Multilingual WSD-like Constraints for Paraphrase Extraction

Wilker Aziz
Research Group in Computational Linguistics
University of Wolverhampton, UK
W.Aziz@wlv.ac.uk

Lucia Specia
Department of Computer Science
University of Sheffield, UK
L.Specia@sheffield.ac.uk

Abstract
The use of pivot languages and word-alignment techniques over bilingual corpora has proved an effective approach for extracting paraphrases of words and short phrases. However, inherent ambiguities in the pivot language(s) can lead to inadequate paraphrases. We propose a novel approach that is able to extract paraphrases by pivoting through multiple languages while discriminating word senses in the input language, i.e., the language to be paraphrased. Text in the input language is annotated with “senses” in the form of foreign phrases obtained from bilingual parallel data and automatic word-alignment. This approach shows 62% relative improvement over previous work in generating paraphrases that are judged both more accurate and more fluent.

1 Introduction
Paraphrases are alternative ways of expressing a given meaning. Generating paraphrases that go beyond morphological variants of the original text is a challenging problem and has been shown to be useful in many natural language applications. These include i) expanding the set of reference translations for Machine Translation (MT) evaluation (Denkowski and Lavie, 2010; Liu et al., 2010) and parameter optimisation (Madnani et al., 2007), where multiple reference translations are important to accommodate for valid variations of system translations; ii) addressing the problem of out-of-vocabulary words or phrases in MT, either by replacing these by paraphrases that are known to the MT system (Mirkin et al., 2009) or by expanding the phrase table with new translation alternatives (Callison-Burch et al., 2006); and iii) expanding queries for improved coverage in question answering (Riezler et al., 2007).

Bannard and Callison-Burch (2005) introduced an approach to paraphrasing which has shown particularly promising results by pivoting through different languages for which bilingual parallel data is available. The approach consists in aligning phrases in the bilingual parallel corpus to find pairs of phrases \((e_1, e_2)\) in the input language, i.e., the language to be paraphrased, which typically align to the same foreign phrases \(F = \{f : e_1 \rightarrow f \rightarrow e_2\}\). This intermediate language is called pivot language and the phrases \(f \in F\) that support the equivalence \((e_1, e_2)\) are called pivot phrases. If there exists a non-empty set of pivots connecting \(e_1\) to \(e_2\), \(e_2\) is said to be a paraphrase of \(e_1\). The paraphrase is scored in terms of the conditional probabilities observed in the parallel corpus\(^1\) by marginalising out the pivot phrases that support the alignment \((e_1, e_2)\) as shown in Equation 1.

\[
p(e_2|e_1) = \sum_{f \in F} p(f|e_1)p(e_2|f) \quad (1)
\]

Equation 1 allows paraphrases to be extracted by using multiple pivot languages such that these languages help discard inadequate paraphrases resulting from ambiguous pivot phrases. However in this formulation all senses of the input phrase are mixed together in a single distribution. For example, for the Spanish input phrase acabar con, both paraphrases superar (overcome) and eliminar (eliminate) may be adequate depending on the context, however they are not generally interchangeable. In (Bannard and Callison-Burch,\(^1\) the distributions \(p(f|e)\) and \(p(e|f)\) are extracted from relative counts in word-aligned parallel corpus.

\(^1\)The distributions \(p(f|e)\) and \(p(e|f)\) are extracted from relative counts in word-aligned parallel corpus.
2005), the distributions learnt from different bilingual corpora are combined through a simple average. This makes the model naturally favour the most frequent senses of the phrases, assigning very low probabilities to less frequent senses. Section 5 shows evidence of how this limitation makes paraphrases with certain senses unreachable.

We propose a novel formulation of the problem of generating paraphrases that is constrained by sense information in the form of foreign phrases, which can be thought of as a quasi-sense annotation. Using a bilingual parallel corpus to annotate phrases with their quasi-senses has proved helpful in building word-sense disambiguation (WSD) models for MT (Carpuat and Wu, 2007; Chan et al., 2007): instead of monolingual senses, possible translations of phrases obtained with word-alignment were used as senses. Our approach performs paraphrase extraction by pivoting through multiple languages while penalising senses of the input that are not supported by these pivots.

Our experiments show that the proposed approach can effectively eliminate inadequate paraphrases for polysemous phrases, with a significant improvement over previous approaches. We observe absolute gains of 15-25% in precision and recall in generating paraphrases that are judged fluent and meaning preserving in context.

This paper is structured as follows: Section 2 describes additional previous work on paraphrase extraction and pivoting. Section 3 presents the proposed model. Section 4 introduces our experimental settings, while Section 5 shows the results of a series of experiments.

2 Related work

In addition to the well-known approach by (Bannard and Callison-Burch, 2005), the following previous approaches using pivot languages for paraphrasing can be mentioned. For a recent and comprehensive survey on a number of data-driven paraphrase generation methods, we refer the reader to (Madnani and Dorr, 2010).

Cohn and Lapata (2007) make use of multiple parallel corpora to improve Statistical Machine Translation (SMT) by triangulation for languages with little or no source-target parallel data available. Translation tables are learnt by pivoting through languages for which source-pivot and pivot-target bilingual corpora can be found. Multiple pivot languages were found useful to preserve the meaning of the source in the triangulated translation, as different languages are likely to realise ambiguities differently. Although their findings apply to generating translation candidates, the input phrases are not constrained to specific senses, and as a consequence multiple translations, which are valid in different contexts but not generally interchangeable, are mixed together in the same distribution. In SMT the target Language Model (LM) helps selecting the adequate translation candidate in context.

Callison-Burch (2008) extends (Bannard and Callison-Burch, 2005) by adding syntactic constraints to the model. Paraphrase extraction is done by pivoting using word-alignment information, as before, but sentences are syntactically annotated and paraphrases are restricted to those with the same syntactic category. This addresses categorial ambiguity by preventing that words with a given category (e.g. a noun) are paraphrased by words with other categories (e.g., a verb). However, the approach does not solve the more complex issue of polysemous paraphrases: words with the same category but different meanings, such as the noun bank as financial institution and land alongside a river/lake.

Marton et al. (2009) derive paraphrases from monolingual data using distributional similarity metrics. The approach has the advantage of not requiring bilingual parallel data, but it suffers from issues typical of distributional similarity metrics. In particular, it produces paraphrases that share the same or similar contexts but are related in ways that do not always characterise paraphrasing, such as antonymy.

3 Paraphrasing through multilingual constraints

Our approach to paraphrasing can be applied to both individual words or sequences of words of any length, conditioned only on sufficient evidence of these segments in a parallel corpus. We use segments as provided by the standard phrase extraction process from phrase-based SMT approaches (see Section 4), which in most cases range from individual words to short sequences of words (up to seven words in our case). Hereafter, we refer to these segments simply as phrases.

A model for paraphrasing under a constrained set of senses should take into account both the input phrase and the sense tag while selecting
Table 1: Size of the bilingual parallel corpora in millions of sentence pairs

<table>
<thead>
<tr>
<th>Paired with</th>
<th>en</th>
<th>de</th>
<th>nl</th>
<th>da</th>
<th>sv</th>
<th>fi</th>
<th>fr</th>
<th>it</th>
<th>pt</th>
<th>cl</th>
</tr>
</thead>
<tbody>
<tr>
<td>es</td>
<td>1.78</td>
<td>1.56</td>
<td>1.62</td>
<td>1.61</td>
<td>1.51</td>
<td>1.58</td>
<td>1.65</td>
<td>1.51</td>
<td>1.60</td>
<td>5.68</td>
</tr>
<tr>
<td>en</td>
<td>-</td>
<td>1.73</td>
<td>1.82</td>
<td>1.78</td>
<td>1.67</td>
<td>1.58</td>
<td>1.62</td>
<td>1.51</td>
<td>1.60</td>
<td>5.68</td>
</tr>
</tbody>
</table>

In order to constrain the extraction of paraphrases such that it complies with a sense repository, in addition to bilingual parallel corpora between the input language and the pivot languages, our model requires bilingual parallel corpora between the input language and the pivot languages and the language that is used for sense annotation.

Callison-Burch (2007) discusses factors affecting paraphrase quality, one of which is word senses. Paraphrasing through pivoting essentially relies on the hypothesis that different pivot phrases can be used to identify synonymy, rather than polysemy (an assumption made in the WSD literature). Callison-Burch (2007) also proposes an extraction procedure that may be conditioned on specific contexts of the input phrase (Bannard and Callison-Burch, 2005), where the context is a given pivot phrase. However, that model is unable to pivot through multiple languages. As we show in Section 5, this makes the model extremely sensitive to ambiguities of the one phrase used as both sense tag and pivot.

The model we propose attempts to perform sense-disambiguated paraphrase extraction, that is, paraphrases are discovered in the context of translation candidates of the input phrases. In addition, it allows the use of multiple pivot languages in the process, capitalising on both the WSD and the paraphrase assumption. While the target phrases discriminate different senses of the input phrases, the pivot phrases coming from multiple languages bring extra evidence to jointly capture the ambiguities introduced by the target phrases themselves.

To illustrate the impact of this contribution, consider the polysemous Spanish word *forma*, and some of its translations into English extracted from our corpus (Section 4): *kind, way, means* and *form*. The English words distinguish three possible senses of *forma*: (a) means/way of doing/achieving something, (b) shape, and (c) type or group sharing common traits. The model presented in (Bannard and Callison-Burch, 2005) cannot discriminate these senses. It mixes valid senses of *forma* and (correctly) proposes the paraphrases *manera* and *modo* for sense (a), and *tipo*...
for sense (c). However, paraphrases for sense (b) are over penalised and account for very little of the probability mass of the candidate paraphrases of *forma*. Their extension which conditions extraction on a given pivot phrase is highly sensitive to the ambiguities of the phrase used as sense annotation. Table 5 shows how this model (CB-wsd in the Table) makes mistakes for most senses of the input due to the ambiguities of the English context *kind, way, means* and *form*. Our approach (multi in the Table) on the other hand successfully separates paraphrases according to the sense annotation provided.

4 Experimental settings

4.1 Resources

The source of bilingual data used in the experiments is the Europarl collection (Koehn, 2005). We paraphrase Spanish (es) phrases using their corresponding English (en) phrases as sense tags and nine European languages as pivots: German (de), Dutch (nl), Danish (da), Swedish (sv), Finnish (fi), French (fr), Italian (it), Portuguese (pt) and Greek (el). The tools provided along with the corpus were used to extract the sentence aligned parallel data as shown in Table 1.

The sentence aligned parallel data is first word-aligned using GIZA++ in both source-target and target-source directions, followed by the application of traditional symmetrisation heuristics (Och and Ney, 2003). These aligned corpora are used for paraphrase extraction, except for a subset of them used in the creation of a test set (Section 4.2).

4.2 Test set creation

Since we are interested in showing the ability of our approach to find adequate paraphrases in the presence of a foreign phrase (the sense tag), it is important that our test set contains polysemous phrases. Like in (Bannard and Callison-Burch, 2005), we use the Spanish WordNet\(^3\) to bias our selection of phrases to paraphrase to contain ambiguous cases. However, rather than biasing selection towards having more multi-word expressions, we chose to have more polysemous cases. From the Spanish WordNet, we selected 50 phrases (with at least one content word) to be paraphrased such that 80% of the samples (40 phrases) had at least 2 senses (with a given part-of-speech tag to avoid selecting simpler, categorial ambiguities). Figure 2 lists the selected words and phrases in their base forms.

The bilingual corpus was queried for sentences containing at least one of the 50 phrases listed in Figure 2, or any of their morphological variants. The resulting sentences were then grouped on the basis of whether or not they shared the same English translation. To find the English phrase (i.e., our sense tag) which constrains the sense of the Spanish phrase, we followed the heuristics used in phrase-based SMT to extract the minimal phrase pair that includes the Spanish phrase and is consistent with the word-alignment\(^4\) (Koehn et al., 2003). We discarded groups containing fewer than five sentence pairs and randomly sampled 2-6 contexts per Spanish phrase. The resulting test set is made of 258 Spanish phrases in context such as the one exemplified in Figure 3.

4.3 Paraphrasing

Nine pivot languages were used to constrain paraphrase extraction following the approach presented in Section 3. The conditional probability distributions over phrase pairs in Equation 2 are estimated using relative frequencies. For each Spanish phrase in the test set, we retrieve their

<table>
<thead>
<tr>
<th>Unambiguous</th>
<th>Ambiguous</th>
</tr>
</thead>
<tbody>
<tr>
<td>concreto, política, fondo, regular, haber, amor propio, sangre fria, dar a luz, dar con, tomar el pelo</td>
<td>derecho, comercial, real, particular, legal, justo, común, cerca, esencial, especial, fuerte, puesto, oficial, figura, informe, parte, cuenta, forma, claro, clave, tiempo, seguro, respuesta, trabajar, responder, garantizar, volver, aumentar, incluir, tratar, ofrecer, establecer, pasar, dejar, realizar, punto de vista, llevar a cabo, dar vueltas, tener que, acabar con</td>
</tr>
</tbody>
</table>

Figure 2: Words and phrases selected to be paraphrased. Ambiguity is determined on the basis of the number of synsets in the Spanish WordNet. We note that this information was only used to bias the selection of the phrases, i.e., WordNet is not used in the proposed approach.

La idea de conceder a la Unión Europea su propia competencia fiscal - la palabra clave es el "impuesto por Europa" - está siendo debatida.

The idea of granting the EU its own tax competence - the keyword is the “Europe tax” - is being discussed.

Figure 3: Example of context selected for the phrase clave.

\(^3\)http://nlp.lsi.upc.edu/freeling/

\(^4\)Note that we did not use gold-standard word-alignments.
paraphrase candidates grouped by sense (English translation) and rank them based on the evidence collected from all bilingual corpora. Evidence from different pivot languages is combined using their average. English itself was not used as a pivot language. It was used only to provide sense tags. The rationale behind this choice is that if the language used to provide sense tags is also used as pivot language, there is no obvious way of estimating $p(q|f)$ in Equation 2. Note that in this case this probability would represent the likelihood of the English phrase aligning to itself.

Similar to (Bannard and Callison-Burch, 2005), we weight our paraphrase probabilities using an LM to adjust it to the context of the input sentence. We use a 5-gram LM trained on the Spanish part of Europarl with the SRILM toolkit (Stolcke, 2002). Paraphrases are re-ranked in context by multiplying the paraphrase probability and the LM score of the sentence.\(^5\)

In order to assess the performance of our model, we compare it to two variants of the models proposed by Bannard and Callison-Burch (2005).

**multi**: the paraphrasing model with multilingual constraints introduced in this paper.

**CCB**: the model in (Bannard and Callison-Burch, 2005) which does not explicitly perform any sense disambiguation.

**CCB-wsd**: an extended model in (Bannard and Callison-Burch, 2005) using English phrases as sense tags for pivoting.

Using each of these three models, we paraphrased the 258 samples in our test set, retrieving the 3-best paraphrases in context for each model. CCB is used with 10 pivot languages (English is included as a pivot) to generate paraphrase candidates. Note that CCB relies solely on the LM component to fit the paraphrase candidate to the context. On the other hand, CCB-wsd and multi both have access to sense annotation, but while multi is able to benefit from multiple pivot languages, CCB-wsd can only pivot through the one English phrase provided as sense annotation.

\(^5\)Given the localised effect of the phrase replacement within a given context in terms of n-gram language modelling, a neighbourhood of $n$-1 words on each side of the selected phrase is sufficient to re-rank paraphrase candidates: $p(w_{-4} \ldots w_{-1}, w_{+1} \ldots w_{+4})$ for our 5-gram LM.

### 4.4 Evaluation

To assess whether the proposed model effectively disambiguates senses of candidate paraphrases, we perform experiments using similar settings to those in (Bannard and Callison-Burch, 2005). Paraphrases are evaluated in context (a sentence) using binary human judgements in terms of the following components:

- **Meaning (M)**: whether or not the candidate conveys the meaning of the original phrase; and
- **Grammar (G)**: whether or not the candidate preserves the fluency of the sentence.

These two components are assessed separately and a paraphrase candidate is considered to be **correct** only when it is judged to be both meaning preserving and grammatical. Our evaluators were presented with one pair of sentences at a time, the original one and its paraphrased version. For every test sample we selected the 3-best paraphrases of each method and distributed them amongst the evaluators. We considered two evaluation scenarios:

- **Gold-standard translations**: the English translation as found in Europarl was taken as sense tag, using automatic word-alignments to identify the English phrase that constrains the sense of the Spanish phrase.
- **SMT translations**: a phrase-based SMT system built using the Moses toolkit (Koehn et al., 2007) and the whole Spanish-English dataset (except the sentences in the test set) was used to translated the Spanish sentences. Instead of gold-standard translations as a *quasi-perfect* sense annotation (*quasi* because the word-alignment is still automatic and thus prone to errors), the phrase-based SMT system plays the role of a sense annotation module predicting the “sense” tags.

Note that models may not be able to produce a paraphrase for certain input phrases, e.g. when the input phrase is not found in the bilingual corpora. Therefore, we assess **precision** (P) and **recall** (R) as the number of paraphrases in context that are judged correct out of the number of cases for which a candidate paraphrase was proposed, and out of the total number of test samples, respectively. To summarise the results, accuracy is expressed in terms of $F_1$. 

---

206
Table 5 shows that our model consistently improves the precision and recall of the predictions. Recall improves by 14% w.r.t. CCB-wsd because multi is able to find more paraphrases, which we believe are only reachable through the additional pivots. For example, in our data the paraphrase *forma → medio* in the sense of way (see Table 5) is only found through the Dutch pivot *middel*, which is not accessible to CCB-wsd. Recall is much lower in CCB because of the model’s strong bias towards the most frequent senses: other senses receive very little of the probability mass and thus rarely feature amongst the top ranked paraphrases. Our multilingual disambiguation model also shows a 25% increase in precision, which must be due to the stronger contribution of the sense discrimination over the LM component in getting the senses of the paraphrases right.

To show the impact of the LM re-ranking component, in Table 3 we remove this component from all models, such that the ranking of paraphrases is done purely based on the paraphrase probabilities. All models are harmed by the absence of the LM component, but to different extents and for different reasons. CCB typically ranks at the top paraphrases that convey the most frequent sense and the LM is the only component with information about the input context. CCB-wsd is impacted the most: typically invalid paraphrases are produced from unrelated senses of the foreign phrase used as sense tag, they do not represent any valid sense of the input but still get ranked at the top. For

<table>
<thead>
<tr>
<th>Method</th>
<th>Top M</th>
<th>G</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCB</td>
<td>33</td>
<td>23</td>
<td>22</td>
</tr>
<tr>
<td>CCB-wsd</td>
<td>19</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>multi</td>
<td>64</td>
<td>43</td>
<td>37</td>
</tr>
</tbody>
</table>

Table 3: Performance (F$_1$) in correctly retrieving the best paraphrase in context using gold-standard translations without the 5-gram LM component.
this model, the LM component is crucial to prune such unrelated paraphrases. Back to Table 2, the superior performance of CCB-wsd over CCB in the presence of the LM component suggest that CCB-wsd assigns less negligible probabilities to the paraphrases that convey a valid sense of the input. Finally, multi’s performance is only truly harmed in terms of grammaticality: sense discrimination is the main responsible for selecting the appropriate sense, while the LM component is responsible for selecting the candidate that makes the sentence more fluent. Further investigation showed that in some cases the most meaning preserving option was down-weighted due to low fluency, and a less adequate option was chosen, explaining the slight improvement under the meaning preservation criterion when no LM re-ranking is performed.

Table 5 lists the 5-best paraphrases of the Spanish phrase *forma* in its different senses. The paraphrases are ranked by CCB-wsd and multi out of context, that is, without LM re-ranking. Note that, because the sense tags are themselves ambiguous in English, most of the top-ranked paraphrases from CCB-wsd are inadequate, that is, they do not convey any valid sense of *forma*.

It is also interesting to observe the impact of the different pivot languages on the performance of our proposed approach. Figure 4 shows CCB-wsd and multi, both using LM re-ranking. For multi we can see the impact of the pivot languages individually and in groups.\(^6\) Except for Finnish when used on its own as pivot all other setups are superior to CCB-wsd. We can also see that putting together languages of different families has a strong positive impact, probably due to the fact that ambiguities are realised differently in languages that are farther from each other, emphasising the potential of sense discrimination by pivoting through multiple languages.

### 5.2 Paraphrasing from machine translations

Finally, we assessed the paraphrasing models using machine translations instead of gold-standard translations from Europarl. In order to have an idea of the quality of the SMT model beforehand, we evaluated the machine translations in terms of BLEU scores (Papineni et al., 2002) using a single reference from Europarl. Our phrase-based SMT model achieved 48.9 BLEU, which can be considered a high score for Europarl data (in-domain evaluation). Table 4 is analogous to Table 2, but with paraphrases extracted from machine translated sentences as opposed to human translations.

We observe that multi still outperforms CCB-wsd by a large margin. On the one hand there is a drop in precision of about 9% for correctness with multi. On the other hand there is an improvement in recall: multi improves from 3% (top-1 guess) to 12% (top-3 guesses). Manual inspection revealed that the tags predicted by the SMT model are more frequent translation options, reducing the chance of finding rare target phrases as sense annotation, for which significant statistics cannot be computed. However, with respect to correctness, the differences between this setting and that with gold-standard translations are not statistically significant.

\(^6\)For a larger version of this figure, we refer the reader to: http://pers-www.wlv.ac.uk/~in1676/publications/2013/conll2013pivots.pdf

<table>
<thead>
<tr>
<th>Method</th>
<th>Top</th>
<th>M</th>
<th>G</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P1</td>
<td>P2</td>
<td>F1</td>
</tr>
<tr>
<td>CCB-wsd</td>
<td>1</td>
<td>71</td>
<td>39</td>
<td>34</td>
</tr>
<tr>
<td>multi</td>
<td>1</td>
<td>69</td>
<td>55</td>
<td>50</td>
</tr>
<tr>
<td>CCB-wsd</td>
<td>2</td>
<td>79</td>
<td>46</td>
<td>40</td>
</tr>
<tr>
<td>multi</td>
<td>2</td>
<td>82</td>
<td>69</td>
<td>63</td>
</tr>
<tr>
<td>CCB-wsd</td>
<td>3</td>
<td>85</td>
<td>50</td>
<td>44</td>
</tr>
<tr>
<td>multi</td>
<td>3</td>
<td>85</td>
<td>74</td>
<td>69</td>
</tr>
</tbody>
</table>

Table 4: Performance in retrieving paraphrases in context using machine translations for sense tags and a 5-gram LM component.

![Figure 4: Impact of pivot languages on correctness.](http://pers-www.wlv.ac.uk/~in1676/publications/2013/conll2013pivots.pdf)
Table 5: Top paraphrases of forma annotated by the English words way, form, means and kind. Starred phrases denote inadequate candidates.

5.3 Potential applications

In what follows we discuss two applications which we believe could directly benefit from the paraphrase extraction approach proposed in this paper.

**MT evaluation** metrics such as METEOR (Denkowski and Lavie, 2010) and TESLA (Liu et al., 2010) already use paraphrases of n-grams in the machine translated sentence in an attempt to match more of the reference translation’s n-grams. TESLA, in particular, uses paraphrases constrained by a single pivot language as sense tag as originally proposed in (Bannard and Callison-Burch, 2005). Metrics like METEOR, which use paraphrases simply as a repository with extra options for the n-gram matching, could be extended to use the word-alignment between the source sentence and the translation to constrain the translated phrases while paraphrasing them with multilingual constraints. In this case the model would attempt to paraphrase the MT, which is not necessarily fluent, therefore potentially compromising its LM component. However, even after completely disregarding the LM re-ranking (see context-insensitive model multi in Table 3), we may be able to improve n-gram matching by paraphrasing.

**Handling out-of-vocabulary words in SMT** by expanding the bilingual phrase-tables (Callison-Burch et al., 2006) is a direct application of the sense constrained paraphrases. We can add translations for a given unknown phrase $f_1$, whose paraphrase $f_2$ is present in the phrase-table and is aligned to the target phrase $e$ (sense tag). We basically expand the phrase table to translate the out-of-vocabulary word $f_1$ using the knowledge associated to its paraphrase $f_2$ in the context of the known translation $e$: $(f_2, e) \rightarrow (f_1, e)$. The multilingual constraints offer more control over ambiguities, therefore potentially leading to more accurate phrase pairs added to the phrase-table.

6 Conclusions and future work

We have proposed a new formulation of the problem of generating “sense” tagged paraphrases for words and short phrases using bilingual corpora and multiple pivot languages to jointly disambiguate the input phrase and the sense tag. Sense tags are phrases in a foreign language of interest, for instance the target language of a phrase-based SMT system.

The approach was evaluated against the state of the art method for paraphrase extraction. Significant improvements were found in particular with respect to two aspects: i) the proposed model has higher recall, since it has access to paraphrases that would receive a negligible probability mass and therefore would never be selected in previous formulations, and ii) the proposed model has higher precision, since it is able to filter out or rank down paraphrases with incorrect senses.

In future work we plan to further evaluate the approach in the two scenarios discussed in Section 5.3: i) to expand the phrase table of SMT systems to address issues such as out-of-vocabulary words and phrases; and ii) to evaluate and optimise parameters of SMT systems using metrics that can accommodate sense disambiguated paraphrases. We also plan to integrate syntactic constraints, as proposed in (Callison-Burch, 2008), to our model to investigate the complementarities between these two ways of constraining paraphrasing.
References


