Distributional Phrasal Paraphrase Generation for Statistical Machine Translation

YUVAL MARTON, University of Maryland, Columbia University, and IBM T.J. Watson Research Center

Paraphrase generation has been shown useful for various natural language processing tasks, including statistical machine translation. A commonly used method for paraphrase generation is pivoting [Callison-Burch et al. 2006], which benefits from linguistic knowledge implicit in the sentence alignment of parallel texts, but has limited applicability due to its reliance on parallel texts. Distributional paraphrasing [Marton et al. 2009a] has wider applicability, is more language independent, but doesn't benefit from any linguistic knowledge. Nevertheless, we show that using distributional paraphrasing can yield greater gains in translation tasks. We report method improvements leading to higher gains than previously published, of almost 2 BLEU points, and provide implementation details, complexity analysis, and further insight into this method.

Categories and Subject Descriptors: I.2.7 [**Natural Language Processing**]: Machine translation/Language generation—*Paraphrasing*

General Terms: Design, Algorithms, Performance

Additional Key Words and Phrases: Semantic similarity, semantic distance, paraphrase generation, statistical machine translation, SMT

ACM Reference Format:
Marton, Y. 2013. Distributional phrasal paraphrase generation for statistical machine translation. ACM Trans. Intell. Syst. Technol. 4, 3, Article 39 (June 2013), 32 pages.
DOI: http://dx.doi.org/10.1145/2483669.2483672

1. INTRODUCTION

Paraphrase generation, or *paraphrasing*, is defined as finding alternative phrasing (words, phrases, sentences, etc.) to convey the same—or nearly the same—meaning as that of a given word or phrasing. Hereafter the paraphrased phrasing is called the *focal* phrasing. For example, given the focal word *deal*, good paraphrases might be *agreement*, or *accord*, but not *sky*; given the focal phrase *to provide any*, good paraphrases might be *to give any*, or *to give further*, but not *yellow submarine*. Paraphrase generation serves various Natural Language Processing (NLP) applications, such as Natural Language Generation (NLG), summarization, Information Retrieval (IR), Question Answering (QA), and Statistical Machine Translation (SMT) [Madnani and Dorr 2010; Androutsopoulos and Malakasiotis 2010]. This work focuses on paraphrasing for SMT, for which it is useful because it increases translation coverage. Untranslated words and phrases, and bad reordering of known words and phrases in unseen larger sequences, remain a major problem [Callison-Burch et al. 2006] for phrase-based SMT systems, flat and hierarchical alike [Koehn et al. 2003, 2007; Koehn 2004a; Chiang 2005, 2007, inter alia], in spite of much progress since statistical word-based translation models were

Author's address: Y. Marton, IBM/Thomas J. Watson Research Center, Building 801 Office 23-140, Route 134/1101 Kitchawan Road, Yorktown Heights, NY 10598; email: yuvalmarton@gmail.com.

DOI: http://dx.doi.org/10.1145/2483669.2483672

introduced [Brown et al. 1993]. This article is based on parts of the author's doctoral dissertation and past publications [Marton 2009, 2010; Marton et al. 2009a]. It extends this previous work with technical improvements, leading to further gains in translation quality, and with implementation details, complexity analysis, optimization issues, and a new set of experiments aimed to explore various ways of incorporating paraphrases in SMT phrase tables.

According to Callison-Burch et al. [2006], an 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 than unigram coverage. 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], METEOR [Banerjee and Lavie 2005], TER [Snover et al. 2006], and human judgment tests such as HTER [Snover et al. 2006; Specia and Farzindar 2010]. Reducing the Out-Of-Vocabulary (OOV) word and phrase rate is therefore important for SMT systems.

Recent work proposed augmenting the training data with paraphrases generated by pivoting through other languages [Bannard and Callison-Burch 2005; Callison-Burch et al. 2006; Callison-Burch 2008]. This indeed alleviates the vocabulary coverage problem, especially for the resource-poor, so-called "low density" languages. However, these approaches require additional parallel texts (or translation tables) where one side contains the original source language. Such parallel texts are uncommon, with the notable exception of the EuroParl corpus [Koehn 2005]. (Moreover, Callison-Burch [2008] also requires parsing information). Most other recent methods require supervised training (see Section 3), resources for which are also scarce in "low density" languages.

To overcome this resource constraint, we subsequently proposed to augment the training data with paraphrases generated by using distributional methods on a large monolingual corpus, a relatively abundant resource [Marton et al. 2009a]. The method constructs monolingual Distributional Profiles (DPs; see Section 2.2) of out-of-vocabulary words and phrases in the source language. It then generates paraphrase candidates from phrases that cooccur in same contexts, and ranks them with semantic distance measures, using cosine of vectors containing log-likelihood ratios. The highest ranking paraphrases are used to augment the translation phrase table (Section 4). However, this approach lacks the human linguistic knowledge that is implicit in the sentence alignment of the parallel texts. Therefore, it is unclear a priori which approach should yield higher gains in translation quality.

The base distributional paraphrasing method was reimplemented or extended in various ways (e.g., Mirkin et al. [2009], Marton [2010], and Marton et al. [2011]). Here we concentrate on providing more implementation details of the base method, mainly the use of a suffix array, augmented with a prefix tree with suffix links [Manber and Myers 1993; Lopez 2007, 2008], and an analysis of the paraphrasing algorithm complexity when using this data structure. We also report further gains in translation quality as a result of limiting the paraphrase length to be in the vicinity of the length of the paraphrased phrase.

In the rest of this article we Describe Distributional Profiles (DPs) and semantic distance measures in Section 2, paraphrasing methods (including Marton et al. [2009a] with its algorithmic complexity) in Section 3, and translation model augmentation methods in Section 4. We report our experiments and results in Section 5, and conclude by discussing the implications and future research directions in Section 6. Since this article brings together various subfields, we discuss related work in each of the relevant sections.

2. SEMANTIC DISTANCE MEASURES

Semantic distance measures aim to detect words (or morphemes) with similar and/or related meaning. Semantic distance measures of larger units—bigrams, phrases, sentences, passages, documents—have been developed too, as an extension or combination of the word-level measures. We use semantic distance measures to recognize paraphrases (see Section 3).[1] These measures are grouped here as follows: lexical resource based, corpus based, and hybrid.

2.1. Measures Based on a Lexical Resource

WordNet [Fellbaum 1998] is a manually created taxonomy,[2] where each node 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 they are 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 2.3). The distance between nodes can be the number of arcs in the connecting path, or it 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, these measures can only be used in languages for which a (sufficiently developed) WordNet exists. 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] introduces a WSD method which relies on maximal word overlap of each of the word's senses' glosses (dictionary definitions) with the glosses of the surrounding words: 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] generalize this approach to a semantic relatedness measure, based on word overlap in the glosses of two words of interest. Hashimoto et al. [2011] scale this approach to paraphrase extraction from the World Wide Web.

2.2. Corpus-Based Measures (Distributional Profiles)

Corpus-based measures of distributional similarity rely on the distributional hypothesis [Harris 1954; Firth 1957]: Words tend to have a typical Distributional Profile (DP); they repeatedly appear next to specific other words in a typical rate of cooccurrence. Moreover, words close in meaning tend to appear in similar surrounding contexts, where *context* is taken to be the surrounding words in some proximity, typically a fixed size sliding window around the word's occurrences. The DP of word u is a feature vector whose dimensions are the surrounding context words (*collocates*), and the values represent Strength-of-Association (SoA) between u and each collocate.[3] Beside simple

[1]For other paraphrase recognition methods, see Androutsopoulos and Malakasiotis [2010].

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

[3]The dimensions of the DP cooccurrence 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].

Table I. Numerical Example of a Distributional Profile (DP) for Word *cord*

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

The DP vector's cells contain cooccurrence counts of *cord* with each of the other words in the vocabulary (see middle column), or a more sophisticated strength-of-association measure such as log-likelihood ratio (rightmost column), incorporating also the frequency of each word separately, and how often it cooccurs with other words.

cooccurrence 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 [Gale and Church 1991], and the log-likelihood ratio [Dunning 1993]. An example DP for the word *cord* is given in Table I.

Profile similarity measures. A DP similarity function $psim(DP_u, DP_v)$ is typically defined as a two-place function, taking vectors as arguments (the DP of some word/phrase u and the DP of some word/phrase v) whose size is the known vocabulary size. These vectors' cell i contains 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 analogs. Similarity can be estimated in several ways, for example, 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 cosine is especially appealing. It is competitive, 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.[4] 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. The formula for the cosine function for similarity measure is given in Eq. (1).

$$psim(DP_u, DP_v) = cos(DP_u, DP_v) = \frac{\sum_{w_i \in V} SoA(u, w_i) SoA(v, w_i)}{\sqrt{\sum_{w_i \in V} SoA(u, w_i)^2} \sqrt{\sum_{w_i \in V} SoA(v, w_i)^2}} \quad (1)$$

In principle, any SoA measure can be used with any profile similarity measure, but in practice only some combinations do well; and finding the 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, in Manning and Schtze [1999]). Other measures are directional (textual entailment) in u and v [Kotlerman et al. 2009]. See Weeds et al. [2004], Curran [2004], Mohammad [2008], and Turney and Pantel [2010] for surveys of distributional measures.

2.3. Hybrid Measures

Resnik [1999] introduced a hybrid model for calculating "information content" by traversing the concept's subtree in WordNet. This measure is hybrid in that it uses

[4]To be precise, their smallest angle, 0–90°, ignoring vector directionality.

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 this subsumer are likely to be more semantically similar than those that are far below it in the WordNet hierarchy.

A corpus-based DP of a word u conflates information about the potentially many senses of u [Mohammad and Hirst 2006]. For example, assume the noun *bank* has two senses WATER/RIVER (as in *riverbank*) and FINANCIAL INSTITUTION, and the noun *wave* has senses WATER/RIVER and PHYSICS/ENERGY. Thus the distributional distance between *bank* and *wave* will be some average of the semantic distance between all their senses. However, for various NLP tasks, what is often needed is the distance between their closest senses, in this case, the WATER/RIVER senses. 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. Mohammad and Hirst [2006] generate separate DPs for the different senses of a word by using the categories in a Roget-style thesaurus as coarse senses or concepts: A word may be found in more than one category if it has multiple meaning. They use a simple unsupervised algorithm to determine concept-based DP(c): a vector containing the SoA between the category c and each of the words w_i in a corpus vocabulary, based on the number of times w_i occurred next to an instance of c in the corpus. We observed that if words u and v appear under the same concept c, the semantic distance between u and v would be indistinguishable (i.e., zero distance), since the concept-based similarity measure returns the semantic distance of the closest sense pair (here: $sim(DP(c), DP(c))$). Therefore, we adopted in Marton et al. [2009b] a hybrid approach with fine-grained soft constraints, discounting cooccurrence counts according to the counts in the concept-based DP of the desired word sense, and then calculating SoA measures over the discounted counts: $f(u_c, w_i) = p(c|w_i) \times f(u, w_i)$ where the conditional probability $p(c|w_i)$ is calculated from the cooccurrence frequencies in the concept-based DPs, and the cooccurrence count $f(u, w_i)$ is calculated from word-based DPs. We also extended this method there to handle any word in the corpus, even if it didn't have a concept-based DP. We do not address word senses here.

Erk and Padó [2008] represent a word sense in context by biasing the word's DP according to the context surrounding a specific occurrence of that word. 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, in a low-density language.

The lexical-resource-based and hybrid semantic distance measures rely on resources that might not exist, or not be sufficiently developed, in a resource-poor "low density" language. Therefore, their applicability is limited; or, in the best case, their quality would degrade to corpus-based measures' level. In the rest of this article we use corpus-based measures, and concentrate on the distributional paraphrasing method. Note, however, that this method may be used with hybrid semantic distance measures as well [Marton 2009, 2010].

3. PARAPHRASE GENERATION

3.1. Paraphrasing Approaches

Paraphrasing is the act of replacing linguistic utterances (typically text) with other linguistic utterances, bearing similar meaning but different form. Hereafter the paraphrased text to be replaced is called the *focal* word or text, or simply, *the focal*.

Paraphrasing research is quite diverse, and can be characterized by many axes, including the following.

Focal unit. This is a word, "phrase" (any word sequence), syntactic constituent, "gappy phrase" a.k.a. "pattern" (as in *X gave Y to Z*), paragraph, sentence, document, etc.

Paraphrased elements. These can be lexical (different words with similar meaning), structural (e.g., switching between active and passive voice), or both.

Use of linguistic knowledge. This can include refer to syntactic information (parses; refer to Lin [1997], Callison-Burch [2008], and Das and Smith [2009]), semantic information (WordNet hierarchy, thesaurus concepts, or sense-annotated corpora as in Shutova [2010]), and/or morphological analysis [Nakov and Ng 2011], or none.

Resource type. This means; refer to sentence-aligned parallel text (a.k.a. bitext), comparable text (e.g., global news from different sources covering the same period, refer to Munteanu and Marcu [2005]), stand-alone text (one monolithic corpus), or crowdsourcing (refer to recently, Resnik et al. [2010]).

Paraphrasing method. These include SMT (translating from and to the same language; Barzilay and McKeown [2001]), pivoting (translating to other languages and back [Bannard and Callison-Burch 2005; Callison-Burch 2008; Madnani et al. 2007]), distributional (relying on similar contexts in which the paraphrases tend to occur [Marton et al. 2009a; Marton 2010]), morphological and character-based analysis (compounds, edit distance), or other (e.g., correlating time-locked bursts of terms such as *earthquake* in several resources).[5]

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

Input and output languages. These can be monolingual (typically), or multilingual (if translation is viewed as an instance of paraphrasing [Fung and Yee 1998; Rapp 1999; Diab and Finch 2000; Haghighi et al. 2008]).

Paraphrasing may be somewhat "lossy" in number of words and/or content, with the extreme cases of summarization and translation. This work uses monolingual, standalone text in order to generate (nongappy) phrasal paraphrases with distributional methods, capable of incorporating lexical resources, and extensible to using syntactic information as well. Out Of Vocabulary (OOV) phrases in the source language are paraphrased and then used to augment an SMT translation model (details are given in Sections 4). Due to space limitations, we only mention here by name the most similar and recent work. For more information on kinds of paraphrasing, see Madnani and Dorr [2010] and Androutsopoulos and Malakasiotis [2010].

Moving along the paraphrasing method axis, Barzilay and McKeown [2001] use direct translation in order to generate paraphrases, in contrast with our method. They extract paraphrases from a monolingual parallel corpus, containing multiple translations of the same source. Their method is limited by the small size of monolingual parallel corpora, if such exist, which are even scarcer than bilingual parallel resources, which are used in the pivoting method described next.

A leading SMT-related paraphrasing method currently is the *pivoting* approach, especially Bannard and Callison-Burch [2005], Callison-Burch et al. [2006], and Callison-Burch [2008], with the latter using syntactic information. "Pivoting" means translating the focal phrases to additional language(s) and back to the source language. The quality of these paraphrases is estimated by marginalizing translation probabilities to and from the additional language side (or sides) e, as follows: $p(f_2|f_1) \approx \sum_e p(e|f_1)p(f_2|e)$,

[5]See also Oard [1997] and the workshops on comparable corpora http://comparable.limsi.fr/, *inter alia*.

where f_1 and f_2 are the phrase and its paraphrase candidate, respectively. A major disadvantage of the approach is that it relies on the availability of parallel corpora in other languages. While this works for English and many European languages, it is far less likely to help when translating from other languages, for which bitexts are scarce or nonexistent. Also, pivoting is inherently noisy in both the paraphrase candidates' correct sense and their translational likelihood, because of the double translation step. (More on that in Section 6). The problem of incorrect sense translation is likely to be exacerbated with out-of-domain translation, that is, when the test set is of a different genre than the bitexts. One of its advantages, however, is the use of linguistic knowledge that is encapsulated in the parallel sentence alignment. Still, we argue in Marton et al. [2009a] that the ability to use much larger resources for paraphrasing should trump or at least match the human knowledge advantage. Moreover, as mentioned in Section 2.3, some human linguistic knowledge can be harnessed from readily available lexical resources, which group words according to common semantic properties. This information can be used to approximate word senses, and model word meaning more accurately, still without relying on parallel text.

Zhao et al. [2008] apply SMT decoding for paraphrasing, using several log-linear weighted resources (phrase table, thesaurus, etc.), while Zhao et al. [2009] filter out paraphrase candidates and weight paraphrase features according to the desired NLP task: sentence compression, simplification, or similarity computation. Malakasiotis [2009] propose paraphrase recognition using machine learning techniques to combine similarity measures. Chevelu et al. [2009] introduce a new paraphrase generation tool based on Monte-Carlo sampling. Mirkin et al. [2009], inter alia, frame paraphrasing as a special, symmetrical case of (WordNet-based) textual entailment.

3.2. Monolingually Derived Distributional Paraphrase Generation

Distributional paraphrase generation relies on the distributional hypothesis much like distributional distance measures do (see Section 2.2). The key idea is that good paraphrases are likely to be found in same (or similar) contexts as the focal word/phrase.

The DIRT system [Lin and Pantel 2001] uses unigram contexts and dependency parsing paths for paraphrasing relations such as "*X* wrote *Y*". Paraphrase candidate gathering is done with a hash table containing for each word in the vocabulary all the paths it serves as *X* or *Y*. Context overlap ratio between paths is used for candidate ranking. This approach requires parsing the entire monolingual text resource.

Pasca and Dienes [2005] use slightly longer, but still predefined context length (2–3 tokens). They find it useful to use contexts with specific templates containing Named Entities (NE) or relative clauses. They discard paraphrase candidates that are clauses, with pronouns, having no verbs, or having only function words, or merely showing upper/lower case alternation from the focal term. Paraphrase candidate gathering is done similarly (via shared contexts). The frequency of the candidate in the same context(s) is used for candidate ranking. This approach requires POS and NE tagging of the entire resource.

Bhagat and Ravichandran [2008] also use POS tagging, and are interested in paraphrasing relations "*X* rel *Y*". They start with a seed set representing each relation, then download top 1000 Google search hits to extend it. They find and rank candidates using nearest neighbors with locality-sensitive hashing (cosine-preserving fingerprint of the DP, so that the cosine can be computed fast).

Like other distributional methods, our paraphrasing method requires a large monolingual corpus of text in the source ("from") language, a relatively abundant resource. This resource is used both for generating paraphrase candidates and for ranking them according to their semantic distance from the focal (paraphrased) phrase. We use

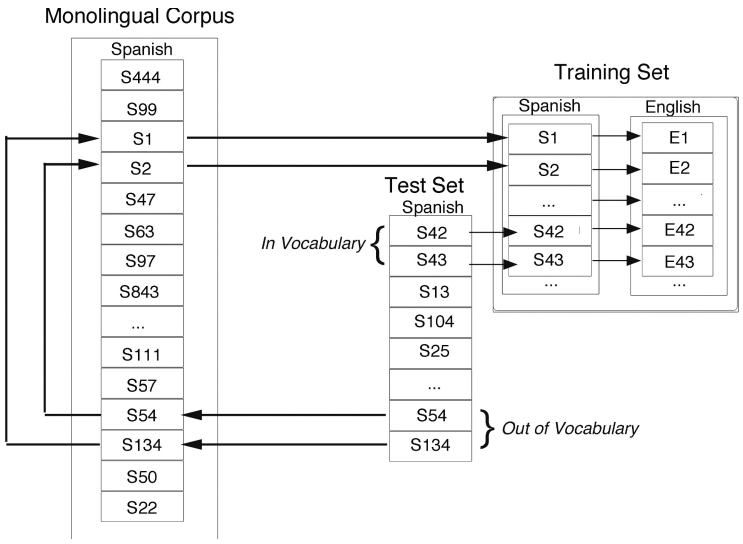

Fig. 1. Monolingual corpus-based distributional paraphrase generation. For a Spanish-to-English translation model, which encounters unknown source language (Spanish) phrases, augment the model by generating distributional paraphrases. This requires a large monolingual corpus, which is a relatively abundant 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 DP of the unknown phrases.

contexts of dynamic length, without relying on POS tagging or parsing (although these can be added in principle), and we calculate vectors (DPs) on-the-fly (so no pruning or huge matrix precalculation needs to take place).

The outline of our method is as follows.

(1) Upon receiving focal phrase *phr*, build distributional profile DP_{phr}.
(2) For each occurrence of *phr*, keep the surrounding (left and right) context $L\_\_R$.
(3) For each such context $L\_\_R$, gather all paraphrase candidates *cand*, such that $L\ cand\ R$ occurs in the training corpus, that is, gather paraphrase candidates occurring in same contexts as *phr*.
(4) Rank all candidates *cand* according to their semantic distance from *phr* by building profile DP_{cand} and measuring profile similarity between DP_{cand} and DP_{phr}.
(5) Optionally: Filter out every candidate *cand* that textually entails *phr*.
(6) Output up to k-best candidates above a certain similarity score threshold.

The input phrases used to evaluate our method are source language phrases unknown to an SMT model (out-of-vocabulary phrases). This is illustrated in Figure 1.

3.2.1. Build Phrasal Profile DP_{phr}. Build a distributional profile of the OOV phrase *phr*, enlisting all collocating words, and their cooccurrence count or strength-of-association with *phr*, as described in Section 2.2. The cooccurrence 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 s occurrences), uniformly sample only s occurrences, multiplying the gathered cocounts by factor of $count(phr)/s$. So if *phr* occurs 30,000 times and the threshold is $s = 10000$,

Table II. Left Half: Example Contexts of the Focal Phrase *to provide any other*, Gathered from Training Set Sentences such as "she declined *to provide any other* information . . . ", and "police refused *to provide any other* details . . . " Right Half: Example of Gathered Paraphrase Candidates for the Same Phrase, Appearing in Identical Contexts

left context	focal phrase	right context	left context	candidate *cand*	right context
. . . declined	to provide any other	details . . .	. . . declined	to give further	details . . .
. . . refused	to provide any other	information . . .	. . . refused	to provide any	information . . .
. . . unable	to provide any other	details . . .	. . . unable	to reveal any	details . . .
. . . failed	to provide any other	explanation . . .	. . . failed	to provide further	explanation . . .
. . .	to provide any other	. . .	. . .	. . .	. . .

than count cooccurring words in a sliding window around only every third occurrence of *phr*, but multiply these cooccurrence counts by 3.

3.2.2. Gather Context. Example contexts are shown in the left half of Table II. The challenge in deciding how much context to keep to the left and right of each occurrence of *phr* is a familiar recall-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, we use the following heuristics. Start small: Start with setting the left context L to be a single word/token to the left of phrase *phr*. If $count(L) > mcc$, where *mcc* is a maximal context count limit, append the next word to the left (now having a bigram left context instead of a unigram), and repeat until the condition is met.[6] Repeat similarly for R, the context to the right of *phr*. Add the resulting L_R context to a context list.

3.2.3. Gather Candidates. For each gathered context in the context list, gather all paraphrase candidate phrases *cand* that connect left-hand side context L with right-hand side context R, that is, gather all *cand* such that the sequence $L\ cand\ R$ occurs in the corpus. Example candidates, appearing in same contexts as the focal phrase *phr*, are shown in the right half of Table II. In practice, to keep search complexity low, limit *cand* to be up to length $MaxPhraseLen$.

3.2.4. Rank Candidates. For each candidate *cand*, build distributional profile DP_{cand} and evaluate $psim(DP_{phr}, DP_{cand})$ as in Section 2.2. Recall that since the DP is represented as a vector, any vector similarity function can be used here, for example, cosine.

3.2.5. Textual Entailment Filtering (Optional). Filter out every candidate *cand* that textually entails *phr*: This is approximated by filtering *cand* if its words all appear in *phr* in the same order. For example, if *phr* is *spoken softly*, then *spoken very softly* would be filtered out. The idea behind this step is not to introduce novel or more specific information that was not present in the source.

3.2.6. Output k-Best Candidates. Output up to k-best paraphrase candidates for phrase *phr*, in descending order of similarity. Filter out paraphrases with score less than $minScore$. For example, let $minScore = .3$ and $k = 20$. Then if the third best paraphrase

[6]In Marton et al. [2009a] and in the first set of experiments reported here, we used a slightly simpler condition: If L is stoplisted, append the next word to the left (now having a bigram left context instead of a unigram), and repeat until L is not in the stoplist. We stoplist "promiscuous" words, that is, those that have more than $StoplistThreshold$ collocates in the training corpus, using the preceding $MaxPos$ parameter value. We also stoplist bigrams which occur more than s 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*.

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} best 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.

3.3. Algorithmic Complexity of Distributional Paraphrasing with (Augmented) Suffix Array

Our paraphrase generation is implemented using a suffix-array-based data structure [Manber and Myers 1993], augmented with pattern matching of word sequences [Lopez 2007, 2008]. [7] A Suffix Array (SA) stores all the suffixes in text corpus T in lexicographic order, where 'suffix' here means a sequence of tokens ending at the end of T (not a sequence of letters ending a word). Populating the SA takes $O(|T|\log|T|)$ time; the cost of sorting the suffixes. This data structure only keeps indices to the beginning of each suffix (the end of each suffix being the end of T by definition), so it requires $O(|T|)$ space.

Searching all occurrences of phrase *phr* in text T takes $O(r+|\text{phr}|+\log|T|)$ time, where r is the number of occurrences of *phr*: Finding the first occurrence of the first word of *phr* takes $O(\log|T|)$; finding the first occurrence of the rest of the phrase takes additional $O(|\text{phr}|)$; the other $r-1$ occurrences of *phr* follow it immediately in the sorted suffix array.[8]

The SA can be augmented with a prefix tree with suffix links [Lopez 2007, 2008]. This additional data structure stores all prefixes of the (potentially gappy) search patterns over the SA—and their suffixes, motivated by the observation that there is no point in searching for pattern *abc* in its entirety if its prefix *ab* was not found in the SA, which in turn should not be searched for if its suffix *b* was not found. The prefix and its suffix are linked, and marked as "dead-ends" if the latter is not found. However, if found, the tree node representing the suffix is annotated with the (contiguous) range of the suffix in the SA. This fact turns out to be very useful for DP calculation, which involves finding the counts (number of occurrences) of many collocates. This is because the size of this range (last index minus first index) is actually the count of the suffix or pattern (or collocate) of interest, and this can be quickly calculated without even having to access the corpus T itself again. The pattern matching code was originally developed for searching and extracting SMT rules in parallel text, but it was adapted here for building phrasal DPs and searching for their paraphrase candidates in similar contexts.[9]

Since using SA frees one from the need to precalculate and hold in memory a huge collocational matrix, we can now compute DPs without pruning the vocabulary (as is often done when using LSA, for example), and are able for the first time to scale up these distributional methods to usage in an end-to-end state-of-the-art SMT system.

The focal phrases and their paraphrase candidates are contiguous (gapless) word sequences in this work. However, the search of candidates in context is done as a gappy pattern matching of the form $L\ldots R$ for the left and right contexts. Whenever the gappy pattern is matched, the token sequence in the gap is taken to be a paraphrase candidate.

In the analysis that follows we assume that text corpus T has a vocabulary V consisting of $|V|$ unique tokens (a.k.a. *types*), and that the maximal phrase (or paraphrase

[7] Suffix arrays have been used in SMT for various gains: eliminating the need to precompute a huge number of grammar rules (many of which might never be used for a given test set), and enabling searching context-sensitive rules by pattern matching [Callison-Burch et al. 2005; Zhang and Vogel 2005; Lopez 2007, 2008].

[8] This search cost can be reduced to $O(|\text{phr}|+\log|T|)$ time [Manber and Myers 1993], and even to the optimal $O(|\text{phr}|)$ time [Abouelhoda et al. 2004], as Lopez 2007, 2008] points out.

[9] The code is used here for paraphrase generation, but it could easily be adapted for paraphrase extraction (for details on this distinction, see Androutsopoulos and Malakasiotis [2010]).

candidate) length is m tokens. The maximal number of occurrences of any phrase is $0 \leq r \leq |T|-|V|+1$; due to the Zipfian nature of natural language, r is, more often than not, very small. The algorithmic complexity of distributional paraphrasing is derived from the following steps.

(1) *Build distributional profile DP_{phr}.* A somewhat naive implementation would traverse T and build a collocation matrix for phrases up to length m in $O(m|T|)$ time but taking $O(|V|^{m+1})$ space, which doesn't scale well and can be prohibitive. If one wishes to save space, one can only keep the collocation information of *phr* and its collocates, in the high price of having to recompute parts of the matrix for each phrase (or many thereof). Alternatively, one could read the text to a suffix-array-based data structure in $O(|T|\log|T|)$ time and $O(|T|)$ space, as described earlier. Then, building DP_{phr} requires finding all occurrences of phrase *phr* in text T, taking $O(r+m+\log|T|)$ time, which can be reduced to $O(m)$ per Footnote 8; for a fixed size sliding window, gathering all collocates and their cooccurrence counts with *phr* takes $O(r)$ time. In order to calculate log-likelihood ratios in the DP cells, one needs to count the collocates' occurrences $c(colloc)$ as well; this can be done without accessing T at all, since it only requires finding the first and last indices of each collocate, taking again $O(c(colloc)+\log|T|)$, reducible to $O(1)$ time per unigram collocate. These counts can be saved in the prefix tree or in a hash table for future reference, so finding the counts of all the collocates in all the DPs would require a one-time cost of $O(|V|\log|T|)$ time. Altogether, beside the one-time cost, building DP_{phr}, including the cooccurrence counts, would require $O(r+m+\log|T|+|V|)$ time, reducible to $O(r+m+|V|)$. Since m, the length of *phr* is typically short, this cost can be simplified to $O(r+|V|)$.

(2) *Gather left and right context $L\_\_R$ for phrase phr.* This step involves, for each occurrence of *phr*, the gradual growing of $|L|$ and $|R|$ until $c(L) < mcc$ and $c(R) < mcc$, where *mcc* is a maximal context count threshold (say, 2000), $|L|$ is the maximal length of left context L, and $c(L)$ is the occurrence count of L; and similarly for R. This can be achieved in $O((|L|+|R|)\log|T|)$ time. In practice, $\max(|L|,|R|) < 7$, so this step practically takes $O(\log|T|)$ time for each occurrence of phrase *phr*, and hence $O(r\log|T|)$ altogether for *phr*.

(3) *Gather paraphrase candidates cand occurring in the afore-mentioned $L\_\_R$ contexts.* This is equivalent to finding all occurrences of gappy phrase $L\ldots R$, which takes $O(|L|+|R|+\log|T|+c(L)c(R))$ for each occurrence (and hence each context) of *phr*, where $|L|$ is the maximal length of left context L, and $c(L)$ is the occurrence count of L, and similarly for R. Note that this step involves potential intersection of the occurrences of L and R, and hence the product $c(L)c(R)$. Lopez [2007, 2008] points out—and it holds here as well—that this time can be improved to $O(|L|+|R|+\log|T|+(|L|+|R|)\log\log|T|)$ by using a stratified tree [Emde Boas et al. 1977] as in Rahman et al. [2006], or to $O(|L|+|R|+\log|T|+c(L)+c(R))$ by using a hash table [Lopez 2007, 2008].

(4) *Rank all candidates cand.* Let d be the number of candidates *cand*. First, build DP_{cand}: as for any DP, this takes $O(r+m+\log|T|+|V|)$, reducible to $O(r+|V|)$ time, where for simplicity, r and m denote here also the count and length of *cand*, respectively. Measuring profile similarity takes $O(|V|)$ for each candidate. Finally, the ranking itself takes the expected sorting time of $O(d\log d)$, but for only k-best, it takes $O(kd)$ time. Altogether this step takes $O(d(r+|V|+\min(k,\log d)))$ time.

(5) *Filter textually entailed candidates.* This is linear in the length of the phrase and the candidate, so for all candidates it takes $O(d(|phr|+|cand|))$ time.

(6) *Output k-best candidates.* This is trivial: $O(1)$ per phrase (assuming k is limited to a small constant, such as 20 or 100).

The total complexity is, then, initial cost: $O(|T|\log|T|)$ time, and per-phrase cost.

$$\begin{aligned} O(&r+|V| \\ &+ r\log|T| \\ &+ d(|L|+|R|+\log|T|+c(L)+c(R)) \\ &+ d(r+|V|+\min(k,\log d)). \\ &+ d(|phr|+|cand|)). \end{aligned} \tag{2}$$

This expression can be simplified, and also pushed further down doing the following.

—Uniformly sample every phrase whose count is above some threshold $s << |V|$, for example, $count(phr) > s = 10{,}000$. This means skipping $count(phr) - s$ indices in the suffix array, and visiting only s loci in T when collecting cooccurrence counts, multiplying each collocate occurrence by $count(phr)/s$.
—As mentioned before, $\max(|L|, |R|)$, $|phr|$, and $|cand|$ are small, negligible constants.
—The maximal context occurrence count is also limited to some constant: $\max(c(L), c(R)) < mcc = 2000$.

These speedups are important for real-world SMT applications, since increasingly, a translation wait time of minutes or even deca-seconds is becoming unacceptable. From our experience, while most phrases are processed in mere milliseconds, others (at the left of the Zipfian curve) may take hours if parameter values are set carelessly. Taking the preceding points into account, we can rewrite the complexity expression as follows.

$$O(d[mcc + s + |V| + \min(k, \log d) + \log|T|]) \tag{3}$$

Note that d, the number of candidates, dominates the complexity, but is dependent in part on mcc, the maximal context count. Besides decreasing mcc, one can devise heuristics for subsampling the number of candidates, perhaps according to cooccurrence or Strength of Association (SoA) of certain contexts with phr. We leave this for future research. Further improvements in both initial and per-phrase runtime can be achieved if T is broken into parts that can be searched in parallel (followed by combining the respective search results).

4. PARAPHRASE-AUGMENTED SMT

4.1. Related SMT Augmentation 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.

—Augment current resources (typically parallel texts or the derived phrase tables) with paraphrases of their elements.
—Create additional resources of same type (additional parallel texts).
—Use alternative resources (lesser or no reliance on parallel texts).

This work belongs to the first category, and therefore we mainly focus here on paraphrasing work. It most resembles Callison-Burch et al. [2006] (and its improved variant in Callison-Burch [2008]) in augmenting translation models with source-side paraphrases of the OOV phrases, using weighted log-linear features.

Habash and Hu [2009] show, with a similar pivoting method 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. Other attempts to reduce the OOV rate by augmenting the source side of a

translation phrase table include Habash [2008, 2009], providing an online tool for paraphrasing OOV phrases by lexical and morphological expansion of known phrases and dictionary terms, and transliteration of proper names.

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. Kuhn et al. [2010], too, do not use external resources. Instead, they pivot within the same phrase table, generating paraphrases this way, and soft clustering semantically similar translation rules (phrase table entries) with new cluster-based probabilistic features.

Also recently (after the publication of this article's core work), source-side paraphrase lattice has been used to augment SMT [Du et al. 2010; Onishi et al. 2010]. Pivot paraphrases were used, although this is not crucial for the lattice core idea. Max [2010] augments SMT models by using paraphrases to improve also the distribution estimates of existing (but mainly infrequent) translation entries.

4.2. Paraphrase-Augmented Translation Models

Each of our paraphrase-augmented models is identical to its corresponding paraphrase-less baseline model (Section 5), with the exception of additional paraphrase-based phrase table entries (translation rules), and extra weighted log-linear feature or features, as in Callison-Burch et al. [2006].

$$h(e, f) = \begin{cases} asim(DP_{f'}, DP_f) & \text{If phrase table entry } (e, f) \text{ is generated from } (e, f') \text{ using monolingually-derived paraphrases.} \\ 1 & \text{Otherwise.} \end{cases} \quad (4)$$

As noted in Marton et al. [2009a] and Marton [2009], 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; for example, 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.

source-side focal 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

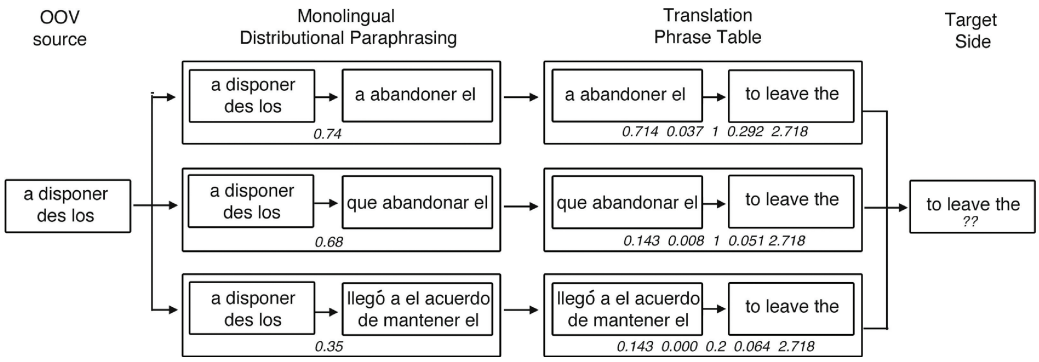

Fig. 2. Multiple paths for augmenting the translation phrase table with a new rule $\langle f, e \rangle$, comprising of the OOV source language phrase f and a certain target language phrase e. Here, for a Spanish-to-English translation model, f = "a disponer de los", and has three paraphrases (with varying quality) appearing in the existing phrase table's source side, all of which translate to e = "to leave the". Paraphrase semantic distance scores and various translation feature scores are in italics. Determining translation feature scores for the new rule is not straightforward (denoted with "*??*").

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 given before. See Figure 2 for a visual example.

There are different possible approaches to the multiple-path phenomenon: A default approach might create a separate new $\langle f, e \rangle$ rule for each path, making these new rules compete with one another in order to enter the final sentence translation derivation during "decoding" time (denoted hereafter as aggregation method NONE); another approach might generate only a single rule from f to e, using only one randomly chosen path (aggregation method RAND); or using only the "best path"—the path with the highest paraphrase similarity score—the upper path in the example previous (with highest similarity score .74; hereafter, aggregation method BEST). 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, either by a simple addition (aggregation method ADD, which is similar to the probability marginalization applied by Callison-Burch et al. [2006], except that the sum here might exceed 1), or by averaging the paths' similarity scores (aggregation method AVRG). Yet another way would be to increase the new rule's associated semantic score in proportion to the paraphrase scores of f to f_1, then f to f_2, and so on (aggregation method UPDT). More formally, for each paraphrase f of some source-side phrases f_i, with similarity scores $sim(f_i, f)$, calculate an aggregate score $asim$: with a "quasi-online-updating" method as

$$asim_i = asim_{i-1} + (1 - asim_{i-1})\, sim(f_i, f), \tag{5}$$

where $asim_0 = 0$. 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. In this aggregation method, we 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.[10] In Section 5.3 we compare the performance of afore-mentioned methods (except for RAND, which seems less principled).

[10]This latter aggregation method was the one used in Marton et al. [2009a] and Marton [2009, 2010].

5. EXPERIMENTS

We examined augmenting translation models with paraphrases generated distributionally and ranked by distributional semantic distance measures. We tested our system on handling unknown phrases when translating from English into Chinese (E2C), and from Spanish into English (S2E).

For all baselines we used the phrase-based Statistical Machine Translation system Moses [Koehn et al. 2007], with the default model features:[11]

—a phrase translation probability,
—a reverse phrase translation probability,
—a lexical translation probability,
—a reverse lexical translation probability,
—a word penalty,
—a phrase penalty,
—six lexicalized reordering features,
—a distortion cost, and
—a Language Model (LM) probability.

The phrase translation probabilities were determined using maximum likelihood estimation over phrases induced from word-level alignments produced by performing intersection of Giza++ training [Och and Ney 2000] on the source and target sides of the parallel training sets. 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; the lower the TER score, the better the result. This is denoted with BLEU↑ and TER↓ in Table IV. When the baseline system encountered unknown words in the test set, its behavior was simply to reproduce the foreign word in the translated output, as in Callison-Burch et al. [2006].

Statistical significance for the BLEU results was calculated using Koehn's paired bootstrap resampling 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 $p < .05$ in the following. Similarly, a 99% CI is denoted as $p < .01$, and so on for other CIs. The word "significant" is used shortly as a shorthand for "statistically significant" (at $p < .05$ unless specified otherwise).

The paraphrase-augmented models were created as described in Section 4, with the UPDT path aggregation method. See Section 5.3 for exploration of alternative path aggregation methods. We used the following parameter settings for the experiments reported throughout this section, unless otherwise specified: For generating the monolingually derived distributional paraphrases, we used a sliding window of size ± 6, a sampling threshold of 10000 occurrences, and a maximal paraphrase length of 6 tokens. Also, we arbitrarily limited the number of occurrences (in which to look for paraphrase candidates) of each context of phrase *phr* to no less than 250 (if there are more than that), and no more than 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*, we output no more than the top $k = 20$ best-scoring paraphrases. We generated paraphrases for phrases up to six tokens in length, with an arbitrary similarity threshold of 0.3.

[11]www.statmt.org/moses.

Table III. English-Chinese (E2C) Training Set Sizes

Set	Millions of tokens (Source+Target)
E2C 29K	0.8 + 0.6
E2C Full	6.4 + 5.1
BNC+APW	187

We experimented with three variants:

—adding a single additional feature for all paraphrases (*1-6grams-distrib*);
—using only paraphrases of unigrams (*1grams-distrib*);
—and adding two features, one only sensitive to unigrams, and the other only to 2–6-grams (*1&2-6grams-distrib*).

All features had the same design as described earlier, and each model's feature weight set, including the baseline's, was tuned using a separate minimum error rate training. We repeated this process with pivot paraphrases for comparison.

5.1. English-to-Chinese Translation

For the English-Chinese (E2C) baseline model, we 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). We trained a trigram language model on the Chinese side, with the SRILM toolkit [Stolcke 2002], using the modified Kneser-Ney smoothing option. We 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 we used the Chinese-English NIST MT 2005 evaluation set. In order to use it for the reverse translation direction (English-Chinese), we arbitrarily chose the first English reference set as the development "source", and the Chinese source as a single "reference translation". For testing we used the English-Chinese NIST MT evaluation 2008 test set with its four reference translations. We calculated the BLEU score using shortest reference length, and using 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.[12]

We augmented the E2C baseline models with paraphrases generated as described earlier, 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 III for training set sizes.

Results are shown in Table IV, and paraphrasing examples in Section 6, Table XIII.

Augmentation with pivot-based paraphrases. In order to compare the monolingual distributional paraphrasing method with the multilingual pivot method, we augmented the translation model with the pivot-generated English paraphrases used in Callison-Burch [2008].[13] Due to memory (RAM) constraints, it was not possible to use the full list. We therefore chose to filter it with a score threshold of $p < .3$, similarly to the one used for the distributional paraphrases. Note, however, that a .3 pivot-based estimated paraphrase probabilistic score is not equivalent to a .3 distributional paraphrase vector similarity score. In addition to using all available lengths of paraphrased phrases

[12]http://www.itl.nist.gov/iad/mig//tests/mt/2008/doc/mt08_official_results_v0.html.
[13]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.

Table IV. E2C Results: Character-Based BLEU and TER Scores

dataset	E2C model	BLEU↑	TER↓
29k	*baseline*	15.2	69.3
	1grams-pivot > .3	15.5	69.4
	1-5grams-pivot > .3	16.1[B]	**69.0**
	1&2-5grams-pivot > .3	**16.2[BI]**	69.1
29k	*1grams-distrib*	**16.9[B]**	**68.8**
	1-6grams-distrib	16.5[B]	69.2
	1&2-6grams-distrib	**16.9[BC]**	**68.8**
232k	*baseline*	**22.2**	**63.6**
	1grams-distrib	21.6[B]	64.2
	1-6grams-distrib	21.8[B]	64.8
	1&2-6grams-distrib	21.4[B]	64.9

All models have one additional feature over baseline, except for the *1&2-5grams* and *1&2-6grams* models that have one feature for unigrams and another feature for longer n-grams. Paraphrases with score < .3 were filtered out. [B], [I],[C] = significantly better than corresponding baseline, *1grams* model, and *1-5grams* or *1-6grams* model, respectively, $p < 0.05$.

(unigram to 5-gram) as done in Callison-Burch [2008], we also experimented with novel *1grams-pivot* and *1&2-5grams-pivot* models, equivalent to *1grams-distrib* and *1&2-6grams-distrib*, respectively. All pivot models showed significant BLEU gains over the baseline. They also showed slight TER gains, except for the unigram pivot model (but recall it was threshold filtered). The novel *1&2-5grams-pivot* model also significantly outperformed its unigram counterpart.

Augmentation with distributional paraphrases. The E2C 29k-line augmented models showed significant gains over the baseline, up to 1.7 BLEU points. All these 29k-line distributional models also showed higher BLEU and TER gains than their pivot counterparts (except for the TER score of *1-6grams-distrib*). For the 232k-line models, results were negative. TER scores generally followed the BLEU pattern. Note that the E2C 232k-line 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),[14] although we 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). For a successful attempt at augmenting the 232k-line model, see Section 5.4.

5.2. Spanish-to-English Translation

In order to permit a more direct comparison with the multilingual pivoting method, we also experimented with Spanish-to-English (S2E) translation, following Callison-Burch et al. [2006]. For baseline we used the Spanish and English sides of the publicly available Europarl multilingual parallel corpus [Koehn 2005], with the standard training, development, and test sets. We 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, we used the same 5-gram language model, development, and test sets: For tuning, we used the Europarl dev2006 Spanish and English sides, and for testing we used the Europarl 2006 test set.[15]

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

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

Table V. Spanish-English (S2E) Training Set Sizes

Set	Millions of tokens (Source+Target)
S2E 10K	0.3 + 0.3
S2E 20K	0.6 + 0.6
S2E 80K	2.3 + 2.3
WMT09+Acquis+AFP	402

Table VI. Spanish-English (S2E) Results: Lowercase BLEU and TER

bitext	model	BLEU↑	TER↓
10k	*(baseline)*	23.8	62.4
	1grams-pivot	**24.4**$^{\mathbf{B}}$	**61.1**
	1-5grams-pivot	24.1$^{\mathrm{B}}$	61.9
	1&2-5grams-pivot	(failed)	
10k	*1grams-distrib*	24.0	62.0
	1&2-6grams-distrib	24.1	61.8
20k	*(baseline)*	24.7	62.3
	1grams-distrib	24.8$^{\times}$	61.3
	1&2-6grams-distrib	**24.9**$^{\mathbf{B}}$	**61.1**
80k	*(baseline)*	**27.9**	58.0
	1grams-distrib	27.8	**57.8**
	1&2-6grams-distrib	**27.9**	57.9

Paraphrases with score $<$ *minScore* were filtered out. Statistical significance: $^{\mathrm{B}}$= significantly better than baseline, $p < 0.05$; $^{\times}$= "almost significantly" better than baseline, $p < 0.1$.

We trained the Spanish paraphrase generation model on the Spanish corpora available from the EACL 2009 Fourth Workshop on Statistical Machine Translation[16] (the Spanish side of the Europarl-v4, news training 2008, and news commentary 2009), the JRC-Acquis-v3 corpus[17], and the AFP part of the LDC Spanish Gigaword (LDC2006T12), and truncating the resulting corpus after the first 150M lines. We lowercased these training sets, tokenized and removed punctuation marks and numbers, and this resulted in training set sizes as detailed in Table V.

Results are shown in Table VI.[18] In order to evaluate the S2E models, we used BLEU [Papineni et al. 2002] over lowercase output. Not recasing the output avoids possible recaser-originated scoring "noise". We used Koehn's [2004b] significance test as before. Paraphrase examples are given in Section 6, Table XII.

The distributional paraphrasing models achieved gains of up .6 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-distrib* was best performer in all cases. In the S2E experiments, the pivot method yielded slightly higher gains. However, note that pivot paraphrases for named entities were filtered out. Note also that the S2E baselines' scores reported here are higher than those of Callison-Burch et al. [2006]. We 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. We concluded from a manual evaluation of the 10,000-line models that the two major weaknesses of the

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

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

[18]For a larger set of experiments, involving varying monolingual corpus size and semantic distance threshold, see Marton [2009].

Table VII. Comparison of Various Path Aggregation Methods Using Different Limits on Context Counts (BLEU)

English to Chinese, *1&2-5grams-distrib*							Spanish to English, *1grams-distrib*			
Method	Maximal context count					Times	Method	Maximal context count		
	256	512	1024	2048	4096	in 3 top		256	512	1024
NONE	**16.6**	16.3	16.4	16.1	16.2	2	**NONE**	24.2	24.1	24.2
ADD	15.6	**16.5**	**16.5**	15.7	16.4	1	**ADD**	24.2	24.1	24.1
BEST	16.1	**16.9**	16.8	15.8	16.8	4	**BEST**	24.2	24.1	24.2
AVRG	15.5	**17.1**	16.7	16.6	17.0	4	**AVRG**	24.1	24.2	24.2
UPDT	15.6	**17.0**	16.9	16.3	16.5	**5**	**UPDT**	24.2	24.1	24.1

Table VIII. Number of Augmentative Rules with Multiple Paths

	Number of paths for same rule					
Model	2	3	4	5	6	7+
E2C 29k *1grams-distrib*	12458	1220	270	92	36	38
E2C 232k *1grams-distrib*	13766	1274	290	74	26	14
E2C 29k *1-5grams-distrib*	238516	13924	2958	1006	440	466
E2C 232k *1-5grams-distrib*	354600	20858	3956	1340	506	462

baseline model were (not surprisingly) number of untranslated (OOV) words/phrases, followed by number of superfluous words/phrases in the translation output.

5.3. Path Aggregation Methods and Limits on Number of Paraphrase Candidates

This section explores two issues that were raised in Sections 3.3 and 4.2: choosing the best path aggregation method, and limiting the maximal context count *mcc* (in order to reduce d, the number of paraphrase candidates, and hence algorithmic complexity and runtime). We therefore explored the interaction of several path aggregation methods (NONE, ADD, BEST, AVRG, UPDT; see Section 4.2) with several *mcc* values (from 256 to 4096 in multiples of 2). We first experimented with the E2C *1&2-5grams-distrib* model (left half of Table VII), since it was the best performer in the experiments reported in Section 5.1 and in Marton [2009]. The difference between the lowest and highest results was 1.6 BLEU. Limiting *mcc* to 512 clearly yielded best results in almost all cases. This is good news, since it means that *mcc* can be lowered from its previous value around 2000 down to around 500 without loss in performance (in fact, this would result in performance gains in addition to faster runtime).

Which is the best path aggregation method? This is less clear from the table. Generally, it seems that NONE and ADD are the worst performers, and that AVRG and UPDT are the best performers. In order to decide which method to use in the future, we counted how many times each method was among the top three performers for each *mcc* value (rightmost column in the table). UPDT was the winner, albeit by a small margin.

In order to get an impression of the magnitude of the multiple path phenomenon, we counted the number of alternative paths per each new $\langle f, e \rangle$ rule[19] in the E2C *1grams-distrib* and *1-5grams-distrib* models, in both 29k and 232k training sizes. An in many other linguistic phenomena, this turned out to be a Zipfian curve (see Table VIII)[20] for each model: high number of rules with two alternative paths, which decreases as the number of alternative paths increases, with a long "tail", in some cases reaching over 20 alternative paths for very few rules. At first thought, one might find the differences between the 29k and 232k models surprising: One might expect that the larger models would have fewer OOV phrases, and therefore fewer augmentative paraphrastic rules,

[19] $\langle f, e \rangle$ are as denoted in Section 4.2.

[20] Counts were performed on models augmented with paraphrases of OOV phrases in tuning and test sets, filtered with a 0.05 semantic distance threshold; see also Section 5.4.

Table IX. E2C Results, Revisited: Character-Based BLEU and TER Scores

bitext sentences	English-Chinese model	BLEU↑	TER↓
29k	*baseline*	15.2	69.3
	1grams-pivot	15.5	69.4
	1-5grams-pivot	16.1^{B}	**69.0**
	1&2-5grams-pivot	$\mathbf{16.2^{BI}}$	69.1
29k	*1grams-distrib*	$\mathbf{16.4^{B}}$	**68.9**
	1-5grams-distrib	16.1^{B}	69.4
	1&2-5grams-distrib	$\mathbf{16.4^{BC}}$	69.4
29k	*1grams-distrib±2*	$\mathbf{17.1^{B}}$	**68.8**
	1-5grams-distrib±2	$\mathbf{17.1^{B}}$	68.9
	1&2-5grams-distrib±2	16.4^{B}	69.2
232k	*baseline*	21.8	63.8
232k	*1grams-distrib*	$\mathbf{22.5^{B}}$	64.4
	1-5grams-distrib	$\mathbf{22.5^{B}}$	66.2
	1&2-5grams-distrib	21.7	**63.9**
232k	*1gram-distrib±2*	$\mathbf{23.0^{BD}}$	**63.6**
	1-5gram-distrib±2	22.5^{B}	64.4
	1&2-5gram-distrib±2	21.8	64.3

All models have one extra feature on top of their baseline model's features, except for the *1&2-5grams...* models, which have one extra feature for unigrams and another for longer n-grams. Statistical significance from corresponding: *baseline* (^{B}), *1grams...* model (^{I}), coarser *1-5grams...* model (^{C}), or from distributional model with up to 6-token long paraphrases (^{D}), $p < .05$.

and smaller values in each alternative path bin. However, the opposite holds: The number of rules in each 232k model path bin is greater than that of the corresponding path bin of the counterpart 29k model. This can be explained by the fact that these augmentative rules depend on finding an existing "anchor" item in the phrase table's source side. Naturally, the larger the model's training set, the larger the phrase table, and hence the higher the chances of finding suitable "anchor" items.

Finally, we repeated a this experiment in a narrower *mcc* range with the S2E *1-gram* model (right half of Table VII). Here, however, the differences between the lowest and highest performers were only about 0.1 BLEU, and therefore, not supporting (but also not undermining) the conclusions from the E2C experiment given before.

5.4. Revisited Experiments

So far, the distributional paraphrases proved useful only in models trained on small datasets. We hypothesized that larger models didn't show gains because they already covered many of the good paraphrases. To test this, we conducted experiments with other paraphrase semantic distance thresholds.[21] We settled on a lowered 0.05 threshold, which helped some models (but not all). For these models, the increased coverage resulted in BLEU score gains, even though the average paraphrasing quality dropped (the additional paraphrases had lower scores by construction). With this setting, two of the 232k models showed for the first time a significant gain over the baseline: 0.7 BLEU (but still worse TER scores). See Table IX.[22] In order to address the concern that in the experiments reported previously in this section, the distributional models

[21]We omit reporting all these experiments since this is not the focus of this article. The interested reader can find more of these experimental results in Marton et al. [2009a] and Marton [2009].

[22]The results reported in this table were achieved using a larger monolingual corpus of over 516M tokens, consisting of the Gigaword documents from 2004 and 2008 (LDC2009T13), preprocessed slightly differently

had an unfair advantage over the pivot models in that they paraphrased unigrams to 6grams, here we experimented with paraphrasing unigrams to 5grams, as in the pivoting models. In preliminary experiments, the longer n-gram models always outperformed their shorter n-gram counterpart models (0.3–0.4 BLEU, and except for one case, about 0.1–0.2 TER), putting this fair comparison concern to rest, and making our claim for superior performance stronger.

A manual inspection of the paraphrases (before augmenting the translation tables) revealed that longer paraphrases tend to have lower quality. Therefore, we repeated some experiments, only allowing the usage of paraphrases whose length difference from the paraphrased phrase was no longer and no shorter than two tokens. We denote these experiments with ± 2 in Table IX. Threshold of 0.05 was used here as well. This time almost all models yielded further gains: The 29k models yielded up to 1.9 BLEU and 0.5 TER over the 29k baseline, and the 232k models yielded up to 1.2 BLEU and 0.2 TER over the 232k baseline. *1grams-distrib±2* was consistently the best performer in both training set sizes.

5.5. English to Arabic: Translation into a Morphologically Rich Language

Following the phrase table augmentation insights gained in Section 5.3 and the revised experiments in Section 5.4, we next turn to translating into a morphologically rich language. Unlike translating into English or Chinese (Sections 5.1–5.2), both of which having a relatively impoverished morphological inflection, Modern Standard Arabic exhibits rich morphology and complex agreement patterns. This fact may pose additional challenges, both because richer morphology implies higher data sparsity (and therefore desired target word forms may be missing from the table), and because it entails more complex syntactic agreement patterns (and therefore some translation derivations with augmented rules may not be used since they would be penalized by elements such as the language model, even though translation with these augmented (but imperfect) rules may be superior to all other derivations available to the model).

The settings in the following experiments were the same as in Section 5.4, including the monolingual corpus for paraphrase generation, consisting of over 500 million tokens. The English-Arabic training parallel corpus consisted of about 135,000 sentences (4 million words) and a subset of 30,000 sentences (1 million words) from Arabic News (LDC2004T17), eTIRR (LDC2004E72), English translation of Arabic Treebank (LDC2005E46), and Ummah (LDC2004T18). Testing was done on the NIST Arabic-English MT05 and MEDAR 2010 English-Arabic four-reference evaluation sets.

English-to-Arabic translation results are given in Table X. Augmenting the larger model seems beneficial, although the improvement in BLEU was found statistically significant only when measured in lemma-based BLEU on MT05. Interestingly, the subset (reduced) model yielded negative results. The lemma-based scores generally portray the augmented models in better light than the respective standard (word-based) BLEU and TER measures, in terms of difference from baseline. This supports the concern that morphologically rich target languages require additional research in order for a method such as ours to work well. For more details on the data, tokenization, evaluation metrics, additional related experiments, and further discussion, see El-Kholy and Habash [2010a, 2010b], and Marton et al. [2011].

6. DISCUSSION AND FUTURE WORK

Schroeder et al. [2009] recently showed that the upper bound for gains by paraphrase augmentation (using human-generated paraphrases in a lattice of the source language)

(conflating numbers, dates, months, days of week, and alphanumeric tokens to their respective classes). See Marton [2010] for details.

Table X. English-Arabic Translation Scores

test set / model	30k-sentence (1M word) training				135k-sentence (4M word) training			
	BLEU↑	Lemma BLEU↑	TER↓	Lemma TER↓	BLEU↑	Lemma BLEU↑	TER↓	Lemma TER↓
MT05 baseline	**23.6**	**31.3**	**57.6**	**47.3**	25.8	33.5	55.7	45.3
aug-1gram	23.2	30.8	58.8	48.4	**26.4**	**34.3^{B}**	**55.1**	**44.7**
MEDAR baseline	**13.6**	**18.7**	**67.6**	**61.3**	17.1	23.1	65.1	58.6
aug-1gram	12.9	18.3	68.9	62.3	**17.2**	**23.5**	65.1	58.6

B = statistically significant w.r.t. (B)aseline.

Table XI. Estimated Probabilities $p(e_2|e_1)$ of English Identity Paraphrases via Pivoting (sample taken from http://www.cs.jhu.edu/∼ccb/howto-extract-paraphrases.html)

Frequency	Phrase e_1	Paraphrase e_2	p
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

Identity paraphrase entries with $p > .75$ are quite rare, and from a short sampling, it seems that almost all of them are named entities.

is high, and has not been not reached yet. We take their work as another validation of this research direction.

6.1. Pivot and Distributional Paraphrasing

Pivot paraphrasing methods (translating to other languages and 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": The more polysemous the focal phrase, the more likely a higher rate of inadequate paraphrase candidates; even if each inadequate candidate score is low, together they might take away substantial probability mass, resulting in lowering the probability estimates for the better candidates, making the paraphrase probability estimate less reliable. Table XI 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 should be a high scoring paraphrase candidate, but in reality, the estimated probabilities are often quite low.[23] In contrast, large monolingual resources are relatively easy to collect, the paraphrasing method described here involves only a single translation/paraphrasing step per focal phrase, and the identity paraphrasing score always equals 1 (unless the focal 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).

It is unclear how to fairly compare the pivoting method to ours. Should the monolingual and bilingual training resources be equivalent in some way? (But large bitexts are

[23]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.

Table XII. Spanish Paraphrases Generated by Pivot and Distributional Methods, Ordered Best First

source	pivot	distributional
baile	danza	el baile
	bailar	baile y
	a	danza
	dans	un baile
	empresa	teatro
	coro	baloncesto el cine
a favor del informe	a favor de este informe	favor del informe
	favor del informe	en contra del informe
	el informe	a favor de este informe
	a favor	en contra de este informe
	por el informe	en contra de la resolución
	al informe	a favor del informe del sr.
	su	en contra del informe del sr.
	del informe	a favor del excelente informe

rare; EuroParl-based pivoting is only applicable to European languages). Should the lengths of the phrase or its paraphrase be the same in both methods? (This was done in Section 5.4). Should pivot paraphrases be threshold filtered as the distributional ones are? (But recall that 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 method should be presented in its best light. But finding the best running parameters for each method is not a simple matter either. Therefore, the comparisons here should be regarded as a first stab at this problem, inviting further research.

One potential advantage of using bitexts for paraphrase generation is the usage of implicit human knowledge, that is, sentence alignments. The concern that not using this knowledge would turn out to be detrimental to the performance of our paraphrase-augmented SMT systems was largely put to rest: In both S2E and E2C language pairs, the original pivot paraphrase augmentation *1-5grams-pivot* models resulted in about 0.3 BLEU gains over the baseline, similar or lower than their distributional counterparts. Our novel improved pivot models yielded higher gains: E2C *1&2-5grams-pivot* yielded 1 BLEU—but its distributional counterpart yielded a higher 1.7 BLEU gain, and S2E *1grams-pivot* yielded 0.6 BLEU—slightly higher than its distributional counterpart; however, this is a rather small gain.

To look at some specific examples, the top part of Table XII shows that for the Spanish *baile*, the top four distributional paraphrases are all appropriately dance related, while an unrelated function word *a* made its way into the top four pivot candidates. However, in the bottom part of the table, for the Spanish source phrase *a favor del informe*, several antonymous paraphrase candidates (containing *contra* instead of *favor*, i.e., *against the report* instead of *for the report*) made their way to top places on the list, while the pivot candidates seem all to carry similar meaning (if perhaps only partially so in some cases), except for the function word *su*. This is typical of both methods: pivoting seems to rank high the alignment-wise and collocationally "promiscuous" function words, while the distributional method tends to rank high antonymous candidates, since they appear in similar contexts as the source phrase.[24]

6.2. Paraphrasing Quality Issues

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

[24]See Marton et al. [2011] and Baker et al. [2012] for first steps to handle antonymous candidates.

Table XIII. Examples of English Distributional Paraphrases of Phrases Unknown to the E2C 29K Baseline Model

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

accord,...), and so is the case for the five best paraphrases of the focal *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 method: 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 focal phrase *to provide any other*. It seems that in general, paraphrases of longer focal phrases are of lower quality than those of unigrams, as can be seen at the bottom right, fourth part of the table. There, only the second-best paraphrase is somewhat semantically close to the focal *we have a situation that*, but the overall quality is clearly lower.

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; Mohammad and Hirst 2006; Marton et al. 2009b; Marton 2010]; using context to help sense disambiguation [Erk and Padó 2008]; improving the semantic distance measure (see also Footnote 24); and optimizing the paraphrase similarity threshold for use in SMT, for example, on a held-out dataset. As for the latter, note that the higher the threshold the lower the coverage, while the lower the threshold the lower the paraphrases and translation quality. We showed gains by lowering the threshold (Section 5.4), but it remains to be seen how these two opposite effects play out and where the optimum lies. We would like to explore ways of incorporating syntactic knowledge that do not sacrifice coverage as much as in Callison-Burch [2008]. We would also like to integrate context with lexical resource-based semantic knowledge.

6.3. Scalability and Data Sufficiency

Scaling up to larger monolingual corpora, although potentially promising in terms of quality and coverage, poses some challenges of disk space, RAM, and processing

time. It remains to be seen if it is feasible, with or without enabling techniques such as sampling, cloud computing, and Map/Reduce (Lin [2008]), refer to, etc., which have been applied in related subfields. Currently, distributional semantic distance measures tend to become less accurate when comparing profiles (DPs) of words (or phrases) 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.

Scaling up our method so that it would improve models trained on larger bitexts is another challenge. Generally, the larger the bitext, the lower the OOV rate; moreover, it seems that the larger the bitext, the harder it is to paraphrase focal phrases, perhaps due to their lower frequency on average, resulting in potentially impoverished DPs, and less reliable cooccurrence statistics and similarity scores. However, for translating low-resourced specialized domains and genres, lowering the OOV rate is likely to remain an important issue.

How much monolingual data would be "sufficient"? Currently we have only little evidence to support the claim that increasing resource size is helpful, although it would be surprising if it were not the case. Giving a more detailed account for this question is not necessarily simple: using artificially reduced resources is not of much interest, since current monolingual data size seems to yield useful paraphrases, but with only modest translation quality gains; using additional monolingual data might involve mixing different domains or genres in an unbalanced manner, which may affect paraphrase quality either way, and which one would have to control and tease apart from mere data size when assessing the contribution of resource size. Future research should further investigate paraphrase quality as a function of monolingual data size.

6.4. Paraphrasing with Suffix Array

One of the advantages of our novel usage of SA for semantic representation that is mentioned in Section 3.3 is not having to precompute a large matrix. Therefore, SA can be useful in a fast-response application such as SMT. Our method can be tuned to be faster (potentially sacrificing paraphrasing quality), and the code can be further optimized for speed. Another advantage is not having to prune vector or matrix size, and this richer representation is likely to contribute to paraphrasing quality. But the potential advantages of using SA for semantic representation go beyond what is mentioned in Section 3.3: SA enables the search and DP construction of any number of arbitrarily long phrases in reasonable time and space, which is either entirely prohibitive in standard methods, or is only partially possible at best. Moreover, several new research efforts apply higher-order matrices (a.k.a. *tensors*) to semantic representation; the memory limitations for such representation are even more acute, which often forces pruning and pairwise projections to two dimensions (matrix) at a time, most likely resulting in a performance degradation (refer to Van de Cruys [2009]). Last, since SA stores all suffixes in lexicographic order, it naturally lends itself to paraphrase extraction for all n-grams of arbitrary length (e.g., 1–5grams), simply by traversing the SA and paraphrasing each desired suffix. Doing this for our purposes here was unnecessary, but arguably it is useful for building lexical resources such as a thesaurus or paraphrase bank. Future research is likely to benefit from further utilizing SA.

6.5. Fine-Grained Soft Constraint Features

Fine-grained features proved advantageous in many cases, that is, having a dedicated feature for paraphrases of unigrams, and another feature for paraphrases of longer phrases, compared with a single (and hence, coarser) feature for all paraphrases. However, paraphrasing unigrams only (*1grams. . .*) was a best performer even more consistently. So it seems that paraphrasing longer phrases is useful, at least in some cases,

but not as reliable as paraphrasing unigrams only. This fact leads us to conjecture that modeling the semantics and semantic distance of larger-than-unigram phrases is still far from being well understood. We are not the first to notice it and take interest in this issue; see for example Sag et al. [2001] and recent MultiWord Expression (MWE) workshops. Further experimentation is required to better understand and tease apart the issues of longer phrase semantics and fine granularity.

When the length of the distributional paraphrases was limited (hard constrained) to the vicinity of the length of their focal phrase ± 2 tokens, further gains were achieved: up to 1.9 BLEU for both *1grams-distrib*± 2 and *1-5grams-distrib*± 2. This novel element can be further developed as a set of additional *soft constraint* features: $|focal| : |cand|$ length ratio, $|focal| - |cand|$ length difference, and/or $|focal|$ and $|cand|$ lengths as separate features.[25] Alternatively, rules with focal and/or paraphrases of certain length can have a dedicated paraphrase similarity feature with a separate weight, similarly to, and extending, the *1&2-5grams* model approach. Yet another alternative (as was suggested by an anonymous reviewer), is to experiment with *1-2grams, 1-3grams,* etc.

Note that there is a trade-off between finer granularity and data sparseness. The longer the unknown phrase, the fewer the generated paraphrases above some similarity threshold. Also, from preliminary experiments, it seems that the number of generated paraphrases per unknown phrase (above a fixed similarity threshold) drops in proportion to the length of the unknown phrase. Therefore, separate soft constraint log-linear features for longer phrases are likely to be of low quality or marginal impact, while increasing runtime. If using the de facto standard MERT [Och 2003]—as opposed to, say, the newer MIRA [Chiang et al. 2008, 2009]—for feature weight optimization, the mere increased number of features might be prohibitive by itself. It remains to be explored what the optimal split to separate features is, in interaction with better modeling of longer phrases. It will most likely also depend on the available monolingual corpus size.

6.6. Additional Future Work Directions

The paraphrasing method presented here is quite general, and therefore different semantic distance measures—including other corpus-based, resource-based, or hybrid measures—can be plugged in to generate phrasal paraphrases. Although our method is largely language independent, English-Arabic results (Section 5.5) indicate that some language-specific issues in paraphrasing or SMT (such as lemmatization of the monolingual paraphrasing corpus, the training bitext, and/or the language model) are needed in order to yield higher gains, at least in such SMT tasks, and likely with any morphologically rich language.

A further goal in the future would be to create a distributional-semantic-distance-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, at least on low-resourced specialized domains and genres.[26] The paraphrases generated by our method can also help in other tasks beside SMT, be it query expansion in information retrieval, document summarization, dialog systems and natural language generation, or other NLP tasks.

[25]By "soft constraint" we mean here that translation rules are scored by various tunable weighted features that take into account the length of the focal phrases and/or their paraphrases, instead of applying hard constraints, such as filtering out paraphrases according to their length.

[26]For interesting first stabs at this direction, see Ravi and Knight [2011] and Klementiev et al. [2012].

7. CONCLUSIONS

We showed that augmenting translation models with monolingually derived paraphrases, generated distributionally, and ranked with distributional semantic distance measures over a large text in the source-side language, yields best improvements over a nonaugmented baseline in almost all cases. Our method is largely language independent, and has the advantage of not relying on bitexts in order to generate the paraphrases, and therefore it clears the way for accessing large amounts of (monolingual) training data, for which creating bitexts of equivalent size is generally unfeasible.

The smaller models, emulating a resource-poor language setting, showed even higher gains than larger models (which were trained on supersets of the smaller models' data), when augmented using the same paraphrase resources. The best smaller model's gain was 1.9 BLEU, compared with 1.2 BLEU for the best larger model's gain.

We also showed that the phenomenon of multiple alternative paths for generating a new translation rule is pervasive, with a Zipfian distribution. We explored several ways of aggregating the multiple alternatives, and settled on the UPDT method.

Last, we provided implementation details of our method, and analyzed its algorithmic complexity. We identified d, the number of paraphrase candidates per phrase, and *mcc*, the related maximal context count, as the main factors driving the complexity. We then explored how much *mcc* (and hence indirectly d as well) can be limited without hurting translation performance.

Our novel use of suffix array for paraphrasing enabled us to generate paraphrases on-the-fly (without collocation matrix precalculation, which would require knowledge of the test set's source side), and to use full (nonpruned) distributional profiles without encountering memory limitation issues. We believe that using suffix array for semantic representation has a great potential that hasn't been fully exploited yet, for example, in building multidimensional (tensor) representation, and searching meaning in context.

ACKNOWLEDGMENTS

The author would like to thank Philip Resnik, Amy Weinberg, and Chris Callison-Burch for their guidance in the early and mid stages of this research; Adam Lopez for making his source code available; Chris Dyer for his help in preprocessing the E2C data; and Saif Mohammad for helpful discussions.

REFERENCES

ABOUELHODA, M. I., KURTZ, S., AND OHLEBUSCH, E. 2004. Replacing suffix trees with enhanced suffix arrays. *J. Discr. Algor. 2*, 1, 53–86.

AGIRRE, E. AND LOPEZ DE LACALLE LEKUONA, O. 2003. Clustering wordnet word senses. In *Proceedings of the 1st International Conference on Recent Advances in Natural Language Processing (RANLP'03)*.

ANDROUTSOPOULOS, I. AND MALAKASIOTIS, P. 2010. A survey of paraphrasing and textual entailment methods. *J. Artif. Intell. Res. 38*, 135–187.

BAKER, K., DORR, B., BLOODGOOD, M., CALLISON-BURCH, C., FILARDO, W., PIATKO, C., LEVIN, L., AND MILLER, S. 2012. Use of modality and negation in semantically-informed syntactic mt. *Comput. Linguist. 38*, 2, 1–48.

BANERJEE, S. AND LAVIE, A. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL'05)*.

BANERJEE, S. AND PEDERSEN, T. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03)*. 805–810.

BANNARD, C. AND CALLISON-BURCH, C. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Association for Computational Linguistics, 597–604.

BARZILAY, R. AND MCKEOWN, K. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the Association for Computational Linguistics (ACL'01)*.

Bhagat, R. and Ravichandran, D. 2008. Large scale acquisition of paraphrases for learning surface patterns. In *Proceedings of the Association for Computational Linguistics - Human Language Technology Conference (ACL-HLT'08)*. 674–682.

Bond, F., Nichols, E., Appling, D. S., and Paul, M. 2008. Improving statistical machine translation by paraphrasing the training data. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT'08)*.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. 1993. The mathematics of statistical machine translation. *Comput. Linguist. 19*, 2, 263–313.

Budiu, R., Royer, C., and Pirolli, P. 2006. Modeling information scent: A comparison of lsa, pmi and glsa similarity measures on common tests and corpora. In *Proceedings of the Conference on Large Scale Semantic Access to Content (Text, Image, Video, and Sound) (RIAO'07)*.

Burnard, L. 2000. *Reference Guide for the British National Corpus World Edition*. Oxford University Computing Services, Oxford, UK.

Callison-Burch, C. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*.

Callison-Burch, C., Bannard, C., and Shroeder, J. 2005. Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In *Proceedings of the Association for Computational Linguistics (ACL'05)*. 255–262.

Callison-Burch, C., Koehn, P., and Osborne, M. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the North American Chapter of the Association for Computational Linguistics Conference (NAACL'06)*.

Chevelu, J., Lavergne, T., Lepage, Y., and Moudenc, T. 2009. Introduction of a new paraphrase generation tool based on monte-carlo sampling. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP'09) Short Papers*. 249–252.

Chiang, D. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the Association for Computational Linguistics (ACL'05)*. 263–270.

Chiang, D. 2007. Hierarchical phrase-based translation. *Comput. Linguist. 33*, 2, 201–228.

Chiang, D., Knight, K., and Wang, W. 2009. 11,001 new features for statistical machine translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technology Conference (NAACL-HLT'09)*. 218–226.

Chiang, D., Marton, Y., and Resnik, P. 2008. Online large-margin training of syntactic and structural translation features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*.

Curran, J. R. 2004. From distributional to semantic similarity. Ph.D. thesis, School of Informatics, University of Edinburgh, Edinburgh, UK.

Dagan, I., Lee, L., and Pereira, F. 1999. Similarity-based models of cooccurrence probabilities. *Mach. Learn. 34*, 1–3, 43–69.

Das, D. and Smith, N. A. 2009. Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (IJCNLP'09)*. 468–476.

Diab, M. and Finch, S. 2000. A statistical word-level translation model for comparable corpora. In *Proceedings of the Conference on Content-Based Multimedia Information Access (RIAO'00)*.

Du, J., Jiang, J., and Way, A. 2010. Facilitating translation using source language paraphrase lattices. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'10)*. 420–429.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist. 19*, 1, 61–74.

El-Kholy, A. and Habash, N. 2010a. Orthographic and morphological processing for English-Arabic statistical machine translation. In *Proceedings of Traitement Automatique du Langage Naturel (TALN'10)*.

El-Kholy, A. and Habash, N. 2010b. Techniques for arabic morphological detokenization and orthographic denormalization. In *Proceedings of the Conference on Language Resources and Evaluation Workshop on Semitic Languages (LREC'10)*.

Emde Boas, P. V., Kaas, R., and Zijlstra, E. 1977. Design and implementation of an efficient priority queue. *Math. Syst. Theory 10*, 2, 99–127.

Erk, K. and Pado, S. 2008. A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*. 897–906.

FELLBAUM, C., ED. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

FINKELSTEIN, L., GABRILOVICH, E., MATIAS, Y., RIVLIN, E., SOLAN, Z., WOLFMAN, G., AND RUPPIN, E. 2002. Placing search in context: The concept revisited. *ACM Trans. Inf. Syst. 20*, 1, 116–131.

FIRTH, J. R. 1957. A synopsis of linguistic theory 1930–55. In *Studies in Linguistic Analysis*, 1–32.

FUNG, P. AND YEE, L. Y. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the International Conference on Computational Linguistics (COLING'98)*. Association for Computational Linguistics, 414–420.

GALE, W. A. AND CHURCH, K. W. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91)*. 177–184.

HABASH, N. 2008. Four techniques for online handling of out-of-vocabulary words in arabic-english statistical machine translation. In *Proceedings of the Association for Computational Linguistics - Human Language Technology Conference (ACL-HLT'08) Short Papers*. 57–60.

HABASH, N. 2009. Remoov: A tool for online handling of out-of-vocabulary words in machine translation. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*.

HABASH, N. AND HU, J. 2009. Improving arabic-chinese statistical machine translation using english as pivot language. In *Proceedings of the 4th EACL Workshop on Statistical Machine Translation*. 173–181.

HAGHIGHI, A., LIANG, P., BERG-KIRKPATRICK, T., AND KLEIN, D. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of the Association for Computational Linguistics - Human Language Technologies Conference (ACL-HLT'08)*. 771–779.

HARRIS, Z. 1954. Distributional structure. *Word 10*, 23, 146–162.

HASHIMOTO, C., TORISAWA, K., SAEGER, S. D., KAZAMA, J., AND KUROHASHI, S. 2011. Extracting paraphrases from definition sentences on the web. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*. 1087–1097.

HIRST, G. AND BUDANITSKY, A. 2005. Correcting real-word spelling errors by restoring lexical cohesion. *Natural Lang. Engin. 11*, 1, 87–111.

JIANG, J. J. AND CONRATH, D. W. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research on Computational Linguistics (ROCLING X'97)*.

KLEMENTIEV, A., IRVINE, A., CALLISON-BURCH, C., AND YAROWSKY, D. 2012. Toward statistical machine translation without parallel corpora. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*.

KOEHN, P. 2004a. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA'04)*.

KOEHN, P. 2004b. Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

KOEHN, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit (MT-Summit'05)*.

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., ZENS, C. M. R., DYER, C., BOJAR, O., CONSTANTIN, A., AND HERBST, E. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics Demonstration Session (ACL)*.

KOEHN, P., OCH, F. J., AND MARCU, D. 2003. Statistical phrase-based translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technology Conference (NAACL-HLT'03)*. 127–133.

KOTLERMAN, L., DAGAN, I., SZPEKTOR, I., AND ZHITOMIRSKY-GEFFET, M. 2009. Directional distributional similarity for lexical expansion. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (IJCNLP) Short Papers*. 69–72.

KUHN, R., CHEN, B., FOSTER, G., AND STRATFORD, E. 2010. Phrase clustering for smoothing tm probabilities—Or, how to extract paraphrases from phrase tables. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*. 608–616.

LANDAUER, T. K., FOLTZ, P. W., AND LAHAM, D. 1998. Introduction to latent semantic analysis. *Discourse Process. 25*, 259–284.

LEE, J. H., KIM, M. H., AND LEE, Y. J. 1993. Information retrieval based on conceptual distance in IS-A hierarchies. *J. Document. 49*, 2, 188–207.

LESK, M. 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*. 24–26.

LIN, D. 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 (EACL'97)*. 64–71.

LIN, D. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*. 296–304.

LIN, D. AND PANTEL, P. 2001. Discovery of inference rules for question answering. *Natural Lang. Engin.* 7, 4, 343–360.

LIN, J. 2008. Scalable language processing algorithms for the masses: A case study in computing word co-occurrence matrices with mapreduce. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*.

LOPEZ, A. 2007. Hierarchical phrase-based translation with suffix arrays. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning ((EMNLP-CoNLL'07)*. 976–985.

LOPEZ, A. 2008. Machine translation by pattern matching. Ph.D. dissertation, University of Maryland. March.

MADNANI, N., AYAN, N. F., RESNIK, P., AND DORR, B. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the Association for Computational Linguistics Workshop on Statistical Machine Translation (ACL'07)*.

MADNANI, N. AND DORR, B. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Comput. Linguit. 36*, 3.

MALAKASIOTIS, P. 2009. Paraphrase recognition using machine learning to combine similarity measures. In *Proceedings 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'09)*. 27–35.

MANBER, U. AND MYERS, G. 1993. Suffix arrays: A new method for on-line string searches. *SIAM J. Comput. 22*, 5, 935–948.

MANNING, C. D. AND SCHTZE, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

MARTON, Y. 2009. Fine-grained linguistic soft constraints on statistical natural language processing models. Ph.D. dissertation, Department of Linguistics, University of Maryland.

MARTON, Y. 2010. Improved statistical machine translation using monolingual text and a shallow lexical resource for hybrid phrasal paraphrase generation. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas (AMTA)*.

MARTON, Y., CALLISON-BURCH, C., AND RESNIK, P. 2009a. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*.

MARTON, Y., MOHAMMAD, S., AND RESNIK, P. 2009b. Estimating semantic distance using soft semantic constraints in knowledge-source/corpus hybrid models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*.

MARTON, Y., EL-KHOLY, A., AND HABASH, N. 2011. Filtering antonymous, trend-contrasting, and polarity dissimilar distributional paraphrases for improving statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing 6th Workshop on Statistical Machine Translation ((EMNLP-WMT'11)*.

MAX, A. 2010. Example-based paraphrasing for improved phrase-based statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'10)*. 656–666.

MIRKIN, S., SPECIA, L., CANCEDDA, N., DAGAN, I., DYMETMAN, M., AND SZPEKTOR, I. S. 2009. Source-language entailment modeling for translating unknown terms. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP'09)*. 791–799.

MOHAMMAD, S. 2008. Measuring semantic distance using distributional profiles of concepts. Ph.D. thesis, Department of Computer Science, University of Toronto, Toronto, Canada.

MOHAMMAD, S. AND HIRST, G. 2006. Distributional measures of concept-distance: A task-oriented evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'06)*.

MUNTEANU, D. S. AND MARCU, D. 2005. Improving machine translation performance by exploiting nonparallel corpora. *Comput. Linguist. 31*, 4, 477–504.

NAKOV, P. AND NG, H. T. 2011. Translating from morphologically complex languages: A paraphrase-based approach. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*. 1298–1307.

NAVIGLI, R. 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*. 105–112.

OARD, D. W. 1997. Alternative approaches for cross-language text retrieval. In *Proceedings of the AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence.

OCH, F. J. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL'03)*. 160–167.

OCH, F. J. AND NEY, H. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00)*. 440–447.

OCH, F. J. AND NEY, H. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the Association for Computational Linguistics (ACL'02)*.

ONISHI, T., UTIYAMA, M., AND SUMITA, E. 2010. Paraphrase lattice for statistical machine translation. In *Proceedings of the Association for Computational Linguistics (ACL'10) Short Papers*. 1–5.

PAPINENI, K., ROUKOS, S., WARD, T., HENDERSON, J., AND REEDER, F. 2002. Corpus-based comprehensive and diagnostic mt evaluation: Initial arabic, chinese, french, and spanish results. In *Proceedings of the Association for Computational Linguistics - Human Language Technology Conference (ACL-HLT'02)*. 124–127.

PASCA, M. AND DIENES, P. 2005. Aligning needles in a haystack: Paraphrase acquisition across the web. In *Proceedings of the International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (IJCNLP'05)*. 119–130.

PATWARDHAN, S. AND PEDERSEN, T. 2006. Using WordNet based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the Making Sense of Sense EACL Workshop*. 1–8.

RADA, R., MILI, H., BICKNELL, E., AND BLETTNER, M. 1989. Development and application of a metric on semantic nets. *IEEE Trans. Syst. Man Cybernet. 19*, 1, 17–30.

RAHMAN, M. S., ILIOPOULOS, C. S., LEE, I., MOHAMED, M., AND SMYTH, W. F. 2006. Finding patterns with variable length gaps or dont cares. In *Proceedings of the 12th Annual International Conference on Computing and Combinatorics (COCOON'06)*.

RAPP, R. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*. 519–525.

RAVI, S. AND KNIGHT, K. 2011. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics - Human Language Technology (ACL-HLT'11)*.

RESNIK, P. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res. 11*, 95–130.

RESNIK, P., BUZEK, O., HU, C., KRONROD, Y., QUINN, A., AND BEDERSON, B. B. 2010. Improving translation via targeted paraphrasing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'10)*. 127–137.

SAG, I. A., BALDWIN, T., BOND, F., COPESTAKE, A., AND FLICKINGER, D. 2001. Multiword expressions: A pain in the neck for nlp. In *Proceedings of 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLING'01)*. 1–15.

SALTON AND MCGILL. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

SCHROEDER, J., COHN, T., AND KOEHN, P. 2009. Word lattices for multi-source translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09)*. 719–727.

SCHUETZE, H. AND PEDERSEN, J. O. 1997. A cooccurrence-based thesaurus and two applications to information retreival. *Inf. Process. Manag. 33*, 3, 307–318.

SHUTOVA, E. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technology Conference (NAACL-HLT)*. 1029–1037.

SNOVER, M., DORR, B., SCHWARTZ, R., MICCIULLA, L., AND MAKHOUL, J. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA'06)*. 223–231.

SPECIA, L. AND FARZINDAR, A. 2010. Estimating machine translation post-editing effort with hter. In *Proceedings of 2nd Joint EM/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry" (JEC)* Held in conjunction with *the 9th Conference of the Association for Machine Translation in the Americas (AMTA'10)*. V. Zhechev, Ed.

STOLCKE, A. 2002. Srilm—An extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*. Vol. 2, 901–904.

TSENG, H., CHANG, P., ANDREW, G., JURAFSKY, D., AND MANNING, C. 2005. A conditional random field word segmenter. In *Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing*.

TURNEY, P. D. AND PANTEL, P. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res. 37*, 141–188.

VAN DE CRUYS, T. 2009. A non-negative tensor factorization model for selectional preference induction. In *Proceedings of the Workshop on Geometric Models for Natural Language Semantics*. 83–90.

WEEDS, J., WEIR, D., AND MCCARTHY, D. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*. 1015–1021.

ZHANG, Y. AND VOGEL, S. 2005. An efficient phrase-to-phrase alignment model for arbitrarily long phrase and large corpora. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT'05)*.

ZHAO, S., LAN, X., LIU, T., AND LI, S. 2009. Application-driven statistical paraphrase generation. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP'09)*. 834–842.

ZHAO, S., NIU, C., ZHOU, M., LIU, T., AND LI, S. 2008. Combining multiple resources to improve smt-based paraphrasing model. In *Proceedings of the Association for Computational Linguistics (ACL)—Human Language Technology (ACL-HLT'08)*. 1021–1029.

Received March 2011; revised July 2011; accepted November 2011