Improved Statistical Machine Translation with Hybrid Phrasal Paraphrases Derived from Monolingual Text and a Shallow Lexical Resource

Yuval Marton*
Center for Computational Learning Systems
Columbia University
ymarton@ccls.columbia.edu

Abstract

Paraphrase generation is useful for various NLP tasks. But pivoting techniques for paraphrasing have limited applicability due to their reliance on parallel texts, although they benefit from linguistic knowledge implicit in the sentence alignment. Distributional paraphrasing has wider applicability, but doesn't benefit from any linguistic knowledge. We combine a distributional semantic distance measure (based on a non-annotated corpus) with a shallow linguistic resource to create a hybrid semantic distance measure of words, which we extend to phrases. We embed this extended hybrid measure in a distributional paraphrasing technique, benefiting from both linguistic knowledge and independence from parallel texts. Evaluated in statistical machine translation tasks by augmenting translation models with paraphrase-based translation rules, we show our novel technique is superior to the non-augmented baseline and both the distributional and pivot paraphrasing techniques. We train models on both a full-size dataset as well as a simulated "low density" small dataset.

1 Introduction

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). This work focuses on paraphrasing for SMT. Paraphrasing is useful for SMT because it increases translation coverage – an inherent problem of SMT, due to the Zipfian and dynamic nature of human language.

Untranslated words and phrases, and bad reordering of known words and phrases in unseen larger sequences, remain a major problem for SMT (Callison-Burch et al., 2006). This is the case for both flat and hierarchical phrase-based SMT systems (Koehn et al., 2007; Chiang, 2007, *inter alia*), in spite of much progress since statistical translation models were introduced (Brown et al., 1993).

Recent work proposes augmenting the training data with paraphrases generated by pivoting through other languages and back (Callison-Burch et al., 2006, and subsequent work). This indeed alleviates the vocabulary coverage problem, especially for the resource-poor, so-called "low density" languages. However, it requires one or more extra parallel texts (or more precisely, 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). Some variants also require syntactic annotation (Callison-Burch, 2008). Most other recent techniques require supervised training (see Section 2), resources for which are also scarce in "low density" languages.

To overcome this resource constraint, the approach in Marton et al. (2009a) proposes augmenting the training data with paraphrases generated by using distributional techniques on a large monolingual corpus – a relatively abundant resource. It constructs monolingual distributional profiles (DPs; see Section 3.1) of words and phrases in the source language that are out-of-vocabulary (OOV) for the translation model. It then generates paraphrase candidates from phrases that co-occur in similar contexts, and estimates their semantic similarity to the paraphrased term by applying distributional semantic distance measures. While this approach alleviates the dependency on scarce and costly resources, it lacks the human linguistic knowledge implicit in the sentence alignment of parallel texts.

The technique in Marton et al. (2009a) is extended here by using human linguistic knowledge, yet still without relying on parallel texts. This is done by replacing the distributional semantic distance measure

*Much of this work was done when the author was at the University of Maryland.

with a hybrid measure (Marton et al., 2009b). This hybrid measure combines a large monolingual corpus of text with a lexical resource in order to approximate word senses without using sense-annotated texts (as these are also scarce and costly). This measure, originally applying to word-pairs, is extended here to apply to phrase-pairs, so it is more useful for augmenting phrase-based SMT. Our hybrid approach benefits from both worlds – generating paraphrases monolingually-distributionally (without parallel texts), while incorporating linguistic knowledge. We show it can out-perform both pivoting and distributional paraphrasing techniques. We present here, to our knowledge for the first time, positive results of integrating unsupervised hybrid paraphrases in an end-to-end state-of-the-art SMT system, trained on a small subset dataset (simulating a "low-density" language). We also present new positive results using both distributional and hybrid paraphrases in models trained on a full-size dataset.

In the rest of this paper we describe paraphrasing techniques in Section 2, distributional and hybrid semantic distance measures in Section 3, and the translation model augmentation technique 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 paper brings together various sub-fields, we discuss related work in each of the relevant sections.

2 Paraphrase Generation

2.1 Paraphrasing Approaches

Paraphrasing is the act of replacing linguistic utterances (typically text) with other linguistic utterances, bearing similar meaning but different form. Paraphrasing research is quite diverse, and can be characterized and classified along many axes, including: paraphrasing unit (word, phrase, sentence, passage), paraphrased elements (lexical synonyms, or structural, such as active/passive voice), required resources (parallel, comparable, or monolingual text), and technique (pivoting or distributional). Paraphrasing may be somewhat "lossy" in number of words and/or content, with the extreme cases of summarization and translation. Due to space limitations, we only mention here by name the most similar and recent work. Madnani and Dorr (2010) give more on different types of paraphrasing.

Previous work largely rely on parallel text or SMT in order to generate paraphrases. Barzilay and McKeown (2001) use direct translation for this. They extract paraphrases from a monolingual parallel corpus, containing multiple translations of the same source. However, monolingual parallel corpora are extremely rare and small. Zhao et al. (2008) apply SMT-style 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.

The leading SMT-related paraphrasing technique is currently the "pivoting" technique, especially Bannard and Callison-Burch (2005) and Callison-Burch et al. (2006). "Pivoting" here means translating the phrases of interest to one or more languages and back to the source language. This is illustrated in Figure 1. The quality of these paraphrases is estimated by marginalizing translation probabilities to and from the additional language side (or sides) e, as follows: $p(f_2|f_1) \approx \sum_e p(e|f_1)p(f_2|e)$, 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 (e.g., with EuroParl), it is far less likely to help when translating from other languages, for which bitexts are scarce or non-existent. Also, the inherent double translation step introduces noise in both the paraphrase candidates' desired sense, and their translational likelihood. More on that in Section 6. The problem of incorrect sense translation is likely to be exacerbated when the test set is of a different genre than the bitexts. One of the advantages of pivoting, however, is the use of linguistic knowledge that is encapsulated in the parallel sentence alignment.

More recently, Callison-Burch (2008) has improved performance of this pivoting technique by imposing syntactic constraints on the paraphrases. In one variant the target phrase and its paraphrase are constrained to have the same parsing tag (e.g., NP), and in another variant, this constraint has been re-

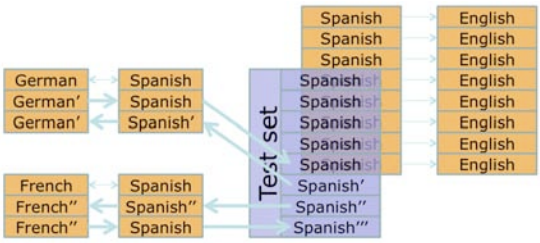

Figure 1: Pivoting technique for paraphrase generation. Paraphrase OOV (Spanish) terms in a SMT model by pivoting through other languages such as French or German.

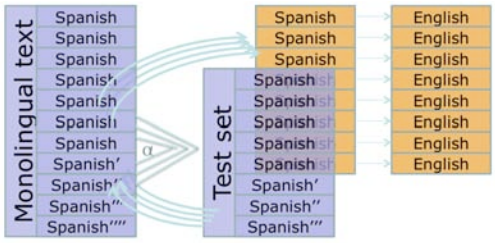

Figure 2: Distributional paraphrase generation. Paraphrase OOV (Spanish) terms in a SMT model using a distributional semantic distance measure and a large monolingual corpus, which is a relatively abundant resource.

laxed so that the phrase and its paraphrase must have the same Combinatory Categorial Grammar (CCG) super-tag sequence, but no longer need to have the same single constituent tag. The limitation of such an approach, in either variant, is the reliance on a good parser (in addition to reliance on bitexts), since a good parser is not available in all languages, especially not in resource-poor languages. Also, parsing large corpora is computationally taxing.

Our work uses monolingual text in order to generate phrasal paraphrases with distributional techniques, combined with semantic information from a lexical resource. OOV phrases in the source language are paraphrased and then used to augment a SMT translation model (Details in Sections 4). Our previous paraphrasing work shows gains in SMT, only using a distributional method over textual resources. Here, when adding linguistic knowledge (from a lexical resource) on top of same amount of textual resources as the distributional method, our hybrid method yields further gains. We argue in Marton et al. (2009a) that the ability to use much larger textual resources for paraphrasing should allow improving also on larger translation models. Using larger textual resources, we show here gains over "full-size" models that we failed to improve there.

2.2 Monolingually Derived Distributional Paraphrase Generation

We have recently introduced a monolingual corpus-based paraphrasing technique (Marton et al., 2009a). This technique makes use of distributional profiles (DPs; see Section 3) of OOV phrases and their paraphrasing candidates. Its outline is this:

1. Upon receiving OOV phrase phr, build distributional profile DP_{phr}.
2. Gather contexts: for each occurrence of phr, keep surrounding (left and right) context $L\_\_R$.
3. For each such context, gather paraphrase candidates $cand$ which occur between L and R in other locations in the training corpus, i.e., all $cand$ such that $L\ cand\ R$ occur in the corpus.
4. For each candidate $cand$, build a profile DP_{cand} and measure profile similarity between DP_{cand} and DP_{phr}. Rank $cand$ according to the profile similarity score.
5. Output k-best candidates above a certain similarity score threshold.

This is illustrated in Figure 2. The technique we now present is different from that in Marton et al. (2009a) in the ranking step 4, since in our model, phr and $cand$ may have multiple senses each. These senses correspond to multiple hybrid sense-aware DPs for each of phr and $cand$, as opposed to a single "vanilla" corpus-based DP for each. We extend the profile similarity function to choose among the different senses, similarly to Mohammad and Hirst (2006). See Section 3.2 for details.

3 Semantic Distance Measures

Various paraphrasing techniques rely on semantic distance measures. We now turn to briefly survey such measures. We group them as follows: lexical resource-based, corpus-based, and hybrid. We do not discuss the widely used WordNet and other lexical resource-based measures here, due to space constraints. We only refer the reader to Hirst and Budanitsky (2005) for a comprehensive survey.

3.1 Corpus-based measures

Corpus-based measures of distributional similarity rely on the distributional hypothesis (Harris, 1954): Words close in meaning tend to appear in similar distribution (surrounding contexts). The distributional profile (DP) of word or phrase 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. Beside simple co-occurrence counts within sliding windows, other SoA measures are based on TF/IDF, mutual information (PMI), conditional probabilities, and the log-likelihood ratio.

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 word/phrase v), whose size is the known vocabulary size. Similarity can be estimated in several ways, e.g., the cosine coefficient, the Jaccard coefficient, the Dice coefficient (all proposed by Salton and McGill, 1983), α-skew divergence (Dagan et al., 1999), and the City-Block measure (Rapp, 1999). The cosine is especially appealing. It is a proven measure, 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. In principle, any SoA can be used with any profile similarity measure, but only some combinations do well. Other measures are directional (textual entailment) in u and v. See Weeds et al. (2004) for surveys of distributional measures.

3.2 Hybrid measures

As Mohammad and Hirst (2006) point out, the DP of a word u conflates information about the senses of u. For example, assume the noun *bank* has two senses: RIVER (as in *riverbank*) and FINANCIAL INSTITUTION, and the noun *wave* has two senses: RIVER and PHYSICS. 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 RIVER senses. Mohammad and Hirst (2006) overcome the sense-conflation problem by generating separate DPs for the different senses of a word, using the categories in a Roget-style thesaurus as coarse senses or concepts. A word may be found in more than one category c if it has multiple meanings. They use a simple unsupervised algorithm to determine concept-based vectors **DPC**(c): each cell of the DPC vector corresponds to each unique word w in a corpus, and contains the SoA between w and the category c, based on the number of times w co-occurred with any of the words associated with c.

In Marton et al. (2009b) we observe that if target words u and v appear under the same concept c, the semantic distance between u and v would be indistinguishable, since the concept-based similarity measure returns the semantic distance of the closest sense pair. In the example above, *bank* and *wave* have two senses each, so there are $2 \times 2 = 4$ DPC pairs to compare:

FINANCIAL INSTITUTION, PHYSICS
FINANCIAL INSTITUTION, RIVER
RIVER, PHYSICS
RIVER, RIVER

The last, identical pair would be returned, falsely representing synonymy between *bank* and *wave*. This issue is addressed in Marton et al. (2009b) using a hybrid approach with fine-grained soft constraints, called "hybrid-sense-proportional". For word u in sense c, the co-occurrence counts in the DP of u are discounted according to the counts in the DPC of c, and then SoA measures in the DP of u_c are calculated 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 co-occurrence frequencies in DPCs, and the co-occurrence count $f(u, w_i)$ is calculated from word-based DPs. The word-sense-biased DP is denoted **DPWS**. A word that has no mapping to any concept c is assumed to have uniform distribution over all concepts (but in practice, we use a single sense), thus freeing us from the thesaurus vocabulary limitations. Figure 3 visualizes a toy DPWS of *bank*, created from the "vanilla" DP of *bank*, biased towards RIVER.

This method is further extended here, to handle not only words but phrases too, evaluated in a SMT setting, although applicable in other NLP settings as well. For phrase phr, define its DP's sliding window of $\pm n$ to include the n tokens immediately preceding the first token of phr, and the n tokens immediately following the last token of phr. DP similarity is calculated as in the traditional word-DP case. With this extended hybrid semantic distance measure, when augmenting translation models, we replace the ranking step 4 (Section 2.2) with:

- Build a sense-aware profile for each candidate $cand$, and measure profile similarity between $DPWS^s_{cand}$ and $DPWS^r_{phr}$ for each sense s of $cand$, and each sense r of phr. Rank can-

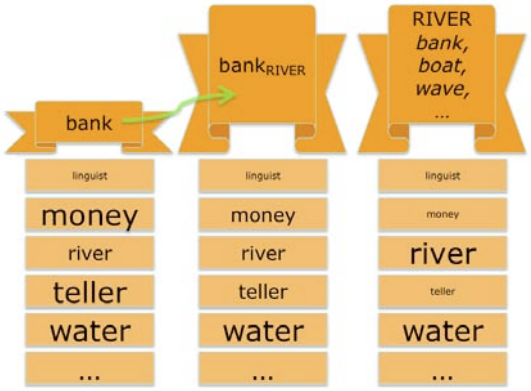

Figure 3: Visual example of a sense-aware distributional profile: the DPWS for the word *bank* in sense RIVER. The *bank*'s strength of association with *money* in the DPWS is decreased relative to the DP, since it is discounted in proportion to its value in the DPC of RIVER, relative to its value in all the DPCs of *bank*.

didates by profile similarity score of the closest $(DPWS^s_{cand}, DPWS^r_{phr})$ pair of each $cand$.

Note that although a thesaurus might not exist for all languages or domains, a coarse thesaurus (which is sufficient for our purpose here) is much more likely to exist than a highly – or even moderately – developed WordNet. Hence our hybrid technique is more applicable than WordNet-based techniques.

Other hybrid measures: 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. But it relies on dependency relations and selectional preferences information, which might be of low quality or unavailable in a low density language.

Resnik (1999) introduces a hybrid model for calculating "information content" by traversing the concept's subtree in WordNet. This measure is hybrid in that it uses both a linguistic knowledge source and a large corpus of text, although it doesn't use the distributional contexts of the words in the corpus. Lin (1997) and Jiang and Conrath (1997) improve on this idea by incorporating the distance of each word from the lowest common subsumer. However, a wordnet might not exist, or not be sufficiently developed, in a low density language.

4 Paraphrase-Augmented SMT

This is not the first attempt to ameliorate the out-of-vocabulary (OOV) words problem in statistical machine translation, and other NLP tasks. Such attempts can be roughly divided as follows:

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

Our work belongs to the first category, and therefore we mainly focus here on this category. More specifically, our work most resembles Callison-Burch et al. (2006) in augmenting translation models with source-side paraphrases of the OOV phrases, using weighted log-linear features. Given an OOV source-side phrase f, if the translation model has a rule $\langle f', e\rangle$ whose source side is a paraphrase f' of f, then a new rule $\langle f, e\rangle$ is added, with an extra weighted log-linear feature(s), whose value for the new rule is the translation probability or similarity score between f and f'. The definition is this:

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

where $asim$ is defined below. As noted there, 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 leverage on these paths and resolve feature value conflicts, we apply an aggregated similarity measure: For each paraphrase f of source-side phrases f_i with similarity scores $sim(f_i, f)$,

$$asim_i = asim_{i-1} + (1 - asim_{i-1})\, sim(f_i, f) \quad (2)$$

where $asim_0 = 0$. 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 $sim(f_i, f)$ was the highest.

Other related work: Habash and Hu (2009) show, pivoting via a trilingual parallel text, that using English as a pivot language between Chinese

and Arabic outperforms translation using a direct Chinese-Arabic bilingual parallel text. They suggest this might be because 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 phrase table's source side include Habash (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 for generating paraphrases. They improve SMT coverage by using a manually crafted monolingual HPSG grammar for generating meaning and grammar-preserving paraphrases. They parse the English side and then convert it to an abstract semantic representation and back to English. This grammar allows for certain word reordering, lexical substitutions, contractions, and "typo" corrections.

5 Experiments

We examined augmenting translation models with paraphrases based on hybrid semantic distance measures. We contrasted these models with models using distributional distance measures, models using pivot-style paraphrases, and non-augmented baseline models. We tested all models in English-to-Chinese translation, augmenting the models with translation rules for unknown English phrases.

For baseline we used the phrase-based SMT system Moses (Koehn et al., 2007), with the default model features: 1. phrase translation probability, 2. reverse phrase translation probability, 3. lexical translation probability, 4. reverse lexical translation probability, 5. word penalty, 6. phrase penalty, 7. six lexicalized reordering features, 8. distortion cost, and 9. language model (LM) probability. We used Giza++ (Och and Ney, 2000) for word alignment. 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 $\uparrow$ and TER$\downarrow$ in Table 1.

The paraphrase-augmented models were created as described in Sections 3.2 and 4. We used cosine distance over DPs of log-likelihood ratios (McDonald, 2000), built with a sliding window of size ± 6, a sampling threshold of 10000 occurrences, and a maximal paraphrase length of 6 tokens. 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 and no more than 2,000 occurrences. For each 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. We experimented with three variants:

- adding an extra single 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 2-6-grams (*1 + 2-6grams*).

All features were designed as described above. Each model's feature weight set was tuned with a separate minimum error rate training. We repeated this process with distributional and pivoting-style paraphrases, for comparison.

5.1 Data

In order to compare the quality of paraphrases generated with pivoting, distributional, and hybrid techniques, we chose English as the source language for the translation task. This is because the new technique requires semantic knowledge base of the source language, and such data, based on the English *Macquaries* thesaurus, was at our disposal (see Marton et al., 2009b). We chose Chinese as the translation target language because it is quite different from English (e.g., in word order), and four reference translation were available from NIST.

For training we used 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

English-Chinese model	BLEU ↑	TER↓
29k-sentence training subset models		
baseline	15.2	69.3
1gram-pivot	15.5	69.4
1-5gram-pivot	16.1[B]	**69.0**
1+2-5gram-pivot	**16.2**[BI]	69.1
1gram-distrib	**16.9**[B]	**68.8**
1-6gram-distrib	16.5[B]	69.2
1+2-6gram-distrib	**16.9**[BC]	**68.8**
1gram-hybrid	16.4[B]	69.0
1-6gram-hybrid	16.7[BD]	68.8
1+2-6gram-hybrid	**17.0**[BCDI]	**68.7**
"full" 232k-sentence training dataset models		
baseline	21.8	**63.8**
1gram-distrib	**22.5**[B]	64.4
1-5gram-distrib	**22.5**[B]	66.2
1+2-5gram-distrib	21.7	**63.9**
1gram-hybrid	**22.7**[BD]	63.9
1-5gram-hybrid	22.3	63.9
1+2-5gram-hybrid	22.3[D]	**63.8**

Table 1: Results: character-based BLEU and TER scores. All models have one extra feature on top of their baseline model's features, except for the "1+2…" models, which have one extra feature for unigrams and another for longer n-grams. Statistical significance from corresponding [B]: baseline, [D]: distributional model, [I]: "1gram" model, or [C]: from coarser "1-5/1-6gram" model, $p < .05$.

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

5.2 Experimental Results

Translation evaluation is given in Table 1. We used the NIST-provided script to split the output words to Chinese characters before evaluation, as is standardly done in the NIST English-Chinese translation task official evaluation.[1] Statistical significance for the BLEU results was calculated using Koehn's paired bootstrap re-sampling test (Koehn, 2004), with a sample size of 2000 pairs; it was determined in case the 95% confidence interval (CI) of the systems' BLEU score difference excluded zero. For conciseness, this is denoted as $p < .05$. We used shortest reference length.

Augmentation with pivot-style paraphrases: Due to memory limitations, it was not possible to use all pivot-style paraphrases.[2] We therefore filtered out paraphrases below a .3 score threshold. Note, however, that this threshold is not equivalent to a .3 score threshold used in the distributional and hybrid paraphrasing methods. In addition to using all available lengths (unigram to 5-gram) of paraphrased phrases, we also experimented with *1grams-pivot* and *1+2-5grams-pivot* models, corresponding to the *1grams** and *1+2-6grams** (distributional or hybrid) models, respectively. The *1-5grams-pivot* and *1+2-5grams-pivot* models showed significant gains up to 1 BLEU point over the baseline, serving as stronger baselines for our technique. The unigram *1grams-pivot* model's TER score was slightly worse than the baseline (but recall it was threshold-filtered).

Augmentation with distributional paraphrases: We repeat here the results in Marton et al. (2009a) for the distributional models, which yielded up to 1.7 BLEU points significant gain over the baseline on the 29,000-line subset. All TER scores were also better than the baseline's. For new results on the full size set, see the end of this section.

Augmentation with hybrid paraphrases: Our claim for the hybrid semantic distance measure' advantage is supported not only by gains in SMT performance over the baseline, but also over the pivot and distributionally-augmented models. The third part in Table 1 shows that for the 29,000-line subset, each hybrid-augmented model did better than the baseline (up to 1.8 BLEU points), its pivot counterpart, and slightly yet significantly better than its distributional counterpart (except for *1gram-hybrid vs. 1gram-distrib*). TER scores follow similar patterns here as well. See Section 6 and Table 2 for further discussion and examples.

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

[2] We used the paraphrases that were not filtered by syntactic criteria, as available from http://www.cs.jhu.edu/~ccb/howto-extract-paraphrases.html

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

Table 2: English-Chinese translation examples on 29k-bitext models. Some translation differences are in bold. Hybrid model correctly translates *can*; distrib. model mis-translates it as *the month of May*; it remains OOV for other models.

Augmentation of full size translation models: Following reviewers' concerns about applicability to the full size model, we next report results with a newer version of our techniques, using a larger monolingual text of over 516M tokens, consisting of the Gigaword documents from 2004 and 2008 (LDC2009T13), pre-processed slightly differently (conflating numbers, dates, months, days of week, and alphanumeric tokens to their respective classes). These techniques have many parameters, and it is impractical to conduct controlled experiments with each parameter. We list the most important differences: a lower score threshold of 0.05, since we observe that many low-score paraphrase candidates still seem good, and are sometimes the only candidates; a dynamic context length (the shortest non-stoplisted left context L occurring less than 512 times in the corpus, and similarly for R); paraphrasing OOV phrases up to 5 tokens in length (to match the maximal length of available pivot paraphrases); excluding paraphrase candidates occurring less than 25 times (inspired by McDonald, 2000); excluding "textually entailed" paraphrase candidates whose words all appear in phr in the same order; and a higher limit on k-best paraphrases, $k = 100$, to compensate for the harsher candidate filtering. The lower half of Table 1 shows the distributional and hybrid models outperformed the baseline by up to .6 and .9 BLEU points, respectively (except for the *1+2-5gram-distrib* model). Each hybrid model almost always outperformed its distributional counterpart as well. This is the first time distributional and hybrid models are reported to yield gains over a full size translation model (although not in TER).

6 Discussion and Future Work

We showed that augmenting translation models with paraphrases that were generated with hybrid semantic distance measures, yielded best improvements in almost all cases, compared with baseline, distributional and pivot-style paraphrases. We also showed for the first time gains over full size baseline model for both distributional and hybrid paraphrases. Still, a natural next step for us would be to use an even larger monolingual corpus and more fine-grained or otherwise effective concepts/senses and hybrid methods. Schroeder et al. (2009) recently showed that the upper bound for gains by parahrase augmentation (using human-generated paraphrases in a lattice of the source language) is high, and has not been not reached yet. We take their work as another validation of this research direction.

Pivot-style paraphrasing methods rely on limited resources (bitexts), and are subject to shifts in meaning and inaccurate translation probability estimation due to their inherent double translation step. A related potential problem is a probability mass "leakage": if some pivot phrase is more polysemous, then there might be more bad paraphrase candidates than with a less polysemous phrase; even if the bad candidates score low, they might result in varyingly lower probability estimates for the better candidates, making the paraphrase probability estimate less reliable. It is unclear how to fairly compare the pivoting technique to ours. Should the monolingual and bilingual training resources be equivalent in some sense? Large bitexts are rare; EuroParl-based pivoting is only applicable to European languages. Should the lengths of the phrase or its paraphrase be limited to the same range in both techniques? Should pivot-

ing paraphrases be threshold-filtered as the distributional and hybrid ones are? Should the number of paraphrases per technique be similar? Perhaps each technique should be presented in its best light. But finding the best running parameters for each technique is not a simple matter either. Therefore, the comparisons here should be regarded as a first stab only, inviting further research.

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), improving the similarity measure; using context to help sense disambiguation (Erk and Padó, 2008); and optimizing the similarity threshold for use in SMT, for example on a held-out dataset: the higher the threshold the lower the coverage, while the lower the threshold the lower the paraphrases and translation quality. It remains to be seen how these two opposite effects play out.

Fine-grained features almost always proved advantageous. Among the subset models, *1+2-6grams-hybrid* was the best hybrid performer, significantly better than both the coarser *1-6grams-hybrid*, and the less informed *1grams-hybrid*. Similarly, *1+2-6grams-distrib* was the best distributional model. This pattern repeated for the *1+2-5grams-pivot* model, although its advantage over the coarse *1-5grams-pivot* did not reach significance.

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. Therefore, separate features for longer phrases are likely to be of low quality or marginal impact, while increasing runtime.

A further goal in the future would be to create a distributional similarity-based, high-performance SMT system (or hybrid-based system when possible), 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.

Conclusions

We have shown that augmenting SMT models with monolingually derived paraphrases, using hybrid (lexical resource / corpus-based) semantic distance measures, out-perfroms using distributional and pivot paraphrases in almost all cases. Our method has the advantage of not relying on parallel or sense-annotated texts for paraphrase generation, and therefore can exploit large amounts of monolingual training data, for which creating bitexts of equivalent size is generally unfeasible. Unlike the distributional method, it also benefits from human linguistic knowledge.

Acknowledgments

Many thanks to Chris Dyer for his help with the bitext; Adam Lopez for his code for pattern matching with Suffix Array; and Chris-Callison-Burch and Philip Resnik for their advice. 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 Euro-MatrixPlus 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

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proc. ACL*, pages 597–604, Ann Arbor, Michigan.

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

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

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

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 *Proc. EMNLP*, Waikiki, Hawai'i.

Jonathan Chevelu, Thomas Lavergne, Yves Lepage, and Thierry Moudenc. 2009. Introduction of a new paraphrase generation tool based on monte-carlo sampling. In *Proc. ACL - IJCNLP Short Papers*, pages 249–252, Suntec, Singapore.

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

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

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proc. EMNLP*, pages 897–906, Honolulu, HI.

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

Nizar Habash. 2009. REMOOV: A tool for online handling of out-of-vocabulary words in machine translation. In *Proc. the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, Cairo, Egypt.

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

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

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

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 *ACL demonstration session*, Prague, Czech Republic.

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

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

Dekang Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proc. EACL*, pages 64–71, Madrid, Spain.

Nitin Madnani and Bonnie Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3).

Prodromos Malakasiotis. 2009. Paraphrase recognition using machine learning to combine similarity measures. In *Proc. ACL - IJCNLP Student Research Workshop*, pages 27–35, Suntec, Singapore.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009a. Improved statistical machine translation using monolingually-derived paraphrases. In *Proc. EMNLP*, Singapore.

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

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

Shachar Mirkin, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor . 2009. Source-language entailment modeling for translating unknown terms. In *Proc. ACL - IJCNLP*, pages 791–799, Suntec, Singapore.

Saif Mohammad and Graeme Hirst. 2006. Distributional measures of concept-distance: A task-oriented evaluation. In *Proc. EMNLP*, Sydney, Australia.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proc. ACL*.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. 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 *Proc. ACL - HLT*, pages 124–127, San Diego, CA.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proc. ACL.*, pages 519–525.

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.

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

Josh Schroeder, Trevor Cohn, and Philipp Koehn. 2009. Word lattices for multi-source translation. In *Proc. EACL*, pages 719–727, Athens, Greece.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. AMTA*, pages 223–231, Cambridge, MA.

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

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*.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proc. COLING*, pages 1015–1021, Geneva, Switzerland.

Shiqi Zhao, Cheng Niu, Ming Zhou, Ting Liu, and Sheng Li. 2008. Combining multiple resources to improve smt-based paraphrasing model. In *Proc. ACL - HLT*, pages 1021–1029, Columbus, Ohio, USA.

Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven statistical paraphrase generation. In *Proc. ACL - IJCNLP*, pages 834–842, Suntec, Singapore.