Example-based Paraphrasing for Improved Phrase-Based Statistical Machine Translation

Aurélien Max
LIMSI-CNRS & Univ. Paris Sud
Orsay, France
aurelien.max@limsi.fr

Abstract

In this article, an original view on how to improve phrase translation estimates is proposed. This proposal is grounded on two main ideas: first, that appropriate examples of a given phrase should participate more in building its translation distribution; second, that paraphrases can be used to better estimate this distribution. Initial experiments provide evidence of the potential of our approach and its implementation for effectively improving translation performance.

1 Introduction

Phrase translation estimation in Statistical Phrase-based Translation (Koehn et al., 2003) is hampered by the availability of both too many and too few training instances. Recent results on tera-scale SMT (Lopez, 2008) show that access to many training examples[1] can lead to significant improvements in translation quality. Also, providing indirect training instances via synonyms or paraphrases for previously unseen phrases can result in gains in translation quality, which are more apparent when little training data is originally available (Callison-Burch et al., 2006; Marton et al., 2009; Mirkin et al., 2009; Aziz et al., 2010). Although there is a consensus on the importance of using more parallel data in SMT, it has never been formally shown that all additional training instances are actually useful in predicting contextually appropriate translation hypotheses.

Attempts at limiting training parallel sentences to those resembling test data through thematic adaptation (Hildebrand et al., 2005) indeed confirm that large quantities of training data cannot compensate for the requirement for contextually appropriate training instances. In fact, it is important that phrase translation models adequatly reflect contextual preferences for each phrase occurrence in a text. A variety of recent works have used dynamically adapted translation models, where each phrase occurrence has its own translation distribution (Carpuat and Wu, 2007; Stroppa et al., 2007; Max et al., 2008; Gimpel and Smith, 2008; Haque et al., 2009) derived from local contextual information in the training examples.[2] These approaches are supported by the study of (Wisniewski et al., 2010) which shows that phrase-based SMT systems are expressive enough to achieve very high translation performance and therefore suggests a better scoring of phrases.

The apparent tradeoff between the number of training examples and their appropriateness in each indivual context naturally asks for means of increasing the number of appropriate examples. Exploiting comparable corpora for acquiring translation equivalents (Munteanu and Marcu, 2005; Abdul-Rauf and Schwenk, 2009) offers interesting prospects to this issue, but so far focus has not been so much on context appropriateness as on globally increasing the number of biphrase examples.

[1] To be more accurate, works such as that of (Lopez, 2008) have recourse to random sampling to build models of a manageable size in a reasonable amount of time.

[2] The study of (Carpuat, 2009) shows that the one translation per discourse hypothesis holds in some cases, but to our knowledge no SMT systems have attempted to exploit it yet. However, in our view, this finding does not contradict the need for estimating translation distributions at the individual phrase level, but they should be integrated as additional information.

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 656–666,
MIT, Massachusetts, USA, 9-11 October 2010.

The approach we take in this article is motivated by the fact that natural language allows for multiple text views on a given content, and that if two phrases are good paraphrases in context, then considering appropriate training examples of one of the phrases could provide larger quantities of training data for translating the other. In other words, we hypothesize that there may be more training data to learn a phrase's translations in a bilingual corpus than what SMT approaches typically use.

In contrast to previous attempts at using paraphrases to improve Statistical Machine Translation, which require external data in the form of additional parallel bilingual corpora (Callison-Burch et al., 2006), monolingual corpora (Marton et al., 2009), lexico-semantic resources (Mirkin et al., 2009; Aziz et al., 2010), or sub-sentential (Resnik et al., 2010) or sentential paraphrases of the input (Schroeder et al., 2009), the approach we take here can be endogenous with respect to the original training data. It also significantly departs from previous work in that paraphrasing is not simply considered as a way of finding alternative wordings that can be translated given the original training data for out-of-vocabulary phrases only (Callison-Burch et al., 2006; Marton et al., 2009; Mirkin et al., 2009; Aziz et al., 2010), but as a means to better estimate translations for *any* possible phrase. Also, as opposed to the work by (Schroeder et al., 2009; Onishi et al., 2010; Du et al., 2010), we do not encode paraphrases into input lattices to have them compete against each other to belong to the source sentential paraphrase that will lead to the highest scoring output sentence[3]. Instead, we make use of all contextually appropriate paraphrases of a source phrase, which *collectively evaluate the quality of each translation for that phrase*.

This work can thus be seen as a contribution towards shifting from global phrase translation distributions to contextual translation distributions for contextually equivalent source units. The remainder of this paper is organized as followed. In section 2 we review relevant previous works and discuss how they differ from our approach. Section 3 provides a description of the details of our approach. We describe an experimental setup in section 4 and comment on our results. Finally, we discuss our future work in section 5.

[3]This highly depends on how well estimated translations for each independent paraphrase are.

2 Relation to previous work

2.1 Contextual estimation of phrase translations

In standard approaches to phrase-based SMT, evidence of a translation is accumulated uniformely every time it is found associated with a source phrase in the training corpus. In addition to the fact that errors in automatic word alignment and non literal translations often produce useless biphrases, this results in rare but appropriate translations being very unlikely to be considered during decoding. Some approaches on source context modelling (Carpuat and Wu, 2007; Stroppa et al., 2007; Max et al., 2008; Haque et al., 2009) build classifiers offline for the phrases in a test set, so that context similarity can for example reinforce scores associated with rare but appropriate translations. However, heavy offline computation makes scaling to larger corpora an issue. Other approaches (Callison-Burch et al., 2005; Lopez, 2008) instead focus on accessing very large corpora. Indexing by suffix arrays is used to allow fast access to phrase instances in the corpus, and random sampling to avoid collecting the full set of examples has been shown to perform well. However, these approaches consider all instances of a phrase as equivalent for the estimation of its translations.

These works converge on the need for accessing a sufficient number of examples that are relevant for any source phrase in context, fast enough to permit on-the-fly phrase table building. This paper proposes an intermediate step: the full set of phrase examples is found efficiently, and a measure of the adequacy of each example with a phrase in context provides evidence for its translation that depends on this value of adequacy. In this way, the translation associated with an example for a different sense of a polysemous word would in the best scenario only be considered marginally when computing the translation distribution. As in most previous works, adequacy can be approximated by context similarity between phrase occurrences and training examples.

Ideally, one would stop extracting examples when enough appropriate examples have been found to estimate a reliable translation distribution. (Callison-

sample size	100	500	1K	5K	10k	50k	unlim.
BLEU score	28.8	28.8	28.8	28.9	29.1	28.9	29.0

Figure 1: Effect of number of samples on translation quality (measured on German to English translation on Europarl data) reported by (Callison-Burch et al., 2005)

Burch et al., 2005) measured the impact on translation quality of the sample size in random sampling of source phrase examples in the training corpus to estimate a phrase's translation probabilities. As Table 1 shows, quality (in terms of BLEU scores) almost remains constant for samples of size 100 or more. This apparent confirmation of the efficiency of random sampling is backed up by the authors with the following possible explanations: 1) the most probable translations remain the same for different sample sizes; 2) misestimated probabilities are ruled out by the target language model; and 3) longer or less frequent phrases, which are not affected by sampling, are preferred. However, as said previously, random sampling cannot guarantee that contextually-appropriate examples are selected. In fact, (Lopez, 2008) points out to using discriminatively trained models with contextual features of source phrases in conjunction with phrase sampling as an open problem. This work does not attempt to directly address it, but instead resorts to complete analysis of the training data to guarantee that all contextually-appropriate examples are considered.

2.2 Using paraphrases for translating

For some phrases, not enough examples can be found in the training corpus to estimate reliable translation probabilities in context. In such cases, one might be interested in finding more appropriate examples, which seems at first impossible using the sole original bilingual corpus. We can in fact consider the set of source phrases that have similar translations in context. This set is roughly made up of a subset of what can be referred to as *paraphrases*. One possible approach to extract local (i.e. phrasal) paraphrases precisely exploits similarity *on the target side* in another language by extracting source phrases that share common translations (Bannard and Callison-Burch, 2005), but recent approaches have combined this approach with

Source phrase	**Paraphrases**
Balkan War	*Balkan war* (0.25) *Balkans War* (0.125) *Balkans* (0.125) *Balkans war* (0.125) *war in the Balkans* (0.125) *Balkan conflict* (0.125)
British forces	*British troops* (0.29) *British armed_forces* (0.19)
Czech president	*President of the Czech Republic* (0.5)
Dalai Lama's	*of the Dalai Lama* (0.27)
I don't see	*I do not believe* (0.18) *I do not think* (0.18) *I do not see* (0.15)

Figure 2: Examples of paraphrases obtained by pivoting via French; values indicate *paraphrase probability* as defined in (Bannard and Callison-Burch, 2005).

similarity computation in the "source" (i.e. original) language (Callison-Burch, 2008; Max, 2008; Kok and Brockett, 2010). Figure 2 provides examples of English paraphrases obtained by automatically pivoting via French. As can be seen, some examples would be clearly useful to better estimate translations of the original source phrase: (*Balkan War* $\leftrightarrow$ *war in the Balkans*) are syntactic variants that can generally substitute with each other, (*Balkan War* $\leftrightarrow$ *Balkans war*) are character-level variants[4]. Other examples, however, clearly illustrate the need for validation in context: (*Dalai Lama's* $\leftrightarrow$ *of the Dalai Lama*) require different syntactic contexts, and (*I don't see* $\leftrightarrow$ *I do not believe*) are only interchangeable in specific semantic contexts.

Previous attempts at exploiting paraphrases in SMT have first concentrated on obtaining translations for phrases absent from the training corpus (Callison-Burch et al., 2006; Marton et al., 2009; Mirkin et al., 2009)[5], with modest gains in translation performance as measured by automatic metrics. (Callison-Burch et al., 2006) obtain paraphrases by pivoting via additional bilingual corpora and use the translations of known paraphrases to translate unseen phrases, which requires that the additional bilingual corpora contain the unseen source phrases and that some of the extracted paraphrases be present in the original corpus. (Marton et al.,

[4]To our knowledge, most implementations of SMT decoders do not integrate flexible matching of phrases.

[5]The work by (Mirkin et al., 2009) in fact considers both paraphrases and entailed texts to increase the number of properly translated texts.

2009) proceed similarly but obtain their paraphrases from comparatively much larger monolingual corpora by following the *distributionality hypothesis*. In both cases, gains are only obtained in very specific conditions where very few training data are available and where useful additional knowledge can be brought in from external resources. Furthermore, the described implementations do not consider acceptability of the paraphrases in context, as their underlying hypothesis is that it might be more desirable to translate some paraphrase than not to translate a given phrase.[6] In contrast, the work by (Mirkin et al., 2009) attempts to model context when using replacements for words (synonyms or hypernyms).

The natural next step that we take here is to exploit the complementarity of the original bilingual training corpus for finding paraphrases and the monolingual (source) side of the same corpus for validating them in context. Furthermore, our focus here is not on paraphrasing unseen phrases[7], but possibly any phrase, or any phrase seen less than a given number of times, or any types of difficult-to-translate phrases (Mohit and Hwa, 2007).

The recent work of (Resnik et al., 2010) uses crowdsourcing to obtain paraphrases for source phrases corresponding to mistranslated target phrases. The spotting of the incorrect target phrases and the paraphrasing of the source phrases can be automated. Promising oracle figures are obtained, validating the claim that some variations of the input sentence might be more easily translated than others by a given system. Paraphrases have also been used to represent alternative inputs encoded in lattices using existing (Schroeder et al., 2009) or automatically built paraphrases (Onishi et al., 2010; Du et al., 2010). In this scenario, paraphrases are in fact competing with each other, whereas in our proposal paraphrases *collectively participate in evaluating the quality of each translation for a source phrase*. We believe that if two phrases are indeed paraphrases in context, then their respective set of translations are both relevant to translate the two phrases. The target language model nevertheless still has an important role to play to select appropriate translations among semantically-compatible translations (i.e., target side paraphrases) in the specific context of a generated target hypothesis.

Lastly, automatic sentential paraphrasing has also been used in SMT to build alternative reference translations for parameter optimization (Madnani et al., 2008) and to build alternative training corpora (Bond et al., 2008).

3 Towards better exploitation of training corpora in phrase-based SMT

In typical phrase-based SMT settings (Koehn et al., 2003), words from the source side of the corpus are first aligned to words on the target side and biphrases are extracted from each training sentence using some heuristics on the word alignments. A source phrase f in a sentence being translated may therefore be aligned to a variety of target phrases. In the example on Figure 3, f is aligned some number of times in the training corpus to target phrases e_1, e_2, e_3 and e_5. Using the number of times f is paired with some target phrase e_i, $count(f, e_i)$, relative frequency estimation can be used to compute the probability of translation e_i given source phrase f:

$$p_{rel}(e_i|f) = \frac{count(f, e_i)}{\sum_j count(f, e_j)} \quad (1)$$

This value, together with other estimates of how appropriate a translation pair (f, e_i) is, are recorded in a phrase table, which typically discards all contextual information.[8] Therefore, the translation distribution of some phrase is globally estimated from a training corpus independently of the actual context of that phrase.[9] On Figure 3, phrase f has at least two distinct senses: one represented by set $\mathcal{E}$, which in our example corresponds to the appropriate sense for a particular occurrence of f in a test sentence; and one which corresponds to translation e_5. A typical problem, due to the lack of context modeling,

[6] The default strategy for most decoders is to copy out-of-vocabulary tokens into the final text.

[7] Doing it in conjunction with our approach for improving the translation of known phrases is part of our future work.

[8] See (Carpuat and Wu, 2007; Stroppa et al., 2007; Max et al., 2008; Gimpel and Smith, 2008; Haque et al., 2009) for notable exceptions.

[9] Context is in fact taken into account to some extent by the target language model, which should score higher translations that are more appropriate given a target translation hypothesis being built. In fact, in this work we consider the target language model as the main source of information for selecting among acceptable target phrases (target language paraphrases).

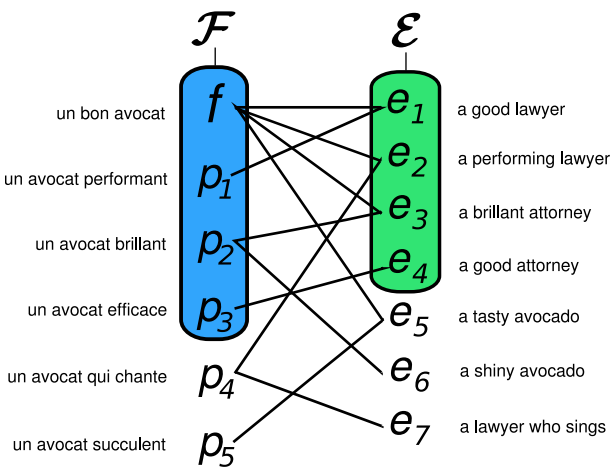

Figure 3: Example of possible source equivalents and translations for phrase occurrence f "*un bon avocat*" in the sentence "*L'embauche d'un bon avocat est cruciale quelle que soit l'activité*" ("*Hiring a good lawyer is crucial to any business"*). Set $\mathcal{E}$ represents target phrase types that are acceptable translations given the particular context of f, and set $\mathcal{F}$ represents source phrase types that can be in a paraphrasing relation to f depending on the context they appear in.

is that in situations such as $\forall e_i \in \mathcal{E}, count(f, e_5) \gg count(f, e_i)$, it is very unlikely that a correct translation will be selected during decoding against the incorrect but much more frequent one. Taking an extreme view on this issue, it is in fact desirable that when estimating phrase translation probabilities for a phrase f, translations of incompatible senses be not considered.[10] Of course, this raises the difficult issue of sense clustering of phrases. We propose here an intermediary solution, which consists in considering each occurrence in the training corpus as counting a number of times that depends on its contextual similarity with the occurrence of f from the test file, through the following additional translation model :

$$p_{cont}(e_i|f) = \frac{\sum_{\langle f_k, e_i \rangle} sim_{cont}(C(f), C(f_k))}{\sum_{\langle f_k, e_j \rangle} sim_{cont}(C(f), C(f_k))} \quad (2)$$

where f is some source phrase to translate and f_k an example of f in the training corpus, $\langle f_k, e_i \rangle$ is a biphrase from the training corpus, $C(f)$ the context of some source phrase, $C(f_k)$ the context of a particular example of f in the training corpus, sim_{phr} a function indicating the contextual similarity between two phrase contexts, and e_j is any possible translation of f.

The problem of modeling phrase translation is however not limited to inappropriate training examples. For various reasons, legitimate occurrences of source phrases may not be considered when building a phrase's translation distribution. We describe those cases by considering the possible source phrases p_i from Figure 3:

- p_1's only translation, e_1, is a common translation with f; each contextually-appropriate example of p_1 should reinforce the probability of e_1 being a translation for f.

- Contextually-appropriate examples of p_2 can reinforce e_3. Translation e_6 should correspond to contextually-inappropriate examples of p_2, so e_6 should not be considered as a new potential translation for f.

- Contrarily to the examples of p_2 translating as e_6, examples of p_3 translating as e_4 are much more likely of being contextually-appropriate with f, meaning that f could be substituted with most p_3 examples. Therefore, e_4, which was not initially considered as a possible translation of f, could now be considered as such.

- p_4 shares a translation with f, e_2, but this is due to the polysemous nature of this translation. Again, all examples of p_4 should be found contextually-inappropriate with f, and their translations should not be considered when estimating the translations of f.

- Lastly, the case of the common translation e_5 between f and p_5 illustrates a consequence of the polysemous nature of the source phrase corresponding to word sequence f: translations corresponding to other senses of f should not get reinforced by paraphrase examples such as those of p_5 as these examples should be found contextually-inappropriate with f.

[10] Put differently, is it more acceptable to copy a source word in the target hypothesis or to incorrectly translate it when the confidence about its being incorrect is high?

We build a separate translation model for translations estimated through paraphrases, defined as follows:

$$p_{para}(e_i|f) = \frac{\sum_{\langle p_k, e_i \rangle} sim_{para}(C(f), C(p_k))}{\sum_{\langle p_k, e_j \rangle} sim_{para}(C(f), C(p_k))} \quad (3)$$

where p_k is a paraphrase of f, $\langle p_k, e_i \rangle$ is a biphrase from the training corpus such that e_i is also a translation of f, $C(f)$ the context of a given source phrase for which we are estimating the translation distribution, $C(p_k)$ the context of a particular example of p_k in the training corpus, sim_{para} a function indicating the contextual similarity between a phrase context and a paraphrase context, and e_j is any possible translation of f.

Several requirements can be drawn from the previous description:

1. **List of potential paraphrases**: some mechanism for finding potential paraphrases for source phrases is required, and several such mechanisms could be combined. Pivoting via bilingual corpora, a natural strategy given the issue at hand, is just one among many different proposed strategies (Madnani and Dorr, 2010).

2. **Contextual similarity measure**: a similarity measure between the contexts of two phrases or two potential local paraphrases is required. This automatic measure should ideally be able to model not only syntactic but also semantic and pragmatic information.

3. **Robust translation evaluation**: our approach is designed to reinforce estimates for any contextually-appropriate translations of a phrase, as shown by set $\mathcal{E}$ on Figure 3. It is therefore important to have some means of accepting them as subparts of valid translations. Robustness in Machine Translation evaluation is an active domain, and potential candidates include using BLEU-like metrics with multiple references, Human-targeted Translation Error Rate (Snover et al., 2006) and the use of paraphrases for reference translations (Kauchak and Barzilay, 2006).

	train		dev.		test	
	# sent.	# tok.	# sent.	# tok.	# sent.	# tok.
en	318K	9.1M	500	14,0K	500	13,6K
fr	318K	10.3M	500	16,1K	500	15,7k

Figure 4: Statistics of the corpora used.

In this paper, we want to evaluate whether an endogenous approach for finding paraphrases can lead to some improvement in translation performance. Note that we will not consider in this initial work the possibility of adding new translations to phrases (such as e_4 for f on Figure 3) as it adds complexity and should be investigated when the other simpler cases can be handled successfully.

In the following section, we describe experiments in which the original bilingual corpus is the only resource used to find potential paraphrases and to estimate phrase translations in context. We chose a very simple measure of similarity, and let to our future work the task of improving context modeling. As regards evaluation, we will resort to various ways to measure the impact of our implementation on translation performance.

4 Experiments and results

4.1 Data and baseline SMT systems

We have conducted our experiments using the MOSES[11] package to build state-of-the-art phrase-based SMT systems for phrases of up to 5 tokens, using standard parameters and MERT for optimizing model weights. We used a subpart of the Europarl corpus[12] in French and English as our training corpus and built baseline MOSES systems (**bsl**) in both directions. The target side of the training corpus was used to train 3-gram target language model with modified Kneser-Ney smoothing. Held-out datasets were used for development and testing. The characteristics of all corpora are described in Figure 4.

4.2 Example-based Paraphrasing SMT systems

We also built systems that exploit phrase and paraphrase context under the form of two additional models p_{cont} and p_{para} described in section 3. These

[11] http://www.statmt.org/moses

[12] http://www.statmt.org/europarl

	phrase table size	num. entries
	en→fr	
baseline	240Mb	2.4M
our systems	5.0Gb	37.5M
	fr→en	
baseline	193Mb	1.9M
our systems	4.0Gb	30.2M

Figure 5: Statistics on the size and the number of entries of the phrase tables filtered on the development set.

models are added to the list of models used to evaluate the various translations of a phrase in the appropriate phrase tables, and are optimized with the other models by standard MERT.

In order to model context, we modified the source texts so that each phrase becomes unique in the phrase table, i.e. it has its own translation distribution. This is done (as in other works (Carpuat and Wu, 2007; Stroppa et al., 2007)) by transforming each token into a unique token, e.g. *token* → *token@337*. This therefore leads to a significant increase in the size of the phrase table, as illustrated on Figure 5, as all occurrences for the same phrase are not factored anymore.[13]

We chose a very simple initial definition of context similarity based on the presence of common *n*-grams in the immediate vicinity of two phrases. Let $length_{left}$ (resp. $length_{right}$) be the length of the longest common *n*-gram in the immediate vicinity on the left (resp. right) of two phrases in context ($C(f)$ and $C(f_i)$). For instance, given the two following contexts (phrases under focus are in bold and common *n*-grams are underlined):

1. *the commission accepts the substance of the* ***amendments@11257 proposed@11258 by@11259*** *the committee on fisheries ...*

2. *this is why we shall support all of the* ***amendments put forward by*** *the committee on agriculture and rural development ...*

$length_{left} = 2$ and $length_{right} = 3$. We further define $length$ as:

$$length = \begin{cases} length_{left} + length_{right} \\ \quad \text{if } length_{left} > 0 \\ \quad \text{and } length_{right} > 0 \\ 0 \text{ otherwise} \end{cases} \quad (4)$$

We can now define the two similarity functions used in Equations 2 and 3 that we used for our experiments:

$$sim_{cont}(C(f), C(f_i)) = (1 + length)^{\alpha} \quad (5)$$
$$sim_{para}(C(f), C(p_i)) = (length)^{\beta} \quad (6)$$

The rationale for these functions is the following. Exact phrase examples add at least a translation count of 1, i.e. their translation is always taken into account to estimate p_{cont}. Paraphrase examples add a translation count of 0 if $length = 0$, i.e. their translation is not taken into account at all if surrounding *n*-gram similarity is too low. We used $\alpha = \beta = 1.5$. Algorithm 1 describes how the two models are estimated from the training data.

foreach *phrase* f *in training file* **do**
  extract $C(f)$;
  /* phrase count */
  foreach *unique phrase* f_i *in* $test(f)$**)** **do**
    extract $C(f_i)$;
    compute $sim_{cont}(C(f), C(f_i))$;
  end
  /* paraphrase count */
  foreach *phrase* p_i *in* $para(f)$ **do**
    foreach *unique phrase* f_j *in* $test(p_i)$ **do**
      extract $C(f_j)$;
      compute $sim_{para}(C(f), C(f_j))$;
    end
  end
end
estimate p_{cont} and $_{para}$;

Algorithm 1: Model estimation for p_{cont} and p_{para}. Function $test(f)$ returns all unique phrases corresponding to phrase f from the test file. Function $para(f)$ return all phrases for which f is a known paraphrase.

We implemented the following strategy to find paraphrases for phrases in the test file. We extract all

[13] These volumes of data and our available hardware facilities for these experiments led us to initially limit the size of our data sets. We will discuss in section 5 how we intend to address this limitation in our future work.

	Left context	phrase/paraphrase	Right context
IS#1	*at the moment it is*	**up to each**	*member state to decide, and practice differs considerably from country to country*
PE#1	*... as regards the terminal portion in the cycle of nuclear fuel, it is*	**the responsability of each**	*member state to define its own policy .*
PT#1		la responsabilité de chaque	
IS#2	*that is why i find it*	**extremely regrettable that**	*the amendment on harmonising the re-registration of cars that have been involved in accidents ...*
PE#2	*for all these reasons and given your most excellent statement , i find it*	**a pity that**	*the new legal base for the daphne programme is so restrictive ...*
PT#2		dommage que	

Figure 6: Examples of paraphrases in context from the development file. The input sentence (IS) contains a source phrase of interest (in bold), the paraphrase example (PE) contains a paraphrase of that source phrase (in bold) for which a paraphrase translation (PT) is known.

paraphrases p for a phrase f by pivot: all target language phrases e aligned to f are first extracted, and all source language phrases p aligned to e are extracted. The following constraints are then applied to define which paraphrases are kept:

- string p is not included in string f and vice versa (in order to minimize the impact of alignment errors in the training corpus);
- the paraphrasing probability is greater than a fixed threshold: $para(f,p) \geq 10^{-2}$, where $para(f,p) = \sum_e p(e|f)p(p|e)$ (Bannard and Callison-Burch, 2005);
- the number of occurrences of phrase f and paraphrase p are equal or less than independent thresholds: $numOccs(f) \leq 100$ and $numOccs(p) \leq 1000$.[14]

Figure 6 shows examples of paraphrases in context with high similarity with some original phrase, and Figure 7 provides various statistics on the paraphrases extracted on the test file.

4.3 Results and analysis

Automatic evaluation results are reported in Table 8 for various configurations. We also wanted to focus our measures on content words, which are known to be important as regards information content in translation. We applied the contrastive lexical evaluation (CLE) methodology described in (Max et al., 2010), which indicates how many times source words grouped into user-defined classes were correctly translated or not across systems. These additional results are reported on Figure 9.

[14]The first threshold value was chosen as (Callison-Burch et al., 2005) report it to be an optimal sample size for estimating phrase translation probabilities. The relatively low value for the second threshold was selected to reduce computation time.

phrase length	**# phrases**		**# paraphrased**		**# paraphrases**	
	en	fr	en	fr	en	fr
1	13,620	15,707	458	725	1,824	2,684
2	13,120	15,207	4,127	4,481	18,871	19,700
3	12,620	14,707	4,782	5,715	24,111	27,377
4	12,120	14,208	2,859	4,078	15,071	20,345
5	11,623	13,711	1,171	2,275	6,077	12,132

Figure 7: Statistics on numbers of phrases, numbers of paraphrased phrases and numbers of paraphrases per phrase length.

On English to French translation, both additional features lead to improvements over the baseline with all metrics, including CLE, and their combination shows a strong improvement in TER (-1.55). CLE on content words reveals that the **para** feature seems particularly effective in reducing the number of words in all categories that only the baseline system translated correctly.

Results on French to English translation are less positive: neither **cont** nor **para** alone improve over the baseline with any metrics. However, their combination improves over the baseline with all metrics except BLEU, including a reduction of -1.07 in TER. Detailed analysis of CLE results shows that the translation of adjectives and nouns benefited

more from using our two additional models. Verbs, whose translation improved slightly, are strongly inflected in French, so finding examples for a given form is more difficult than for less inflected word categories, as is finding paraphrases with the appropriate inflection. Also, pivoting via English is one reason why paraphrases obtained via a low-inflected language can be of varying quality. Furthermore, the simplicity of our context modeling may have been ineffective in filtering out some bad examples. Overall, **para** was more effective with the low-inflected English as the source language, improving over the baseline with all metrics.

These results confirm that translation performance can be improved by exploiting context and paraphrases in the original training corpus only. We next attempted to measure whether some improvement in the quality of the paraphrases used would have some measurable impact on translation performance. To this end, we devised a semi-oracle experiment in the following way: the source and target test files were automatically aligned, and for each source phrase possible target phrases (i.e., reference translations) were extracted, and used as pivots to extract potential paraphrases, which were then filtered with the same constraints as previously. In this way, we exploit the information that paraphrases can at least produce the desired translation, but they may also propose other incorrect translations and/or be present in very few examples. Results appear in the **inf** rows of Tables 8 and 9. We obtain the most important improvement over the baseline in BLEU for the two language pairs (resp. +0.99 and +0.44), though the results for the other metrics for French to English translation are more difficult to interpret. For this language pair, possible reasons include again that the pivot language may not be appropriate, and also that the limitation to a single pivot[15] may not have produced more monolingual variation that might have proved useful. CLE on English to French, however, reveals significant gains with a relative improvement over the baseline of +116 content words. Under this condition, this result shows that the higher the quality of the paraphrases used, the more translation quality can be improved, which is in line with works that make use of human-made paraphrases to improve translation quality (Schroeder et al., 2009; Resnik et al., 2010).

Table 10 presents a typology of paraphrases found in our development set and classifies the impact of using them for phrase translation estimation. As can be seen, more work is needed to better understand the characteristics of the phrases that should be paraphrased and of their paraphrases.

[15]Several pivot phrases may in fact have been automatically extracted for a given phrase, some of which being possible bad candidates.

	BLEU		NIST		TER		METEOR	
	en→fr							
bsl	30.28	-	6.66	-	57.86	-	54.79	-
+cont	**31.11**	+0.83	6.77	+0.11	57.24	-0.62	55.22	+0.43
+para	30.97	+0.69	6.74	+0.08	57.38	-0.48	**55.39**	+0.60
all	30.93	+0.65	**6.84**	+0.18	**56.31**	-1.55	55.28	+0.49
inf	31.27	+0.99	6.78	+0.12	57.22	-0.64	55.80	+1.01
	fr→en							
bsl	**29.90**	-	6.90	-	54.64	-	61.36	-
+cont	29.56	-0.34	6.89	-0.01	54.95	+0.31	60.98	-0.38
+para	29.70	-0.20	6.92	+0.02	54.64	+0.00	61.10	-0.26
all	29.75	-0.15	**7.03**	+0.13	**53.57**	-1.07	**61.63**	+0.27
inf	30.34	+0.44	6.93	+0.03	54.90	+0.26	60.99	-0.37

Figure 8: Automatic scores for the MOSES baseline systems (**bsl**), systems additionnally using the contextual feature (**+cont**), systems additionnally using the paraphrasing feature (**+para**), systems using both features (**all**), and pivot-informed systems (**inf**).

		Adj	Adv	Noun	Verb	∑	
		en→fr					
+cont	-	74	28	113	60	275	
	+	55	35	114	85	289	+14
+para	-	62	12	82	46	202	
	+	58	32	111	78	279	+77
all	-	72	25	91	72	260	
	+	50	37	118	97	302	+42
inf	-	58	20	108	56	242	
	+	65	43	147	103	358	+116
		fr→en					
+cont	-	30	16	80	69	195	
	+	15	21	69	46	151	-44
+para	-	32	19	72	60	183	
	+	12	18	65	43	138	-45
all	-	21	18	67	61	167	
	+	30	18	94	48	190	+23
inf	-	38	21	83	66	208	
	+	31	23	106	57	217	+9

Figure 9: Contrastive lexical evaluation results per part-of-speech measured on the test file. '-' (resp. '+') rows indicate the number of source words that only **bsl** (resp. the compared system) correctly translated.

Type	Impact	Examples
Morphological variants	+/-	(yugoslav republic ↔ yugoslavian republic), (go far ↔ goes far)
Synonymy	+	(duties ↔ obligations), (to look into ↔ to study)
Grammatical word substitution	?/-	(states in the ↔ the states of the), (amendments by ↔ amendments to)
Word deletion or insertion	?/-	(first reading, the → first reading the), (amendments by ↔ amendments proposed by)
Syntactic rewritings	+	(approval of the majority ↔ majority support), (capacity of the european union ↔ european union's ability)
Phrasal idiomatic substitutions	+	(must be said that the ↔ goes without saying that the), (is fully in line ↔ is totally coherent), (is amazing ↔ strikes me)
Context-dependent substitutions	+/-	(is not right ↔ is unacceptable), (offer my ↔ express my)
Alignment and translation problems	-	(unnecessary if ↔ vital if), (the crime ↔ organized), (ill-advised ↔ wise), (to begin by thanking ↔ to begin by congratulating)

Figure 10: Main types of paraphrase pairs found in our dev. and training corpora. Pairs shown have $length > 0$.

5 Conclusion and future work

We have introduced an original way of exploiting both context and paraphrasing for the estimation of phrase translations in phrase-based SMT. To our knowledge, this is the first time that paraphrases acquired in an endogenous manner have been shown to improve translation performance, which shows that bilingual corpora can be better exploited than they typically are. Our experiments further showed the promises of our approach when paraphrases of higher quality are available.

In the light of our results and our initial typology of paraphrases presented on Figure 10, as well as previous work on paraphrasing for SMT, the difficult question of what units should be paraphrased for what success should be addressed, taking into account parameters such as language pairs, quantity of training data and availability of external resources.

Our future work includes three main areas: first, we want to improve the modeling of context, by notably working on techniques inspired from Information Retrieval to quickly access contextually-similar examples of source phrases in bilingual corpora. Such *contextual sampling* on large bilingual corpora for phrases and their paraphrases, which could integrate more complex linguistic information, will allow us to assess our approach on more challenging conditions. This would also allow us to build contextual models on-the-fly, and experiment with using lattices to encode contextually estimated paraphrases. Second, we will combine paraphrases obtained via different techniques and resources, which will allow us to also learn translation distributions for phrases absent from the original corpus. Lastly, we want to also exploit paraphrases for the *additional* translations that they propose (such as e_4 on Figure 3) and that would be contextually similar in the target language to other existing translations of a given phrase or that could even represent a new sense of the original phrase.

Acknowledgements

This work was partly supported by ANR project Trace (ANR-09-CORD-023). The author would like to thank the anonymous reviewers for their helpful questions and comments.

References

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the Use of Comparable Corpora to Improve SMT performance. In *Proceedings of EACL*, Athens, Greece.

Wilker Aziz, Marc Dymetman, Shachar Mirkin, Lucia Specia, Nicola Cancedda, and Ido Dagan. 2010. Learning an Expert from Human Annotations in Statistical Machine Translation: the Case of Out-of-Vocabulary Words. In *Proceedings of EAMT*, Saint-Raphael, France.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of ACL*, Ann Arbor, USA.

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, USA.

Chris Callison-Burch, Colin Bannard, and Josh Schroeder. 2005. Scaling Phrase-Based Statis-

tical Machine Translation to Larger Corpora and Longer Phrases. In *Proceedings of ACL*, Ann Arbor, USA.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved Statistical Machine Translation Using Paraphrases. In *Proceedings of NAACL*, New York, USA.

Chris Callison-Burch. 2008. Syntactic Constraints on Paraphrases Extracted from Parallel Corpora. In *Proceedings of EMNLP*, Hawai, USA.

Marine Carpuat and Dekai Wu. 2007. Context-Dependent Phrasal Translation Lexicons for Statistical Machine Translation. In *Proceedings of Machine Translation Summit XI*, Copenhagen, Denmark.

Marine Carpuat. 2009. One Translation Per Discourse. In *Proceedings of the NAACL-HLT Workshop on Semantic Evaluations*, Boulder, USA.

Jinhua Du, Jie Jiang, and Andy Way. 2010. Facilitating Translation Using Source Language Paraphrase Lattices. In *Proceedings of EMNLP*, Cambridge, USA.

Kevin Gimpel and Noah A. Smith. 2008. Rich Source-Side Context for Statistical Machine Translation. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, Columbus, USA.

Rejwanul Haque, Sudip Kumar Naskar, Yanjun Ma, and Andy Way. 2009. Using Supertags as Source Language Context in SMT. In *Proceedings of EAMT*, Barcelona, Spain.

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adaptation of the Translation Model for Statistical Machine Translation Based on Information Retrieval. In *Proceedings of EAMT*, Budapest, Hungary.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. In *Proceedings of NAACL HLT*, New York, USA.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of NAACL HLT*, Edmonton, Canada.

Stanley Kok and Chris Brockett. 2010. Hitting the Right Paraphrases in Good Time. In *Proceedings of NAACL*, Los Angeles, USA.

Adam Lopez. 2008. Tera-Scale Translation Models via Pattern Matching. In *Proceedings of COLING*, Manchester, UK.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating Phrasal & Sentential Paraphrases: A Survey of Data-Driven Methods. *Computational Linguistics*, 36(3).

Nitin Madnani, Philip Resnik, Bonnie J. Dorr, and Richard Schwartz. 2008. Are Multiple Reference Translations Necessary? Investigating the Value of Paraphrased Reference Translations in Parameter Optimization. In *Proceedings of AMTA*, Waikiki, USA.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved Statistical Machine Translation Using Monolingually-derived Paraphrases. In *Proceedings of EMNLP*, Singapore.

Aurélien Max, Rafik Makhloufi, and Philippe Langlais. 2008. Explorations in using grammatical dependencies for contextual phrase translation disambiguation. In *Proceedings of EAMT*, Hamburg, Germany.

Aurélien Max, Josep M. Crego, and François Yvon. 2010. Contrastive Lexical Evaluation of Machine Translation. In *Proceedings of LREC*, Valletta, Malta.

Aurélien Max. 2008. Local rephrasing suggestions for supporting the work of writers. In *Proceedings of GoTAL*, Gothenburg, Sweden.

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 ACL*, Singapore.

Behrang Mohit and Rebecca Hwa. 2007. Localization of Difficult-to-Translate Phrases. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-parallel Corpora. *Computational Linguistics*, 31(4).

Takashi Onishi, Masao Utiyama, and Eiichiro Sumita. 2010. Paraphrase Lattice for Statistical Machine Translation. In *Proceedings of ACL, short paper session*, Uppsala, Sweden.

Philip Resnik, Olivia Buzek, Chang Hu, Yakov Kronrod, Alex Quinn, and Benjamin B. Bederson. 2010. Improving Translation via Targeted Paraphrasing. In *Proceedings of EMNLP*, Cambridge, USA.

Josh Schroeder, Trevor Cohn, and Philipp Koehn. 2009. Word Lattices for Multi-Source Translation. In *Proceedings of EACL*, Athens, Greece.

Matthew Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*, Boston, USA.

Nicolas Stroppa, Antal van den Bosch, and Andy Way. 2007. Exploiting Source Similarity for SMT using Context-Informed Features. In *Proceedings of TMI*, Skovde, Sweden.

Guillaume Wisniewski, Alexandre Allauzen, and François Yvon. 2010. Assessing Phrase-based Translation Models with Oracle Decoding. In *Proceedings of EMNLP*, Cambridge, USA.