UEdin: Translating L1 Phrases in L2 Context using Context-Sensitive SMT

Eva Hasler
ILCC, School of Informatics
University of Edinburgh
e.hasler@ed.ac.uk

Abstract

We describe our systems for the SemEval 2014 Task 5: *L2 writing assistant* where a system has to find appropriate translations of L1 segments in a given L2 context. We participated in three out of four possible language pairs (English-Spanish, French-English and Dutch-English) and achieved the best performance for all our submitted systems according to word-based accuracy. Our models are based on phrase-based machine translation systems and combine topical context information and language model scoring.

1 Introduction

In the past years, the fields of statistical machine translation (SMT) and word sense disambiguation (WSD) have developed largely in parallel, with each field organising their own shared tasks aimed at improving translation quality (Bojar et al., 2013) and predicting word senses, e.g. Agirre et al. (2010). Because sense disambiguation is a central problem in machine translation, there has been work on integrating WSD classifiers into MT systems (Carpuat and Wu, 2007a; Carpuat and Wu, 2007b; Chan et al., 2007). However, one problem with the direct integration of WSD techniques into MT has been the mismatch between word predictions of WSD systems and the phrase segmentations of MT system. This problem was adressed in Carpuat and Wu (2007b) by extending word sense disambiguation to phrase sense disambiguation. The relation between word sense distinctions and translation has also been explored in past SemEval tasks on *cross-lingual word sense disambiguation*, where senses are not defined in terms of WordNet senses as in previous tasks, but instead as translations to another language (Lefever and Hoste, 2010; Lefever and Hoste, 2013).

This year's *L2 writing assistant* task is similar to the cross-lingual word sense disambiguation task but differs in the context provided for disambiguation and the length of the fragments (source phrases instead of words). While in other translation and disambiguation tasks the source language context is given, the *L2 writing assistant* task assumes a given target language context that constrains the possible translations of L1 fragments. This is interesting from a machine translation point-of-view because it allows for a direct comparison with systems that exploit the target context using a language model. As language models have become more and more powerful over the years, mostly thanks to increased computing power, new machine translation techniques are also judged by their ability to improve performance over a baseline system with a strong language model. Another difference to previous SemEval tasks is the focus on both lexical and grammatical forms, while previous tasks have mostly focused on lexical selection.

2 Translation Model for L1 Fragments in L2 Context

Our model for translating L1 fragments in L2 context is a phrase-based machine translation system with an additional context similarity feature. We aim to resolve lexical ambiguities by taking the entire topical L2 context of an L1 fragment into account rather than only relying on the phrasal L1 context. We do not explicitly model the grammaticality of target word forms but rather use a standard 5-gram language model to score target word sequences. We describe the context similarity feature in the following section.

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 688–693,
Dublin, Ireland, August 23-24, 2014.

2.1 Context Similarity Feature

The context similarity feature is derived from the phrase pair topic model (PPT) described in Hasler et al. (2014). At training time, this model learns topic distributions for all phrase pairs in the phrase table in an unsupervised fashion, using a variant of Latent Dirichlet Allocation (LDA). The underlying assumption is that all phrase pairs share a set of global topics of predefined size, thus each phrase pair is assigned a distribution over the same set of global topics. This is in contrast to Word Sense Induction (WSI) methods which typically learn a set of topics or senses for each word type, for example in Lau et al. (2010).

The input to the model are distributional profiles of words occurring in the context of each phrase pair, thus, the model learns lower-dimensional representations of the likely context words of a phrase pair. While in a normal machine translation setup the source sentence context is given, it is straightforward to replace source language words with target language words as given in the L2 context for each test example. At test time, the topic model is applied to the given L2 context to infer a topic distribution of the test context. The topic distribution of an applicable phrase pair is compared to the topic distribution of a given test context (defined as all L2 words in the same sentence as the L1 fragment, excluding stop words) using cosine similarity.

To adapt the translation system to the context of each test sentence, the phrase table is filtered per test sentence and each applicable phrase pair receives one additional feature that expresses its topical similarity with the test context. While the baseline system (the system without similarity feature) translates the entire test set with the same translation model, the context-sensitive system loads an adapted phrase table for each test sentence. While the phrase pair topic model can also deal with document-level context, here we consider only sentence-level context as no wider context was available. We evaluate three variations of the context similarity feature on top of a standard phrase-based MT system:

- **50-topics** The cosine similarity according to the PPT model trained with 50 topics (submitted as run1)
- **mixture:geoAvg** The geometric average of the cosine similarities according to PPT models trained with 20, 50 and 100 topics (submitted as run2)
- **mixture:max** For each source phrase, the cosine similarity according to the PPT model that yields the lowest entropy (out of the models with 20, 50 and 100 topics) when converting the similarities into probabilities (submitted as run3)

2.2 Language Model Scoring of L2 Context

On top of using the words in the L2 context for computing the similarity feature described above, we introduce a simple method for using a language model to score the target sequence that includes the translated L1 segments and the words to the left and right of the translated segments. In order to use the language model scoring implemented in the Moses decoder, we present the decoder with an input sentence that contains the L1 fragment as well as the L2 context with XML markup. While the L1 fragments are translated without special treatment, the L2 tokens are passed through untranslated by specifying the identity translation as markup. The XML markup also includes reordering walls to prevent the decoder from reordering the L2 context. An example input sentence with markup (French-English) is shown below:

```
<wall/>les manifesteurs<wall/>
<np translation=”want”>want</np><wall/>
<np translation=”to”>to</np><wall/>
<np translation=”replace”>replace</np><wall/>
<np translation=”the”>the</np><wall/>
<np translation=”government”>government</np><wall/>
<np translation=”.”>.</np><wall/>
```

3 Experimental Setup

Although the task is defined as building a *translation assistance system* rather than a full machine translation system, we use a standard machine translation setup to translate L1 phrases. We used the Moses toolkit (Koehn et al., 2007) to build phrase-based translation systems for the language pairs English-Spanish, French-English and Dutch-English[1]. For preprocessing, we applied punctuation normalisation, truecasing and tokenisation using the scripts provided with the Moses toolkit. The model contains the following standard features: direct and inverse phrase translation probabilities, lexical weights, word and phrase penalty, lexicalised reordering and distortion features and a 5-gram language model with modified Kneser-Ney smoothing. In addition, we add the context similarity feature described in Section 2.1.

[1] We left out the English-German language pair for time reasons.

Training data	English-Spanish	French-English	Dutch-English
Europarl	1.92M	1.96M	1.95M
News Commentary	192K	181K	n/a
TED	157K	159K	145K
News	2.1G	2.1G	2.1G
Commoncrawl	50M	82M	-

Table 1: Overview of parallel and monolingual training data (top/bottom, in number of sentences/words).

3.1 Training Data

Most of the training data was taken from the WMT13 shared task (Bojar et al., 2013), except where specified otherwise. For the English-Spanish and French-English systems, we used parallel training data from the Europarl and News Commentary corpora, as well as the TED corpus (Cettolo et al., 2012). For Dutch-English, we used parallel data from the Europarl and TED corpus. The language models were trained on the target sides of the parallel data and additional news data from the years 2007-2012. For English-Spanish and French-English, we used additional language model data from the Commoncrawl corpus[2]. Separate language models were trained for each corpus and interpolated on a development set. An overview of the training data is shown in Table 1.

3.2 Tuning Model Parameters

The parameters of the baseline MT excluding the similarity feature were tuned with kbest-mira (Cherry and Foster, 2012) on mixed development sets containing the trial data (500 sentence pairs with XML markup) distributed for the task as well as development data from the news and TED corpora for the English-Spanish and French-English systems and development data from the TED corpus for the Dutch-English system. Because the domain(s) of the test examples was not known beforehand, we aimed for learning model weights that would generalise across domains by using rather diverse tuning sets. In total, the development sets consisted of 3435, 3705 and 3516 sentence pairs, respectively. We did not tune the weight of the similarity feature automatically, but set it to an empirically determined value instead.

3.3 Simulating Ambiguous Development Data

When developing our systems using the trial data supplied by the task organisers, we noticed that the context similarity feature did not add much to the overall performance, which we attributed to the small number of ambiguous examples in the trial data. Therefore, we extracted a set of 1076 development instances containing 14 ambiguous French words and their English translations from a mixed corpus containing data from the News Commentary, TED and Commoncrawl corpora as used in Hasler et al. (2014). Examples of ambiguous source words and their translations in that de-

Source words	Translations
chaîne	chain, string, channel, station
matière	matter, material, subject
flux	stream, flow, feed
démon	demon, daemon, devil
régime	regime, diet, rule

Table 2: Examples of ambiguous source words and their different translations in the simulated development set.

System	French-English
Baseline	0.314
+ LM context	0.726
20-topics	0.603
+ LM context	0.845
50-topics	**0.674**
+ LM context	**0.886**
100-topics	0.628
+ LM context	0.872
mixture:arithmAvg	0.650
+ LM context	0.869
mixture:geoAvg	**0.670**
+ LM context	**0.883**
mixture:max	**0.690**
+ LM context	**0.889**

Table 3: Word accuracy (best) on the simulated development set for the smaller baseline system and the systems with added context similarity feature, with and without language model context.

[2] For the Dutch-English system, the Commoncrawl data did not seem to improve performance.

velopment set are shown in Table 2.

Input: There are many ways of cooking <f>des œufs</f> for breakfast.
Reference: There are many ways of cooking <f>eggs</f> for breakfast.

Input: I loved animals when I was <f>un enfant</f>.
Reference: I loved animals when I was <f>a kid<alt>a child</alt></f>.

Figure 1: Examples of official test instances.

Translating the L1 fragments in the simulated development set using a smaller baseline system trained on this mixed data set yields the results at the top of Table 3. Note that even though the instances were extracted from the training set, this does not affect the translation model since the L1 fragments contain only the ambiguous source words and no further source context that could be memorised.

The bottom part of Table 3 shows the performance of the three context similarity features described in Section 2.1 plus some further variants (the models with 20 and 100 topics as well as the arithmetic average of the cosine similarities of models trained with 20, 50 and 100 topics). First, we observe that each of the features clearly outperforms the baseline system without language model context. Second, each context similarity feature together with the language model context still outperforms the *Baseline + LM context*. Even though the gain of the context similarity features is smaller when the target context is scored with a language model, the topical context still provides additional information that improves lexical choice. We trained versions of the three best models from Table 3 (in bold) for our submissions on the official test sets.

4 Results and Discussion

In this section we report the experimental results of our systems on the official test sets. The results without scoring the L2 context with a language model are shown in Table 4 and including language model scoring of L2 context in Table 5. We limit the reported scores to word accuracy and do not report recall because our systems produce output for every L1 phrase.

In Table 4, we compare the performance of the baseline MT system to systems including one of three variants of the similarity feature as described in Section 2.1, according to the 1-best translation (best) as well as the 5-best translations (out-of-five) in a distinct n-best list. For five out of the six tasks, at least one of the systems including the similiary feature yields better performance than the baseline system. Only for French-English *best*, the baseline system yields the best word accuracy. Among the three variants, 50-topics and mixture:geoAvg perform slightly better than mixture:max in most cases.

Table 5 shows the results of our submitted runs (run1-run3) as well as the baseline system, all with language model scoring of L2 context via XML markup. The first thing to note in comparison to Table 4 is that providing the L2 context for language model scoring yields quite substantial improvements (0.165, 0.101 and 0.073, respectively). Again, in five out of six cases at least one of the systems with context similarity feature performs better than the baseline system. Only for Spanish-English *best*, the baseline system yields higher word accuracy than the three submitted runs. As before, 50-topics and mixture:geoAvg perform slightly better than mixture:max, with a preference for 50-topics. For comparison, we also show the word accuracies of the 2nd-ranked system for both tasks and each language pair. We note that the distance to the respective runner-up system is largest for French-English and on average larger for the *out-of-five* task than for the *best* task.

As a general observation, we can state that although the similarity feature improves performance in most cases, the improvements are small compared to the improvements achieved by scoring the L2 language model contexts. We suspect two reasons for this effect: first, we do not explicitly model grammaticality of word forms. Therefore, our system relies on the language model to choose the best word form for those test examples that do not contain any lexical ambiguity. Second, we have noticed that for some of the test examples, the correct translations do not depend particularly on words in the L2 context, as shown in Figure 1 where the most common translations of the source phrases without context would match the reference translations. These are cases where we do not expect much of an improvement in translation by taking the L2 context into account.

Since in Section 3.3 we have provided evidence that topical similarity features can improve lexical choice over simply using a target language model, we believe that the lower performance of the similarity features on the official test set is caused by

System	English-Spanish		French-English		Dutch-English	
	best	oof	best	oof	best	oof
Baseline	0.674	0.854	**0.722**	0.884	0.613	0.750
50-topics	**0.682**	0.860	0.719	**0.896**	0.616	**0.759**
mixture:geoAvg	0.677	**0.863**	0.715	**0.896**	**0.619**	0.756
mixture:max	0.679	0.860	0.712	0.887	0.618	0.753

Table 4: Word accuracy (best and out-of-five) of the baseline system and the systems with added context similarity feature. All systems were run without scoring the language model context.

System	English-Spanish		French-English		Dutch-English	
	best	oof	best	oof	best	oof
Baseline + LM context	**0.839**	0.944	0.823	0.934	0.686	0.809
run1: 50-topics + LM context	0.827	0.946	**0.824**	0.938	**0.692**	**0.811**
run2: mixture:geoAvg + LM context	0.827	0.944	0.821	**0.939**	0.688	0.808
run3: mixture:max + LM context	0.820	**0.949**	0.816	0.937	0.688	0.808
2nd-ranked systems	0.809[1]	0.887[2]	0.694[2]	0.839[2]	0.679[3]	0.753[3]

Table 5: Word accuracy (best and out-of-five) of all submitted systems (runs 1-3) as well as the baseline system without the context similarity feature. All systems were run with the language model context provided via XML input. Systems on 2nd rank: [1]UNAL-run2, [2]CNRC-run1, [3]IUCL-run1

different levels of ambiguity in the simulated development set and the official test set. For the simulated development set, we explicitly selected ambiguous source words in contexts which trigger multiple different translations, while the official test set also contains examples where the focus is on correct verb forms. It further contains examples where the baseline system without context information could easily provide the correct translation, as shown above. Thus, the performance of our topical context models should ideally be evaluated on test sets that contain a sufficient number of ambiguous source phrases in order to measure its ability to improve lexical selection.

Finally, in Figure 2 we show some examples where the 50-topics system (with LM context) produced semantically better translations than the baseline system and where words in the L2 context would have helped in promoting them over the choice of the baseline system.

Input:	Why has Air France authorised <f>les appareils électroniques</f> at take-off?
Baseline:	.. <f>the electronics</f> ..
50-topics:	.. <f>electronic devices</f> ..
Reference:	.. <f>electronic devices</f> ..
Input:	This project represents one of the rare advances in strenghtening <f>les liens</f> between Brazil and the European Union.
Baseline:	.. <f>the links</f> ..
50-topics:	.. <f>the ties</f> ..
Reference:	.. <f>the ties<alt>relations</alt><alt> the bonds</alt></f> ..

Figure 2: Examples of improved translation output with the context similarity feature.

5 Conclusion

We have described our systems for the SemEval 2014 Task 5: *L2 writing assistant* which achieved the best performance for all submitted language pairs and both the *best* and *out-of-five* tasks. All systems are based on phrase-based machine translation systems with an added context similarity feature derived from a topic model that learns topic distributions for phrase pairs. We show that the additional similarity feature improves performance over our baseline models and that further gains can be achieved by passing the L2 context through the decoder via XML markup, thereby producing language model scores of the sequences of L2 context words and translated L1 fragments. We also provide evidence that the relative performance of the context similarity features depends on the level of ambiguity in the L1 fragments.

Acknowledgements

This work was supported by funding from the Scottish Informatics and Computer Science Alliance and funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 287658 (EU BRIDGE).

References

Eneko Agirre, Oier Lopez De Lacalle, Christiane Fellbaum, Maurizio Tesconi, Monica Monachini, Piek Vossen, and Roxanne Segers. 2010. SemEval-2010 Task 17: All-words Word Sense Disambiguation on a Specific Domain. In *Proceedings of the 5th International Workshop on Semantic Evaluation*.

Ondrej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 workshop on statistical machine translation. In *Proceedings of WMT 2013*.

Marine Carpuat and Dekai Wu. 2007a. How Phrase Sense Disambiguation outperforms Word Sense Disambiguation for Statistical Machine Translation. In *International Conference on Theoretical and Methodological Issues in MT*.

Marine Carpuat and Dekai Wu. 2007b. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proceedings of EMNLP*, pages 61–72.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of EAMT*.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of ACL*.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of NAACL*.

Eva Hasler, Barry Haddow, and Philipp Koehn. 2014. Dynamic Topic Adaptation for SMT using Distributional Profiles. In *Proceedings of the 9th Workshop on Statistical Machine Translation*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for SMT. In *Proceedings of ACL: Demo and poster sessions*.

Jey Han Lau, Paul Cook, Diana Mccarthy, David Newman, and Timothy Baldwin. 2010. Word Sense Induction for Novel Sense Detection. In *Proceedings of EACL*.

Els Lefever and Veronique Hoste. 2010. SemEval-2010 Task 3: Cross-lingual Word Sense Disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*.

Els Lefever and Veronique Hoste. 2013. SemEval-2013 Task 10: Cross-lingual Word Sense Disambiguation. In *Proceedings of the 7th International Workshop on Semantic Evaluation, in Conjunction with the Second Joint Conference on Lexical and Computational Semantics*.