Improved Statistical Machine Translation Using Monolingual Paraphrases

Preslav Nakov[1,2,3]

Abstract. We propose a novel monolingual sentence paraphrasing method for augmenting the training data for statistical machine translation systems "for free" – by creating it from data that is already available rather than having to create more aligned data. Starting with a syntactic tree, we recursively generate new sentence variants where noun compounds are paraphrased using suitable prepositions, and vice-versa – preposition-containing noun phrases are turned into noun compounds. The evaluation shows an improvement equivalent to 33%-50% of that of doubling the amount of training data.

1 Introduction

Most modern Statistical Machine Translation (SMT) systems rely on aligned bilingual corpora (bi-texts) from which they learn how to translate small pieces of text. In many cases, these pieces are semantically equivalent but syntactically different from translation-time text, and thus the potential for high-quality translation can be missed.

In this paper, we describe a method for expanding the training bi-text using paraphrases that are nearly-equivalent semantically but different syntactically. In particular, we apply sentence-level paraphrasing on the source-language side, focusing on noun compounds (NCs) and noun phrases (NPs), which have been reported to be very frequent in English written text: 2.6% of the tokens in the *British National Corpus* and 3.9% in the *Reuters corpus* are covered by NCs [2], and about half of the words in news texts are part of an NP [13].

The proposed approach is novel in that it augments the training corpus with paraphrases of the original sentences, thus augmenting the training bi-text without increasing the number of training translation pairs needed. It is also monolingual; other related approaches map from the source language to other languages in order to obtain paraphrases. Finally, while our paraphrasing rules are English-specific, the method is general enough to be domain-independent.

2 Related Work

Recent work in automatic corpus-based paraphrasing includes using bi-texts as a source of alternative expressions for the same term [4, 20], or using multiple expressions of the same concept in one language [22]. In a more recent work, [3] propose using phrases in a second language as pivots. For example, if in a parallel English-German corpus, the English phrases *under control* and *in check* happen to be aligned (in different sentences) to the same German phrase *unter controlle*, they would be hypothesised to be paraphrases of each other with some probability.

Recently, paraphrases have been used to improve machine translation *evaluation* . For example, [9] argue that automated evaluation measures like *Bleu* [21] end up comparing n-gram overlap rather than semantic similarity with the reference text. Having performed an experiment asking two human translators to translate the same set of 10,000 sentences, they found that less than .2% of the translations were identical, and 60% differed by more than ten words. Therefore, they proposed an evaluation method which paraphrases the machine-produced translations and yields improved correlation with human judgements compared to *Bleu*. In a similar spirit, [25] use a paraphrase table extracted from a bilingual corpus in order to improve the evaluation of automatic summarization algorithms.

Another related research direction is in translating units of text smaller than a sentence, e.g., NCs [8, 23, 2], NPs [7, 13], named entities [1], and technical terms [15]. While we focus on paraphrasing NPs/NCs, unlike these approaches, we paraphrase and translate full sentences, as opposed to working with small text units in isolation.

The approach we propose below is most closely related to that of [6], who translate English sentences into Spanish and French by substituting unknown source phrases with suitable paraphrases. Our paraphrases, however, are quite different. Theirs are extracted with the above-mentioned bilingual method of [3] using eight additional languages from the *Europarl corpus* [12] as pivots. These paraphrases are incorporated in the machine translation process by adding them as additional entries in the phrase table and pairing them with the foreign translation of the original phrase. Finally, the system is tuned using minimum error rate training [17] with an extra feature penalising the low-probability paraphrases. This yielded dramatic increases in coverage (from 48% to 90% of the test word types when 10,000 training sentences were used), and notable increase on *Bleu* (up to 1.5%). However, the method requires large multi-lingual parallel corpora, which makes it domain-dependent and most likely limits its applicability to Chinese, Arabic, and the languages of the EU, for which such large resources are likely to be available.

3 Method

We propose a novel general approach for improving SMT systems using monolingual paraphrases. Given a sentence from the source (English) side of the training corpus, we generate conservative meaning-preserving syntactic paraphrases of that sentence. Each paraphrase is paired with the foreign (Spanish) translation that is associated with the original source sentence in the training bi-text. This augmented training corpus is then used to train an SMT system.

[1] Linguistic Modeling Department of the Institute for Parallel Processing at the Bulgarian Academy of Sciences, 25A, Acad. G. Bonchev St., 1113 Sofia, Bulgaria, email nakov@lml.bas.bg

[2] Department of Mathematics and Informatics, Sofia University, 5, James Bourchier blvd. 1164 Sofia, Bulgaria

[3] Part of this research was done while the author was a PhD student at the EECS department, CS division, University of California at Berkeley, USA.

We further introduce a variation on this idea that can be used with a *phrase-based* SMT. In this alternative, the source-language *phrases* from the phrase table are paraphrased, but again using the target source-language phrase only, as opposed to requiring a third parallel pivot language as in [6]. We also try to combine these approaches.

4 Paraphrasing

Given a sentence like "*I welcome the Commissioner's statement about the progressive and rapid lifting of the beef import ban.*", we parse it using the *Stanford parser* [10], and we recursively apply the following syntactic transformations:

1. **[NP NP$_1$ P NP$_2$] ⇒ [NP NP$_2$ NP$_1$].**
   the lifting of the beef import ban ⇒ the beef import ban lifting
2. **[NP NP$_1$ *of* NP$_2$] ⇒ [NP NP$_2$ gen NP$_1$].**
   the lifting of the beef import ban ⇒ the beef import ban's lifting
3. **NP**$_{gen}$ **⇒ NP.**
   Commissioner's statement ⇒ Commissioner statement
4. **NP**$_{gen}$ **⇒ NP**$_{PP_{of}}$**.**
   Commissioner's statement ⇒ statement of (the) Commissioner
5. **NP**$_{NC}$ **⇒ NP**$_{gen}$**.**
   inquiry committee chairman ⇒ inquiry committee's chairman
6. **NP**$_{NC}$ **⇒ NP**$_{PP}$**.**
   the beef import ban ⇒ the ban on beef import

where: **gen** is a genitive marker: ' or 's; **P** is a preposition; **NP**$_{PP}$ is an NP with an internal PP-attachment; **NP**$_{PP_{of}}$ is an NP with an internal PP headed by *of*; **NP**$_{gen}$ is an NP with an internal genitive marker; **NP**$_{NC}$ is an NP that is a noun compound.

The resulting paraphrases are shown in Table 2. In order to prevent transformations (1) and (2) from constructing awkward NPs, we impose certain limitations on NP$_1$ and NP$_2$. They cannot span a verb, a preposition or a quotation mark (although they can contain some kinds of nested phrases, e.g., an ADJP in case of coordinated adjectives, as in *the progressive and controlled lifting*). Therefore, the phrase *reduction in the taxation of labour* is not transformed into *taxation of labour reduction* or *taxation of labour's reduction*. We further require the head to be a noun and we do not allow it to be an indefinite pronoun like *anyone*, *everybody*, and *someone*.

Transformations (1) and (2) are more complex than they may look. In order to be able to handle some hard cases, we apply additional restrictions. First, some determiners, pre-determiners and possessive adjectives must be eliminated in case of conflict between NP$_1$ and NP$_2$, e.g., *the lifting of this ban* can be paraphrased as *the ban lifting*, but not as *this ban's lifting*. Second, in case both NP$_1$ and NP$_2$ contain adjectives, these adjectives have to be put in the right order, e.g., *the first statement of the new commissioner* can be paraphrased as *the first new commissioner's statement*, but not *the new first commissioner's statement*. There is also the option of not re-ordering them, e.g., *the new commissioner's first statement*. Third, further complications are due to scope ambiguities of modifiers of NP$_1$. For example, in *the first statement of the new commissioner*, the scope of the adjective *first* is not *statement* alone, but *statement of the new commissioner*. This is very different for the NP *the biggest problem of the whole idea*, where the adjective *biggest* applies to *problem* only, and therefore it cannot be transformed to *the biggest whole idea's problem* (although we do allow for *the whole idea's biggest problem*).

While the first four transformations are purely syntactic, (5) and (6) are not. The algorithm must determine whether a genitive marker is feasible for (5) and must choose the correct preposition for (6). In either case, for noun compounds of length three or more, we also need to choose the correct position to modify, e.g., *inquiry's committee chairman* vs. *inquiry committee's chairman*.

In order to improve the accuracy of the paraphrases, we use the Web as a corpus, generating and testing the paraphrases in the context of the preceding and the following words in the sentence. First, we split the noun compound into two sub-parts N_1 and N_2 in all possible ways, e.g., *beef import ban lifting* would be split as: (a) N_1="*beef*", N_2="*import ban lifting*", (b) N_1="*beef import*", N_2="*ban lifting*", and (c) N_1="*beef import ban*", N_2="*lifting*". For each split, we issue exact phrase queries to *Google* using the following patterns:

"lt N_1 gen N_2 rt"
"lt N_2 prep det N_1' rt"
"lt N_2 that be det N_1' rt"
"lt N_2 that be prep det N_1' rt"

where: N_1' can be a singular or a plural form of N_1; `lt` is the word preceding N_1 in the original sentence, if any; `rt` is the word following N_2 in the original sentence, if any; `gen` is a genitive marker ('s or '); `that` is *that*, *which* or *who*; `be` is *is* or *are*; `det` is *the*, *a*, *an*, or none; and `prep` is one of the prepositions used by [14] for NC interpretation: *about*, *at*, *for*, *from*, *in*, *of*, *on*, and *with*.

Given a particular split, we find the number of page hits for each instantiation of the above paraphrase patterns, filtering out the ones whose page hit counts are less than ten. We then calculate the total number of page hits H for all paraphrases (for all splits and all patterns), and we retain the ones whose page hits counts are at least 10% of H, which allows for multiple paraphrases (possibly corresponding to different splits) for a given noun compound. If no paraphrases are retained, we repeat the above procedure with `lt` set to the empty string. If there are still no good paraphrases, we set `rt` to the empty string. If this does not help either, we make a final attempt, by setting both `lt` and `rt` to the empty string. For example, *EU budget* is paraphrased as *EU's budget* and *budget of the EU*; also *environment policy* becomes *policy on environment*, *policy on the environment*, and *policy for the environment*; *UN initiatives* is paraphrased as *initiatives of the UN*, *initiatives at the UN*, and *initiatives in the UN*, and *food labelling* becomes *labelling of food* and *labelling of foods*.

We apply the same algorithm to paraphrasing English *phrases* from the phrase table, but without transformations (5) and (6). See Table 1 for sample paraphrases.

1	**% of members of the irish parliament** % of irish parliament members % of irish parliament 's members
2	**universal service of quality .** universal quality service . quality universal service . quality 's universal service .
3	**action at community level** community level action
4	**, and the aptitude for communication and** , and the communication aptitude and
5	**to the fall-out from chernobyl .** to the chernobyl fall-out .
6	**flexibility in development - and quick** development flexibility - and quick
7	**, however , the committee on transport** , however , the transport committee
8	**and the danger of infection with aids** and the danger of aids infection and the aids infection danger and the aids infection 's danger

Table 1. Sample English phrases from the phrase table and corresponding automatically generated paraphrases.

I welcome the Commissioner 's statement about the progressive and rapid beef import ban lifting .
I welcome the progressive and rapid beef import ban lifting Commissioner 's statement .
I welcome the Commissioner 's statement about the beef import ban 's progressive and rapid lifting .
I welcome the beef import ban 's progressive and rapid lifting Commissioner 's statement .
I welcome the Commissioner 's statement about the progressive and rapid lifting of the *ban on beef imports* .
I welcome the Commissioner statement about the progressive and rapid lifting of the beef import ban .
I welcome the Commissioner statement about the progressive and rapid beef import ban lifting .
I welcome the progressive and rapid beef import ban lifting Commissioner statement .
I welcome the Commissioner statement about the beef import ban 's progressive and rapid lifting .
I welcome the beef import ban 's progressive and rapid lifting Commissioner statement .
I welcome the Commissioner statement about the progressive and rapid lifting of the *ban on beef imports* .
I welcome the statement of Commissioner about the progressive and rapid lifting of the beef import ban .
I welcome the statement of Commissioner about the progressive and rapid beef import ban lifting .
I welcome the statement of Commissioner about the beef import ban 's progressive and rapid lifting .
I welcome the statement of Commissioner about the progressive and rapid lifting of the *ban on beef imports* .
I welcome the statement of the Commissioner about the progressive and rapid lifting of the beef import ban .
I welcome the statement of the Commissioner about the progressive and rapid beef import ban lifting .
I welcome the statement of the Commissioner about the beef import ban 's progressive and rapid lifting .
I welcome the statement of the Commissioner about the progressive and rapid lifting of the *ban on beef imports* .

Table 2. Sample sentences and their automatically generated paraphrases. Paraphrased noun compounds are in italics.

5 Experiments and Evaluation

5.1 Europarl Corpus

We trained and evaluated several English→Spanish phrase-based statistical machine translation systems using the *Europarl corpus* [12] and the standard training/tuning/testing dataset splits.

First, we built English→Spanish and Spanish→English directed word alignments using IBM model 4 [5], we combined them using the *intersect+grow heuristic* [18], and we extracted phrase-level translation pairs using the *alignment template approach* [19]. We thus obtained a *phrase table* where each translation pair is associated with five parameters: forward phrase translation probability, reverse phrase translation probability, forward lexical translation probability, reverse lexical translation probability, and phrase penalty.

We then trained a log-linear model using the following feature functions: language model probability, word penalty, distortion cost, and the above-mentioned parameters from the phrase table. We set the feature weights by optimising the *Bleu* score directly using *minimum error rate training* (MERT) [17] on the first 500 sentences from the development set. We then used these weights in a beam search decoder [11] to translate the 2,000 test sentences, and we compared the translations to the gold standard using *Bleu* [21].

Baseline. Using the above procedure, we built and evaluated a baseline system S, trained on the original training corpus.

Sentence-Level Paraphrasing. We further built S_{pW}, which uses a version of the training corpus augmented with syntactic paraphrases of the *sentences* from the English side paired with the corresponding Spanish translations. In order to see the effect of not breaking NCs and not using the Web, we built S_p, which does not use transformations (5) and (6).

Phrase Table Paraphrasing. System $S^\star$ paraphrases and augments the *phrase table* of the baseline system S using syntactic transformations (1)-(4), similarly to S_p, i.e., without NC paraphrasing. Similarly, $S^\star_{pW}$ is obtained by paraphrasing the *phrase tabl*e of S_{pW}.

Combined Systems. Finally, we merged the phrase tables for some of the above systems, which we designate with a "+", e.g., $S + S_{pW}$ and $S^\star + S^\star_{pW}$. In these merges, the phrases from the first phrase table are given priority over those from the second one in case a phrase pair is present in both phrase tables. This is important since the parameters estimated from the original corpus are more reliable.

Following [3], we also performed an experiment with an additional feature F_{pW} for each phrase: its value is 1 if the phrase is in the phrase table of S, and 0.5 if it comes from the phrase table of S_{pW}. As before, we optimised the weights using MERT. For $S^\star + S^\star_{pW}$, we also tried using two features: in addition to F_{pW}, we introduced $F_\star$, whose value is 0.5 if the phrase comes from paraphrasing a phrase table entry, and 1 if it was in the original phrase table.

The evaluation results are shown in Tables 3 and 4. The differences between the baseline and the remaining systems shown in Table 3 are statistically significant, which was tested using bootstrapping [24].

Gain of 33%–50% compared to doubling the training data. As Table 4 shows, neither paraphrasing the training sentences, S_{pW}, nor paraphrasing the phrase table, $S^\star$, lead to notable improvements. For 10k training sentences, the systems are comparable and improve *Bleu* by .3, while for 40k sentences, $S^\star$ matches the baseline, and S_{pW} even drops below it. However, merging the phrase tables of S and S_{pW}, yields an improvement of almost .7 for 10k and 20k sentences, and about .3 for 40k sentences. While this improvement might look small, it is comparable to that of [3], who achieved .7 improvement for 10k sentences, and 1.0 for 20k (translating in the reverse direction: Spanish→English). Note also that the .7 improvement in *Bleu* for 10k and 20k sentences is about 1/3 of the 2 *Bleu* point improvement achieved by the baseline system by doubling the training size. Note also that the .3 gain on *Bleu* for 40k sentences is equal to half of what would have been gained if we had trained on 80k sentences.

Improved precision for all n-grams. Table 3 compares different systems trained on 10k sentences. In addition to the *Bleu* score, we give its elements: n-gram precisions, BP (brevity penalty), and ration. Comparing the baseline with the last four systems, we can see that all n-gram precisions are improved by about .4-.7 *Bleu* points.

Importance of noun compound splitting. S_p is trained on the training corpus augmented with paraphrased sentences, where the

System	Bleu	n-gram precision 1-gr.	2-gr.	3-gr.	4-gr.	Bleu BP	ration	# of phrases gener.	used
S (baseline)	22.38	55.4	27.9	16.6	10.0	0.995	0.995	181k	41k
S_p	21.89	55.7	27.8	16.5	10.0	0.973	0.973	193k	42k
S_{pW}	22.57	55.1	27.8	16.7	10.2	1.000	1.000	202k	43k
$S^\star$	22.58	55.4	28.0	16.7	10.1	1.000	1.001	207k	41k
$S + S_p$	22.73	55.8	28.3	16.9	10.3	0.994	0.994	262k	54k
$S + S_{pW}$	**23.05**	55.8	28.5	17.1	10.6	0.995	0.995	280k	56k
$S + S_{pW}\dagger$	**23.13**	55.8	28.5	17.1	10.5	1.000	1.002	280k	56k
$S^\star + S^\star_{pW}$	**23.09**	56.1	28.7	17.2	10.6	0.993	0.993	327k	56k
$S^\star + S^\star_{pW}\ddagger$	**23.09**	55.8	28.4	17.1	10.5	1.000	1.001	327k	56k

Table 3. *Bleu* scores and n-gram precisions for 10k training sentences. The last two columns show the total number of entries in the phrase table and the number of phrases that were usable at testing time, respectively.

System	# of training sentences 10k	20k	40k	80k
S (baseline)	**22.38**	**24.33**	**26.48**	**27.05**
S_{pW}	22.57	24.41	25.96	
$S^\star$	22.58	25.00	26.48	
$S + S_{pW}$	**23.05**	**25.01**	**26.75**	

Table 4. *Bleu* scores for different number of training sentences.

NC splitting rules (5) and (6) are not used. We can see that the results for this system go below the baseline: while there is a .3 gain on *Bleu* on unigram precision, bigram and trigram precision go down by about .1. More importantly, BP decreases as well: since the sentence-level paraphrases (except for genitives, which are infrequent) convert NPs into NCs, the resulting sentences are shorter, and thus the translation model learns to generate shorter sentences. This is different in S_{pW}, where transformations (5) and (6) counter-weight (1)-(4), thus balancing BP to 1. A somewhat different kind of argument applies to $S + S_p$, which is worse than $S + S_{pW}$, but not because of BP. In this case, there is no improvement for unigrams, but a consistent .2-.3 drop for bigrams, trigrams and fourgrams. The reason is shown in the last column of Table 4: omitting rules (5) and (6) results in fewer training sentences, which means fewer phrases in the phrase table and therefore fewer ones usable at translation time.

More usable phrases. The last two columns of Table 3 show that, in general, having more phrases in the phrase table implies more usable phrases at translation time. A notable exception is $S^\star$, whose phrase table is bigger than those of S_p and S_{pW}, but yields less utile phrases. Therefore, we can conclude that the additional phrases extracted from paraphrased sentences are more likely to be usable at test time than the ones generated by paraphrasing the phrase table.

Paraphrasing sentences vs. paraphrasing the phrase table. As Tables 3 and 4 show, paraphrasing the phrase table, as in $S^\star$ (*Bleu* score 22.58), cannot compete against paraphrasing the training corpus followed by merging the resulting phrase table with the phrase table for the original corpus[4], as in $S + S_{pW}$ (*Bleu* score 23.05). We also tried to paraphrase the phrase table of $S + S_{pW}$, but the resulting system $S^\star + S^\star_{pW}$ yielded little improvement: 23.09 *Bleu* score. Adding the two extra features, $F_\star$ and F_{pW}, did not yield improvements as well: $S^\star + S^\star_{pW}\ddagger$ achieved the same *Bleu* score as $S^\star + S^\star_{pW}$. This shows that extracting additional phrases from the augmented corpus is a better idea than paraphrasing the phrase table, which can result in erroneous splitting of noun phrases. Paraphrasing whole sentences as opposed to paraphrasing the phrase table could potentially improve the approach of [6] as well: while low probability and context dependency could be problematic, a language model could help filter the bad sentences out. Such filtering could potentially improve our results as well. Finally, note that different paraphrasing strategies could be used when paraphrasing phrases vs. sentences. For example, paraphrasing the phrase table can be done more aggressively: if an ungrammatical phrase is generated in the phrase table, it would most likely have no negative effect on translation quality since it would be unlikely to be observed at translation time.

Quality of the paraphrases and comparison to [6]. An important difference between our syntactic paraphrasing and the multilingual approach of [6] is that their paraphrases are only contextually synonymous and often depart significantly from the original meaning. As a result, they could not achieve improvements by simply augmenting the phrase table: this introduced too much noise and yielded accuracy that was below their baseline by 3-4 *Bleu* points. In order to achieve an improvement, they had to introduce an extra feature penalising the low probability paraphrases and promoting the original phrase table entries. In contrast, our paraphrases are meaning-preserving and less context-dependent. For example, introducing feature F_{pW} which penalises phrases coming from the paraphrased corpus in system $S + S_{pW}\dagger$ yielded a tiny improvement on *Bleu* score (23.13 vs. 23.05), i.e., the phrases extracted from our augmented corpus are almost as good as the ones from the original corpus. Finally, note that our paraphrasing method is *complementary* to that of [6] and therefore the two can be combined: the strength of our approach is in improving the *coverage of longer phrases* using syntactic paraphrases, while the strength of theirs is in improving the *vocabulary coverage* with words extracted from additional corpora (although they do get some gain from using longer phrases as well).

Paraphrasing the target side. We also tried paraphrasing the target language side, i.e., translating into English, which resulted in decreased performance. This is not surprising: the set of available source phrases remains the same, and a possible improvement could only come from producing a more fluent translation, e.g., from transforming an NP with an internal PP into an NC. However, unlike the original translations, the extra ones are a priori less likely to be judged correct since they were not observed on training.

5.2 News Commentary & Domain Adaptation

We further applied the proposed paraphrasing method to domain adaptation using the data from the ACL'07 Workshop on SMT: 1.3M words (64k sentences) of *News Commentary* data and 32M words of *Europarl* data. We used the standard training/tuning/testing splits, and we tested on *News Commentary* data.

This time we used two additional features with MERT (indicated with the $\prec$ operation): for the original and for the augmented phrase table, which allows extra weight to be given to phrases appearing in both. With the default distance reordering, for 10k sentences we had 28.88 Bleu for $S + S_{pW}$ vs. 28.07 for S, and for 20k we had 30.65 vs. 30.34. However, for 64k sentences, there was almost no difference: 32.77 vs. 32.73. Using a different tokenizer and a lexicalized reordering model, we got 32.09 vs. 32.34, i.e., the results were worse.

However, as Table 5 shows, using a second language trained on *Europarl*, we were able to improve *Bleu* to 34.42 (for $S + S_{pW}$) from 33.99 (for S). Using S_{pW} lead to even bigger improvements (0.64 Bleu) when added to $S^{news} \prec S^{euro}$, where an additional phrase table from *Europarl* was used. See [16] for further details.

[4] Note that $S^\star$ does not use rules (5) and (6). However, as $S + S_p$ shows, the claim holds even if these rules are excluded when paraphrasing whole sentences: the *Bleu* score for $S + S_p$ is 22.73 vs. 22.58 for $S^\star$.

Model	Language Models	
	News Only	News+Euro
S^{news}	32.27	**33.99**
$S^{news} \prec S^{news}_{pW}$	32.09	**34.42**
$S^{news} \prec S^{euro}$		**34.05**
$S^{news} \prec S^{news}_{pW} \prec S^{euro}$		34.25
$S^{news} \prec S^{euro} \prec S^{news}_{pW}$		**34.69**

Table 5. Bleu scores on the *News Commentary* data (64k sentences).

6 Problems and Limitations

Error analysis has revealed that the major problems for the proposed method are incorrect PP-attachments in the parse tree, and, less frequently, wrong POS tags (e.g., JJ instead of NN). Using a syntactic parser further limits the applicability of the approach to languages for which such parsers are available. In fact, for our purposes, it might be enough to use a shallow parser or just a POS tagger. This would cause problems with PP-attachment, but these attachments are often assigned incorrectly by parsers anyway. The main target of our paraphrases are noun compounds – we turn NPs into NCs and vice versa – which limits the applicability of the approach to languages where noun compounds are a frequent phenomenon, e.g., Germanic, but not Romance or Slavic. From a practical viewpoint, an important limitation is that the size of the phrase table and/or of the training corpus increases, which slows down both training and translation, and limits the applicability to relatively small corpora for computational reasons. Last but not least, as Table 4 shows, the improvements get smaller for bigger training corpora, which suggests it becomes harder to generate useful paraphrases that are not already in the corpus.

7 Conclusion and Future Work

We presented a novel domain-independent approach for improving statistical machine translation by augmenting the training corpus with monolingual source-language side paraphrases, thus increasing the training data "for free", by creating it from data that is already available rather than having to create more aligned data.

While in our experiments we used phrase-based SMT, any machine translation approach that learns from parallel corpora could potentially benefit from the idea of syntactic corpus augmentation. At present, our paraphrasing rules are English-specific, but they could be easily adapted to other Germanic languages, which make heavy use of noun compounds; the general idea of automatically generating nearly equivalent source-side syntactic paraphrases can in principle be applied to any language. The current version of the method should be considered preliminary, as it is limited to NPs; still, the results are already encouraging, and the approach is worth considering when building MT systems from small corpora, e.g., in case of resource-poor language pairs, in specific domains, etc.

Better use of the Web could be made for paraphrasing noun compounds (e.g., using verbal paraphrases), and other syntactic transformations could be tried (e.g., adding/removing complementisers like *that* and commas from nonmandatory positions).

Even more promising, but not that simple, would be using a tree-to-tree syntax-based SMT system and learning suitable syntactic transformations that can make the source-language trees structurally closer to the target-language ones. For example, the English sentence "*Remember the guy who you are with!*" would be transformed into "*Remember the guy with whom you are!*", whose word order is closer to the Spanish "*¡Recuerda al individuo con quien estás!*", which might facilitate the translation process.

Finally, the process could be made part of the decoding, which would eliminate the need of paraphrasing the training corpus and might allow dynamically generating paraphrases both for the phrase table entries and for the target sentence that is being translated.

ACKNOWLEDGEMENTS

This research was supported in part by NSF DBI-0317510 and by FP7-REGPOT-2007-1 SISTER.

REFERENCES

[1] Y. Al-Onaizan and K. Knight, 'Translating named entities using monolingual and bilingual resources', in *Proc. of ACL*, pp. 400–408, (2001).
[2] T. Baldwin and T. Tanaka, 'Translation by machine of compound nominals: Getting it right', in *Proceedings of ACL'04 Workshop on Multiword Expressions: Integrating Processing*, pp. 24–31, (2004).
[3] C. Bannard and C. Callison-Burch, 'Paraphrasing with bilingual parallel corpora', in *Proceedings of ACL*, pp. 597–604, (2005).
[4] R. Barzilay and K. McKeown, 'Extracting paraphrases from a parallel corpus', in *Proceedings of ACL*, pp. 50–57, (2001).
[5] P. Brown, V. Della Pietra, S. Della Pietra, and R. Mercer, 'The mathematics of statistical machine translation: parameter estimation', *Computational Linguistics*, **19**(2), 263–311, (1993).
[6] C. Callison-Burch, P. Koehn, and M. Osborne, 'Improved statistical machine translation using paraphrases', in *HLT*, pp. 17–24, (2006).
[7] Y. Cao and H. Li, 'Base noun phrase translation using web data and the EM algorithm', in *Proc. of Computational Linguistics*, pp. 1–7, (2002).
[8] G. Grefenstette, 'The World Wide Web as a resource for example-based machine translation tasks', in *Translating and the Computer 21*, (1999).
[9] D. Kauchak and R. Barzilay, 'Paraphrasing for automatic evaluation', in *Proceedings of HLT*, pp. 455–462, (2006).
[10] D. Klein and C. Manning, 'Accurate unlexicalized parsing', in *Proceedings of ACL*, pp. 423–430, (2003).
[11] P. Koehn, 'Pharaoh: a beam search decoder for phrase-based statistical machine translation models', in *Proc. of AMTA*, pp. 115–124, (2004).
[12] P. Koehn, 'Europarl: A parallel corpus for evaluation of machine translation', in *Proceedings of MT Summit*, pp. 79–86, (2005).
[13] P. Koehn and K. Knight, 'Feature-rich statistical translation of noun phrases', in *Proceedings of ACL*, pp. 311–318, (2003).
[14] Mark Lauer, *Designing Statistical Language Learners: Experiments on Noun Compounds*, Ph.D. dissertation, Department of Computing Macquarie University NSW 2109 Australia, 1995.
[15] Masaaki Nagata, Teruka Saito, and Kenji Suzuki, 'Using the web as a bilingual dictionary', in *Proceedings of the workshop on Data-driven methods in machine translation*, pp. 1–8, (2001).
[16] P. Nakov, 'Improving English-Spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing', in *Proceedings of Workshop on SMT*, (2008).
[17] F. J. Och, 'Minimum error rate training in statistical machine translation', in *Proceedings of ACL*, pp. 160–167, (2003).
[18] F. J. Och and H. Ney, 'A systematic comparison of various statistical alignment models', *Computational Linguistics*, **29**(1), 19–51, (2003).
[19] F. J. Och and H. Ney, 'The alignment template approach to statistical machine translation', *Computat. Linguistics*, **30**(4), 417–449, (2004).
[20] B. Pang, K. Knight, and D. Marcu, 'Syntax-based alignment of multiple translations: extracting paraphrases and generating new sentences', in *Proceedings of NAACL*, pp. 102–109, (2003).
[21] K. Papineni, S. Roukos, T. Ward, and W. Zhu, 'Bleu: a method for automatic evaluation of machine translation', in *Proceedings of ACL*, pp. 311–318, (2001).
[22] Y. Shinyama, S. Sekine, and K. Sudo, 'Automatic paraphrase acquisition from news articles', in *Proceedings of HLT*, pp. 313–318, (2002).
[23] T. Tanaka and T. Baldwin, 'Noun-noun compound machine translation: a feasibility study on shallow processing', in *Proceedings of ACL'03 workshop on Multiword expressions*, pp. 17–24, (2003).
[24] Y. Zhang and S. Vogel, 'Measuring confidence intervals for the machine translation evaluation metrics', in *Proceedings of TMI*, pp. 4–6, (2004).
[25] L. Zhou, C. Lin, and E. Hovy, 'Re-evaluating machine translation results with paraphrase support', in *Proc. of EMNLP*, pp. 77–84, (2006).