dependency relations. In this chapter I concentrate on paraphrasing word sequences (phrases in the non-linguistic sense), with no gaps, for SMT. Paraphrasing targets with gaps may be also helpful in SMT, but I leave this for future research. For more information on various forms and types of paraphrasing, see Madnani (to appear).

4.3 Phrasal Distributional Profiles

Collocational distributional profiles (DPs), traditionally capturing the context words with which a single word (the *target* word) appears, are detailed in Section 3.2.2.1. These traditional word DPs can be generalized to the phrase case: the target, or collocates (which constitute the dimensions of the DP), or both, may be redefined to be longer than a single token. In preliminary experiments I found no gain in using phrasal collocates (bigrams or trigrams) as vector dimensions / features, instead of unigrams. Therefore, and since phrasal collocates are not the focus of this doctoral work, I will concentrate hereafter on DPs of phrasal targets, or *phrasal DPs*.

Word DPs can be generalized to phrasal DPs, simply by counting words that co-occur within a sliding window around the target phrase's occurrences (e.g., count-

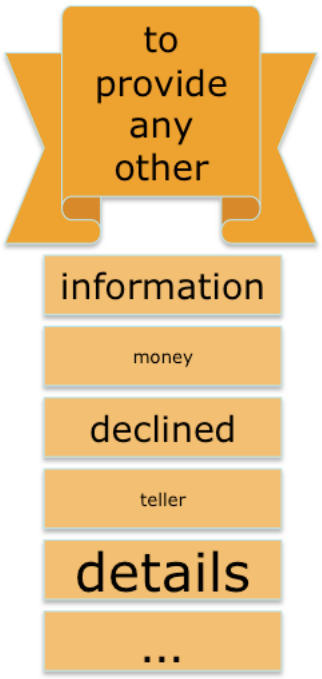

Figure 4.2: Visual example of a distributional profile for the phrase *to provide any other*. It is comprised of collocates found in training set sentences such as "she declined *to provide any other* information . . . ", "unable *to provide any other* details . . . ", and "police refused *to provide any other* details . . . ".

ing occurrences of words up to 6 words before or after the target phrase). For example, when building a DP for the target phrase *counting words* in the previous sentence, then *simply* is in relative position -2, and *sliding* is in relative position 5. Searching for similar phrasal DPs poses an additional computational challenge over the word DP case (see Section 4.4), but there is no additional difficulty in building the phrasal profile itself as described above. A phrasal DP is illustrated in Figure 4.2. Examples of sliding window contexts used to construct this DP are shown in Figure 4.4.

The individual words that make up the target phrase do not effect the DP in this model. Conceivably, however, following the example above, the DP of *counting*

words could benefit from the distributional information of the individual *counting* and *words*. But the individual words' distributional information could also introduce noise, e.g., in the case of idioms (the meaning and distribution of *kicked the bucket* is not similar to that of either *kicked* or *bucket*). Even within semantically compositional phrases such as *counting words*, it is not immediately clear how to combine the individual words' contributions:

- How to model the head of the phrase, the complement, or adjuncts? E.g., what would be the difference between the DPs of *student town* and *town student*?
- The phrase can be structurally ambiguous. For example, is *counting words* a verb phrase headed by the verb *counting*, or a noun phrase headed by the noun *words* (as in *words that are used for counting*)?
- How to model the contribution of function words such as *of, the, in* to the phrasal DP?
- How to model adjectives and adverbs? For example, simply adding all the contexts (collocates) in which the adjective *quick* can occur in, might introduce more noise than helpful information to the DP of *quick fox*. A similar challenge exists for the context contribution of *very*.
- How to model phrases with a gap? This issue is orthogonal to the semantic compositionality issue, as this challenge exist for compositional phrases such as *gave X the book*, as well as idioms such as *threw X to the wolves*. Particularly,

should the occurrences of X be modeled as part of the phrase or part of the context?

Without discounting the importance and potential gains of modeling individual words' contributions, I assume that with sufficiently large monolingual training corpora , this issue will be marginalized. For simplicity, and similarity to the traditional word DP, the target phrase is treated here as an atom. I leave further improvements along these lines to future research.

4.4 Phrasal Distributional Paraphrase Generation

The paraphrase generation process is as follows: upon receiving OOV phrase phr, build distributional profile DP_{phr}. Next, gather contexts: for each occurrence of phr, keep surrounding (left and right) context $L\_\_R$. For each such context, gather paraphrase candidates X which occur between L and R in other locations in the training corpus, i.e., all X such that LXR occur in the corpus. Finally, rank all candidates X, by building distributional profile DP_X and measuring profile similarity between DP_X and DP_{phr}, for each X. Output k-best candidates above a certain similarity score threshold. This is illustrated in Figure 4.3. The rest of this section describes this approach in more detail.

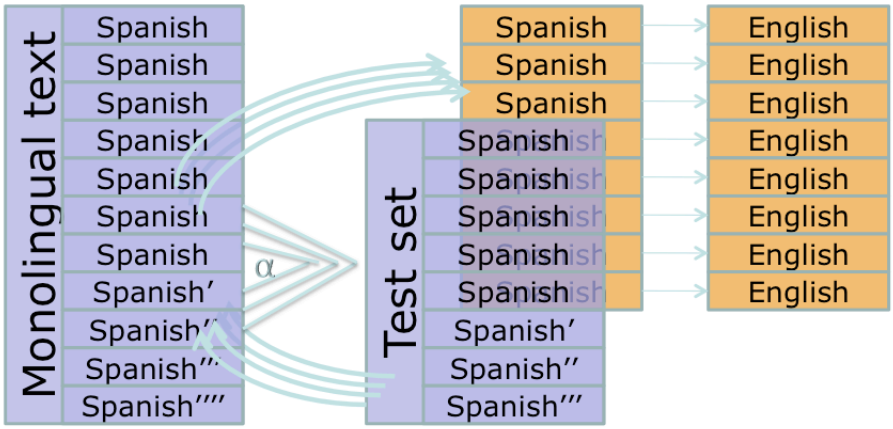

Figure 4.3: Monolingual corpus-based distributional paraphrase generation. For a Spanish-to-English translation model, which encounters unknown source language (Spanish) phrases, augment the model by generating distributional paraphrases. This requires a large monolingual corpus, which is a relatively abundant resource. It then requires building DPs for the unknown phrases, gathering the contexts in which they appear, gathering paraphrase candidates that also appear in these contexts, and selecting those candidates whose DPs are most similar to the unknown phrases.

4.4.1 Build phrasal profile DP_{phr}.

Build a distributional profile of the target phrase *phr*, enlisting all collocating words, and their co-occurrence count or strength-of-association with *phr*, as described in Section 3.2.2.1. The co-occurrence counts are collected using a sliding window of size $MaxPos$ tokens to each side of each occurrence of *phr* in the monolingual training corpus. If *phr* is very frequent (above some threshold of t occurrences), uniformly sample only t occurrences, multiplying the gathered co-counts by factor of $count(phr)/t$. So if *phr* occurs $30,000$ times, and the threshold is $t = 10000$, than count co-occurring words in a sliding window around only every third occurrence of *phr*, but multiply these co-occurrence counts by 3.

4.4.2 Gather context

Example contexts are shown in Figure 4.4. The challenge in deciding how much context to keep to the left and right of each occurrence of the target is a familiar recall vs. precision tension: if the context is very short and/or very frequent (e.g., "the __ is"), then it might not be very informative, in the sense that many words can appear in that context (in this example, practically any noun); however, if the context is too long (too specific), then it might not occur enough times elsewhere (or not at all) in the training corpus. Therefore, to balance between these two extremes, I use the following heuristics. Start small: Start with setting the left part of the context, L, to be a single word/token to the left of phrase phr. If it is stoplisted, append the next word to the left (now having a bigram left context instead of a unigram), and repeat until the left context is not in the stoplist. Repeat similarly for R, the context to the right of phr. Add the resulting $L\_\_R$ context to a context list. I stoplist "promiscuous" words, i.e., those that have more than $StoplistThreshold$ collocates in the training corpus, using the above $MaxPos$ parameter value. I also stoplist bigrams which occur more than t times and comprise solely from stoplisted unigrams. This typically results in filtering out function words such as *the, and, in, before*, and bigrams such as *in the*.

4.4.3 Gather candidates

For each gathered context in the context list, gather all paraphrase candidate phrases X that connect left hand side context L with right hand side context R, i.e., gather all X such that the sequence LXR occurs in the corpus. Example candidates, appearing in same contexts as the target phrase, are shown in Figure 4.5. In practice, to keep search complexity low, limit X to be up to length $MaxPhraseLen$. Also, to further speed up runtime, I uniformly sample the context occurrences as follows: Let $contextCount$ be the number of occurrences of the current context, $allContextsCount$ be the sum of the former count over all contexts of phr, and t the sampling threshold as above. Then only look at $fraccontextCountallContextsCount * t$ occurrences of the current context, but no less then $minContextCount$ (if there are more than that), and no more than $maxContextCount$ occurrences.

4.4.4 Rank candidates

For each candidate X, build distributional profile DP_X, and evaluate $psim(DP_{phr}, DP_X)$ as in Section 3.2.2.3. I remind the reader that since the DP is represented as a vector, any vector similarity function can be used here, e.g., cosine.

Left context (L)	__	Right context (R)
declined	to provide any other	details
refused	to provide any other	information
unable	to provide any other	details
failed	to provide any other	explanation
...	to provide any other	...

Figure 4.4: Example of gathered context of the target phrase *to provide any other*. It is comprised of training set sentences such as "she declined *to provide any other* information . . . ", and "police refused *to provide any other* details . . . ".

Left context (L)	__	Right context (R)
declined	**to give further**	details
refused	**to provide any**	information
unable	**to reveal any**	details
failed	**to provide further**	explanation
...	to provide any other	...

Figure 4.5: Example of gathered paraphrase candidates for the target phrase *to provide any other*. These are phrases appearing in identical contexts to those surrounding the target phrase (See Figure 4.4).

4.4.5 Output k-best candidates

Output k-best paraphrase candidates for phrase phr, in descending order of similarity. Filter out paraphrases with score less than $minScore$. For example, suppose we set $minScore = .3$ and $k = 20$. Then if the third best paraphrase has a similarity score .25, it will be filtered out because its score is too low, even though it is in the top 20 list. Conversely, if the 25^{th} paraphrase has score .76, it will be filtered out because it is not in the top 20, even though its score is above the threshold.

4.5 Experiments

I examined the application of the engine's paraphrases to handling unknown phrases when translating from English into Chinese (E2C) and from Spanish into English (S2E). Following Callison-Burch et al. (2006), for all baselines I used the phrase-based statistical machine translation system Moses (Koehn et al., 2007), with the default model features:[3]

- a phrase translation probability,
- a reverse phrase translation probability,
- a lexical translation probability,
- a reverse lexical translation probability,
- a word penalty,

[3] `www.statmt.org/moses`

- a phrase penalty,
- six lexicalized reordering features,
- a distortion cost, and
- a language model (LM) probability.

All features were weighted in a log-linear framework (Och and Ney, 2002). Feature weights were set with minimum error rate training (Och, 2003) on a development set using BLEU (Papineni et al., 2002) as the objective function. Test results were evaluated using BLEU and TER (Snover et al., 2006): The higher the BLEU score, the better the result (basically, it indicates higher n-gram overlap between the test and translation references); the *lower* the TER score, the better the result (basically, it indicates less translation errors). This is denoted with BLEU ↑ and TER↓ in the tables below. The phrase translation probabilities were determined using maximum likelihood estimation over phrases induced from word-level alignments produced by performing Giza++ training (Och and Ney, 2000) on both source and target sides of the parallel training sets. (Uni-directional alignment data are deleted prior to /bleu scoring). When the baseline system encountered unknown words in the test set, its behavior was simply to reproduce the foreign word in the translated output.

4.5.1 Paraphrase-Augmented Translation Models

The paraphrase-augmented models were identical to the corresponding baseline model, with the exception of additional (paraphrase-based) phrase-table entries (translation rules), and additional feature or features, described below. Similarly to Callison-Burch et al. (2006), I added the following feature:

$$h(e,f) = \begin{cases} psim(DP_{f'}, DP_f) & \text{If phrase table entry } (e,f) \text{ is generated from } (e,f') \\ & \text{using monolingually-derived paraphrases.} \\ 1 & \text{Otherwise.} \end{cases} \tag{4.1}$$

Note that it is possible to construct a new translation rule from f to e via more than one pair of source-side phrase and its paraphrase; e.g., if f_1 is a paraphrase of f, and so is f_2, and both f_1, f_2 translate to the same e, then both lead to the construction of the new rule translating f to e, but with potentially different feature scores. To illustrate this, suppose a Spanish-English phrase-table has the following rules, all with the same target-side translation:

source-side phrase ||| target-side phrase ||| word alignment info ... ||| feature scores

a abandonar el ||| to leave the ||| (0) (1) (2) ||| (0) (1) (2) ||| 0.714286 0.0365803 1 0.291936 2.718

que abandonar el ||| to leave the ||| (0) (1) (2) ||| (0) (1) (2) ||| 0.142857 0.00794395 1 0.0508198 2.718

llegó a el acuerdo de mantener el ||| to leave the ||| (1) (0) () () () () (2) ||| (1) (0) (6) ||| 0.142857 7.32192e-12 0.2 0.0636951 2.718

and suppose further that the source-side phrase *a disponer de los* is unknown (not in the table), and that among its top paraphrases, are the following:

phrase	**paraphrase**	**score**
a disponer de los	a abandonar el	.74
a disponer de los	que abandonar el	.68
a disponer de los	llegó a el acuerdo de mantener el	.35

Then there are three paths to construct a new translation rule from *a disponer de los* to *to leave the*, each going through one of the phrase-table entries above:

1. a disponer de los → a abandonar el → to leave the
2. a disponer de los → que abandonar el → to leave the
3. a disponer de los → llegó a el acuerdo de mantener el → to leave the

There are different possible approaches to the multiple path phenomenon: A default approach might create a separate new rule for each path, making these new rules compete with one another in order to enter the final sentence translation derivation during "decoding" time; another approach might generate only a single

rule from f to e, using only one randomly chosen path; or using only the "best path" – the path with the highest paraphrase similarity score – path 1 in the example above (with highest similarity score .74). However, it is also possible to have *all paths* reinforce the model's confidence in using a single new translation rule from f to e, by increasing the new rule's associated semantic score in proportion to the paraphrase scores of f to f_1, f to f_2, and so on. Preliminary experiments showed that while the default approach resulted in negative results for SMT, the latter resulted in significant improvements. Therefore, all reported results in this chapter are based on this latter approach (using semantic similarity evidence to reinforcement the confidence in certain augmented translation rules). A more thorough comparison of alternative approaches to handling multiple paths is left for future research. The details and example of the semantic reinforcement approach are given next.

For each paraphrase f of some source-side phrases f_i, with respective similarity scores $sim(f_i, f)$, I calculated an aggregate score $asim$ with a "quasi-online-updating" method as follows:

$$asim_i = asim_{i-1} + (1 - asim_{i-1})\ sim(f_i, f), \quad \text{where } asim_0 = 0 \tag{4.2}$$

The aggregate score $asim$ is updated in an "online" fashion with each pair f_i, f as they are processed, but only the final $asim_k$ score is used, after all k pairs have been processed. Simple arithmetics can show that this method is insensitive to the order in which the paraphrases are processed. I only augment the phrase table with

a single rule from f to e, and in it are the feature values of the phrase f_i for which the score $sim(f_i, f)$ was the highest. Continuing the example above, we see that $<a$ *disponer de los*, *a abandonar el*$>$ has the highest similarity score, and so we use the corresponding phrase-table entry (the top entry in the example) as the base for the new entry. To calculate the aggregated similarity score for the added feature, we start with $asim_0 = 0$, and then iteratively process each of the above entries:

$$\begin{aligned} asim_1 &= 0 + (1 - 0) \times .74 &&= .74 \\ asim_2 &= .74 + (1 - .74) \times .68 &&= .92 \\ asim_3 &= .92 + (1 - .92) \times .35 &&= .95 \end{aligned}$$

and use the final score .95 as an added feature in the entry:

a disponer de los ||| to leave the ||| (0) (1) (2) ||| (0) (1) (2) ||| 0.714286 0.0365803 1 0.291936 **.95** 2.718

Note that the score (and quality) of the third paraphrase is low, and so its contribution to the aggregated score is proportionally small.

For generating the monolingually-derived distributional paraphrases, I used a sliding window of size $MaxPos = 6$, a sampling threshold $t = 10000$, and a maximal gap $MaxPhraseLen = 6$ between the left and right contexts of paraphrase candidates. Also, I arbitrarily limited the number of occurrences (in which to look for paraphrase candidates) of each context of phrase *phr* to no less than $minContextCount = 250$ (if there are more than that), and no more than

$maxContextCount = 2,000$ occurrences, in order to keep the runtime short, but still give a reasonable chance to any context to contribute candidates. For each phrase *phr*, I output no more than the top $k = 20$ best-scoring paraphrases.

4.5.2 English-to-Chinese Translation

In order to compare the quality of paraphrases generated with "pure" distributional and hybrid semantic similarity measures, I chose English as the source language for the translation task. This is because an English semantic knowledge base (the *Macquaries* thesaurus; see Chapter 3) was at my disposal, and the new technique augments the phrase table by paraphrasing the source side. I chose Chinese as the translation target language because it is quite different from English (e.g., in word order), and four reference translation were available from NIST (see below).

For the English-Chinese (E2C) baseline model, I trained on the LDC Sinorama and FBIS tests (LDC2005T10 and LDC2003E14), and segmented the Chinese side with the Stanford Segmenter (Tseng et al., 2005). After tokenization and filtering, this bitext contained 231,586 lines (6.4M + 5.1M tokens). I trained a trigram language model on the Chinese side, with the SRILM toolkit (Stolcke, 2002), using the modified Kneser-Ney smoothing option. I then split the bitext into 32 even slices, and constructed a reduced set of about 29,000 sentence pairs by using only every eighth slice. The purpose of creating this subset model was to simulate a resource-poor language.

For development I used the Chinese-English NIST MT 2005 evaluation set. In order to use it for the reverse translation direction (English-Chinese), I arbitrarily chose the first English reference set as the development "source", and the Chinese source as a single "reference translation". For testing I used the English-Chinese NIST MT evaluation 2008 test set with its four reference translations.

I augmented the E2C baseline models with paraphrases generated as described above, training on the British National Corpus (BNC) v3 (Burnard, 2000) and the first 3 million lines of the English Gigaword v2 APW, totaling 187M tokens after tokenization, and number and punctuation removal. See Table 4.1 for training set sizes.

Set	**# Tokens Source+Target**
E2C 29K	0.8 + 0.6
E2C Full	6.4 + 5.1
bnc+apw	187

Table 4.1: English-Chinese (E2C) training set sizes (million tokens).

I generated paraphrases for phrases up to six tokens in length, and used an arbitrary similarity threshold of $minScore = 0.3$. I experimented with three variants: adding a single additional feature for all paraphrases (*1-6grams*); using only paraphrases of unigrams (*1grams*); and adding two features, one only sensitive to unigrams, and the other only to the rest (*1 + 2-6grams*). All features had the same design as described in Section 4.5, and all feature weights in each model, including the baseline, were tuned using a separate minimum error rate training for each model.

Results are shown in Table 4.2. For the E2C models, for which I had four reference translations for the test set, I used shortest reference length, and used the NIST-provided script to split the output words to Chinese characters before evaluation, as is standardly done in the NIST English-Chinese translation task official evaluation.[4] Statistical significance for the BLEU results was calculated using Koehn's paired bootstrap re-sampling test (Koehn, 2004b), with a sample size of 2000 pairs. Statistical significance was determined in case the 95% confidence interval (CI) of the systems' BLEU score difference did not include zero. For conciseness, this is denoted as $p < .05$ below. Similarly, a 99% CI is denoted as $p < .01$, and so on for other CIs. The word "significant" is used below as a shorthand for "statistically significant" (at $p < .05$ unless specified otherwise). The associated t-test p-value for the significant cases was always $p < 0.0001$. Paraphrasing and translation examples are given in Section 4.6, Tables 4.7 and 4.10.

Augmentation with "pure" distributional paraphrases. On the E2C 29,000-line subset, the augmented model had a significant 1.67 BLEU points gain over its baseline. On the full size model, results were negative. TER scores generally follow the same patterns. Note that the E2C full size baseline is reasonably strong: Its character-based BLEU score is slightly higher than the JHU-UMD system that participated in the NIST 2008 MT evaluation (constrained training track),[5] although

[4] http://www.itl.nist.gov/iad/mig//tests/mt/2008/doc/mt08_official_results_v0.html

[5] http://www.itl.nist.gov/iad/mig//tests/mt/2008/doc/mt08_official_results_v0.html

dataset	E2C model	BLEU ↑	TER↓
29k	baseline	15.21	69.285
29k	1grams-pivot > .3	15.50	69.365
29k	1-5grams-pivot >.3	16.10*	**68.956**
29k	1 + 2-5grams-pivot >.3	**16.17*^{I}**	69.069
29k	1grams	16.87*	**68.784**
29k	1-6grams	16.54*	69.236
29k	1 + 2-6grams	**16.88*^{C}**	68.790
29k	1grams-hybrid	16.44*	68.987
29k	1-6grams-hybrid	16.65*^{P}	68.802
29k	1 + 2-6grams-hybrid	**16.95*^{PIC}**	**68.742**
Full	baseline	**22.17**	**63.557**
Full	1grams	21.64*	64.235
Full	1-6grams	21.75*	64.751
Full	1 + 2-6grams	21.39*	64.929

Table 4.2: E2C Results: character-based BLEU and TER scores. All models have one additional feature over baseline, except for the "1 + 2-5" and "1 + 2-6" models that have one feature for unigrams and another feature for bigrams to 6-grams. Paraphrases with score < .3 were filtered out. *,^{P},^{I},^{C} = significantly better than corresponding baseline, non-hybrid "pure-augmented" model, "1gram" model, and "1-5gram" or "1-6gram" model, respectively, $p < 0.05$, using Koehn's (2004b) pairwise bootstrap resampling test

I used a subset of that system's training materials, and a smaller language model. Results there ranged from 15.69 to 30.38 BLEU (ignoring a seeming outlier of 3.93).

Augmentation with hybrid knowledge/corpus-based paraphrases. After experimenting with "pure" corpus-based paraphrases, I then experimented with hybrid knowledge/corpus-based paraphrases: These paraphrases were generated exactly as their "pure" distributional counterparts above, except for the semantic similarity measure used for candidate ranking. The semantic similarity measure used here is precisely the *hybrid-sense-proportional* method described in Section 3.3. I took the E2C 29,000-line subset baseline model, and the 29,000-line subset models that were augmented with "pure" distributional paraphrases, as strong double baselines: The "pure-augmented" models did better than the baseline, and therefore, the claim for the hybrid semantic distance measure' advantage is strongly supported not only by gains in SMT performance over both the baseline, but also over the "pure-augmented" models. The middle section in Table 4.2 shows that all the hybrid-augmented models did better than baseline (up to 1.74 BLEU points), and all but one did better than their "pure-augmented" counterparts, by small but still significant gains. TER scores generally follow the same patterns here as well. See Section 4.6 and Table 4.10 for further discussion and examples.

Augmentation with pivot-based paraphrases. I also attempted to augment the translation model with the pivot-style English paraphrases used in Callison-

Burch (2008).[6] Due to memory (RAM) constraints, it was not possible to use the full list. I therefore chose to filter it with a score threshold, similarly to the one used for the distributional paraphrases. I filtered out paraphrases with a threshold $p < .3$, since using a lower threshold still encountered insufficient memory problems. Note, however, that this .3 pivot-based estimated paraphrase probability threshold is not equivalent to a .3 distributional paraphrase vector similarity score. In addition to using all available lengths (unigram to 5-gram) of paraphrased phrases, as done in Callison-Burch (2008), I also experimented with *1grams-pivot* and *1+2-5grams-pivot* models, equivalent to the *1grams* and *1+2-6grams* models mentioned above, respectively. The pivot-style unigram paraphrase-augmented model showed significant BLEU gains over the baseline, but was out-performed by its "pure-augmented" counterpart. Its TER score was slightly worse than the baseline (but recall it was threshold-filtered). The other two pivot-style paraphrase-augmented models also showed significant gains over the baseline, but were out-performed by both "pure-augmented" and hybrid counterparts.

The *1+2-5grams-pivot* model was the best pivot-style performer, and similarly, the *1+2-6grams-hybrid* was the best hybrid performer, and *1+2-6grams* was the best "pure-augmented" performer. These results suggest again that a finer feature granularity is advantageous over using only a single feature for all paraphrases (*1-6grams, 1+2-6grams-hybrid, 1+2-5grams-pivot*), or using only partial data as paraphrases of certain phrase lengths (*1grams, 1grams-hybrid, 1grams-pivot*).

[6]The baseline paraphrases that were not filtered by syntactic criteria, available from Chris Callison-Burch's site: `http://www.cs.jhu.edu/~ccb/howto-extract-paraphrases.html`

4.5.3 Spanish-to-English Translation

In order to to permit a more direct comparison with the standard bilingual pivoting technique, I also experimented with Spanish to English (S2E) translation, following Callison-Burch et al. (2006). For baseline I used the Spanish and English sides of the Europarl multilingual parallel corpus (Koehn, 2005), with the standard training, development, and test sets. I created training subset models of 10,000, 20,000, and 80,000 aligned sentences, as described in Callison-Burch et al. (2006). For better comparison with their pivoting system, I used the same 5-gram language model, development and test sets: For development, I used the Europarl dev2006 Spanish and English sides, and for testing I used the Europarl 2006 test set.[7]

I trained the Spanish paraphrase generation model on the Spanish corpora available from the EACL 2009 Fourth Workshop on Statistical Machine Translation:[8] the Spanish side of the Europarl-v4, news training 2008, and news commentary 2009. I also re-trained adding the JRC-Acquis-v3 corpus[9] to the paraphrase training set, and then adding also the LDC Spanish Gigaword (LDC2006T12) and truncating the resulting corpus after the first 150M lines. I lowercased these training sets, tokenized and removed punctuation marks and numbers, and this resulted in training set sizes as detailed in Table 4.3.

[7]These data were obtained from Chris Callison-Burch's site: `http://www.cs.jhu.edu/~ccb/howto-extract-paraphrases.html` and personal communication.

[8]http://www.statmt.org/wmt09

[9]http://wt.jrc.it/lt/Acquis

Set	# Tokens Source+Target
S2E 10K	0.3 + 0.3
S2E 20K	0.6 + 0.6
S2E 80K	2.3 + 2.3
wmt09	84
wmt09+acquis	139
wmt09+acquis+afp	402

Table 4.3: Spanish-English (S2E) training set sizes (million tokens).

I generated paraphrases for phrases up to six tokens in length, and used two arbitrary similarity thresholds of $minScore = 0.3$ (as in the E2C experiments), and 0.6, for enforcing only higher precision paraphrasing. With $minScore = 0.3$, I experimented with these variants (as in the E2C experiments): adding a single feature for only paraphrases of unigrams (*1grams*); and adding two features, one only sensitive to unigrams, and the other only to the rest (*1 + 2-6grams*). With $minScore = 0.6$, I experimented with adding a single feature for only paraphrases of unigrams (*1grams*); adding a single feature for all paraphrase (*1-4grams*); and adding two features: one only sensitive to unigrams and bigrams, and the other to the rest (*1-2 + 3-4grams*). Each feature had an associated weight, and all feature weights in each model were tuned using a separate minimum error rate training, as in the baseline.

Results are shown in Table 4.4. In order to evaluate the S2E models, I used Bleu (Papineni et al., 2002) over lowercase output. Not re-casing the output avoids possible re-caser-originated scoring "noise". I used Koehn's (2004b) significance test as above. Translation examples are given in Section 4.6, Table 4.9.

For $minScore = 0.3$, paraphrasing achieved gains of up to .63 BLEU points on the S2E 10,000-line subset (not all significant), and diminishing gains on the 20,000-line and 80,000 subsets. *1 + 2-6grams* was best performer.

Using higher scoring paraphrases ($minScore = 0.6$), to be more restrictive in paraphrasing quality, doesn't seem to result in higher gains, possibly due to excessive loss of coverage. I concluded from a manual evaluation of the 10,000-line models that the two major weaknesses of the baseline model were (not surprisingly) number of untranslated (OOV) words / phrases, followed by number of superfluous words / phrases.

On the larger subset models, no model significantly outperformed the baseline. Note that the S2E baselines' scores reported here are higher than those of Callison-Burch et al. (2006). I attribute this to evaluating lowercased outputs instead of recased ones, and also possibly due to improvements in the Moses decoder over the three years separating the experiments reported in Callison-Burch et al. (2006) and those reported here.

4.6 Discussion and Future Work

I have shown that monolingually-derived paraphrases, based on distributional semantic similarity measures over a source-language corpus, can improve the performance of statistical machine translation (SMT) systems. Moreover, when using hybrid semantic distance measures (Sections 3.3 and 4.3) for the paraphrase gen-

bitext	mono.corp.	features	minScore	BLEU ↑	TER↓
10k	(baseline)	–	–	23.78	62.382
10k	(pivoting)	1grams-pivot	–	**24.42***	**61.121**
10k	(pivoting)	1-5grams-pivot	–	24.08*	61.859
10k	(pivoting)	1+2-5grams-pivot	–	(failed)	
10k	wmt09+aquis	1grams	.3	24.11	61.979
10k	wmt09+aquis+afp	1grams	.3	23.97	61.974
10k	wmt09+aquis	1+2-6grams	.3	**24.21**$^{\times}$	**61.813**
10k	wmt09+aquis+afp	1+2-6grams	.3	24.10	61.834
10k	wmt09+aquis	1grams	.6	24.11	61.979
10k	wmt09+aquis+afp	1grams	.6	24.06	62.048
10k	wmt09	1-4grams	.6	23.81	62.023
10k	wmt09+aquis	1-4grams	.6	24.13*	61.739
10k	wmt09	1-2+3-4gr	.6	23.92	62.202
10k	wmt09+aquis	1+2-6grams	.6	**24.15***	**61.690**
10k	wmt09+aquis+afp	1+2-6grams	.6	24.12$^{\times}$	61.911
20k	(baseline)	–	–	24.68	62.333
20k	wmt09+aquis+afp	1grams	.3	24.77$^{\times}$	61.276
20k	wmt09+aquis+afp	1+2-6grams	.3	**24.89***	**61.126**
20k	wmt09+aquis	1-4grams	.6	**24.75**$^{\times}$	61.528
20k	wmt09+aquis+afp	1+2-6grams	.6	24.73*	**61.140**
80k	(baseline)	–	–	**27.89**	57.977
80k	wmt09+aquis+afp	1grams	.3	27.84	**57.781**
80k	wmt09+aquis+afp	1+2-6grams	.3	27.87	57.901
80k	wmt09+aquis	1-4grams	.6	27.82	**57.906**
80k	wmt09+aquis+afp	1+2-6grams	.6	27.77	58.222

Table 4.4: Spanish-English (S2E) Results: Lowercase BLEU and TER. Paraphrases with score $<$ *minScore* were filtered out. * = significantly better than baseline, $p < 0.05$, using Koehn's (2004b) pair-wise bootstrap resampling test. $^{\times}$= "almost significantly" better than baseline, $p < 0.1$.

eration, instead of "pure" corpus-based measures (Sections 3.2.2 and 4.3), further improvements are achieved in almost all cases. The presented method has the advantage of not relying on bitexts in order to generate the paraphrases, and therefore gives access to large amounts of monolingual training data, for which creating bitexts of equivalent size is generally unfeasible. I haven't trained this system on nearly as large a corpus as it can handle on current machines: most of this work was done on 8GB RAM linux machines; current machines typically come now with at least 32GB. Indeed I see this as a natural next step.

Results are inconclusive with respect to the assumption that a larger monolingual paraphrase training set yields better paraphrases: As summarized in Table 4.5, most cases of using a larger monolingual training corpus for paraphrase generation resulted in modest losses in performance. However, all losses are confounded with adding the AFP corpus. It is possible that this corpus is not suitable for this task due to genre or domain differences, or other reasons. The most pronounced difference is shown in the last row, where adding the Aquis corpus resulted in a gain of .32 Bleu points and reduction of TER by .284 points. (Note again that the higher the Bleu score and the lower the TER score, the better the quality). Additional support of using larger corpora (and particularly the AFP) comes from looking at specific examples (even if not from a representative sample), as discussed in the next paragraph. Interestingly, the losses are minimized when using a higher paraphrase score threshold (.6), which suggests that the losses were largely caused by addition of low-scoring new paraphrases. More research is required in order to better under-

stand the potential contribution of larger monolingual corpora and the conditions for yielding gains from doing that, such as genre differences.

model	minScore	smaller text	larger text	BLEU ↑	TER↓
1gram	.3	wmt09+aquis	wmt09+aquis+afp	-0.14	-0.005
1gram	.6	wmt09+aquis	wmt09+aquis+afp	-0.05	+0.069
1+2-6gram	.3	wmt09+aquis	wmt09+aquis+afp	-0.11	+0.021
1+2-6gram	.6	wmt09+aquis	wmt09+aquis+afp	-0.03	+0.221
1-4gram	.6	wmt09	wmt09+aquis	+0.32	-0.284

Table 4.5: Gains from using larger monolingual corpora for paraphrasing (S2E 10,000-line subset, summarized from Table 4.4). Adding the AFP corpus slightly hurts SMT performance in almost all cases; adding Aquis helps.

To look at some specific examples, the two rightmost columns in Table 4.6 show that although Spanish monolingual paraphrases for the unigram *baile* improve when using the larger corpus, (e.g., *danza* and *un balie* become the third and fourth top candidates, pushing much worse candidates far down the list), the two top paraphrase candidates remained unchanged. However, for the 4gram *a favor del informe*, antonymous candidates, which are bad and misleading for translation, are pushed down from the top first and third spots by synonymous, better candidates. I use "synonymous candidates" to refer to candidates with a meaning close to *a favor del informe (for the report)*, and "antonymous candidates" to refer to candidates with a meaning close to the contrary: *en contra del informe (against the report)*, although for arbitrary word sequences it may not always possible to define an antonymous phrase.

Table 4.7 contains additional examples of good and bad top paraphrase candidates, in English. All top paraphrases of *deal* are semantically close to it (*agreement,*

pivot	**wmt09+acquis**	**wmt09+acquis+afp**
Source: *baile*		
danza	el baile	el baile
bailar	baile y	baile y
a	de david palomar y la	danza
dans	viejo como quien se acomoda una	un baile
empresa	por julián estrada el tercero de	teatro
coro	al baile a la	baloncesto el cine
Source: *a favor del informe*		
a favor de este informe	en contra del informe	favor del informe
favor del informe	a favor de este informe	en contra del informe
el informe	en contra de este informe	a favor de este informe
a favor	a favor de la resolución	en contra de este informe
por el informe	a favor de esta resolución	en contra de la resolución
al informe	a favor del informe del señor	a favor del informe del sr.
su	a favor del informe del sr.	en contra del informe del sr.
del informe	en contra de la propuesta	a favor del excelente informe
de este informe	contra el informe	a favor del informe deprez

Table 4.6: Comparison of Spanish paraphrases: by pivoting, and by two monolingual corpora. Ordered from best to worst score.

accord, ...), and so is the case for the five best paraphrases of *fall*, except for the one-best (*rise*). This is another example of the tendency of distributional measures to rank antonyms high – which is undesired for SMT. The sixth-best paraphrase (*fall tokyo ap stock prices fell*) demonstrates another weakness of this technique: This paraphrase seems to have been ranked high due to the collapsing of two separate paraphrase candidates at its edges (*fall* and *fell*), benefitting from the context to the left of *fall tokyo...* and the context to the right of *...fell*. Such cases can be ameliorated with incorporation of syntactic parsing information (Callison-Burch, 2008) or other structural cues that would help filter out these cases. The third part of the table shows semantically close top paraphrases of the phrase *to provide any other*. But it seems that in general, paraphrases of phrases are of lower quality than those

of unigrams, as can be seen at the bottom, fourth part of the table. There only the second-best paraphrase is somewhat semantically close to *we have a situation that*, but the overall quality is clearly lower.

These results suggest that the new monolingual distributional paraphrasing method is especially useful in settings involving low-density languages or special domains: The smaller subset models, emulating a resource-poor language situation, show higher gains than larger models (which are supersets of the smaller subset models), when augmented with paraphrases derived from the same paraphrase training set. This was validated in two very different language pairs: English to Chinese, and Spanish to English. I believe that larger monolingual training sets for paraphrasing can help languages with richer resources, and I intend to explore this, too. Schroeder et al. (2009) recently showed that the upper bound for gains by parahrase augmentation (using human-generated paraphrases in a lattice of the source language) is high, and has not been not reached yet. I take their work as another validates of this research direction.

Although the gains in the Spanish-English subsets are somewhat smaller than the pivoting technique reported in Callison-Burch et al. (2006) – e.g., .7 BLEU for the 10k subset there, and only .4 BLEU here – I take these results as a proof of concept that can yield better gains with larger same-genre monolingual training sets. Used in their entirety (*1-5grams-pivot*), as in Callison-Burch et al. (2006), the gain from pivoting with the 10k subset was similar, and even slightly lower than the gains using the distributional paraphrases. However, when applying finer granularity, and

Paraphrase	**Score**
Source: *deal*	
agreement	0.56
accord	0.53
talks	0.45
contract	0.42
peace deal	0.33
merger	0.32
agreement is	0.30
Source: *fall*	
rise	0.87
slip	0.82
tumbled today	0.68
fell today	0.67
tumble	0.65
fall tokyo ap stock prices fell	0.56
are mixed	0.54
Source: *to provide any other*	
to give any	0.74
to give further	0.70
to provide any	0.68
to give any other	0.62
to provide further	0.61
to provide other	0.53
to reveal any	0.52
to provide any further	0.48
to disclose any	0.47
to publicly discuss the	0.43
Source: *we have a situation that*	
uncontroversial question about our	0.66
obviously with the developments this morning	0.65
community staffing of community centres	0.64
perhaps we are getting rather impatient	0.63
er around the inner edge	0.60
interested in going to the topics	0.60
and that is the day that	0.60
as a as a final point	0.59
left which it may still have	0.56

Table 4.7: Examples of English paraphrases of phrases unknown to the E2C 29K-bitext baseline model, generated with "pure" distributional semantic similarity measure.

using only pivot paraphrases for unigrams (as was done with the distributional paraphrases), the gains were the highest for that subset (24.42 BLEU points). Pivoting techniques (translating and then translating back) rely on limited resources (bitexts), and are subject to shifts in meaning and inaccurate translation probability estimation due to their inherent double translation step. A related potential problem is a probability mass "leakage": if some pivot phrase is more polysemous, then there might be more bad paraphrase candidates than with a less polysemous phrase; even if the bad candidates score low, they might result in varyingly lower probability estimates for the better candidates, making the paraphrase probability estimate less reliable. Table 4.8 demonstrates the potential pivot-related problems in the extreme case of identity paraphrases, for which one might intuitively expect a relatively high probability: trivially, the phrase itself is highly likely to appear where it appears in the text. But in reality, the estimated probabilities are often quite low.[10] In contrast, large monolingual resources are relatively easy to collect, the paraphrasing engine described here involves only a single translation/paraphrasing step per target phrase, and the identity paraphrasing score is always 1 (unless the target phrase is not in the monolingual corpus). In addition, the Callison-Burch et al. (2006) paraphrases were reported to filter out named entities and numbers, while here named entities were not filtered out (but digits and punctuation were).

[10]In fact, identity paraphrase entries with $p > .75$ are quite rare, and from a short sampling, it seems that almost all of the higher scoring cases are named entities. Obviously, these identity paraphrases are of no use in augmenting translation models. They are brought here merely to illustrate the potential inaccuracy of the translation probability estimation via pivoting.

What is a fair comparison? Should the monolingual and bilingual training resources be equivalent in some sense? Should the lengths of the phrase or its paraphrase be limited to the same values in both techniques? Should pivoting paraphrases be threshold-filtered as the distributional ones are? (But a .3 vector similarity score is not equivalent to a .3 probability score). Or should the number of augmentative paraphrases be similar in both? Perhaps each technique should be presented in its best light. But finding the best running parameters for each technique is not a simple matter either. Therefore, the comparisons here should be regarded as a first stab at this problem, which further research is likely to shed more light on.

Phrase e_1	**Paraphrase** e_2	**Estim. prob.** $p(e_2\|e_1)$
Typical:		
abandon	abandon	0.12
abandon the idea of	abandon the idea of	0.37
deal between the two	deal between the two	0.48
zagreb	zagreb	0.65
Rare:		
jimmy	jimmy	0.87
john	john	0.74
larry	larry	0.91

Table 4.8: Estimated probabilities of English identity paraphrases via pivoting (sample taken from `http://www.cs.jhu.edu/~ccb/howto-extract-paraphrases.html`). Identity paraphrase entries with $p > .75$ are quite rare, and from a short sampling, it seems that almost all of them are named entities.

Table 4.6 also shows an exemplar comparison with the pivoting paraphrases used in Callison-Burch et al. (2006). It seems that the pivoting paraphrases might suffer more from having frequent function words as top candidates, which might be a by-product of their alignment "promiscuity". However, the top antonymous can-

didate problem seems to mainly plague the monolingual distributional paraphrases (but improves with larger corpora).

Tables 4.9 and 4.10 present paraphrase-augmented translation examples. The baseline English translation contains a few untranslated (OOV) words. The pivot model succeeds in translating *escucho* to *hear*, and *limitar* to *limit*, but omits translation for *afirman*. The distributional models fail to improve on *escucho*, but offer semantically close translations to the other two OOV words: The models trained with larger monolingual corpora for paraphrasing produce better translations: (*reducing vs. reduce* and *considered vs. say*). The baseline Chinese translation omits translation for *men* and *may*. All other models contain translation for *men (man)*. The pivot model and the "pure-augmented" *1 + 2-6grams* model do worse than baseline in omitting correct translation for *reap*, but the hybrid model is as good as the baseline there. In addition, the hybrid model is the only model to have semantically close translation for *may (can)*: the baseline and pivot models omit it, and the "pure-augmented" model translates it as *the month of May*.

One potential advantage of using bitexts for paraphrase generation is the usage of implicit human knowledge, i.e., sentence alignments. The concern that not using this knowledge would turn out detrimental to the performance of SMT systems augmented by paraphrases as described here was largely put to rest, as the new method improved the tested subset SMT systems' quality.

system / origin	example
source	cuando escucho las distintas intervenciones , creo que quienes afirman que deberíamos analizar nuestras prioridades y limitar el número de objetivos que queremos conseguir , están en lo cierto .
reference	when i listen to the various comments made , i find myself agreeing with those who recommend that we take a look at our priorities and then limit the number of aims we want to achieve
baseline	**escucho** when the various speeches, i believe that those who **afirman** that we should our environmental **limitar** priorities and the number of objectives we want to achieve, are in this way.
pivoting (MW)	when i can **hear** the various speeches , i believe that those people that we should look at our priorities and to **limit** the number of objectives we want to achieve , are in fact .
wmt09+acquis .1-4grams	**escucho** when the various speeches, i believe that those who **claiming** that we should environmental **limitar** our priorities and the number of objectives we want to achieve, are on the way.
wmt09+acquis .1grams	**escucho** when the various speeches, i believe that those who **considered** that we should our environmental priorities and **reducing** the number of objectives we want to achieve, are on the way.
wmt09+acquis+afp .1grams	**escucho** when the various speeches, i believe that those who **say** that we should our environmental priorities and **reduce** the number of objectives we want to achieve, are on the way.

Table 4.9: Spanish-English (S2E) translation examples on 10k-bitext models. Some translation differences are in bold.

model/source	example
source	men , too , may reap protection from exercise .
reference	**男** 人 也 **可 以 从** 锻炼 中 获**得** 保护 。
baseline	, 太 多 **得 到** 保 护 , **从** 演 习 。
gloss	, too many **reap** protection , **from** maneuver .
1 + 2-5grams-pivot >.3	**男 性** , 太 , 果 保 护 演 习 。
gloss	**man** , too , fruit protection maneuver .
1 + 2-6grams	**男 性** , 太 , 五 月 果 保 护 **从** 演 习 。
gloss	**man** , too , May fruit protection **from** maneuver .
1 + 2-6grams-hybrid	**男 性** , 太 多 **得 到** 保 护 , **可 以 从** 演 习 。
gloss	**man** , too much **reap** protection , **can** **from** maneuver .

Table 4.10: English-Chinese (E2C) translation examples on 29k-bitext models. Some translation differences are in bold.

Some of the experiments presented here differed only in the similarity score threshold used (.3 or .6). As can be seen in Table 4.11, the effect of such a switch is hard to predict for these threshold values.

subset	**mono.corp.**	**features**	**BLEU ↑**	**TER↓**
10k	wmt09+aquis	1grams	0.00	0.000
10k	wmt09+aquis+afp	1grams	+0.09	+0.074
10k	wmt09+aquis	1+2-6grams	-0.06	-0.123
10k	wmt09+aquis+afp	1+2-6grams	+0.02	+0.077
20k	wmt09+aquis+afp	1+2-6grams	-0.16	-0.014
80k	wmt09+aquis+afp	1+2-6grams	-0.10	+0.321

Table 4.11: Gain differences when switching from .3 to .6 similarity score threshold

The paraphrase quality remains an issue with this method (as with all other paraphrasing methods). Some possible ways of improving it, besides using larger corpora, are: using syntactic information (Callison-Burch, 2008), using semantic knowledge such as thesaurus or WordNet to perform word sense disambiguation (WSD; Resnik, 1999), improving the similarity measure, and refining the similarity threshold. I would like to explore ways of incorporating syntactic knowledge that do not sacrifice coverage as much as in Callison-Burch (2008); incorporating semantic knowledge to disambiguate phrasal senses; using context to help sense disambiguation (Erk and Padó, 2008); and optimizing the similarity threshold for use in SMT, for example on a held-out dataset: the higher the threshold the lower the coverage, while the lower the threshold the lower the paraphrases and translation quality. It remains to be seen how these two opposite effects play out.

Scaling up to larger monolingual corpora, e.g., one billion (1G) words or more, although potentially promising in terms of quality and coverage, poses some chal-

lenges. If pre-loading the corpus to working memory (RAM), loading time becomes non-negligible, and if using data structures such as a suffix array for pattern matching, then memory capacity becomes an issue. Searching all occurrences and contexts of some phrase from disk, even with a dedicated data structure, becomes too slow, when this has to be done millions of times. Sampling techniques may help, and in fact are already in place. However, when the sampling size ratio is too small, inaccuracies become non-negligible too. Splitting the corpus and searching in parallel, for example with a Map/Reduce paradigm over a Hadoop cluster, is one way to handle larger corpora. Similar approaches have been applied successfully for similar cases such as word co-occurrence counting (Lin, 2008). Currently, distributional semantic distance measures tend to become less accurate when comparing profiles (DPs) of targets with a large difference in occurrence frequency in the monolingual corpus. This problem is expected to exacerbate with larger corpora, and needs to be taken up in future research. Augmenting the phrase-table with the paraphrase-based translation rules, which is done now in memory using a hash table, also poses memory capacity problems, since using larger corpora results in generating more paraphrases, which in turn results in augmenting the table with more translation rules. This problem is even more pronounced when augmenting with more than one feature. The hash table size problem, however, can be ameliorated using a disk (trie) grammar instead.

Fine-grained feature granularity proved advantageous here too, as was shown in the previous chapters: The *1+2-6grams-hybrid* model was the best hybrid per-

former, significantly better than both the coarser *1-6grams-hybrid*, and the less informed *1grams-hybrid*. Similarly, the *1+2-6grams* model was the best "pure-augmented" performer, significantly better than the coarser *1-6grams*. This pattern seems to have held also for the *1+2-5grams-pivot* model, which was the best pivot-style performer, although its advantage over the coarse *1-5grams-pivot* did not reach statistical significance. This pattern further supports the claim that a finer feature granularity is advantageous over using only a single feature for all paraphrases (*1-6grams, 1+2-6grams-hybrid, 1+2-5grams-pivot*), and over using only partial data as paraphrases of certain phrase lengths (*1grams, 1grams-hybrid, 1grams-pivot*).

Note that there is a trade-off between finer granularity and data sparseness. The number of generated paraphrases of unknown phrases, especially above a certain similarity score threshold (.3 in most experiments here), drops in proportion to the length of the unknown phrases. Therefore, separate soft constraint features for longer phrases is likely to be of low quality or marginal impact, while increasing runtime. If using the *de facto* standard MERT (as opposed to, say, the newer MIRA) for feature weight optimization, the mere increased number of features might be prohibitive by itself. In order to show the fine granularity advantage, it was sufficient to split paraphrases of unigrams from those of longer phrases. It remains to be explored what is the optimal split, which is probably dependent on monolingual corpus size.

The paraphrasing method presented here is quite general, and therefore different similarity measures, including other corpus-based or hybrid measures, can be

plugged in to generate phrasal paraphrases. These, in turn, regardless of generation technique, yielded better results when used in finer granularity of associated log-linear features. Scaling up is an issue, but there are clear and promising research directions to tackle this issue. A further goal in the future would be to create a distributional similarity-based, high-performance SMT system, with reduced or even no dependency on manually-aligned parallel texts. Such a system would be especially beneficial to the "low-density", resource-poor languages, but has potential to benefit all languages and language pairs.

Chapter 5

A Unified Statistical NLP Model with Linguistic Soft Constraints

5.1 Introduction

This is a technical chapter, offering a unified framework, which (a) generalizes both the syntax-aware translation models (Chapter 2) and the hybrid knowledge / corpus-based semantic similarity models (Chapters 3 and 4), so that each can be viewed as an instance of the generalized framework, and which (b) in principle allows combining both syntactic and semantic soft constraints in a single tunable unified statistical NLP model with soft constraints.

I start below with discussing potential benefits in defining a unified model, continue in Section 5.2 with describing a log-linear model, go in section 5.3 through the definition of soft constraints and how they are added to a model, and end with showing how the soft syntactic constraints (Section 5.4) and the soft semantic constraints (Section 5.5) can each be viewed as an instance of a general unified model. I leave the actual implementation and evaluation of such a framework for future research.

There are several benefits in defining a unified model :

1. relations and similarity among the specific cases are formalized, and defined more precisely;

2. new such relations might be discovered, potentially benefitting the specific sub-fields / cases;

3. insights in one sub-field may become applicable to other sub-fields that fit the generalized unified model; and

4. techniques developed for one sub-field may become applicable to other sub-fields that fit the generalized unified model.

The emphasis in this dissertation on finer-grained constraints, in both the syntactic and semantic cases, falls under points (1) and (3) above: The positive results in the syntactic case served as an additional motivation to try finer granularity in the semantic case, too. Currently, there is no weight tuning in the semantic work described here. Applying a task-specific weight tuning algorithm – MERT (Och, 2003) or MIRA (Crammer and Singer, 2003; Crammer et al., 2006; Watanabe et al., 2007; Chiang et al., 2008) – to the semantic constraints, as is the case for the syntactic constraints, falls under point (4), and is a natural next step, which I leave for the future.

5.2 Log-Linear Model

A common case in natural language processing (NLP) is a problem that involves estimating many factors. A typical example would be finding the most likely word sequence, which involves estimating probabilities of encountering (or a source generating) series of words, where each word or short sequence of words (n-gram) would have an associated probability (or a probability approximation) based on past observation. The likelihood of each point in such a space can be expressed as a product of all these probability factors, i.e., a non-linear model, which is slow to compute and often results in underflow errors. A search problem in a non-linear model may be defined as follows:

$$\arg\max_{\bar{x}} \prod_{i=1}^{m} g_i(\bar{x})^{\lambda_i},\ \forall \bar{x} \in \bar{X} \tag{5.1}$$

where each g_i is called a feature of the model, and is defined over some domain $\bar{X}$, e.g., all strings in some language. The vector notation of $\bar{X}$ denotes possible multiple dimensions for each string, e.g., lemmatized form or syntactic information, in addition to the surface form. The contribution of each feature g_i is weighted by a power-weight λ_i.

In order to speed up calculations and avoid underflow errors, these models are often taken the log of, resulting in a simple sum of weighted log terms. A log-linear

model (equivalent to the model above) would be:

$$\arg\max_{\bar{x}} \sum_{i=1}^{m} \lambda_i h_i(\bar{x}), \quad \forall \bar{x} \in \bar{X} \tag{5.2}$$

where each $h_i = \log g_i$. Weights are typically optimized using a development set and an optimization algorithm such as minimum error rate training (MERT; Och, 2003) or Margin Infused Relaxed Algorithm (MIRA; Crammer and Singer, 2003; Crammer et al., 2006; Watanabe et al., 2007; Chiang et al., 2008).

For the purposes of this exposition, I will use a more specific notation, assuming the input consists of two vectors, $\bar{x}_1, \bar{x}_2$, where $\bar{x}_1$ is given, and can be viewed as a source language string in a SMT setting, and the model searches $\bar{x}_2$ values, which can be viewed as target language strings in such a setting:

$$\arg\max_{\bar{x}_2} \sum_{i=1}^{m} \lambda_i h_i(\bar{x}_1, \bar{x}_2), \ \bar{x}_1 \in \bar{X}_1, \forall \bar{x}_2 \in \bar{X}_2 \tag{5.3}$$

5.3 Constraints

5.3.1 Hard Constraints

A constraint, and more specifically, a *hard constraint*, can be defined or viewed as some feature g_i, for which exist some range $\bar{r}$, outside of which input values $\bar{x}$ give zero. The feature g_i can be defined as a binary feature $g(\bar{x}) = 1$ if $\bar{x} \in \bar{r}$, or 0 otherwise. The value 0 will zero the whole product in Equation (5.1), even if the

particular $\bar{x}$ scores high with many other features, while the value 1 will have no effect on the product.[1]

An alternative way of defining a hard constraint would be to define it as a *partial feature function* g, which is defined only for the range $\bar{r} \subset \bar{X}$. Therefore the whole model too is not defined for input $\bar{x}$ outside range $\bar{r}$ – again, even if $\bar{x}$ scores high with many other features.

Either way, hard constraints typically allow for speed ups and shortcuts in calculation, since the search algorithm can take into account the zeros and not attempt to look in the corresponding areas of the search space. This kind of *a-priori* constraint is often *theory-driven*. For example, in syntax-directed SMT, a hard constraint design might be not to consider translation units (source word sequences) that are not syntactic constituents (e.g., Yamada and Knight, 2001). In the example in Figure 2.2, a model with a hard syntactic constraint will not consider translating *minster gave a* as a unit. While it might seem as a good constraint in this case, it turns out that it is too restrictive in other cases, e.g., the German word sequence *es gibt*, which is not a syntactic constituent, translates very naturally to *there is* (Koehn, 2003).

[1] g_i may be defined as returning any other non-zero value instead of 1, but since all inputs that do not result in g_i returning zero result in returning the same other value, it can be canceled out when comparing all non-zero products of Equation (5.1). Hence this is equivalent to contributing 1 to the product in this equation.

5.3.2 Soft Constraints

In contrast to the above, a *soft constraint* can be viewed as a fully defined non-binary feature function $g : \bar{X} \rightarrow \Re$, where $\Re$ denotes the Real numbers. Define $g : \bar{X} \rightarrow [0..1]$.[2] A soft constraint can be viewed as biasing the model towards certain ranges. In the German example above, a soft syntactic constraint might discourage the model from translating a non-syntactic constituent such as *es gibt*, but the model would still be able to translate it as a unit, if the total contributions from all features warrant it. This is in contrast to a hard constraint that would rule it out as a possible translation unit.

Adding a (soft) constraint to a model is realized simply by adding a feature function to the log-linear sum in Equation (5.2) or (5.3) above.

The advantage of soft constraints is the consideration of solutions that might be dispreferred by some constraints or features, but still be potentially globally optimal when taking data-driven patterns and all weighted constraints and features into account. A key difference between the soft and hard cases is that the soft constraints can be realized as tunable biases, i.e., the constraints' weights are tuned during a weight optimization step. They do not exclude any solution *a-priori*, while the hard constraints simply narrow down the search space in a non-tunable fashion (e.g., with binary values that may zero out a score product).

[2] g_i may be defined as returning any range, but for practical reasons, if the range is [0..1], an associated weight $lambda_i$ can scale the feature's overall influence up or down. A negative weight inverts the influence from a reward-type feature to a penalty-type feature.

In the following sections I will re-describe my soft syntactic constraints (Section 5.4), and my hybrid word/concept-based semantic similarity measure (Section 5.5). I will argue that this semantic similarity measure can be viewed as a model having a concept-based soft semantic constraint over word-based distributional profiles. I will show that models containing either of these soft syntactic and semantic constraints can be viewed as special cases of a more general log-linear NLP model with soft linguistic constraints.

5.4 Soft Syntactic Constraints

The *de facto* standard in SMT is using weighted features (functions of the source and target language strings), combined in a log-linear framework (Och, 2003):

$$\arg\max_{e} \sum_{i=1}^{m} \lambda_i h_i(e, f),\ f \in F, \forall e \in E \tag{5.4}$$

where f is a given string in F, F is the set of all foreign (source) language strings, E is the set of target language strings, h_i are feature functions over strings from F and E, and λ_i are their corresponding tunable weights. I introduce the following additional features / constraints, defined in Section 2.4:

- Reward for using a phrase translation rule whose source side precisely matches the boundaries of a certain syntactic constituent:

  $h'(e, f) = isMatchingConstituent(f)$, and

- Penalty for using a phrase trasnlation rule whose source side crosses the boundaries of a certain syntactic constituent:

  $h''(e, f) = isCrossingConstituentBoundaries(f)$.

These soft syntactic constraints were implemented by adding the above binary feature functions as weighted terms to the weighted sum, and re-training the model to find new optimal weights λ_i. In contrast, the corresponding hard syntactic constraints can be viewed as considering only the partial domains

$\{f|f \in F \ \wedge \ isMatchingConstituent(f) == 1\}$

and/or

$\{f|f \in F \ \wedge \ isCrossingConstituentBoundaries(f) == 1\}$,

instead of the full domain of all $f \in F$.

Equation (5.4) is a special case of the log-linear model in Equation (5.3), where $\bar{x}_1 = f, \bar{X}_1 = F, \bar{x}_2 = e, \bar{X}_2 = E$. The feature functions above were simply added as weighted terms to the sum.

5.5 Soft Semantic Constraints

The task of finding a best paraphrase to a target word or phrase e can be concisely formalized as:

$$\arg\max_{e'} sim(e, e') \quad \forall e, e' \in E \tag{5.5}$$

where E=all strings/phrases in English (or any other language).

Corpus-based distributional semantic similarity measures often collect distributional profiles (DPs), a.k.a. distributional vectors, for the target words or phrases (denoted by e and e' above). As mentioned in Section 3.2.2.1, a DP of some target word / phrase e is a set of ordered pairs $\langle collocate, SoA \rangle$, where *collocate* are the words or phrases that co-occur in the vicinity of e (usually occurring within a small fixed-size sliding window around the occurrences of e in a training corpus), although in principle *collocate* can be any word in the training vocabulary E; SoA is a strength-of-association measure between e and collocate, such as a co-occurrence count, conditional probability $p(collocate|e)$, point-wise mutual information (PMI), log-likelihood ratio, etc. The DP similarity measure is implemented simply as a similarity function over such vectors (where the collocates serve as the vectors' dimensions). A typical vector similarity function, which is also used in this work, is the well-known cosine function (Equation (3.6), repeated here for convenience).

$$\begin{aligned} & sim(e, e') = \\ & = psim(\mathrm{DP}_e, \mathrm{DP}_{e'}) = \\ & = Cos(\mathrm{DP}_e, \mathrm{DP}_{e'}) = \\ & = \frac{\sum\limits_{w_i \in E} SoA(e, w_i)\ SoA(e', w_i)}{\sqrt{\sum\limits_{w_i \in E} SoA(e, w_i)^2}\ \sqrt{\sum\limits_{w_i \in V} SoA(e', w_i)^2}} \end{aligned} \tag{5.6}$$

Mohammad and Hirst (2006), hereafter MH06, argued that the correlation of corpus-based similarity scores with human judgments can be improved if one

teases apart the different senses of each target word. Using a thesaurus, they assign each word as many senses as there are concepts under which it is listed in the thesaurus. Then, they collect distributional profiles for each such concept/sense, denoted DPCs. Next, given two target words e, e', they measure similarity of each DPC of e to each DPC of e', and return the smallest distance (largest similarity score) of all these pairs. For example, here is one way of expressing the MH06 semantic similarity formula, using cosine:

$$sim_{MH06}(e, e') = \underset{\substack{s \in senses(e), \\ s' \in senses(e')}}{\arg\max} \cos(\mathrm{DPC}_s, \mathrm{DPC}_{s'}) \tag{5.7}$$

In a setting of paraphrase generation (Chapter 4), the goal is to find the most similar word/phrase e' to the given word/phrase e. Still using the cosine example above, one can take its argmax:

$$\begin{aligned} &\arg\max_{e'} \underset{\substack{s \in senses(e), \\ s' \in senses(e')}}{\arg\max} \cos(\mathrm{DPC}_s, \mathrm{DPC}_{s'}) = \\ &= \arg\max_{e'} \underset{\substack{s \in senses(e), \\ s' \in senses(e')}}{\arg\max} \frac{\sum_{w_i \in E} SoA(s, w_i)\ SoA(s', w_i)}{\sqrt{\sum_{w_i \in E} SoA(s, w_i)^2}\ \sqrt{\sum_{w_i \in E} SoA(s', w_i)^2}} \end{aligned} \tag{5.8}$$

where SoA can be any strength-of-association measure between the concept or coarse sense s and any word w_i. If using cosine over vectors of conditional probabilities, which MH06 denote Cos_{CP}, then SoA would be $p_C(w_i|s)$, the conditional probability of any word w_i given concept s.

The MH06 method has two main weaknesses: (1) if e is not in the known vocabulary, the method is inapplicable, and (2) it is inherently coarse, since a DPC models an aggregated "concept" (word grouping) target, and not an individual word, let alone a sense-disambiguated word (see Section 3.2.3). For example, if *wizard, warlock*, and *wand* are listed under the same concept, there is no way of telling which two of the three are closer in meaning; if *bank* and *wave* are listed under the same thesaurus category – say, RIVER – they will be reported as perfect synonyms, even if one is also listed under categories that do not include the other. In order to overcome these limitations, I introduced hybrid models (Chapter 3), which can be viewed as a finer-grained generalization of the MH06 model. In essence, the MH06 DPCs serve as soft semantic constraints on a corpus-based word-based DP model. Equation (5.8) takes the following form when used with the hybrid models of Chapter 3:

$$\arg\max_{e'} \arg\max_{\substack{s\in senses(e),\\ s'\in senses(e')}} \cos(\mathrm{DPWC}_{e,s}, \mathrm{DPWC}_{e',s'}) = \arg\max_{e'} \arg\max_{\substack{s\in senses(e),\\ s'\in senses(e')}} \frac{\sum_{w_i\in E} SoA_s(e,w_i)\ SoA_{s'}(e',w_i)}{\sqrt{\sum_{w_i\in E} SoA_s(e,w_i)^2}\ \sqrt{\sum_{w_i\in E} SoA_{s'}(e',w_i)^2}} \tag{5.9}$$

where $\mathrm{DPWC}_{e,s}$ is a word/concept hybrid distributional profile of target word e in sense s, whose SoA may be calculated as follows:

$$\begin{aligned} SoA_s(e,w_i) &= \lambda\ q_s(e,w_i) + (1-\lambda)\ count_W(e,w_i), \quad \text{where} \\ q_s(e,w_i) &= p_C(s|w_i)\ count_W(e,w_i) \end{aligned} \tag{5.10}$$

Here $count_W$ is the "pure" word-based co-occurrence count, q_s is the concept-based sense-proportional co-occurrence count, p_C is the conditional probability calculated using the concept-based co-occurrence matrix, and $0 \leq \lambda \leq 1$ is the interpolation weight of the discounted model with the "pure" word-based model. When $w_i \notin V_\theta$, where V_θ is the concept-based matrix vocabulary (thesaurus vocabulary), define $p_C(s|w_i)$ to be uniform over all senses s: $p_C(s|w_i) = 1/\sum_s 1$. This way, these conditional probabilities will sum to 1, and therefore, the sense-aware word/concept hybrid co-occurrence counts of w_i over all senses will sum to the word-based sense-unaware count: $\sum_s q_s(e,w_i) = count_W(e,w_i)$.

Intuitively, q_s can be viewed as a discounted co-occurrence count: The collocate's "pure" word-based co-occurrence count or SoA is discounted in proportion to its strength of association with sense s (relative to all senses). The interpolation weight λ can be interpreted as the degree of confidence in each model (the pure word-based model and the concept-based sense-proportional discount model); on the one hand, Mohammad and Hirst showed that sense information helps in word pair ranking task, but on the other hand, their concept-based collocation matrix was calculated using heuristics, and therefore is noisy. In addition, one might believe (as part of a cognitive theory) that even collocates of other senses play some role (as small as it might be) in the mental representation of the target word, and therefore also influence similarity judgments – in which case, the word-based collocates should not be totally discounted if not co-occurring with the current sense s.

Note that although such an interpolation may have a smoothing effect *de facto*, (for example, in case that the thesaurus vocabulary is too small and does not contain the collocate), the interpolation is different than smoothing. A typical smoothing here would move some "count mass" among the collocates, but will generally preserve their relative strengths; however, the interpolation may well result in increasing the SoA value of some collocate w_i so that $SoA(e, w_i) > SoA(e, w_j)$ for some other collocate w_j, while in the discounted model before interpolation it was the case that $SoA(e, w_i) < SoA(e, w_j)$. Note also that a non-interpolated model (Equation (5.10) with $\lambda = 1$) is simpler and more elegant, since it does not require estimating the λ parameter. In practice, my reported results in Chapters 3 and 4 were based on

$\lambda = 1$. However, an optimized (tuned) value for λ is potentially more accurate, and I see it as a future research direction.

The difference between Equations (5.8) and (5.9) is this: MH06 apply a hard semantic constraint, where $SoA_s(e, w_i) = p_C(w_i|s)$, and $SoA_{s'}(e', w_i) = p_C(w_i|s')$. In other words, they have a non-tunable component, which always ignores the identity of the target word e once its concept/sense s is retrieved (and similarly for e' and s'). The soft semantic constraints in this dissertation do not abstract away from e and e', and allow for optimizing the discount weight. Another limitation of the MH06 approach is that due to the small size of V_θ, many of their SoA_s values might end up being zero. By introducing the interpolated variant of SoA_s , my proposed model ameliorates this problem for any $\lambda \neq 1$, i.e., $0 \leq \lambda < 1$.

In order to see more clearly the structure of the proposed soft semantic constraints in a unified model framework, Equation (5.9) can be rewritten as follows (explanation below), starting with defining Z_C to be the denominator in Equation (5.9):

$$Z_C = \sqrt{\sum_{w_i \in E} SoA_s(e, w_i)^2} \sqrt{\sum_{w_i \in E} SoA_{s'}(e', w_i)^2} \tag{5.11}$$

Next, the double arg max argument can be rewritten as follows:

$$\frac{\sum_{w_i \in E} SoA_s(e, w_i)\ SoA_{s'}(e', w_i)}{Z_C} = \\ = \frac{\sum_{w_i \in E} [\lambda q_s(e, w_i) + (1-\lambda) count_W(e, w_i)]\ [\lambda q_{s'}(e', w_i) + (1-\lambda) count_W(e', w_i)]}{Z_C} \tag{5.12}$$

$$= \sum_{i=1}^{4} \delta_i h_i(e, s, e', s') \tag{5.13}$$

The transition to Equation (5.12) above comes from substituting the SoA_s formula from Equation (5.10). Equation (5.13) simply breaks down the parentheses and renames all the terms from Equation (5.12) as follows:

$$\delta_1 = \lambda^2$$

$$\delta_2 = \lambda(1-\lambda)$$

$$\delta_3 = \lambda(1-\lambda)$$

$$\delta_4 = (1-\lambda)^2$$

$$h_1(e, s, e', s') = \frac{\sum_{w_i \in E} q_s(e, w_i)\, q_{s'}(e', w_i)}{Z_C} = \cos(\mathrm{DPWC}_{e,s}, \mathrm{DPWC}_{e',s'})$$

$$h_2(e, s, e', s') = \frac{\sum_{w_i \in E} q_s(e, w_i)\, count_W(e', w_i)}{Z_C}$$

$$h_3(e, s, e', s') = \frac{\sum_{w_i \in E} count_W(e, w_i)\, q_{s'}(e', w_i)}{Z_C}$$

$$h_4(e, s, e', s') = \frac{\sum_{w_i \in E} count_W(e, w_i) count_W(e', w_i)}{Z_C} = \frac{Z_W}{Z_C} \cos(\mathrm{DP}_e, \mathrm{DP}_{e'})$$

where similarly to the concept-related denominator Z_C, I define a shorthand symbol for the word-related denominator:

$$Z_W = \sqrt{\sum_{w_i \in E} count_W(e, w_i)^2} \sqrt{\sum_{w_i \in E} count_W(e', w_i)^2} \tag{5.14}$$

The original (pre-bias) concept-based sense-proportional model is expressed by h_1 in the formula above, and is weighted by δ_1; the original (pre-bias) word-based model is expressed by h_4 , weighted by δ_4 and the ratio of the denominators Z_W and Z_C , which depend on e and e'. The other h_m and δ_m, for $m = 2, 3$, can be interpreted as "cross-term" hybrid models consisting of some distance (or relation) between DP_e and $\mathrm{DPC}_{e'}$, or some distance/relation between DPC_e and $\mathrm{DP}_{e'}$ – weighted by δ_2 or δ_3, respectively.

The semantic model as expressed in Equation (5.13) is, similarly to the syntactic model in Section 5.4, an instance of the linear model in Equation (5.3), with $\bar{x}_1 = \langle e, s \rangle, \bar{x}_2 = \langle e', s' \rangle, \bar{X}_1 = \bar{X}_2 = \langle E, senses \rangle$.

5.6 Discussion and Conclusion

Both syntactic and semantic models and soft constraints described above can be framed as instances of Equation (5.2) or (5.3). But their resemblance does not end there: Translation can be viewed as a special case of paraphrase generation (and hence, a semantic distance problem). Therefore one can define Equations (5.4) and (5.9)-(5.13) as special cases of a more general similarity : Define the paraphrase function $par(u)$ whose domain U is a set of phrases (e.g., the set of all English phrases), and whose range V is also a set of phrases (same set as U or a different one, e.g., the set of all French phrases). Let s and r denote the senses of $u \in U$ and $v \in V$, respectively. Then:

$$par(u) = \arg\max_{v} \arg\max_{\substack{s \in senses(u), \\ r \in senses(v)}} \sum_{m} \delta_m h_m s(u, s, v, r) \tag{5.15}$$

The translation model described in Section 5.4 can be viewed as a special (somewhat degenerate) case of this formula, where one only knows of one sense of u and one sense of v, and hence can omit the second argmax;[3] $U = F, V = E$, and $h_m(u, s, v, r) = h_m(u, v)$ feature functions of the SMT model. The soft syntactic constraints would be $h_i(u, v) = isMatchingConstituent(u)$ and/or $h_j(u, v) = isCrossingConstituentBoundaries(u)$ as described above. The semantic similarity model and soft constraints described in Section 5.5 can be viewed as an almost trivial special case, where $U = V$ (and both =English in my experiments), $r = s', v = e'$, and the h feature functions are as above.

Beside their common form as additional function terms in a linear sum, and their being special cases of $par()$, the soft syntactic and semantic constraints also share another characteristic: They draw their bias from human linguistic knowledge, syntactic or semantic, respectively, that is currently non-extractable from a non-annotated corpus. But rather than limit the translation/paraphrase search space according to the respective linguistic theory used (as done with hard constraints), they enable corpus-based patterns to emerge even if these patterns do not fit the theoretical bias.

[3]Models that perform WSD or phrase-sense disambiguation (such as Carpuat and Wu, 2007) might fit into the more general formula, using (instead of omitting) the second argmax.

Chapter 6

Conclusion

6.1 Overview and Summary of Contributions

This dissertation presented effective ways of combining statistical data-driven approaches to natural language processing with linguistic knowledge sources that are based on manual text annotation or word grouping according to semantic commonalities. This was achieved via the use of linguistic resource-based constraints – of syntactic or semantic nature – on statistical NLP models. The key properties of these constraints were that they were (a) *soft*, and (b) *fine-grained*, both of which are discussed below. I showed how to gainfully apply and evaluate each of these knowledge / corpus-based hybrid models in state-of-the-art end-to-end SMT settings. I presented a generalized unified model – a statistical NLP model with (linguistic) soft constraints – and showed how the seemingly different hybrid models with syntactic or semantic constraints can be viewed as instances of the generalized model. This unified framework opens the door, in principle, to combining these different linguistic soft constraints – and potentially other constraints, too – in a single model.

In Chapter 2, I showed that fine-grained soft syntactic constraints can significantly improve SMT quality. Use of syntactic parsing information in NLP tasks, and especially in SMT, is wide-spread (see Section 2.2 and Lopez, 2008b), and needs no introduction. Use of soft constraints applying syntactic information in SMT has also been introduced before, even if previously without positive results (Chiang, 2005). However, use of the newly introduced semantic soft constraints required first initial investigation of their properties on a basic word (unigram) level, with an intrinsic evaluation of their performance. In Chapter 3, I addressed this by testing models with and without these soft semantic constraints on word-pair similarity ranking tasks. The hybrid models (with soft semantic constraints) out-performed or equaled their non-hybrid corresponding models. In Chapter 4, I extended these semantic models from modeling words to modeling phrases, and from measuring phrase similarity to *finding* similar phrases – i.e., generating paraphrases. I presented a novel, distributional paraphrase generation technique, employing these semantic models, and used it to augment SMT models, evaluating paraphrase quality on translation tasks, similarly to the evaluation of the soft syntactic constraints. The SMT model augmentation with this paraphrasing technique significantly improved translation quality of models trained with smaller training sets, in different language pairs. Hybrid semantic models out-performed their non-hybrid corresponding models, as was the case in the previous chapter. Fine-grained use of linguistic information proved beneficial in each of these chapters, and in many cases significantly so. In Chapter 5, I showed that these two types of soft linguistic constraints are more similar than

what first meets the eye, and that they can be combined, in principle, in a single model.

My main contributions in this doctoral research were:

- Showing the advantage of soft constraints with fine-grained linguistic information, relative to "pure" corpus-based baseline and coarse-grained soft constraints, in SMT. (Chapter 2)
- Showing the advantage of soft constraints with fine-grained linguistic information, relative to "pure" corpus-based baseline, hard constraints and coarse-grained soft constraints, in lexical semantics and paraphrase generation. (Chapter 3)
- Evaluating both syntactic and semantic (paraphrastic) contributions in state-of-the-art end-to-end phrase-based SMT systems, showing statistically significant gains in BLEU score. (Chapters 2 and 4)
- Introducing a novel paraphrase generation technique, using a monolingual corpus-based distributional approach, independent of commonly used sentence-aligned parallel texts, which are limited, human labor-intensive resources. (Chapter 4)
- Introducing a novel semantic reinforcement component (evidence from similar paths or rules) for scoring paraphrase-based translation rules, and using these scores to augment translation models. (Chapter 4)

- Showing the advantage of fine-grained scoring of paraphrase-based translation rules. (Chapter 4)

- Proposing a unified linear statistical NLP model with linguistic resource-based soft constraints, which, in principle, can be tuned using standard parameter optimization techniques, and of which the syntactic and semantic constraints models can be viewed as instances. (Chapter 5)

6.2 Soft Linguistic Constraints

I showed in Chapter 2 that soft syntactic constraints can, in fact, improve data-driven SMT models – in contract to previous attempts (Chiang, 2005). This was done both by introducing a new type of constraint (the penalty for crossing syntactic constituent boundaries), and by using fine-grained constraints (discussed below). Models including the new constraint type did better than the replication of the original Chiang (2005) model with the old constraint type (reward for matching syntactic constituent boundaries) more often than not. For example, *all-labels_* did significantly better than *Chiang-05* on the Chinese-English translation task in both test sets. But results of the *all-labels_* and *all-labels2* models on the Arabic-English translation task were inconclusive. Comparison of the two constraint types in fine-grained features were inconclusive as well. However, using the new constraint type with fine-grained features and feature combinations yielded significant gains over both the *Chiang-05* and the syntax-unaware baseline models, of up to 1.65 BLEU

points on the Chinese-English task, and up to 1.94 BLEU points on the Arabic-English task.

I showed in Chapter 3 that models with soft semantic constraints (the hybrid models) perform better than, or equal to, models with hard semantic constraints (the concept-based models) or with no semantic constraints (the word-based models). For example, on the Rubenstein and Goodenough (1965) noun-pair similarity task, the *hybrid-filtered*-cos-ll* model achieved a Spearman rank correlation of .77, compared to .64 and .73 by *concept*-cos-ll* and *word-cos-ll*, respectively. On the Resnik and Diab (2000) verb-pair task, *hybrid-proportional*-cos-pmi* achieved a correlation of .71, compared to .28 and .57 by *concept*-cos-pmi* and *word-cos-pmi*, respectively.

Soft constraints were also used in Chapter 4, with weighted log-linear features for semantic scoring of the paraphrase-based translation rules. The hard constraint equivalent (not including scoring features for the new translation rules) was shown to perform badly in Callison-Burch et al. (2006).

Soft constraints come with a price. As mentioned in Section 5.3, their disadvantage compared with hard constraints is that the latter narrow the search space and hence allow for speeding up the calculation, and potentially applying more efficient algorithms in both memory and runtime complexity. However, soft constraints may offer gains in output quality thanks to the consideration of solutions that might be completely ruled out by their hard constraint counterparts. Such solutions might still be optimal when taking all data patterns, weighted constraints and features into account. The potential benefits of using linguistic theoretical and/or resource-based

soft constraints on data-driven (corpus-based) models are both empirical, and to some extent, theoretical (or pertaining to the use of linguistic theory in NLP).

Empirical, in the sense that soft constraints enable better coverage of the data than hard constraints (as pointed out earlier). Other researchers, using hard constraints, e.g., in syntax-aware SMT, found it beneficial to increase coverage by hybridizing their syntax-driven or syntax-directed models with "pure" data-driven models, or otherwise relaxing the hard constraints, e.g., by binarizing parsing trees. Models with linguistic soft constraints are also more informed than "pure" corpus-based models. Therefore, such models can yield better performance, as I have shown in my experiments.

Theoretical, in the sense that current syntactic theory or its usage in NLP tends to be too coarse, or neglect to cover certain phenomena, that are nevertheless frequent in the language. For example, Koehn (2003) pointed out that the use of syntactic constituents as translation units is problematic: while useful in some cases (e.g., the German-English pair *das Haus – the house*), only translating constituents leads to loss of coverage (e.g., *es gibt – there is*). Soft syntactic constraints have the benefit of biasing and guiding the model to translate constituents, and yet, allow for translation of emerging non-constituent patterns such as *es gibt*, if frequent enough in the training data. The benefits of using soft constraints are potentially two-way: Such cases of emerging patterns can also potentially alert the theory side about certain overlooked phenomena.

6.3 Fine Granularity

Fine granularity was found to be key in the successful combination of these soft constraints:

For syntactic constraints, previous attempts to constrain SMT models by adding a single weighted feature, preferring translation of all syntactic constituents over other word sequences, yielded negative results. In contrast, the work described in Chapter 2, Marton and Resnik (2008) and Chiang et al. (2008), produced positive results: The soft constraints were applied using the syntactic parsing information with finer granularity – to each parsing label separately, with dedicated weighted features. Each such fine-grained constraint was implemented with an additional, cross-constituent boundary penalty variant, in addition to the previously attempted syntactic constituency reward variant (Chiang, 2005). Some new fine-grained features yielded significant gains over both the coarse *Chiang-05* and the syntax-unaware baseline models. For example, the fine-grained $NP^{=}$ model yielded up to 1.53 BLEU points over the baselines on the Chinses-English translation task. The fine-grained $AdvP^{=}$ model yielded up 1.46 BLEU points over the baselines on the Arabic-English translation task. Some new feature combinations yielded even significantly higher gains, up to 1.94 BLEU points – especially VP- and IP-related combinations, although in these experiments it was hard to find a precise consistent pattern cross-linguistically. These translation models remain essentially data-driven (corpus-based), but are constrained, or biased, by syntactic parsing information.

Feature selection, which was a problem when using minimum error rate training (MERT) for feature weight optimization, was no longer a problem when switched to using the newer Margin-Infused Relaxed Algorithm (MIRA) instead.

For semantic constraints, previous related work, attempting to create word sense-aware models (Mohammad and Hirst, 2006), created only coarser models of linguistic resource-based "concepts" – aggregated models of groups of related words according to the resource, and not models of individual words. The work described in Chapters 3 and 4, Marton et al. (2009b) and Marton et al. (2009a), applied soft constraints on distributional semantic models of words to effectively create word-sense-disambiguated models. These models are non-aggregated word-based models that remain essentially corpus-based, but are biased towards each of the linguistic resource's concepts that contain the model's target word – achieving, in fact, a word-sense resolution (whose optimal granularity is out of the scope of this work). These hybrid models resulted in most cases in higher gains over the "pure" corpus-based (word-based) and coarse concept-based baselines, as mentioned in the previous section.

Fine-grained semantic scoring of paraphrase-based translation rules yielded similar or additional significant gains as well, on the English-Chinese translation task (Table 4.2). This pattern repeated for both distributional and pivot paraphrasing techniques: The "1 + 2-5grams" and "1 + 2-6grams" models out-performed the respective coarser "1-5grams" and "1-6grams" models and the less informative "1grams" models, in most cases significantly so.

6.4 Novel Distributional Paraphrasing Technique

The distributional paraphrasing technique, presented in Chapter 4, was evaluated in automatic translation metrics (BLEU and TER), and yielded significant gains in BLEU , using a "pure" distributional semantic distance measure. Even greater gains, slightly but significantly better than the former gains, were achieved using the hybrid semantic models presented in Chapter 3. Manual observation of several sentence translations increased the confidence in the advantage of the hybrid models. The main advantage of the distributional monolingual corpus-based technique presented here over current pivoting techniques for paraphrasing is independence from parallel texts, which are a more limited resource than monolingual text. Although not conclusively shown here, I believe that the use of a sufficiently large same-genre monolingual corpus for paraphrasing can outperform pivoting techniques, in addition to being available also where parallel texts might not exist at all.

A noteworthy novelty in the paraphrase generation technique is the use of semantic reinforcement: the use of alternative paths of generating a particular paraphrastic translation rule as reinforcing evidence for the goodness of that rule (e.g., translating f to e both via f-f_1 + f_1-e and via f-f_2 + f_2-e; see Section 4.5.1). Preliminary experiments showed that not only the use of this semantic reinforcement resulted in memory-slimmer models, but it also enabled significant SMT gains in

BLEU , whereas the models that added a new rule for each path did not result in significant BLEU gains.

6.5 Unified Framework

In addition to evaluating the soft syntactic and semantic constraints in end-to-end state-of-the-art SMT settings, I also showed in Chapter 5 how they can all be viewed as instances of a unified statistical NLP model with soft constraints. In this unified framework, each of the linguistic soft constraints can in principle be added to the model linearly as weighted terms.

I took this analogy even further, and extended the *de facto* standard model to explicitly include the target sense of the translated or paraphrased word or phrase: Given a word, or generally a phrase u, potentially in context, return the semantically closest phrase v, under certain restrictions, taking potentially different senses of u and v into account. Sense-aware shortest semantic distance means that for the target sense s of the target phrase u, return a phrase v that has sense r, such that v in sense r is semantically closest to u in sense s.[1] The difference between tasks lies in the restrictions, which are task-specific: In a translation task, v must be in the target language; in a paraphrasing task, v must be in the same language, and formally non-identical to u.

[1]If context cannot be used to determine the current sense of u, then v must have a sense that is closest to one of the senses of u, closer than any sense of any other phrase v' to any sense of u.

6.6 Future Work

The various issues that this dissertation touched lead to many new questions and research directions:

Syntactic constraints:

- The NP-related models were salient in their absence from the top performing models in Arabic-English translation, although NPs seem intuitively natural translation units. Why is that?
- Interestingly, in some cases BLEU gains were observed even in the presence of few or no tags, which a feature was sensitive to, and which spanned more than a single token in the test set. Why is that?

Semantic constraints:

- The log-likelihood ratio-based semantic distance measures worked best for the noun-noun pairs test sets, while point-wise mutual information (PMI) worked best for the verb-verb test set. I would like to explore what measure, or measure combination, would work best for adjective–adjective, adverb–adverb, and cross-part-of-speech pairs, by exploiting specific information pertaining to these parts of speech in lexical resources, such as dictionaries and thesauri.
- Evaluate distributional and hybrid measures on phrase-pair test sets. Constructing a balanced phrase-pair set is not a trivial problem: Should all phrases

be of same length? Even if limited to bigram pairs, should they all belong to the same syntactic constituent, e.g., noun phrases? What about non-syntactic word sequences, such as *there is*? Should the heads of the phrases repeat in other phrases (e.g., *big balloon, tiny balloon*), and if so, how often? Should the complements repeat (e.g., *big balloon, big party*)? Should the test set include different types of complements (intersective, sub-sective/gradable, non-intersective, anti-intersective, etc., e.g., *green, big, alleged, fake*, respectively)? Should the test set include idioms? And so on.

- Infuse the co-occurrence-based models with linguistic information: e.g., instead of counting all collocates in a small sliding window, count collocates that are in specific syntactic relations with the target word or phrase, as in Lin (1997). However, here the syntactic dependency trees will be used for modeling semantic distance instead of word sense disambiguation as in Lin (1997). Optionally augment a sparse phrase with the distributional profile of its head (e.g., the verb in a verb phrase). Use such models for paraphrase generation, as well.

- The hybrid semantic models are currently restricted to languages such as English, that are not poor in lexical resources. This is because these hybrid models rely on lexical resources such as a thesaurus in order to construct the sense-aware concept / word co-occurrence matrix. I would like to extend the applicability of these hybrid models to resource-poor languages, as well. Since these models already makes use of the Mohammad and Hirst DPCs, one

straightforward way to extend them would be to make use of their cross-lingual DPCs (Mohammad et al., 2007).

Distributional paraphrasing technique and semantic reinforcement:

- Intrinsically evaluate phrasal paraphrasing, with test set(s) as described above, and human-rated gold standard (e.g., the first paraphrase that most people suggest for each phrase would be the top rank paraphrase for that phrase in the gold standard).

- Find or construct a sufficiently large, balanced or same-genre monolingual corpus that will help showing that distributional techniques can outperform pivoting techniques.

- To further reduce the dependency on parallel texts, extract translation rules from distributional profiles (DPs) in each language, with a bilingual bridging seed lexicon to measure the semantic distance cross-lingually. So far, work in this approach has concentrated on unigram translations (Fung and Yee, 1998; Rapp, 1999; Diab and Finch, 2000), and has not been evaluated in an end-to-end SMT system. item I believe the notion of semantic reinforcement (evidence from similar paths or rules) has further potential beyond scoring translation rules for unknown phrases. For example, it could be used to reinforce the confidence in automatically learned (standard, non-paraphrastic) translation rules that are similar to one another. Simple "hard" clustering and merging of these rules results in loss of information of the variations encapsulated in the

different rules; however, confidence reinforcement offers benefits of similarity detection with more information retention.

Unified framework:

- The experiments with the hybrid sense-proportional semantic models used an arbitrary weight for interpolating the concept-based and word-based information. However, the models and the unified framework, as presented, allow for optimizing these weights automatically. It would be interesting to see if task-specific optimization, e.g., for SMT, yields significant improvements.

  item The unified framework, described in Chapter 5, suggest incorporating all the above-mentioned linguistic soft constraints in a single SMT model, in the hope of yielding additional gains. Using a formally syntactic (hierarchical) phrase-based SMT system such as Hiero seems a natural choice for this. However, augmenting hierarchical translation rules poses additional challenges, e.g., should rules with gaps ("X") be paraphrased? If so, how?

References

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 16–23, Athens, Greece.

Eneko Agirre and Oier Lopez de Lacalle Lekuona. 2003. Clustering WordNet word senses. In *Proceedings of the 1st International Conference on Recent Advances in Natural Language Processing (RANLP-2003)*, Borovets, Bulgaria.

Masayuki Asahara, Chooi-Ling Goh, Kenta Fukuoka, Yotaro Watanabe, Ai Azuma, Yuji Matsumoto, and Takashi Tsuzuki. 2007. Combination of machine learning methods for optimum chinese word segmentation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

Nguyen Bach, Stephan Vogel, and Colin Cherry. 2009. Cohesive constraints in a beam search phrase-based decoder. In *Proceedings of NAACL HLT 2009: Short Papers*, pages 1–4, Boulder, Colorado.

Satanjeev Banerjee and Ted Pedersen. 2003. Extended Gloss Overlaps as a Measure of Semantic Relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*, pages 805–810, Acapulco, Mexico.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL05)*, pages 597–604, Ann Arbor, Michigan. Association for Computational Linguistics.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of NAACL-HLT*.

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of ACL-2001*.

John R. L. Bernard, editor. 1986. *The Macquarie Thesaurus*. Macquarie Library, Sydney, Australia.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2007. CCG supertags in factored statistical machine translation. In *Proceedings of the ACL Workshop on Statistical Machine Translation 2007*.

Francis Bond, Eric Nichols, Darren Scott Appling, and Michael Paul. 2008. Improving statistical machine translation by paraphrasing the training data. In *Proceedings of IWSLT*, Hawai'i, USA.

Fabiano C. Botelho, Yoshiharu Kohayakawa, and Nivio Ziviani. 2005. A practical minimal perfect hashing method. In *4th International Workshop on Efficient and Experimental Algorithms (WEA05)*.

Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–313.

Raluca Budiu, Christiaan Royer, and Peter Pirolli. 2006. Modeling information scent: A comparison of lsa, pmi and glsa similarity measures on common tests and corpora. In *Proceedings of RIAO'07*, Pittsburgh, PA.

Lou Burnard. 2000. *Reference Guide for the British National Corpus.* Oxford University Computing Services, Oxford, England, world edition edition.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings NAACL-2006.*

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of EMNLP 2008*, Waikiki, Hawai'i.

Marine Carpuat and Dekai Wu. 2007. How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, pages 43–52, Skövde, Sweden.

Eugene Charniak, Kevin Knight, and Kenji Yamada. 2003. Syntax-based language models for machine translation. In *Proceedings of MT Summit IX.*

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Tech. Report TR-10-98, Comp. Sci. Group, Harvard U.

Colin Cherry and Dekang Lin. 2006. Soft syntactic constraints for word alignment through discriminative training. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, Association for Computational Linguistics*, Sydney, Australia.

Colin Cherry. 2008. Cohesive phrase-based decoding for statistical machine translation. In *Proceedings of ACL-08: HLT*, pages 72–80, Columbus, Ohio, USA.

David Chiang, Adam Lopez, Nitin Madnani, Christof Monz, Philip Resnik, and Michael Subotin. 2005. The hiero machine translation system: Extensions, evaluation, and analysis. In *Proceedings of HLT/EMNLP*, pages 779–786.

David Chiang, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Procedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, Waikiki, Honolulu, Hawaii.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL-05*, pages 263–270.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL-05*.

Brooke Cowan, Ivona Kucerova, and Michael Collins. 2006. A discriminative model for tree-to-tree translation. In *Proc. EMNLP*.

Koby Crammer and Yoram Singer. 2003. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–991.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

D. Allen Cruse. 1986. *Lexical semantics*. Cambridge University Press, Cambridge, UK.

James R. Curran. 2004. *From Distributional to Semantic Similarity*. Ph.D. thesis, School of Informatics, University of Edinburgh, Edinburgh, UK.

Ido Dagan, Lillian Lee, and Fernando Pereira. 1999. Similarity-based models of cooccurrence probabilities. *Machine Learning*, 34(1–3):43–69.

Vera Demberg. 2007. A language-independent unsupervised model for morphological segmentation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

Steve DeNeefe and Kevin Knight. 2009. Synchronous tree adjoining machine translation. In *Proceedings of ACL*, Singapore.

Steve DeNeefe, Kevin Knight, Wei Wang, and Daniel Marcu. 2007. What can syntax-based MT learn from phrase-based MT? In *Proceedings of EMNLP-CoNLL*.

Mona Diab and Steve Finch. 2000. A statistical word-level translation model for comparable corpora. In *Proceedings of the Conference on Content-Based Multimedia Information Access (RIAO)*.

Mona Diab, Kadri Hacioglu, and Daniel Jurafsky. 2004. Automatic tagging of Arabic text: From raw text to base phrase chunks. In *Proceedings of HLT/NAACL*, volume Companion volume, pages 149–152.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics of the Association for Computational Linguistics*, Geneva, Switzerland.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing word lattice translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Columbus, Ohio.

Jason Eisner. 2003. Learning non-isomorphic tree mappings for machine translation. In *ACL Companion Vol.*

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2086)*, pages 897–906, Honolulu, HI.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

John R. Firth. 1957. A synopsis of linguistic theory 1930–55. *Studies in Linguistic Analysis*, (special volume of the Philological Society):1–32. Distributional Hypothesis.

Heidi Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proc. EMNLP 2002*.

Pascale Fung and Percy Cheung. 2004. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1051, Geneva, Switzerland. Association for Computational Linguistics.

Pascale Fung and Lo Yuen Yee. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of COLING-ACL98*, pages 414–420, Montreal, Canada.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of COLING/ACL-06*.

Daniel Gildea. 2003. Loosely tree-based alignment for machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 80–87.

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of EMNLP*.

Iryna Gurevych and Michael Strube. 2004. Semantic similarity applied to spoken dialogue summarization. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 764–770, Geneva, Switzerland.

Nizar Habash and Jun Hu. 2009. Improving arabic-chinese statistical machine translation using english as pivot language. In *Proceedings of the 4th EACL Workshop on Statistical Machine Translation*, pages 173–181, Athens, Greece.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, pages 771–779, Columbus, Ohio, USA.

Greg Hanneman and Alon Lavie. 2009. Decoding with syntactic and non-syntactic phrases in a syntax-based machine translation system. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation at the 2009 Meeting of the North-American Chapter of the Association for Computational Linguistics (NAACL-HLT-2009)*, Boulder, CO.

Zellig S. Harris. 1940. Review of louis h. gray, foundations of language (new york: Macmillan, 1939). *Language*, 16(3):216–231.

Zellig Harris. 1954. Distributional structure. *Word*, 10 (2)(3):146–162.

Hany Hassan, Khalil Sima'an, and Andy Way. 2007. Integrating supertags into phrase-based statistical machine translation. In *Proc. ACL-07*, pages 288–295.

Graeme Hirst and Alexander Budanitsky. 2005. Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, 11(1):87–111.

Chu-Ren Huang, Petr Simon, Shu-Kai Hsieh, and Laurent Prevot. 2007. Rethinking chinese word segmntation: Tokenization, character classification, or wordbreak identification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

Zhongqiang Huang, Denis Filimonov, and Mary Harper. 2008. Accuracy enhancements for mandarin parsing. Tech. report, University of Maryland.

Mario Jarmasz and Stan Szpakowicz. 2003. Roget's Thesaurus and semantic similarity,. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003),*, pages 212–219, Borovets, Bulgaria.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research on Computational Linguistics (ROCLING X)*, Taiwan.

Dan Klein and Christopher D. Manning. 2003a. Accurate unlexicalized parsing. In *Proceedings of ACL-03*, pages 423–430.

Dan Klein and Christopher D. Manning. 2003b. Fast exact inference with a factored model for natural language parsing. *Advances in Neural Information Processing Systems*, 15(NIPS 2002):3–10.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proc. EMNLP+CoNLL*, pages 868–876, Prague.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of EACL*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL*, pages 127–133.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic.

Philipp Koehn. 2003. *Noun Phrase Translation.* Phd thesis, University of Southern California.

Philipp Koehn. 2004a. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA*.

Philipp Koehn. 2004b. Statistical significance tests for machine translation evaluation. In *Proc. EMNLP*.

Philipp Koehn. 2005. A parallel corpus for statistical machine translation. In *Proceedings of MT-Summit*.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25:259 – 284.

Joon Ho Lee, Myoung Ho Kim, and Yoon Joon Lee. 1993. Information retrieval based on conceptual distance in IS-A hierarchies. *Journal of Documentation*, 49(2):188–207.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26, Toronto, Canada.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360.

Dekang Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 64–71, Madrid, Spain.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, San Francisco, CA.

Jimmy Lin. 2008. Scalable language processing algorithms for the masses: A case study in computing word co-occurrence matrices with mapreduce. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, Honolulu, Hawaii.

Fernando López-Ostenero, Julio Gonzalo, and Felisa Verdejo. 2005. Noun phrases as building blocks for cross-language search assistance. *Information Processing and Management*, 41:549–568.

Adam Lopez. 2007. Hierarchical phrase-based translation with suffix arrays. In *Proceedings of EMNLP-CoNLL*, pages 976–985.

Adam Lopez. 2008a. *Machine Translation by Pattern Matching.* Ph.d. dissertation, University of Maryland College Park.

Adam Lopez. 2008b. Statistical machine translation. *ACM Computing Surveys*, 40(3):1–49. Earlier version: A Survey of Statistical Machine Translation. U. of Maryland, UMIACS tech. report 2006-47. Apr 2007.

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie Dorr. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the ACL Workshop on Statistical Machine Translation.*

Nitin Madnani. to appear. A survey of automatic paraphrase generation.

Christopher D. Manning and Hinrich Schütze. 1999. Foundations of statistical natural language processing. page 304. MIT Press, Cambridge, MA.

Daniel Marcu, Wei Wang, Abdessamad Echihabi, and Kevin Knight. 2006. SPMT: Statistical machine translation with syntactified target language phrases. In *Proc. EMNLP*, pages 44–52.

Yuval Marton and Philip Resnik. 2008. Soft syntactic constraints for hierarchical phrased-based translation. In *Proceedings of ACL-08: HLT*, pages 1003–1011, Columbus, Ohio, USA.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009a. Improved statistical machine translation using monolingually-derived paraphrases, with. In *Procedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore.

Yuval Marton, Saif Mohammad, and Philip Resnik. 2009b. Estimating semantic distance using soft semantic constraints in knowledge-source / corpus hybrid models. In *Procedings of EMNLP*, Singapore.

Aurelien Max. 2009. Sub-sentential paraphrasing by contextual pivot translation. In *Proceedings of the 2009 Workshop on Applied Textual Inference (ACL-IJCNLP)*, pages 18–26, Singapore. Suntec.

Diana McCarthy. 2006. Relating WordNet senses for word sense disambiguation. In *Proceedings of the European Chapter of the Association for Computational Linguistics Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, pages 17–24, Trento, Italy.

Scott McDonald. 2000. *Environmental determinants of lexical processing effort.* Ph.D. thesis, University of Edinburgh.

Coskun Mermer, Hamza Kaya, and Mehmet Ugur Dogan. 2007. The tÜb tak-uekae statistical machine translation system. In *IWSLT 2007.*

Haitao Mi, Liang Huang, and Qun Liu. 2008. Forest-based translation. In *Proceedings of ACL-08: HLT*, pages 192–199, Columbus, Ohio, USA.

Saif Mohammad and Graeme Hirst. 2006. Distributional measures of concept-distance: A task-oriented evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, Sydney, Australia.

Saif Mohammad, Iryna Gurevych, Graeme Hirst, and Torsten Zesch. 2007. Cross-lingual distributional profiles of concepts for measuring semantic distance. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL-2007)*, Prague, Czech Republic.

Saif Mohammad. 2008. *Measuring Semantic Distance using Distributional Profiles of Concepts*. Ph.D. thesis, Department of Computer Science, University of Toronto, Toronto, Canada.

Jane Morris and Graeme Hirst. 2004. Non-classical lexical semantic relations. In *Proceedings of the Workshop on Computational Lexical Semantics, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 46–51, Boston, Massachusetts.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Roberto Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association*, pages 105–112, Sydney, Australia.

Rebecca Nesson, Stuart M. Shieber, and Alexander Rush. 2006. Induction of probabilistic synchronous tree-insertion grammars for machine translation. In *Proceedings of AMTA*.

Sonja Nissen and Hermann Ney. 2004. Statistical machine translation with scarce resources using morpho-syntatic analysis. *Computational Linguistics*, 30(2):181–204.

Doug Oard, David Doermann, Bonnie Dorr, Daqing He, Phillip Resnik, William Byrne, Sanjeeve Khudanpur, David Yarowsky, Anton Leuski, Philipp Koehn, and Kevin Knight. 2003. Desperately seeking cebuano. In *Proceedings of HLT-NAACL.*

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the ACL*, pages 440–447.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL.*

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the ACL*, pages 160–167.

Marian Olteanu, Pasin Suriyentrakorn, and Dan Moldovan. 2006. Language models and reranking for machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 150–153, New York City, NY.

Karolina Owczarzak, Bart Mellebeek, Declan Groves, Josef Van Genabith, and Andy Way. 2006. Wrapper syntax for example-based machine translation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 148–155.

Kishore Papineni, Salim Roukos, Todd Ward, John Henderson, and Florence Reeder. 2002. Corpusbased comprehensive and diagnostic MT evaluation: Initial Arabic, Chinese, French, and Spanish results. In *Proceedings of the ACL Human Language Technology Conference*, pages 124–127, San Diego, CA.

Siddharth Patwardhan and Ted Pedersen. 2006. Using WordNet based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of Making Sense of Sense EACL Workshop*, pages 1–8.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2006)*, pages 192–199, New York, NY.

Chris Quirk and Arul Menezes. 2006. Dependency treelet translation: The convergence of statistical and example-based machine translation? *Machine Translation*, 20:43–65.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal smt. In *Proceedings of the ACL*.

Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. 1989. Development and application of a metric on semantic nets. *IEEE Transaction on Systems, Man, and Cybernetics*, 19(1):17–30.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th Annual Conference of the Association for Computational Linguistics.*, pages 519–525.

Philip Resnik and Mona Diab. 2000. Measuring verb similarity. In *Twenty Second Annual Meeting of the Cognitive Science Society (COGSCI2000)*, Philadelphia.

Philip Resnik and Noah Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Philip Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research (JAIR)*, 11:95–130.

Stefan Riezler and John Maxwell. 2006. Grammatical machine translation. In *Proc. HLT-NAACL*, New York, NY.

Sheldon Ross. 1976. *A First Course in Probability*. Macmillan.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Salton and McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

Frank Schilder and Bridget Thomson McInnnes. 2006. Word and tree-based similarities for textual entailment. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, pages 140–145, Venice, Italy.

Josh Schroeder, Trevor Cohn, and Philipp Koehn. 2009. Word lattices for multi-source translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL.*, pages 719–727, Athens, Greece.

Hinrich Schuetze and Jan O. Pedersen. 1997. A cooccurrence-based thesaurus and two applications to information retreival. *Information Processing and Management*, 33(3):307–318.

Hendra Setiawan, Min-Yen Kan, and Haizhou Li. 2007. Ordering phrases with function words. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 712–719.

Hendra Setiawan, Min Yen Kan, Haizhou Li, and Philip Resnik. 2009. Topological ordering of function words in hierarchical phrase-based translation. In *Proceedings of the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, Singapore.

Stuart Shieber and Yves Schabes. 1990. Synchronous tree-adjoining grammars. In *Proceedings of the 13th International Conference on Computational Linguistics*.

David A. Smith and Jason Eisner. 2006. Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies. In *Proceedings of the Workshop on Statistical Machine Translation, Association for Computational Linguistics*, pages 23–30, New York City, NY.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901–904.

David Talbot and Thorsten Brants. 2008. Randomized language models via perfect hash functions. In *Proc. ACL-08: HLT*.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter. In *Fourth SIGHAN Workshop on Chinese Language Processing*.

Giannis Varelas, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides G.M. Petrakis, and Evangelos E. Milios. 2005. Semantic similarity methods in WordNet and their application to information retrieval on the web. In *Proceedings of the 7th Annual Association of Computing Machinery International Workshop on Web Information and Data Management*, pages 10–16, Bremen, Germany.

Ashish Venugopal, Andreas Zollmann, Noah Smith, and Stephan Vogel. 2009. Preference grammars softening syntactic constraints to improve statistical machine translation. In *Proceedings of NAACL*, Boulder, CO.

David Vilar, Jan-T. Peter, and Hermann Ney. 2007. Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Chao Wang, Michael Collins, and Phillip Koehn. 2007a. Chinese syntactic reordering for statistical machine translation. In *Proceedings of EMNLP*.

Wei Wang, Kevin Knight, and Daniel Marcu. 2007b. Binarizing syntax trees to improve syntax-based machine translation accuracy. In *Proc. EMNLP+CoNLL 2007*.

Taro Watanabe, Jun Suzuki, Hajime Tsukuda, and Hideki Isozaki. 2007. Online large-margin training for statistical machine translation. In *Proc. EMNLP*.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*, pages 1015–1021, Geneva, Switzerland.

Hua Wu and Haifeng Wang. 2008. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.

Hua Wu and Ming Zhou. 2003. Synonymous collocation extraction using translation information. In *Proceedings of the ACL Workshop on Multiword Expressions: Integrating Processing*, pages 120–127.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23:377–404.

Deyi Xiong, Min Zhang, Aiti Aw, and Haizhou Li. 2009. A syntax-driven bracketing model for phrase-based translation. In *Proceedings of the 47th Annual Conference of the Association for Computational Linguistics (ACL)*, Singapore.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting of the ACL*, pages 523–530.

Min Zhang, Hongfei Jiang, Aiti Aw, Haizhou Li, Chew Lim Tan, and Sheng Li2. 2008. A tree sequence alignment-based tree-to-tree translation model. In *Proceedings of ACL-08: HLT*, pages 559–567, Columbus, Ohio, USA.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings of the SMT Workshop, HLT-NAACL*.