NRC: A Machine Translation Approach to Cross-Lingual Word Sense Disambiguation (SemEval-2013 Task 10)

Marine Carpuat
National Research Council
Ottawa, Canada
Marine.Carpuat@nrc.gc.ca

Abstract

This paper describes the NRC submission to the Spanish Cross-Lingual Word Sense Disambiguation task at SemEval-2013. Since this word sense disambiguation task uses Spanish translations of English words as gold annotation, it can be cast as a machine translation problem. We therefore submitted the output of a standard phrase-based system as a baseline, and investigated ways to improve its sense disambiguation performance. Using only local context information and no linguistic analysis beyond lemmatization, our machine translation system surprisingly yields top precision score based on the best predictions. However, its top 5 predictions are weaker than those from other systems.

1 Introduction

This paper describes the systems submitted by the National Research Council Canada (NRC) for the Cross-Lingual Word Sense Disambiguation task at SemEval 2013 (Lefever and Hoste, 2013). As in the previous edition (Lefever and Hoste, 2010), this word sense disambiguation task asks systems to disambiguate English words by providing translations in other languages. It is therefore closely related to machine translation. Our work aims to explore this connection between machine translation and cross-lingual word sense disambiguation, by providing a machine translation baseline and investigating ways to improve the sense disambiguation performance of a standard machine translation system.

Machine Translation (MT) has often been used indirectly for SemEval Word Sense Disambiguation (WSD) tasks: as a tool to automatically create training data (Guo and Diab, 2010, for instance) ; as a source of parallel data that can be used to train WSD systems (Ng and Chan, 2007; van Gompel, 2010; Lefever et al., 2011); or as an application which can use the predictions of WSD systems developed for SemEval tasks (Carpuat and Wu, 2005; Chan et al., 2007; Carpuat and Wu, 2007). This SemEval shared task gives us the opportunity to compare the performance of machine translation systems with other submissions which use very different approaches. Our goal is to provide machine translation output which is representative of state-of-the-art approaches, and provide a basis for comparing its strength and weaknesses with that of other systems submitted to this task. We submitted two systems to the Spanish Cross-Lingual WSD (CLWSD) task:

1. BASIC, a baseline machine translation system trained on the parallel corpus used to define the sense inventory;
2. ADAPT, a machine translation system that has been adapted to perform better on this task.

After describing these systems in Sections 2 and 3, we give an overview of the results in Section 4.

2 BASIC: A Baseline Phrase-Based Machine Translation System

We use a phrase-based SMT (PBSMT) architecture, and set-up our system to perform English-to-Spanish translation. We use a standard SMT system set-up, as for any translation task. The fact that this PBSMT system is intended to be used for CLWSD only influences data selection and pre-processing.

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 188–192, Atlanta, Georgia, June 14-15, 2013. ©2013 Association for Computational Linguistics

2.1 Model and Implementation

In order to translate an English sentence e into Spanish, PBSMT first segments the English sentence into phrases, which are simply sequences of consecutive words. Each phrase is translated into Spanish according to the translations available in a translation lexicon called phrase-table. Spanish phrases can be reordered to account for structural divergence between the two languages. This simple process can be used to generate Spanish sentences, which are scored according to translation, reordering and language models learned from parallel corpora. The score of a Spanish translation given an English input sentence e segmented into J phrases is defined as follows: $score(s,e) = \sum_i \sum_j \lambda_i log(\phi_i(s_j, e_j)) + \lambda_{LM}\phi_{LM}(s)$

Detailed feature definitions for phrase-based SMT models can be found in Koehn (2010). In our system, we use the following standard feature functions ϕ to score English-Spanish phrase pairs:

- 4 phrase-table scores, which are conditional translation probabilities and HMM lexical probabilities in both directions translation directions (Chen et al., 2011)
- 6 hierarchical lexicalized reordering scores, which represent the orientation of the current phrase with respect to the previous block that could have been translated as a single phrase (Galley and Manning, 2008)
- a word penalty, which scores the length of the output sentence
- a word-displacement distortion penalty, which penalizes long-distance reorderings.

In addition, fluency of translation is ensured by a monolingual Spanish language model ϕ_{LM}, which is a 5-gram model with Kneser-Ney smoothing.

Phrase translations are extracted based on IBM-4 alignments obtained with GIZA++ (Och and Ney, 2003). The λ weights for these features are learned using the batch lattice-MIRA algorithm (Cherry and Foster, 2012) to optimize BLEU-4 (Papineni et al., 2002) on a tuning set. We use PORTAGE, our internal PBSMT decoder for all experiments. PORTAGE uses a standard phrasal beam-search algorithm with cube pruning. The main differences between this set-up and the popular open-source Moses system (Koehn et al., 2007), are the use of hierarchical reordering (Moses only supports non-hierarchical lexicalized reordering by default) and smoothed translation probabilities (Chen et al., 2011).

As a result, disambiguation decisions for the CLWSD task are based on the following sources of information:

- **local source context**, represented by source phrases of length 1 to 7 from the translation and reordering tables
- **local target context**, represented by the 5-gram language model.

Each English sentence in the CLWSD task is translated into Spanish using our PBSMT system. We keep track of the phrasal segmentation used to produce the translation hypothesis and identify the Spanish translation of the English word of interest. When the English word is translated into a multi-word Spanish phrase, we output the Spanish word within the phrase that has the highest IBM1 translation probability given the English target word.

For the BEST evaluation, we use this process on the top PBSMT hypothesis to produce a single CLWSD translation candidate. For the Out-Of-Five evaluation, we produce up to five CLWSD translation candidates from the top 1000 PBSMT translation hypotheses.

2.2 Data and Preprocessing

Training the PBSMT system requires a two-step process with two distinct sets of parallel data.

First, the translation, reordering and language models are learned on a large parallel corpus, the **training set**. We use the sentence pairs extracted from Europarl by the organizers for the purpose of selecting translation candidates for the gold annotation. Training the SMT system on the exact same parallel corpus ensures that the system "knows" the same translations as the human annotators who built the gold standard. This corpus consists of about 900k sentence pairs.

Second, the feature weights λ in the PBSMT are learned on a smaller parallel corpus, the **tuning set**. This corpus should ideally be drawn from the test domain. Since the CLWSD task does not provide parallel data in the test domain, we construct the tuning set using corpora publicly released for the WMT2012 translation task[1]. Since sentences provided in the trial data appeared to come from a wide variety of genres and domains, we decided to build our tuning set using data from the news-commentary domain, rather then the more narrow Europarl domain used for training. We selected the top 3000 sentence pairs from the WMT 2012 development test sets, based on their distance to the CLWSD trial and test sentences as measured by cross-entropy (Moore and Lewis, 2010).

All Spanish and English corpora were processed using FreeLing (Padró and Stanilovsky, 2012). Since the CLWSD targets and gold translations are lemmatized, we lemmatize all corpora. While FreeLing can provide a much richer linguistic analysis of the input sentences, the PBSMT sytem only makes use of their lemmatized representation. Our systems therefore contrast with previous approaches to CLWSD (van Gompel, 2010; Lefever et al., 2011, for instance), which use richer sources of information such as part-of-speech tags.

3 ADAPT: Adapting the MT system to the CLWSD task

Our ADAPT system simply consists of two modifications to the BASIC PBSMT system.

First, it uses a shorter maximum English phrase length. Instead of learning a translation lexicons for phrases of length 1 to 7 as in the BASIC system, the ADAPT system only uses phrases of length 1 and 2. While this dramatically reduces the amount of source side context available for disambiguation, it also reduces the amount of noise due to incorrect word alignments. In addition, there is more evidence to estimate reliable translation probabilities for short phrase, since they tend to occur more frequently than longer phrases.

Second, the ADAPT system is trained on larger and more diverse data sets. Since MT systems are known to perform better when they can learn from larger amounts of relevant training data, we augment our training set with additional parallel corpora from the WMT-12 evaluations. We learn translation and reordering models for (1) the Europarl subset used by the CLWSD organizers (900k sentence pairs, as in the BASIC system), and (2) the news commentary corpus from WMT12 (which comprises 150k sentence pairs). For the language model, we use the Spanish side of these two corpora, as well as that of the full Europarl corpus from WMT12 (which comprises 1.9M sentences). Models learned on different data sets are combined using linear mixtures learned on the tuning set (Foster and Kuhn, 2007).

We also attempted other variations on the BASIC system which were not as successful. For instance, we tried to update the PBSMT tuning objective to be better suited to the CLWSD task. When producing translation of entire sentences, the PBSMT system is expected to produce hypotheses that are simultaneously fluent and adequate, as measured by BLEU score. In contrast, CLWSD measures the adequacy of the translation of a single word in a given sentence. We therefore attempted to tune for BLEU-1, which only uses unigram precision, and therefore focuses on adequacy rather than fluency. However, this did not improve CLWSD accuracy.

4 Results

Table 1 gives an overview of the results per target word for both systems, as measured by all official metrics (see Lefever and Hoste (2010) for a detailed description.) According to the BEST Precision scores, the ADAPT system outperforms the BASIC system for almost all target words. Using only the dominant translation picked by the human annotators as a reference (Mode), the precision for BEST scores yield more heterogeneous results. This is not surprising since the ADAPT system uses more heterogeneous training data, which might make it harder to learn a reliable estimate of a single dominant translation. When evaluating the precision out of the top 5 candidates (OOF), all systems improve, indicating that PBSMT systems can usually produce some correct alternatives to their top hypothesis.

Table 2 lets us compare the average performance of the BASIC and ADAPT systems with other participating systems. The ADAPT system surprisingly yields the top performance based on the Precision BEST evaluation setting, suggesting that, even with relatively poor models of context, a PBSMT sys-

[1] http://www.statmt.org/wmt12/translation-task.html

Precision:	Best	Best	Best Mode	Best Mode	OOF	OOF	OOF Mode	OOF Mode
Systems:	**BASIC**	**ADAPT**	**BASIC**	**ADAPT**	**BASIC**	**ADAPT**	**BASIC**	**ADAPT**
coach	22.30	60.10	13.64	59.09	38.30	66.30	31.82	63.64
education	36.07	38.01	73.08	84.62	42.36	42.80	84.62	84.62
execution	41.07	41.07	32.00	32.00	41.57	41.57	36.00	36.00
figure	23.43	29.02	33.33	37.04	31.15	36.12	37.04	44.44
job	13.45	24.26	0.00	37.23	26.52	37.57	27.27	54.55
letter	35.35	37.23	66.67	64.10	37.22	41.20	66.67	66.67
match	15.07	16.53	2.94	2.94	20.70	20.90	5.88	8.82
mission	67.98	67.98	85.29	85.29	67.98	67.98	85.29	85.29
mood	7.18	8.97	0.00	0.00	26.99	29.90	11.11	11.11
paper	31.33	44.59	29.73	40.54	50.45	55.61	45.95	51.35
post	32.26	33.72	23.81	19.05	50.67	53.28	57.14	42.86
pot	34.20	36.63	35.00	32.50	36.12	37.13	32.50	25.00
range	5.41	7.56	10.00	0.00	10.39	17.47	10.00	20.00
rest	20.91	23.44	12.00	8.00	27.44	25.89	16.00	16.00
ring	15.87	10.10	18.92	10.81	42.80	43.14	48.65	45.95
scene	15.86	23.42	43.75	62.50	38.35	37.53	81.25	81.25
side	24.63	33.14	13.04	17.39	36.84	44.03	21.74	39.13
soil	43.88	43.63	66.67	66.67	51.73	57.15	66.67	66.67
strain	24.00	26.24	35.71	35.71	38.37	36.58	42.86	35.71
test	34.45	37.51	50.00	28.57	43.61	40.86	50.00	28.57
Average	27.24	**32.16**	32.28	**36.20**	37.98	**41.65**	42.92	**45.38**

Table 1: Precision scores by target word for the BASIC and ADAPT systems

Precision:	Best	Best Mode	OOF	OOF Mode
System				
Best	32.16	37.11	61.69	57.35
ADAPT	32.16	36.20	41.65	45.38
BASIC	27.24	32.28	37.98	42.92
Baseline	23.23	27.48	53.07	64.65

Table 2: Overview of official results: comparison of the precision scores of the ADAPT and BASIC systems with the best system according to each metric and with the official baseline

tem can succeed in learning useful disambiguating information for its top candidate. Despite the problems stemming from learning good dominant translations from heterogeneous data, ADAPT ranks near the top using the Best Mode metric. The rankings in the out-of-five settings are strikingly different: the difference between BEST and OOF precisions are much smaller for BASIC and ADAPT than for all other participating systems (including the baseline.) This suggests that our PBSMT system only succeeds in learning to disambiguate one or two candidates per word, but does not do a good job of a estimating the full translation probability distribution of a word in context. As a result, there is potentially much to be gained from combining PBSMT systems with the approaches used by other systems, which typically use richer feature representation and context models. Further exploration of the role of context in PBSMT performance and a comparison with dedicated classifiers trained on the same word-aligned parallel data can be found in (Carpuat, 2013).

5 Conclusion

We have described the two systems submitted by the NRC to the Cross-Lingual Word Sense Disambiguation task at SemEval-2013. We used phrase-based machine translation systems trained on lemmatized parallel corpora. These systems are unsupervised and do not use any linguistic analysis beyond lemmatization. Disambiguation decisions are based on the local source context available in the phrasal translation lexicon and the target n-gram language model. This simple approach gives top performance when measuring the precision of the top predictions. However, the top 5 predictions are interestingly not as good as those of other systems.

(Carpuat, 2013)

References

Marine Carpuat and Dekai Wu. 2005. Word Sense Disambiguation vs. Statistical Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 387–394, Ann Arbor, Michigan.

Marine Carpuat and Dekai Wu. 2007. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 61–72, Prague, June.

Marine Carpuat. 2013. A semantic evaluation of machine translation lexical choice. In *Proceedings of the 7th Workshop on Syntax, Semantics and Structure in Statistical Translation*, Atlanta, USA, May.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word Sense Disambiguation improves Statistical Machine Translation. In *45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, Prague, June.

Boxing Chen, Roland Kuhn, George Foster, and Howard Johnson. 2011. Unpacking and transforming feature functions: New ways to smooth phrase tables. In *Proceedings of Machine Translation Summit*.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada, June.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135, Prague, Czech Republic, June.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 848–856.

Weiwei Guo and Mona Diab. 2010. COLEPL and COLSLM: An unsupervised wsd approach to multilingual lexical substitution, tasks 2 and 3 semeval 2010. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 129–133, Uppsala, Sweden, July.

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 Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic, June.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Els Lefever and Véronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20, Uppsala, Sweden, July.

Els Lefever and Véronique Hoste. 2013. Semeval-2013 task 10: Cross-lingual word sense disambiguation. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, USA, May.

Els Lefever, Véronique Hoste, and Martine De Cock. 2011. Parasense or how to use parallel corpora for word sense disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 317–322, Portland, Oregon, USA, June.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, Stroudsburg, PA, USA.

Hwee Tou Ng and Yee Seng Chan. 2007. SemEval-2007 Task 11: English Lexical Sample Task via English-Chinese Parallel Text. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007)*, pages 54–58, Prague, Czech Republic. SIGLEX.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–52.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May.

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 of the Association for Computational Linguistics*, Philadelphia, PA, July.

Maarten van Gompel. 2010. Uvt-wsd1: A cross-lingual word sense disambiguation system. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 238–241, Uppsala, Sweden, July.