Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases

Yuval Marton,* **Chris Callison-Burch,†** and **Philip Resnik***

*Department of Linguistics and the CLIP Lab
at the Institute for Advanced Computer Studies (UMIACS)
University of Maryland College Park, MD 20742-7505, USA
{ymarton,resnik}@umiacs.umd.edu

†Computer Science Department, Johns Hopkins University
3400 N. Charles Street (CSEB 226-B) Baltimore, MD 21218
ccb@cs.jhu.edu

Abstract

Untranslated words still constitute a major problem for Statistical Machine Translation (SMT), and current SMT systems are limited by the quantity of parallel training texts. Augmenting the training data with paraphrases generated by pivoting through other languages alleviates this problem, especially for the so-called "low density" languages. But pivoting requires additional parallel texts. We address this problem by deriving paraphrases monolingually, using distributional semantic similarity measures, thus providing access to larger training resources, such as comparable and unrelated monolingual corpora. We present what is to our knowledge the first successful integration of a collocational approach to untranslated words with an end-to-end, state of the art SMT system demonstrating significant translation improvements in a low-resource setting.

1 Introduction

Phrase-based systems, flat and hierarchical alike (Koehn et al., 2003; Koehn, 2004b; Koehn et al., 2007; Chiang, 2005; Chiang, 2007), have achieved a much better translation coverage than word-based ones (Brown et al., 1993), but untranslated words remain a major problem in SMT. For example, according to Callison-Burch *et al.* (2006), a SMT system with a training corpus of 10,000 words learned only 10% of the vocabulary; the same system learned about 30% 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-gram levels is even harder. This problem plays a major part in reducing machine translation quality, as reflected by both automatic measures such as BLEU (Papineni et al., 2002) and human judgment tests. Improving translation coverage accurately is therefore important for SMT systems.

The first solution that might come to mind is to use larger parallel training corpora. However, current state-of-the-art SMT systems cannot learn from non-aligned corpora, while sentence-aligned parallel corpora (bitexts) are a limited resource (See Section 2 for discussion of automatically-compiled bitexts). Another direction might be to make use of non-parallel corpora for training. However, this requires developing techniques to extract alignments or translations from them, and in a sufficiently fast, memory-efficient, and scalable manner. One approach that can, in principle, better exploit both alignments from bitexts and make use of non-parallel corpora is the distributional collocational approach, e.g., as used by Fung and Yee (1998) and Rapp (1999). However, the systems described there are not easily scalable, and require pre-computation of a very large collocation counts matrix. Related attempts propose generating bitexts from comparable and "quasi-comparable" bilingual texts by iteratively bootstrapping documents, sentences, and words (Fung and Cheung, 2004), or by using a maximum entropy classifier (Munteanu and Marcu, 2005). Alignment accuracy remains a challenge for them.

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

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 381–390,
Singapore, 6-7 August 2009.

one side contains the original source language.

The paradigm described in this paper involves constructing monolingual distributional profiles (DPs; a.k.a. word association profiles, or co-occurrence vectors) of out-of-vocabulary words and phrases in the source language; then, generating paraphrase candidates from phrases that co-occur in similar contexts, and assigning them similarity scores. The highest ranking paraphrases are used to augment the translation phrase table. The table augmentation idea is similar to Callison-Burch *et al.*'s (Callison-Burch et al., 2006), but our proposed paradigm does not require using a limited resource such as parallel texts in order to generate paraphrases. Moreover, our proposed paradigm can, in principle, achieve large-scale acquisition of paraphrases with high semantic similarity. However, using parallel training texts in pivoting techniques offers the potential advantage of implicit translational knowledge, in the form of sentence alignments, while our approach is unguided in this respect. Therefore, we conducted experiments to find out how these relative advantages play out. We present here, to our knowledge for the first time, positive results of integrating distributional monolingually-derived paraphrases in an end-to-end state-of-the-art SMT system.

In the rest of this paper we discuss related work in Section 2, describe the distributional hypothesis and distributional profiles in Section 3, and present the monolingually-derived paraphrase generation system 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.

2 Related Work

This is not the first to attempt to ameliorate the out-of-vocabulary (OOV) words problem in statistical machine translation, and other natural language processing tasks. This work is most closely related to that of Callison-Burch *et al.* (2006), who also translate source-side paraphrases of the OOV phrases. There, paraphrases are generated from bitexts of various language pairs, by "pivoting": translating the OOV phrases to an additional language (or languages) and back to the source language. The quality of these paraphrases is estimated by marginalizing translation probabilities to and from the additional language side(s) e, as follows: $p(f_2|f_1) = \sum_e p(e|f_1)p(f_2|e)$. A major disadvantage of their approach is that it relies on the availability of parallel corpora in other languages. While this works for English and many European languages, it is far less likely to help when translating from other source languages, for which bitexts are scarce or non-existent. Also, the pivoting approach is inherently noisy (in both the paraphrase candidates' correct sense, and their translational likelihood), and it is likely to fare poorly with out-of-domain translation. One advantage of the bitext-dependent pivoting approach is the use of the additional human knowledge that is encapsulated in the parallel sentence alignment. However, we argue that the ability to use much larger resources for paraphrasing should trump the human knowledge advantage.

More recently, Callison-Burch (2008) has improved performance of this pivoting technique by imposing syntactic constraints on the paraphrases. The limitation of such an approach is the reliance on a good parser (in addition to reliance on bitexts), but a good parser is not available in all languages, especially not in resource-poor languages. Another approach using a pivoting technique augments the human reference translation with paraphrases, creating additional translation "references" (Madnani et al., 2007). Both approaches have shown gains in BLEU score.

Barzilay and McKeown (2001) extract paraphrases from a monolingual parallel corpus, containing multiple translations of the same source. In addition to the parallel corpus usage limitations described above, this technique is further limited by the small size of such materials, which are even scarcer than the resources in the pivoting case.

Dolan *et al.* (2004) explore generating paraphrases by edit-distance and headlines of time- and topic-clustered news articles; they do not address the OOV problem directly, as their focus is sentence-level paraphrases; although they use a standard SMT measure, alignment error rate (AER), they only report results of the alignment quality, and not of an end-to-end SMT system. Much of the previous research largely focused on morphological analysis in order to reduce type sparseness; Callison-Burch *et al.* (2006) list some of the influential work in that direction.

Work that relies on the distributional hypothesis using bilingual comparable corpora (without the need for bitexts), typically uses a seed lexicon for "bridging" source language phrases

with their target languages paraphrases (Fung and Yee, 1998; Rapp, 1999; Diab and Finch, 2000). This approach is sometimes viewed as, or combined with, an information retrieval (IR) approach, and normalizes strength-of-association measures (see Section 3) with IR-related measures such as TF/IDF (Fung and Yee, 1998). To date, reported implementations suffer from scalability issues, as they pre-compute and hold in memory a huge collocation matrix; we know of no report of using this approach in an end-to-end SMT system.

Another approach aiming to reduce OOV rate concentrates on increasing parallel training set size without using more dedicated human translation (Resnik and Smith, 2003; Oard et al., 2003).

3 Collocational Profiles

The distributional hypothesis and distributional profiles. Natural language processing (NLP) applications that assume the distributional hypothesis (Harris, 1940; Firth, 1957) typically keep track of word co-occurrences in *distributional profiles* (a.k.a. *collocation vectors*, or *context vectors*). Each distributional profile DP_u (for some word u) keeps counts of co-occurrence of u with all words within a usually fixed distance from each of its occurrences (a *sliding window*) in some training corpus. More advanced profiles keep "strength of association" (SoA) information between u and each of the co-occurring words, which is calculated from the counts of u, the counts of the other word, their co-occurrence count, and the count of all words in the corpus (corpus size). The information on the other words with respect to u is typically kept in a vector whose dimensions correspond to all words in the training corpus. This is described in Equation (1), where V is the training corpus vocabulary:

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

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

$$\begin{aligned} &\forall u, v, w \in V, \\ &[semsim(u, v) > semsim(u, w)] \implies \\ &[psim(DP_u, DP_v) > psim(DP_u, DP_w)], \end{aligned} \tag{2}$$

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

The sliding window and word association (SoA) measures. Some researchers count *positional* collocations in a sliding window, i.e., the co-counts and SoA measures are calculated per relative position (e.g., for some word/token u, position 1 is the token immediately after u; position -2 is the token preceding the token that precedes u) (Rapp, 1999); other researchers use *non-positional* (which we dub here *flat*) collocations, meaning, they count all token occurrences within the sliding window, regardless of their positions in it relative to u (McDonald, 2000; Mohammad and Hirst, 2006). We use here flat collocations in a 6-token sliding window. Beside simple co-occurrence counts within sliding windows, other SoA measures include functions based on TF/IDF (Fung and Yee, 1998), mutual information (PMI) (Lin, 1998), conditional probabilities (Schuetze and Pedersen, 1997), chi-square test, and the log-likelihood ratio (Dunning, 1993).

Profile similarity measures. A profile similarity function $psim(DP_u, DP_v)$ is typically defined as a two-place function, taking vectors as arguments, each vector representing a distributional profile of some word u and v, respectively, and whose cells contain the SoA of u (or v) with each word ("collocate") in the known vocabulary. Similarity can be (and have been) estimated in several ways, e.g., the cosine coefficient, the Jaccard coefficient, the Dice coefficient, and the City-Block measure. The formula for the cosine function for similarity measure is given in Eq. (3):

$$psim(DP_u, DP_v) = cos(DP_u, DP_v) = \frac{\sum_{w_i \in V} SoA(u, w_i) SoA(v, w_i)}{\sqrt{\sum_{w_i \in V} SoA(u, w_i)^2} \sqrt{\sum_{w_i \in V} SoA(v, w_i)^2}} \quad (3)$$

In principle, any SoA can be used with any profile similarity measure. However, in practice, only some SoA/similarity measure combinations do well, and finding the 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}). We use here cosine of log-likelihood vectors (McDonald, 2000).

Phrasal distributional profiles. Word DPs can be generalized to phrasal DPs, simply by counting words that co-occur within a sliding window around the target phrase's occurrences (i.e., counting occurrences of words up to 6 words before or after the target phrase). For example, when building a DP for the target phrase *counting words* in the previous sentence, then *simply* is in relative position -2, and *sliding* is in relative position 5. Searching for similar phrasal DPs poses an additional challenge over the word DP case (see Section 4), but there is no additional difficulty in building the phrasal profile itself as described above. In preliminary experiments we found no gain in using phrasal collocates (i.e., count how many times a *phrase* of more than one word co-occurs in a sliding window around the target word/phrase).

4 Searching and Scoring Phrasal Paraphrases

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

Build phrasal profile DP_{phr}. Build a profile of all word collocates, as described in Section 3. Use sliding window of size $MaxPos = 6$. If phr is very frequent (above some threshold of t occurrences), uniformly sample only t occurrences, multiplying the gathered co-counts by factor of $count(phr)/t$. We set $t = 10000$.

Gather context. The challenge in choosing the relevant context is this: if it 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 it 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 part of the context L to be a single word/token to the left of phrase phr. If it is stoplisted, append the next word to the left (now having a bigram left context instead of a unigram), and repeat until the left context is not in the stoplist. Repeat similarly for R, the context to the right of phr. Add the resulting $L\_\_R$ context to a context list. We stoplist "promiscuous" words, i.e., those that have more than $StoplistThreshold$ collocates in the training corpus, using the above $MaxPos$ parameter value. We also stoplist bigrams which occur more than t times and comprise solely from stoplisted unigrams.

Gather candidates. For each gathered context in the context list, gather all paraphrase candidate phrases X that connect left hand side context L with right hand side context R, i.e., gather all X such that the sequence LXR occurs in the corpus. In practice, to keep search complexity low, limit X to be up to length $MaxPhraseLen$. Also, to further speed up runtime, we uniformly sample the context occurrences.

Rank candidates. For each candidate X, build distributional profile DP_X, and evaluate $psim(DP_{phr}, DP_X)$.

Output k-best candidates. Output k-best paraphrase candidates for phrase phr, in descending order of similarity. We set $k = 20$. Filter out paraphrases with score less than $minScore$.

5 Experiment

We examined the application of the system's paraphrases to 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, 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., 2005). The phrase translation probabilities were determined using maximum likelihood estimation over phrases induced from word-level alignments produced by performing Giza++ training (Och and Ney, 2000) on both source and target sides of the parallel training sets. When the baseline system encountered unknown words in the test set, its behavior was simply to reproduce the foreign word in the translated output.

The paraphrase-augmented systems were identical to the corresponding baseline system, with the exception of additional (paraphrase-based) translation rules, and additional feature(s). Similarly to Callison-Burch *et al.* (2006), we added the following feature:

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

Note that it is possible to construct a new translation rule from f to e via more than one pair of source-side phrase and its paraphrase; e.g., if f_1 is a paraphrase of f, and so is f_2, and both f_1, f_2 translate to the same e, then both lead to the construction of the new rule translating f to e, but with potentially different feature scores.

In order to eliminate this duplicity and leverage over these alternate paths which can be used to increase our confidence level in the new rule, we did the following: For each paraphrase f of some source-side phrases f_i, with respective similarity scores $sim(f_i, f)$, we calculated an aggregate score $asim$ with a "quasi-online-updating" method as follows: $asim_i = (1 - asim_{i-1})sim(f_i, f)$, 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. 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.

5.1 English-to-Chinese Translation

For the English-Chinese (E2C) baseline system, we trained on the LCD Sinorama and FBIS tests (LCD2005T10 and LCD2003E14), 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. We then split the bitext to 32 even slices, and constructed a reduced set of about 29,000 lines (sentences) by using only every eighth slice. The purpose of creating this subset model was to simulate a resource-poor language. See Table 1.

Set	# Tokens Source+Target
E2C 29K	0.8 + 0.6
E2C Full	6.4 + 5.1
bnc+apw	187
S2E 10K	0.3 + 0.3
S2E 20K	0.6 + 0.6
S2E 80K	2.3 + 2.3
wmt09	84
wmt09+acquis	139
wmt09+acquis+afp	402

Table 1: Training set sizes (million tokens).

For development, we used the Chinese-English NIST MT 2005 evaluation set, taking one of the English references as source, and the Chinese source as a single reference translation. We tested the system using the English-Chinese NIST MT evaluation 2008 test set with its four reference translations.

We augmented the E2C baseline models with paraphrases generated as described above, training on the British National Corpus (BNC) v3 (Burnard, 2000) and the first 3 million lines of the English Gigaword v2 APW, totaling 187M terms after tokenization, and number and punctuation removal. We generated paraphrases for phrases up to six tokens in length, and used an ar-

bitrary similarity threshold of $minScore = 0.3$. We experimented with three variants: adding a single additional feature for all paraphrases (*1-6grams*); using only paraphrases of unigrams (*1grams*); and adding two features, one only sensitive to unigrams, and the other only to the rest (*1 + 2-6grams*). All features had the same design as described in Section 5, each had an associated weight (as all other features), and all feature weights in each system, including the baseline, were tuned using a separate minimum error rate training for each system.

Results are shown in Table 2. For the E2C systems, for which we had four reference translations for the test set, we used shortest reference length, and used the NIST-provided script to split the output words to Chinese characters before evaluation. Statistical significance for the BLEU results were calculated using Koehn's (Koehn, 2004) pair-wise bootstrapping test with 95% confidence interval.

On the E2C 29,000-line subset, the augmented system had a significant 1.7 BLEU points gain over its baseline. On the full size model, results were negative. Note that our E2C full size baseline is reasonably strong: Its character-based BLEU score is slightly higher than the JHU-UMD system that participated in the NIST 2008 MT evaluation (constrained training track), 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).

5.2 Spanish-to-English Translation

In order to to permit a more direct comparison with the pivoting technique, 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 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 development, we used the Europarl dev2006 Spanish and English sides, and for testing we used the Europarl 2006 test set.

We trained the Spanish paraphrase generation system on the Spanish corpora available from

dataset	E2C model	BLEU	TER
29k	baseline	15.21	90.354
29k	1grams	16.87***	90.370
29k	1-6grams	16.54***	90.376
29k	1 + 2-6grams	**16.88*** **	**90.349**
Full	baseline	**22.17**	**90.398**
Full	1grams	21.64***	90.459
Full	1-6grams	21.75	90.421
Full	1 + 2-6grams	21.39***	90.433

Table 2: E2C Results: character-based BLEU and TER scores. All models have one additional feature over baseline, except for the "1 + 2-6" models that have one feature for unigrams and another feature for bigrams to 6-grams. Paraphrases with score $<$ .3 were filtered out. *** = significance test over baseline with $p < 0.0001$, using Koehn's (2004) pair-wise bootstrap resampling test for BLEU with 95% confidence interval.

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

Table 3: English paraphrases from E2C 29K-bitext systems.

the EACL 2009 Fourth Workshop on Statistical Machine Translation:[1] the Spanish side of the Europarl-v4, news training 2008, and news commentary 2009. We also re-trained adding the JRC-Acquis-v3 corpus[2] to the paraphrase training set, and then adding also 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 1. We generated paraphrases for phrases up to four tokens in length, and used two arbitrary similarity thresholds of $minScore = 0.3$ (as in the E2C experiments), and 0.6, for enforcing only higher precision paraphrasing.

We experimented with these variants: a single feature for all paraphrase (*1-4grams*); using only paraphrases of unigrams (*1grams*); and using two features: one only sensitive to unigrams and bigrams, and the other to the rest (*1-2 + 3-4grams*).

Results are shown in Table 4. We used BLEU over lowercased outputs to evaluate all S2E systems, and Koehn's significance test as above.

On the S2E 10,000-line subset, both the *1grams* and *1-4grams* models achieved significant gains of .4 BLEU points over the baseline. We concluded from a manual evaluation of the 10,000-line models that the two major weaknesses of the baseline system were (not surprisingly) number of untranslated (OOV) words / phrases, followed by number of superfluous words / phrases.

On the larger subset models, no system significantly outperformed the baseline. Note that our S2E baselines' scores are higher than those of Callison-Burch *et al.* (2006), since we evaluate lowercased outputs, instead of recased ones.

6 Discussion and Future Work

We have shown that monolingually-derived paraphrases, based on distributional semantic similarity measures over a source-language corpus, can improve the performance of statistical machine translation (SMT) systems. Our proposed method has the advantage of not relying on bitexts in order to generate the paraphrases, and therefore gives access to large amounts of monolingual training data, for which creating bitexts of equivalent size is generally unfeasible. We haven't trained our system on nearly as large a corpus as it can handle, and indeed we see this as a natural next step.

Results support the assumption that a larger monolingual paraphrase training set yields better paraphrases: our S2E *1-4grams* model performed significantly better than baseline when using *wmt09+aquis* for paraphrasing, but when only using *wmt09*, the model had a smaller advantage that did not reach significance. However, for the S2E *1grams* model, there was a slight decrease in performance when switching paraphrasing corpus from *wmt09+aquis* to *wmt09+aquis+afp*. This effect might be due to the genre or unbalanced content of the additional corpus, or perhaps it is the case that in this corpus size, paraphrases of higher-level ngrams benefitted from the additional text much more than paraphrases of unigrams did. The two rightmost columns in Table 5 show that although Spanish monolingual paraphrases for the unigram *baile* improve when using the larger corpus, (e.g., *danza* and *un balie* become the third and fourth top candidates, pushing much worse candidates far down the list), the two top paraphrase candidates remained unchanged. However, for the 4gram *a favor del informe*, antonymous candidates, which are bad and misleading for translation, are pushed down from the top first and third spots by synonymous, better candidates. Table 3 contains additional examples of good and bad top paraphrase candidates, also in English. Paraphrases of phrases seem to be of lower quality than those of unigrams, as can be seen at the bottom of the table.

These results also show that our method is especially useful in settings involving low-density languages or special domains: The smaller subset models, emulating a resource-poor language situation, show higher gains than larger models (which are supersets of the smaller subset models), when augmented with paraphrases derived from the same paraphrase training set. This was validated in two very different language pairs: English to Chinese, and Spanish to English. We believe that larger monolingual training sets for paraphrasing can help languages with richer resources, and we intend to explore this too.

Although the gains in the Spanish-English subsets are somewhat smaller than the pivoting technique reported in Callison-Burch *et al.* (2006), e.g., .7 BLEU for the 10k subset, we take these results as a proof of concept that can yield better

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

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

bitext	mono.corp.	features	minScore	BLEU	TER
10k	(baseline)	–	–	23.78	62.382
10k	wmt09	1-4grams	.6	23.81	
10k	wmt09	1-2+3-4gr	.6	23.92	62.202
10k	wmt09+aquis	1-4grams	.6	**24.13*****	**61.739**
10k	wmt09+aquis	1grams	.6	24.11	61.979
20k	(baseline)	–	–	24.68	62.333
20k	wmt09+aquis	1-4grams	.6	**24.75**	**61.528**
80k	(baseline)	–	–	**27.89**	57.977
80k	wmt09+aquis	1-4grams	.6	27.82	**57.906**
10k	wmt09+aquis	1grams	.3	**24.11**	61.979
10k	wmt09+aquis+afp	1grams	.3	23.97	61.974
20k	wmt09+aquis+afp	1grams	.3	**24.77**	61.276
80k	wmt09+aquis+afp	1grams	.3	27.84***	**57.781**

Table 4: S2E Results: Lowercase BLEU and TER. Paraphrases with score < *minScore* were filtered out. *** = significance test over baseline with $p < 0.0001$, using Koehn's (2004) pair-wise bootstrap test for BLEU with 95% confidence interval.

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

Table 5: Comparison of Spanish paraphrases: by pivoting, and by two monolingual corpora. Ordered from best to worst score.

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

Table 6: S2E translation examples on 10k-bitext systems. Some translation differences are in bold.

gains with larger monolingual training sets. Pivoting techniques (translating back and forth) rely on limited resources (bitexts), and are subject to shifts in meaning due to their inherent double translation step. In contrast, large monolingual resources are relatively easy to collect, and our system involves only a single translation/paraphrasing step per target phrase. Table 5 also shows an exemplar comparison with the pivoting paraphrases used in Callison-Burch *et al.* (2006). It seems that the pivoting paraphrases might suffer more from having frequent function words as top candidates, which might be a by-product of their alignment "promiscuity". However, the top antonymous candidate problem seems to mainly plague the monolingual distributional paraphrases (but improves with larger corpora). See also Table 6.

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), improving the similarity measure, and refining the similarity threshold. We would like to explore ways of incorporating syntactic knowledge that do not sacrifice coverage as much as in Callison-Burch (2008); incorporating semantic knowledge to disambiguate phrasal senses; using context to help sense disambiguation (Erk and Padó, 2008); and optimizing the similarity threshold for use in SMT, for example on a held-out dataset: too high a threshold reduces coverage, while too low a threshold results in bad paraphrases and translation.

The method presented here is quite general, and therefore different similarity measures, including other corpus-based ones, can be plugged in to generate paraphrases. We are looking into using DPs with word-sense disambiguation: Since it has been shown that similarity is often judged by the semantic distance of the closest senses of the two target words (Mohammad and Hirst, 2006), and that paraphrases generated this way are likely to be of higher quality (Marton et al., 2009), hence it is also likely that the overall performance of an SMT system using them will also improve further.

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

Acknowledgments

Many thanks to Chris Dyer for his help with the E2C set, and to Adam Lopez for his implementation of pattern matching with Suffix Array. This research was partially supported by the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-2-001 and NSF award 0838801, by the EuroMatrixPlus project funded by the European Commission, and by the US National Science Foundation under grant IIS-0713448. The views and findings are the authors' alone.

References

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

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

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

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

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

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

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

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

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

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

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

John R. Firth. 1957. A synopsis of linguistic theory 1930Ű55. *Studies in Linguistic Analysis*, (special volume of the Philological Society):1–32. Distributional Hypothesis.

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

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

Zellig S. Harris. 1940. Review of louis h. gray, foundations of language (new york: Macmillan, 1939). *Language*, 16(3):216Ű231.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hinrich Schuetze and Jan O. Pedersen. 1997. A cooccurrence-based thesaurus and two applications to information retreival. *Information Processing and Management*, 33(3):307Ű318.

Matthew Snover, Bonnie J. Dorr, Richard Schwartz, John Makhoul, Linnea Micciulla, and Ralph Weischedel. 2005. A study of translation error rate with targeted human annotation. Technical Report LAMP-TR-126, CS-TR-4755, UMIACS-TR-2005-58, University of Maryland, July, 2005.

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