Filtering Antonymous, Trend-Contrasting, and Polarity-Dissimilar Distributional Paraphrases for Improving Statistical Machine Translation

Yuval Marton
T.J. Watson Research Center
IBM
yymarton@us.ibm.com

Ahmed El Kholy and **Nizar Habash**
Center for Computational Learning Systems
Columbia University
{akholy,habash}@ccls.columbia.edu

Abstract

Paraphrases are useful for statistical machine translation (SMT) and natural language processing tasks. Distributional paraphrase generation is independent of parallel texts and syntactic parses, and hence is suitable also for resource-poor languages, but tends to erroneously rank antonyms, trend-contrasting, and polarity-dissimilar candidates as good paraphrases. We present here a novel method for improving distributional paraphrasing by filtering out such candidates. We evaluate it in simulated low and mid-resourced SMT tasks, translating from English to two quite different languages. We show statistically significant gains in English-to-Chinese translation quality, up to 1 BLEU from non-filtered paraphrase-augmented models (1.6 BLEU from baseline). We also show that yielding gains in translation to Arabic, a morphologically rich language, is not straightforward.

1 Introduction

Paraphrase recognition and generation has proven useful for various natural language processing (NLP) tasks, including statistical machine translation (SMT), information retrieval, query expansion, document summarization, and natural language generation. We concentrate here on phrase-level (as opposed to sentence-level) paraphrasing for SMT. Paraphrasing is useful for SMT as it increases translation coverage – a persistent problem, even in large-scale systems. Two common approaches are "pivot" and distributional paraphrasing. Pivot paraphrasing translates phrases of interest to other languages and back (Callison-Burch et al., 2006; Callison-Burch, 2008). It relies on parallel texts (or translation phrase tables) in various languages, which are typically scarce, and hence limit its applicability. Distributional paraphrasing (Marton et al., 2009) generates paraphrases using a distributional semantic distance measure computed over a large monolingual corpus.[1] Monolingual corpora are relatively easy and inexpensive to collect, but distributional semantic distance measures are known to rank antonymous and polarity-dissimilar phrasal candidates high. We therefore attempt to identify and filter out such ill-suited paraphrase candidates.

A phrase pair may have a varying degree of antonymy, beyond the better-known complete opposites (*hot / cold*) and contradictions (*did / did not*), e.g., weaker contrasts (*hot / cool*), contrasting trends (*covered / reduced coverage*), or sentiment polarity (*happy / sad*). Information extraction, opinion mining and sentiment analysis literature has been grappling with identifying such pairs (Pang and Lee, 2008), e.g., in order to distinguish positive and negative reviews or comments, or to detect contradictions (Marneffe et al., 2008; Voorhees, 2008). We transfer some of the insights, data and techniques to the area of paraphrasing and SMT. We distributionally expand a small seed set of antonyms in an unsupervised manner, following Mohammad et al. (2008). We then present a method for filtering antonymous and polarity-dissimilar distributional paraphrases using the expanded antonymous list and a list of negators (e.g., *cannot*) and trend-decreasing words (*reduced*). We evaluate the impact of our approach in a SMT setting, where non-

[1] Other variants use a lexical resource in conjunction with the monolingual corpus (Mirkin et al., 2009; Marton, 2010).

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 237–249,
Edinburgh, Scotland, UK, July 30–31, 2011.

baseline translation models are augmented with distributional paraphrases. We show gains of up to 1 BLEU relative to non-filtered models (1.6 BLEU from non-augmented baselines) in English-Chinese models trained on small and medium-large size data, but lower to no gains in English-Arabic. The small training size simulates resource-poor languages.

The rest of this paper is organized as follows: We describe distributional paraphrase generation in Section 2, antonym discovery in Section 3, and paraphrase-augmented SMT in Section 4. We then report experimental results in Section 5, and discuss the implications in Section 6. We survey related work in Section 7, and conclude with future work in Section 8.

2 Distributional Paraphrases

Our method improves on the method presented in Marton et al. (2009). Using a non-annotated monolingual corpus, our method constructs distributional profiles (DP; a.k.a. context vectors) of focal words or phrases. Each DP_{phr} is a vector containing log-likelihood ratios of the focal phrase phr and each word w in the corpus. Given a paraphrase candidate phrase $cand$, the semantic distance between phr and $cand$ is calculated using the cosine of their respective DPs (McDonald, 2000). For details on DPs and distributional measures, see Weeds et al. (2004) and Turney and Pantel (2010).

The search of the corpus for paraphrase candidates is performed in the following manner:

1. For each focal 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}.
5. Rank all $cand$ according to the profile similarity score.
6. 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.
7. Filter out every candidate $cand$ that is antonymous to phr (See Algorithm 1 below).
8. Output k-best remaining candidates above a certain similarity score threshold t.

Most of the steps above are similar to, and have been elaborated in, Marton et al. (2009). Due to space limitations, we concentrate on the main novel element here, which is the antonym filtering step, detailed below. Antonyms (largely speaking) are opposites, terms that contrast in meaning, such as *hot / cold*. Negators are terms such as *not* and *lost*, which often flip the meaning of the word or phrase that follows or contains them, e.g., *confidence / lost confidence*. Details on obtaining their definitions and on obtaining the antonymous pair list and the negator list are given in Section 3.

Algorithm 1 Antonymous candidate filtering

```
Given an antonymous pair list, a negator list, and a phrase-paraphrase candidate ($phr$-$cand$) pair list,
for all $phr$-$cand$ pairs do
  for all words $w$ in $phr$ do
    if $w$ is also in $cand$, and there is a negator up to two words before it in either $phr$ or $cand$ (but not both!) then
      filter out this pair
    if $w, ant$ is an antonymous pair, and $ant$ is in $cand$, and there is no negator up to two words before $w$ and $ant$, or there is such a negator before both then
      filter out this pair
```

3 Antonyms, Trends, Sentiment Polarity

Native speakers of a language are good at determining whether two words are antonyms (*hot–cold, ascend–descend, friend–foe*) or not (*penguin–clown, cold–chilly, boat–rudder*) (Cruse, 1986; Lehrer and Lehrer, 1982; Deese, 1965). Strict antonyms apart, there are also many word pairs that exhibit some degree of contrast in meaning, for example, *lukewarm–cold, ascend–slip,* and *fan–enemy* (Mohammad et al., 2008). Automatically identifying such contrasting word pairs has many uses including detecting and generating paraphrases (*The lion* ***caught*** *the gazel / The gazel could* ***not escape*** *the lion*) and detecting contradictions (Marneffe et al., 2008; Voorhees, 2008) (*The inhabitants of Peru are* ***well off*** */ the inhabitants of Peru are* ***poor***). Of course, such "contradictions" may be a result of differing sentiment, new information, non-coreferent mentions, or genuinely contradictory statements. Identifying paraphrases and contradictions are in turn useful in effectively re-ranking target language hypotheses in machine translation, and for re-ranking query responses in information retrieval. Identifying contrasting word pairs (or short phrase pairs) is also useful for detecting humor (Mihalcea and Strapparava, 2005), as satire and jokes tend to have contradictions and oxymorons. Lastly, it is useful to know which words contrast a focal word, even if only to filter them out. For example, in the automatic creation of a thesaurus it is necessary to distinguish near-synonyms from contrasting word pairs. Distributional similarity measures typically fail to do so.

Instances of strong contrast are recorded to some extent in manually created dictionaries, but hundreds of thousands of other contrasting pairs are not. Further, antonyms can be of many kinds such as those described in Section 3.1 below. We use the Mohammad et al. (2008) method to automatically generate a large list of contrasting word pairs, which are used to identify false paraphrases. Their method is briefly described in Section 3.2.

3.1 Kinds of antonyms

Antonyms can be classified into different kinds. A detailed description of one such classification can be found in Cruse (1986) (Chapters 9, 10, and 11), where the author describes complementaries (*open–shut, dead–alive*), gradable adjective pairs (*long–short, slow–fast*) (further classified into polar, overlapping, and equipollent antonyms), directional opposites (*up–down, north–south*), (further classified into antipodals, counterparts, and reversives), relational opposites (*husband–wife, predator–prey*), indirect converses (*give–receive, buy–pay*), congruence variants (*huge–little, doctor–patient*), and pseudo opposites (*black–white*). It should be noted, however, that even though contrasting word pairs and antonyms have long been studied by linguists, lexicographers, and others, experts do not always agree on the scope of antonymy and the kinds of contrasting word pairs. Some lexical relations have also received attention at the Educational Testing Services (ETS). They classify antonyms into contradictories (*alive–dead, masculine–feminine*), contraries (*old–young, happy-sad*), reverses (*attack–defend, buy–sell*), directionals (*front–back, left–right*), incompatibles (*happy–morbid, frank–hypocritical*), asymmetric contraries (*hot–cool, dry–moist*), pseudoantonyms (*popular–shy, right–bad*), and defectives (*default–payment, limp–walk*) (Bejar et al., 1991).

As mentioned earlier, in addition to antonyms, there are other meaning-contrasting phenomena, or other ways to classify them, such as contrasting trends and sentiment polarity. They all may have varying degrees of contrast in meaning. Hereafter we sometime broadly refer to all of these as *antonymous phrases*. The antonymous phrase pair generation algorithm that we use here does not employ any antonym-subclass-specific techniques.

3.2 Detecting antonyms

Mohammad et al. (2008) used a Roget-like thesaurus, co-occurrence statistics, and a seed set of antonyms to identify the degree of antonymy between two words, and generate a list of antonymous words. The thesaurus divides the vocabulary into about a thousand coarse categories. Each category has, on average, about a hundred closely related words. (A word with more than one sense, is listed in more than one category.) Mohammad et al. first determine pairs of thesaurus categories that are contrasting in meaning. A category pair is said to be contrasting if it has a seed antonym pair. A list of seed antonyms is compiled using 16 affix patterns such as X and unX (*clear–unclear*) and X and disX (*honest–dishonest*). Once a contrasting category pair is identified, all the word pairs across the two categories are considered to have contrasting meaning. The strength of co-occurrence (as measured by pointwise mutual information) between two contrasting word pairs is taken to be the degree of antonymy. This is based on the *distributional hypothesis of antonyms*, which states that antonymous pairs tend to co-occur in text more often than chance. Co-occurrence counts are made from the *British National Corpus (BNC)* (Burnard, 2000). The approach attains more than 80% accuracy on GRE-style closest opposite questions.

3.3 Detecting negators

The General Inquirer (GI) (Stone et al., 1966) has 11,788 words labeled with 182 categories of word tags, such as positive and negative semantic orientation, pleasure, pain, and so on.[2] Two of the GI categories, NOTLW and DECREAS, contain terms that negate the meaning of what follows (Choi and Cardie, 2008; Kennedy and Inkpen, 2005). These terms (with limited added inflection variation) form our list of negators.

4 Paraphrase-Augmented SMT

Augmenting the source side of SMT phrase tables with paraphrases of out-of-vocabulary (OOV) items was introduced by Callison-Burch et al. (2006), and was adopted practically ‘as-is’ in consequent work (Callison-Burch, 2008; Marton et al., 2009; Marton, 2010). 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, whose value for the new rule is the similarity score between f and f' (computed as a function of the pivot translation probabilities or the distributional semantic distance of the respective DPs). We follow the same line here:

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

where the definition of $asim$ is repeated below. As noted in that previous work, 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, an aggregated similarity measure was applied: For each paraphrase f of source-side phrases f_i with similarity scores $sim(f_i, f)$,

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

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.

5 Experiment

5.1 System and Parameters

We augmented translation models with paraphrases based on distributional semantic distance measures, with our novel antonym-filtering, and without it. We tested all models in English-to-Chinese and English-to-Arabic translation, augmenting the models with translation rules for unknown English phrases. We also contrasted these models with non-augmented baseline models.

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 tuning 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 1. Statistical significance of model output differences was determined using Koehn (2004)’s test on the objective function (BLEU).

The paraphrase-augmented models were created as described in Section 4. We used the same data and parameter settings as in Marton (2010).[3] We used cosine distance over DPs of log-likelihood ratios (McDonald, 2000), built with a sliding win-

[2] http://www.wjh.harvard.edu/∼inquirer

[3] Data preprocessing and paraphrasing code slightly differ from those used in Marton et al. (2009) and Marton (2010), and hence scores are not exactly the same across these publications.

dow of size ±6, a sampling threshold of 10000 occurrences, and a maximal paraphrase length of 6 tokens. We applied a paraphrase score threshold $t = 0.05$; a dynamic context length (the shortest non-stoplisted left context L occurring less than 512 times in the corpus, and similarly for R); paraphrasing of OOV unigrams; filtering paraphrase candidates occurring less than 25 times in the corpus (inspired by McDonald, 2000); and allowing up to $k = 100$ best paraphrases per phrase. We tuned the weights of each model (non-augmented baseline, unigram-augmented, and unigram-augmented-filtered) with a separate minimum error rate training.

We explored here augmenting OOV unigrams, although our paraphrasing and antonym filtering methods can be applied to longer n-grams with no further modifications. However, preliminary experiments showed that longer n-grams require additional provisions in order to yield gains.

5.2 Data

In order to take advantage of the English antonym resource, we chose English as the source language for the translation task. We chose Chinese as the translation target language in order to compare with Marton (2010), and for the same reasons it was chosen there: It is quite different from English (e.g., in word order), and four reference translation were available from NIST. We chose Arabic as another target language, because it is different from both English and Chinese, and richer morphologically, which introduces additional challenges.

English-Chinese: 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 followed the split in Marton (2010), and constructed the reduced set of about 29,000 sentence pairs. The purpose of creating this subset model was to simulate a resource-poor language. We trained separate translation models, using either the subset or the full-size training dataset.

For weight tuning 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 tuning "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.

English-Arabic: We use an English-Arabic parallel corpus of about 135k sentences (4 million words) and a subset of 30K sentences (one million words) for the translation models' training data. The sentences were extracted from Arabic News (LDC2004T17), eTIRR (LDC2004E72), English translation of Arabic Treebank (LDC2005E46), and Ummah (LDC2004T18).[4] For Arabic preprocessing, we follow previously reported best tokenization scheme (TB)[5] and orthographic word normalization condition (Reduced) when translating from English to Arabic (El Kholy and Habash, 2010b). MADA (Habash and Rambow, 2005) is used to pre-process the Arabic text for the translation model and 5-gram language model (LM). As a post-processing step, we jointly denormalize and detokenize the text to produce the final Arabic output. Following El Kholy and Habash (2010a), we use their best detokenization technique, T+R+LM. The technique crucially utilizes a lookup table (T), mapping tokenized forms to detokenized forms, based on our MADA-fied LM. Alternatives are given conditional probabilities, $P(detokenized|tokenized)$. Tokenized words absent from the tables are detokenized using deterministic rules (R), as a backoff strategy. We use a 5-gram untokenized LM and the `disambig` utility in the SRILM toolkit to decide among different alternatives. Word alignment is done using GIZA++, as in English-Chinese system. We use lemma-based alignment which consistently yields superior results to surface-based alignment (El Kholy and Habash, 2010b). For LM, we use 200M words from the Arabic Gigaword Corpus (LDC2007T40) together with the Arabic side of our training data.

All experiments were conducted using Moses here as well. We used a maximum phrase length

[4] All are available from the Linguistic Data Consortium (LDC) http://www.ldc.upenn.edu

[5] TB: Penn Arabic Tree Bank tokenization scheme

of size 8 tokens. Weight optimization was done using a set of 300 sentences from the NIST MT 2004 Arabic-English evaluation test set (MT04). The tuning was based on tokenized Arabic without detokenization. Testing was done on the NIST Arabic-English MT05 and MEDAR 2010 English-Arabic four-reference evaluation sets. For both tuning on MT04 and testing on MT05, since we need the reverse English-Arabic direction, we chose one English reference translation as the "source", and the Arabic as a single "reference". We evaluated using BLEU and TER here too.

English paraphrases: We augmented the baseline models with paraphrases generated as described above, using a monolingual text of over 516M tokens, consisting of the BNC and English Gigaword documents from 2004 and 2008 (LDC2009T13), pre-processed to remove punctuation and to conflate numbers, dates, months, days of week, and alphanumeric tokens to their respective classes.

5.3 Results

English-Chinese: Results are given in Table 1. Augmenting SMT phrase tables with paraphrases of OOV unigrams resulted in gains of 0.6-0.7 BLEU points for both subset and full models, but TER scores were worse (higher) for the full model. Augmenting same models with same paraphrases filtered for antonyms resulted in further gains of 1.6 and 1 BLEU points for both subset and full models, respectively, relative to the respective baselines. The TER scores of the antonym filtered models were also as good or better (lower) than those of the baselines.

	reduced size		*large size*	
model	**BLEU↑**	**TER↓**	**BLEU↑**	**TER↓**
baseline	15.8	69.2	21.8	63.8
aug-1gram	16.4^{B}	68.9	22.5^{B}	64.4
aug-1gram-ant-filt	$\mathbf{17.4}^{BD}$	**68.7**	$\mathbf{22.8}^{BD}$	**63.7**

Table 1: English-Chinese scores. B/D = statistically significant w.r.t. (B)aseline or (D)istributional 1gram model, using Koehn (2004)'s statistical significance.

English-Arabic: Results are given in columns 1-7 of Table 2. On the MT05 test set, the 135k-sentence aug-1gram model outperformed its baseline in both BLEU and TER scores. The lemmatized variants of the scores showed higher or same gains. Since only one entry was antonym-filtered here, we do not provide separate scores for aug-1gram-ant-filt. Surprisingly, for the reduced 30k models, all scores (BLEU, TER, and even their lemmatized variants) of the augmented 1gram model were somewhat worse than the baseline's, and those of the antonym-filtered model were the worst. we also ran a 4-reference test (Medar) to see whether the single MT05 reference was problematic, but results were similar. We examine possible reasons for this in the next section.

6 Discussion

Filtering quality: Our filtering technique is based on antonymous pair and negator lists that were expanded distributionally from seed sets. Therefore, they are noisy. From a small random sample (Table 3) it seems that only about 10% of filtered cases should not have been filtered; of the rest, 50% were strongly antonymous, 25% mildly so, and 15% were siblings (*co-hypernyms*) in a natural categorical hierarchy or otherwise noisy paraphrases filtered due to a noisy antonym pair. Negators in the unigrams' paraphrase candidates were rare.

English-Chinese: Our paraphrase filtering technique yielded an additional 1 BLEU point gain over the non-filtered paraphrase-augmented reduced model (totaling 1.6 BLEU over baseline). The reduced and large augmented models' phrase table size increased by about 27% and 4%, respectively – and antonym filtering did not change these numbers by much (see left side of Table 4). Therefore, the difference in performance between the filtered and non-filtered systems is unlikely to be quantitative (phrase table size). The out of vocabulary (OOV) rate of the 29k subset model is somewhat high (see Table 4), especially for the test set; but only after these experiments were completed did we peek at the test set for calculating these statistics, and in any case, we should not be guided by such information in choosing the test set. At first glance it may seem surprising that only 0.4% of the paraphrase candidates of the English OOV unigrams (248 candidates) were filtered by our procedure, and that it accounted for as much as 1 BLEU in the reduced set. (For English-Arabic only 0.6%, or 23 candidates, were filtered). Leaving the estimation of antonymous phrase detection recall for the future, we note that these num-

	BLEU	Lemm.	Brev.	Ref/Sys	TER	Lemm.	Unigram Lemma Match Analysis			
	↑	BLEU	penal.	ratio	↓	TER	Exact Match	Lemma-only	Unmatchable	Total
30k-sentence (1M word) training dataset models										
MT05 baseline	**23.6**	**31.3**	99.2	1.008	**57.6**	**47.3**	15614 55.4%	4055 14.4%	8550 30.3%	28219
aug-1gram	23.2	30.8	99.9	1.001	58.8	48.4	15387 54.2%	**4195** 14.8%	8831 31.1%	28413
aug-1gram-ant-filt	23.2	30.8	99.9	1.001	58.8	48.3	15387 54.2%	**4195** 14.8%	8831 31.1%	28413
MEDAR baseline	**13.6**	**18.7**	93.6	1.066	**67.6**	**61.3**	4924 53.0%	1563 16.8%	2800 30.1%	9287
aug-1gram	12.9	18.3	94.2	1.060	68.9	62.3	4894 52.0%	1710 18.2%	2815 29.9%	9419
aug-1gram-ant-filt	12.9	18.3	94.2	1.060	69.0	62.3	4891 51.9%	**1715** 18.2%	2815 29.9%	9421
135k-sentence (4M word) training dataset models										
MT05 baseline	25.8	33.5	99.2	1.008	55.7	45.3	16115 57.1%	3999 14.2%	8128 28.8%	28242
aug-1gram	**26.4**	**34.3**^{B}	99.5	1.005	**55.1**	**44.7**	16156 57.1%	4068 14.4%	8089 28.6%	28313
aug-1gram-ant-filt	**26.4**	**34.3**^{B}	99.5	1.005	**55.0**	**44.6**	16153 57.1%	**4090** 14.5%	8068 28.5%	28311
MEDAR baseline	17.1	23.1	94.7	1.054	65.1	58.6	5483 57.7%	1577 16.6%	2438 25.7%	9498
aug-1gram	**17.2**	**23.5**	95.3	1.048	65.1	58.6	5586 58.1%	**1606** 16.7%	2424 25.2%	9616
aug-1gram-ant-filt	**17.2**	**23.5**	95.3	1.048	65.1	58.6	5586 58.1%	**1606** 16.7%	2424 25.2%	9616

Table 2: English-Arabic translation scores and analysis for NIST MT05 and MEDAR test sets. B = statistically significant w.r.t. (B)aseline using Koehn (2004)'s statistical significance test.

bers from English are not directly comparable to the Chinese side: they relate to paraphrase candidates and not phrase table entries; they relate to types and not tokens; each OOV English word may translate to one or more Chinese words, each of which may comprise of one or more characters; and last but not least, the BLEU score we use is character-based.

phrase	**||| paraphrase**	**||| score**	**comments**
absence	||| occupation	||| 0.06	mild
absence	||| presence	||| 0.33	good
backwards	||| forwards	||| 0.21	good
wooden	||| plastic lawn	||| 0.12	sibling
dump	||| dispose of	||| 0.41	bad
cooler	||| warm	||| 0.45	mild
diminished	||| increased	||| 0.23	good
minor	||| serious	||| 0.42	good
relic	||| youth activist in the	||| 0.12	harmless
dive	||| rise	||| 0.15	good
argue	||| also recognize	||| 0.05	mild
bother	||| waste time	||| 0.79	bad
dive	||| climb	||| 0.17	good
moonlight	||| spring	||| 0.05	harmless
sharply	||| slightly	||| 0.60	good
substantial	||| meager	||| 0.14	good
warmer	||| cooler	||| 0.72	good
tough	||| delicate	||| 0.07	good
tiny	||| mostly muslim	||| 0.06	mild
softly	||| deep	||| 0.06	mild

Table 3: Random filtering examples

While individual unigram to 4gram scores for the augmented models were lower than the baseline's, filtered model's unigram and bigram scores were lower or similar to the baseline's, and their trigram and 4gram scores were higher than the baseline's. We intend to further investigate the cause for this pattern, and its effect on translation quality, with the help of a native Chinese speaker – and on BLEU, together with the brevity penalty – in the future.

English-Arabic: The most striking fact is the set of differences between the language pairs: In English-Chinese, we see gains with distributional paraphrase augmentation, and further gains when antonymous and contrasting paraphrase candidates are filtered out. But in the 30k-sentence English-Arabic models, paraphrase augmentation actually degrades performance, even in lemma scores. It has been observed before that BLEU (and similarly TER) is not ideal for evaluation of contributions of this sort (Callison-Burch et al., 2006). Therefore we conducted both manual and focused automatic analysis, including OOV statistics and unigram lemma match analysis[6]

[6] Unigram lemma match analysis is a classification of all the words in the translation hypothesis (against the translation reference) into: (a) exact match, which is equal to simple unigram precision, (b) lemma-only match, which counts words that can only be matched at the lemma level, and (c) unmatchable.

between the system output and the reference translation.

Table 4 shows that the OOV rates for English-Arabic are lower than English-Chinese. But if they were negligible, we would not expect to see gains (or in fact any change) in either model size, contrary to fact. It is interesting to point out that our translation model augmentation technique handles about 50% of the (non-digit, non-punctuation) OOV words in all models (except for only half that in the 135k model, which still showed gains).

Another concern is that the current maximal paraphrase length (6 tokens) may be too far from the paraphrasee's length (unigram), resulting in lower quality. However, a closer examination of the length difference evident through the BLEU brevity penalty and the reference:system-output length ratio (columns 4-5 of Table 2), reveals that the differences are small and inconsistent; on average, the brevity penalty difference accounts for roughly 0.1 absolute BLEU points and 0.2 absolute lemmatized BLEU points of the respective differences.[7]

Last, Modern Standard Arabic is a morphologically rich language: It has many inflected forms for most verbs, and several inflected forms for nouns, adjectives and other parts of speech – and complex syntactic agreement patterns showing these inflections. It might be the case that the inflected Arabic LM model might not serve well the augmented models, since they include translation rules that are more likely to be "off" inflection-wise (e.g., showing ungrammatical syntactic agreement or simply an acceptable choice that differs from the reference). Presumably, the smaller the training set, the larger this problem, since there would be fewer rules and hence smaller variety of inflected forms per similar core meaning. The unigram lemma match analysis and lemma scores' statistics (Table 2) support this concern. In the 30k model, lemma-only match seems to even further increase, at the expense of the exact word-form match. Possible solutions include using a lemma-based LM, or another LM that is adjusted to this sort of inflection-wise "off" text.

Error Analysis We conducted an error analysis of our Arabic 30k system using part of the MT05 test set. That set had 571 OOV types, out of which, we were able to augment phrases for 196 OOV types. The majority of OOV words were proper nouns (67.8%), with the rest being mostly nouns, adjectives and verbs (in the order of their frequency). Among the OOVs for which we augmented phrases, the proper noun ratio was smaller than the full set (45.4% relative). We selected a random sample of 50 OOV words, and examined their translations in the MT05 test set. The analysis considered all the OOV word occurrences (96 sentences). We classified each OOV translation in the augmented system and the augmented-filtered system as follows:

a1 correct (and in reference)
a2 correct (morphological variation)
a3 acceptable translation into a synonym
a4 acceptable translation into a hypernym

b1 wrong translation into a hypernym
b2 co-hypernym: a sibling in a psychologically natural category hierarchy
b3 antonymous, trend-contrasting, or polarity dissimilar meaning

c1 wrong proper-noun translation (sibling)
c2 wrong proper-noun translation (other)

d wrong translation for other reasons

Both the augmented and augmented-filtered system had 27.1% correct cases (category **a**). Only one-quarter of these were exact matches with the reference (category **a1**) that can be captured by BLEU. Incorrect proper-noun translation (category **c**) was the biggest error (augmented model: 33.3%, filtered model: 37.5%); within this category, sibling mistranslations (category **c1**), e.g., *Buddhism* is translated as *Islam*, were the majority (over half in augmented model, and about two-thirds in the filtered model). Proper nouns seem to be a much bigger problem for translation into Arabic than into Chinese in our sets. Category **b** mis-translations appeared in 20.8% of the time (equally in augmented and filtered). Almost half of these were sibling mistranslations (category **b2**), e.g., *diamond* translated as *gold*. Only two OOV translations in our sample were antonymous (category **b3**). It is possible, therefore, that our Arabic sets do not give room for our filtering method to be effective. In one case, the verb *deepen* (reference translation تعمق) is mis-

[7] These values are computed by subtracting the difference between two BLEU scores from the difference between the same two BLEU scores without the effect of brevity penalty (i.e., each divided by its brevity penalty).

translated as *summit* (قمة). In the other case, the adjective *cool (political relations)*, whose reference translation uses a figure of speech *periods of tension* (فترات من التوتر), is mistranslated as *good* (جيدة), which carries the opposite sentiment. The rest of category **b** involve hypernyms (**b1**), such as translating the OOV word *telecom* into *company* (الشركة).

Overall, the filtered model did not behave significantly differently from its augmented counterpart.

Chinese-Arabic score difference: We conjecture that another possible reason for the different score gain patterns between the two language pairs is the fact that in Chinese, many words that are siblings-in-meaning share a character, which doesn't necessarily have a stand-alone meaning; therefore, character-based BLEU was able to give credit to such paraphrases on the Chinese side, which was not case for the word-based BLEU on the Arabic side.

7 Related Work

This paper brings together several sub-areas: SMT, paraphrase generation, distributional semantic distance measures, and antonym-related work. Therefore we can only briefly survey the most relevant work here. Our work can be viewed as an extension of the line of research that seeks to augment translation tables with automatically generated paraphrases of OOV words or phrases in a fashion similar to Section 4: Callison-Burch et al. (2006) use pivoting technique (translating to other languages and back) in order to generate paraphrases, and the pivot translation probability as their similarity score; Callison-Burch (2008) filters such paraphrases using syntactic parsing information; Marton et al. (2009) use distributional paraphrasing technique that applies distributional semantic distance measure for the paraphrase score; Marton (2010) applies a lexical resource / corpus-based hybrid semantic distance measure for the paraphrase score instead, approximating word senses; here, we apply a distributional semantic distance measure that is similar to Marton et al. (2009), with the main difference being the filtering of the resulting paraphrases for antonymity.

Other work on augmentating SMT: 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. 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 pivot for paraphrasing. They improve SMT coverage by using a manually crafted monolingual HPSG grammar for generating meaning and grammar-preserving paraphrases. This grammar allows for certain word reordering, lexical substitutions, contractions, and "typo" corrections.

Onishi et al. (2010), Du et al. (2010), and others, pivot-paraphrase the input, and represent the paraphrases in a lattice format, decoding it with Moses.

Work on paraphrase generation: Barzilay and McKeown (2001) extract paraphrases from a monolingual parallel corpus, containing multiple translations of the same source. However, monolingual parallel corpora are extremely rare and small. Dolan et al. (2004) use edit distance for paraphrasing. Max (2009) and others take the context of the paraphrased word's occurrence into account. Zhao et al. (2008) apply SMT-style decoding for paraphrasing, using several log linear weighted resources while Zhao et al. (2009) filter out paraphrase candidates and weight paraphrase features according to the desired NLP task. 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. See Madnani and Dorr (2010) for a good paraphrasing survey.

Work on measuring distributional semantic distance: For one survey of this rich topic, see Weeds et al. (2004) and Turney and Pantel (2010). We use here cosine of log-likelihood ratios (McDonald, 2000). A recent paper (Kazama et al., 2010) advocates a Bayesian approach, making rare terms have lower strength of association, as a by-product of relying on their probabilistic Expectation.

Work on detecting antonyms: Our work with antonyms can be thought of as an application-based extension of the (Mohammad et al., 2008) method. Some of the earliest computational work in this area is by Lin et al. (2003) who used patterns

model	e2z:29k		e2z:232k		e2a:30k		e2a:135k	
phrase table baseline vocab. (# source-side types)	13916		34825		24371		49854	
phrase table entries: baseline	1996k		13045k		2606k		12344k	
phrase table entries: aug-1gram	2543k	127.38%	13615k	104.37%	2635k	101.09%	*12373k	100.23%
phrase table entries: aug-1gram-ant-filt	2542k	127.35%	13615k	104.37%	2635k	101.09%	*12373k	100.23%
OOV types in tune (% tune types)	1097	21.58%	451	8.87%	141	7.31%	84	4.35%
OOV tokens in tune (% tune tokens)	2138	6.10%	917	2.62%	193	2.18%	115	1.30%
OOV types in test (% test types)	2473	33.59%	1227	16.66%	574	12.42%	339	7.34%
OOV tokens in test (% test tokens)	4844	10.40%	2075	4.46%	992	2.83%	544	1.55%
tune OOV token decrease in aug-1gram/ant-filt	1343	27.73%	510	24.58%	79	7.96%	28	5.15%
tune OOV type decrease in aug-1gram/ant-filt	646	58.89%	203	45.01%	60	42.55%	22	26.19%
test OOV token decrease in aug-1gram /ant-filt	2776	57.31%	996	48.00%	460	46.37%	127	23.35%
test OOV type decrease in aug-1gram/ant-filt	1394	56.37%	585	47.68%	246	42.86%	76	22.42%

Table 4: Out-of-vocabulary (OOV) word rates and phrase table sizes for all model sizes and language pairs. e2z = English-Chinese; e2a = English-Arabic. The statistics marked with * in the top-right cell are identical, see §5.3.

such as "from *X* to *Y*" and "either *X* or *Y*" to distinguish between antonymous and similar word pairs. Harabagiu et al. (2006) detected antonyms by determining if their WordNet synsets are connected by the hypernymy–hyponymy links and exactly one antonymy link. Turney (2008) proposed a supervised method to solve word analogy questions that require identifying synonyms, antonyms, hypernyms, and other lexical-semantic relations between word pairs.

8 Conclusions and Future Work

We presented here a novel method for filtering out antonymous phrasal paraphrase candidates, adapted from sentiment analysis literature, and tested in simulated low- and mid-resourced SMT tasks from English to two quite different languages. We used an antonymous word pair list extracted distributionally by extending a seed list. Then, the extended list, together with a negator list and a novel heuristic, were used to filter out antonymous paraphrase candidates. Finally, SMT models were augmented with the filtered paraphrases, yielding English-Chinese translation improvements of up to 1 BLEU from the corresponding non-filtered paraphrase-augmented model (up to 1.6 BLEU from the corresponding baseline model). Our method proved effective for models trained on both reduced and mid-large English-Chinese parallel texts. The reduced models simulated "low density" languages by limiting the amount of the training text.

We also showed for the first time translation gains for English-Arabic with paraphrase-augmented (non-filtered) models. However, Arabic, and presumably other morphologically rich languages, may require more complex models in order to benefit from our filtering method.

Our antonym detection and filtering method is distributional and heuristic-based; hence it is noisy. We suspect that OOV terms in larger models tend to be harder to paraphrase (judging by the difference from the reduced models, and the lower OOV rate), and also harder to filter paraphrase candidates of (due to the lower paraphrase quality, which might not even include sufficiently distributionally similar candidates, antonymous or otherwise). In the future, we intend to improve our method, so that it can be used to improve also the quality of models trained on even larger parallel texts.

Last, we intend to extend our method beyond unigrams, limit paraphrase length to the vicinity of the paraphrasee's length, and improve our inflected Arabic generation technique, so it can handle this novel type of augmented data well.

Acknowledgments

Part of this work was done while the first author was at Columbia University. The second author was funded through a Google research award. The authors wish to thank Saif Mohammad for providing his data and for useful discussion, and also thank the anonymous reviewers for their useful feedback.

References

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Isaac I. Bejar, Roger Chaffin, and Susan Embretson. 1991. *Cognitive and Psychometric Analysis of Analogical Problem Solving*. Springer-Verlag, New York, NY.

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

Lou Burnard. 2000. *Reference Guide for the British National Corpus (World Edition)*. Oxford University Computing Services.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (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 *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) - the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (IJCNLP) Short Papers*, pages 249–252, Suntec, Singapore.

Yejin Choi and Claire Cardie. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, Waikiki, Hawaii.

David A. Cruse. 1986. *Lexical semantics*. Cambridge University Press.

James Deese. 1965. *The structure of associations in language and thought*. The Johns Hopkins Press.

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

Jinhua Du, Jie Jiang, and Andy Way. 2010. Facilitating translation using source language paraphrase lattices. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 420–429, MIT, Massachusetts, USA.

Ahmed El Kholy and Nizar Habash. 2010a. Techniques for Arabic Morphological Detokenization and Orthographic Denormalization. In *Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta.

Ahmed El Kholy and Nizar Habash. 2010b. Orthographic and Morphological Processing for English-Arabic Statistical Machine Translation. In *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, Montreal, Canada.

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

Nizar Habash and Owen Rambow. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 573–580, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Nizar Habash. 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)*, Cairo, Egypt.

Sanda M. Harabagiu, Andrew Hickl, and Finley Lacatusu. 2006. Lacatusu: Negation, contrast and contradiction in text processing. In *Proceedings of the 23rd National Conference on Artificial Intelligence (AAAI)*, Boston, MA.

Junichi Kazama, Stijn De Saeger, Kow Kuroda, Masaki Murata, and Kentaro Torisawa. 2010. A Bayesian method for robust estimation of distributional similarities. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 247–256, Uppsala, Sweden.

Alistair Kennedy and Diana Inkpen. 2005. Sentiment classification of movie and product reviews using contextual valence shifters. *COMPUTATIONAL INTELLIGENCE*, pages 110–125.

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 *the 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 *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Adrienne Lehrer and K. Lehrer. 1982. Antonymy. *Linguistics and Philosophy*, 5:483–501.

Dekang Lin, Shaojun Zhao, Lijuan Qin, and Ming Zhou. 2003. Identifying synonyms among distributionally similar words. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1492–1493, Acapulco, Mexico.

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

Marie-Catherine de Marneffe, Anna Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, Columbus, OH.

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

Yuval Marton. 2010. Improved statistical machine translation using monolingual text and a shallow lexical resource for hybrid phrasal paraphrase generation. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA)*, Denver, Colorado.

Aurelien Max. 2009. Sub-sentential paraphrasing by contextual pivot translation. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) - the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (IJCNLP) - Workshop on Applied Textual Inference*, pages 18–26, Singapore. Suntec.

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

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, Canada.

Shachar Mirkin, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor . 2009. Source-language entailment modeling for translating unknown terms. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) - the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (IJCNLP)*, pages 791–799, Suntec, Singapore.

Saif Mohammad, Bonnie Dorr, and Codie Dunn. 2008. Computing word-pair antonymy. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 982–991, Waikiki, Hawaii.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (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.

Takashi Onishi, Masao Utiyama, and Eiichiro Sumita. 2010. Paraphrase lattice for statistical machine translation. In *Proceedings of the Association for Computational Linguistics (ACL) Short Papers*, pages 1–5, Uppsala, Sweden.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

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.

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

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

Philip Stone, Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie, and associates. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

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*.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Articial Intelligence Research*, 37:141–188.

Peter Turney. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference* *on Computational Linguistics (COLING)*, pages 905–912, Manchester, UK.

Ellen M Voorhees. 2008. Contradictions and justifications: Extensions to the textual entailment task. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, Columbus, OH.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, 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 *Proceedings of the Association for Computational Linguistics (ACL)Human Language Technology (HLT)*, pages 1021–1029, Columbus, Ohio, USA.

Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven statistical paraphrase generation. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) - the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (IJCNLP)*, pages 834–842, Suntec, Singapore.