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.

$$p(e_2|e_1) = \sum_{f \in F} p(f|e_1)p(e_2|f) \qquad (1)$$

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.

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

Paired with	en	de	nl	da	sv	fi	fr	it	pt	el
es	1.78	1.56	1.62	1.61	1.51	1.58	1.65	1.51	1.60	5.68
en	-	1.73	1.82	1.78	1.67	1.74	1.82	1.73	1.78	1.06

Table 1: Size of the bilingual parallel corpora in millions of sentence pairs

the pivot phrases that will lead to adequate paraphrases. In our approach a sense tag consists in a phrase in a foreign language, that is, a valid translation of the input phrase in a language of interest, here referred to as *target language*. Treating the target language vocabulary as a sense repository is a good strategy from both theoretical and practical perspectives: it has been shown that monolingual sense distinctions can be effectively captured by translations into second languages, especially as language family distance increases (Resnik and Yarowsky, 1999; Specia et al., 2006). These translations can be easily captured given the availability of bilingual parallel data and robust automatic word-alignment techniques (Carpuat and Wu, 2007; Chan et al., 2007).

Figure 1 illustrates the proposed model to produce sense tagged paraphrases. We start the process at e_1 and we need to make sure that the pivot phrases $f \in F$ align back to the input language, producing the paraphrase e_2, and to the target language, producing the sense tag q. To avoid computing the distribution $p(e_2, q|f)$ – which would require a trilingual parallel corpus – we assume that e_2 and q are conditionally independent on f:

$$p(e_2, q|f) \overset{e_2 \perp\!\!\!\perp q|f}{=} p(e_2|f)p(q|f)$$

In other words, we assume that pivot phrases generate paraphrases and sense tags independently. Equation 2 shows how paraphrase probabilities are computed by marginalising out the pivot phrases under this assumption.

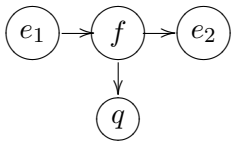

Figure 1: Pivot phrases must align back to target phrases (sense annotation).

$$p(e_2|e_1, q) = \frac{1}{z} \sum_{f \in F} p(e_2|f)p(q|f)p(f|e_1) \quad (2)$$

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 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.[2] 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*

[2] A paraphrase is scored in the context of a given pivot phrase f: $p(e_2|e_1, f) = p(e_2|f)p(f|e_1)$.

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.

Unambiguous	**Ambiguous**
concreto, política, fondos, regular, haber, amor proprio, sangre fría, dar a luz, dar con, tomar el pelo	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

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*.

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

[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: $p(w_{-4} \dots w_{-1} e_2 w_{+1} \dots w_{+4})$ 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.

Method	Top	M	G	Correct		
		F_1	F_1	P	R	F_1
CCB	1	32	28	25	25	25
CCB-wsd	1	61	38	34	28	30
multi	1	62	55	59	42	49
CCB	2	41	37	33	33	33
CCB-wsd	2	68	44	40	33	36
multi	2	71	64	66	47	55
CCB	3	46	42	37	37	37
CCB-wsd	3	71	47	45	36	40
multi	3	74	67	71	50	59

Table 2: Performance in retrieving paraphrases in context using gold-standard translations for sense tags and a 5-gram LM component.

In the following section we present results on whether the best candidate (Top-1) or at least one of the two (Top-2) or three (Top-3) best candidates satisfies the criterion under consideration (meaning/grammar).

5 Results

The evaluation was performed by seven native speakers of Spanish who judged a total of $5,110$ sentences containing one paraphrased input phrase each. We used 40 overlapping judgements across annotators to measure inter-annotator agreement. The average inter-annotator agreement in terms of Cohen's Kappa (Cohen, 1960) is 0.54 ± 0.15 for *meaning* judgements, 0.63 ± 0.16 for *grammar* judgements and 0.62 ± 0.20 for *correctness* judgements. These figures are similar or superior to those reported in (Bannard and Callison-Burch, 2005; Callison-Burch, 2008), which we consider particularly encouraging as in our case we have seven instead of only two annotators. In Tables 2, 3 and 4 we report the performance of the three models in terms of precision, recall and F_1, with p-values < 0.01 based on the *t-test* for statistical significance.

5.1 Paraphrasing from human translations

We first assess the paraphrasing models using gold-standard translations, that is, the English phrases were selected via automatic word-alignments between the input text and its corresponding human translation from Europarl. Table 2 shows the performance in terms of F_1 for our three criteria: meaning preservation, grammaticality, and correctness. Our method (**multi**) outperforms the best performing alternative (**CCB-wsd**) by a large margin. It is 19% more effective at selecting the 1-best candidate in terms of correctness. A consistent gain is also observed when more guesses are allowed (top 2–3), showing that our model is better at ranking the top candidates as well. **CCB-wsd** and **multi** are close in terms of paraphrases that are meaning preserving, however their differences become more obvious as more guesses are allowed, again showing that **multi** is better at ranking more adequate paraphrases first. Moreover, **multi** consistently chooses more grammatical paraphrases.

Method	M	G	Correct
CCB	33	23	22
CCB-wsd	19	9	8
multi	64	43	37

Table 3: Performance (F_1) in correctly retrieving the best paraphrase in context using gold-standard translations **without** the 5-gram LM component.

Table 2 also shows that our model consistently improves both 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* $\rightarrow$ *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

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.

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

Method	Top	M	G	Correct		
		F_1	F_1	P	R	F_1
CCB-wsd	1	71	39	34	32	33
multi	1	69	55	50	45	48
CCB-wsd	2	79	46	40	38	39
multi	2	82	69	63	57	60
CCB-wsd	3	83	50	44	41	42
multi	3	85	74	69	62	65

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

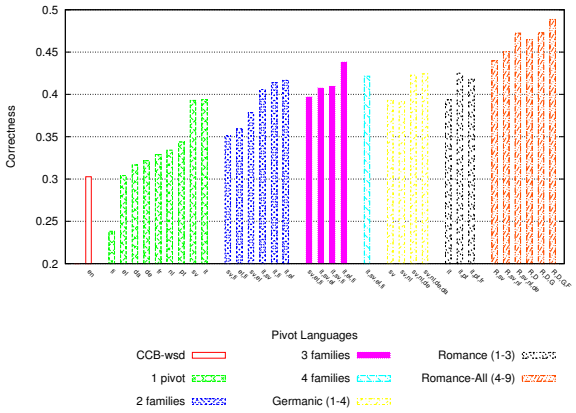

Figure 4: Impact of pivot languages on correctness. Language codes follow the convention presented in Section 4.1. Additionally R stands for Romance languages, D for Germanic languages, G for Greek and F for Finnish.

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.

multi: English as sense annotation and nine other pivot languages							
forma → *way*		*forma* → *form*		*forma* → *means*		*forma* → *kind*	
forma	0.34	forma	0.64	medio	0.64	tipo	0.37
manera	0.24	tipo	0.10	través	0.23	forma	0.23
modo	0.23	forma de	0.05	instrumento	0.13	especie	0.06
forma de	0.02	formas	0.03			especie de	0.03
medio	0.02	modo	0.02			tipo de	0.03
CCB-wsd: English as sense annotation and sole evidence for pivoting							
forma → *way*		*forma* → *form*		*forma* → *means*		*forma* → *kind*	
*way	0.08	*formulario	0.18	*significa contar	0.07	*amables	0.16
*vía por	0.08	de sus formas	0.10	medios que tiene	0.07	*kind	0.12
*camino que hay	0.07	*formulario de	0.07	*significa	0.06	especie	0.09
*camino que hay que	0.07	modalidad	0.06	*significa contar con	0.06	*amable	0.08
*vía por la	0.07	aspecto formal	0.05	*anterior significa	0.06	tipo	0.07

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

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604, Ann Arbor, Michigan.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 17–24, New York, New York.

Chris Callison-Burch. 2007. *Paraphrasing and Translation*. Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 196–205, Honolulu, Hawaii.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '07, pages 61–72, Prague, Czech Republic.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40, Prague, Czech Republic.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, April.

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic.

Michael Denkowski and Alon Lavie. 2010. METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support For Five Target Languages. In *Proceedings of the ACL 2010 Joint Workshop on Statistical Machine Translation and Metrics MATR*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Edmonton, Canada.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics: Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *The Proceedings of the Tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2010. Tesla: Translation evaluation of sentences with linear-programming-based analysis. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 354–359, Uppsala, Sweden.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie J. Dorr. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 120–127, Prague, Czech Republic.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 381–390, Suntec, Singapore.

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

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19–51, March.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Philip Resnik and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Nat. Lang. Eng.*, 5(2):113–133.

Stefan Riezler, Er Vasserman, Ioannis Tsochantaridis, Vibhu Mittal, and Yi Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 464–471, Prague, Czech Republic.

Lucia Specia, Mark Stevenson, Maria das Graças Volpe Nunes, and Gabriela C.B. Ribeiro. 2006. Multilingual versus monolingual WSD. In *Proceedings of the EACL Workshop "Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together"*, pages 33–40, Trento, Italy.

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