Towards Semantic-based Hybrid Machine Translation between Bulgarian and English

Kiril Simov and **Petya Osenova** and **Alex Popov**
Linguistic Modelling Department, IICT-BAS
kivs|petya|alex.popov@bultreebank.org

Abstract

The paper focuses on the creation of a semantic-based hybrid Machine Translation system between Bulgarian and English in the domain of Information Technology. The preprocessing strategies are presented. A method for the substitution of English word forms with the synsets or Bulgarian representative lemmas is discussed. Finally, the creation of a factored model in the Moses system is described.

1 Introduction

In this paper we present first results from the implementation of a hybrid machine translation system between Bulgarian and English (en↔bg) using Word Sense Annotation (WSD) of the source language. There is an existing line of research that aims to combine the advantages of competing approaches to machine translation in a hybrid framework. (Thurmair, 2009) summarized several different architectures of hybrid systems using SMT and RBMT systems. Some widely explored ones are: 1) using an SMT to post-edit the outputs of an RBMT; 2) selecting the best translations from several hypotheses coming from different SMT/RBMT systems; and 3) selecting the best segments (phrases or words) from different hypotheses. In our case, after WSD, we use the alignment between Bulgarian and English WordNets for the generation of rules for word substitution. On the basis of these rules we generate a corpus with factors, on which the Moses system (Koehn et al., 2007) is trained: word forms, lemmas and POS tags as factors. Although we have not improved the baseline, the substitution rules have proven helpful. We plan to generate more sophisticated rules in our future work.

The structure of the paper is as follows: in the next section the related work is presented. Section 3 describes the workflow of the Parallel Corpora Processing. Section 4 outlines the experiments that have been conducted. Section 5 introduces some discussion on the results and concludes the paper.

2 Related Work

Previous work on using WSD for SMT has yielded mixed results. (Carpuat and Wu, 2005) report a negative impact on BLEU scores. They used a supervised WSD system to select translation candidates for the SMT system, but, contrary to common sense expectations, this only made the translation model perform worse. Several reasons for this are suggested, chiefly that the SMT model works well enough on its own and state-of-the-art WSD systems cannot really boost it in a significant number of cases, and also that SMT architectures might not be well-adapted to make use of the output of WSD systems. (Cabezas and Resnik, 2005) present an approach to using WSD for SMT, whereby target language lexical items are treated as "sense tags", given as soft translation alternatives to the translation model, which chooses the final version in accordance with its language model. The study reported a small gain against a base-

Proceedings of the 2nd Workshop on Semantics-Driven Machine Translation, pages 22–26,
San Diego, California, June 16, 2016.

line that is, according to the authors, stronger than the one used in (Carpuat and Wu, 2005). (Vickrey et al., 2005) recasts WSD as a translation task, defining the different sense options for the separate words as the words or phrases aligned to them in a parallel corpus. The authors demonstrate that this approach is successful, as tested on word translation and blank-filling, thus showing that WSD and SMT have a lot in common and improving one should be helpful for improving the other.

(Chan et al., 2007) present another study in which WSD is beneficial to SMT. Disambiguation is performed between the possible translations of each source phrase. Translations are selected so as to maximize the length of the chunk proposed by the WSD model; the score provided by the WSD model is also taken into consideration. This approach yields a statistically significant improvement in terms of BLEU score. In a study that builds on their previously discouraging results, (Carpuat and Wu, 2007) show how a deeper integration of WSD into SMT systems can help systematically and significantly. Instead of performing disambiguation on single words, their system performs multi-word phrasal disambiguation, thus achieving improvements over the baseline, as measured by eight different translation metrics. The rich context provided by the supervised WSD system helps rank correct translations higher than erroneous ones suggested by the baseline SMT system; also, it helps the decoder pick longer translation sequences, which often results in better translations.

3 Parallel Corpora Processing

We are building a machine translation system between Bulgarian and English that can support the automatic identification of appropriate answers to user questions in a multilingual question/answering system. The domain of interest is related to information technology, smart phones and related devices and technologies. We have three domain-specific corpora of a thousand real-world question-answer pairs each - called Batch1, Batch2 and Batch3, respectively. In the experiments reported here, we exploit Batch1 for tuning the translation model and Batch3 for testing. For some experiments we have divided the questions from the answers. Then the two subcorpora have been denoted with "a" and "q" subscripts: Batch3a and Batch3q.

As training data we used the following corpora: the Setimes parallel corpus, the Europarl parallel corpus and a corpus created on the basis of the documentation of LibreOffice. The corpora are linguistically processed with the IXA[1] pipeline for the English part and the BTB pipeline for the Bulgarian. The analyses include POS tagging, lemmatization and WSD, using the UKB system[2], which provides graph-based methods for Word Sense Disambiguation and measuring lexical similarity. The tool uses the Personalized PageRank algorithm, described in Agirre and Soroa (2009). It has been used to perform Named Entity Disambiguation as well (Agirre et al., 2015). We have exploited the mapping between Bulgarian WordNet (BTB-WordNet) and Princeton English WordNet (PWN) in order to perform the Bulgarian WSD task — (Simov et al., 2015a).

For the baseline MT system, the following factors have been constructed: WordForm|Lemma|POStag. We have trained the Moses system using these factors. The results are presented in Table 1.

We also explored the impact of the bilingual morphological lexicons in the translation process, due to the occurrences of the so-called out-of-training word forms in the texts. See more in (Simov et al., 2015b). The bilingual lexicon was constructed by exploiting the following resources: BTB-Morphological lexicon containing all word-forms for more than 110 000 Bulgarian lemmas; BTB-bilingual Bulgarian-English lexicons (with about 8000 entries); English Wiktionary. From it the English word-forms were extracted for the English lemmas. Then we mapped the word-form lexicons for both languages to the corresponding part of the bilingual

[1] http://ixa2.si.ehu.es/ixa-pipes/
[2] http://ixa2.si.ehu.es/ukb/

lexicon. Afterwards, the corresponding word-forms were aligned on the basis of their morphological features like *number* (singular, plural); *degree* (comparative, superlative); *definiteness* (definite, indefinite), etc.

The lexicon represents more than 70 000 aligned word-forms. It was added to the training data. The results show the positive impact of the wordform-aligned parallel lexicon on the translation in both directions.

Our goal was to check the impact of WSD on this type of factor-based MT. The very first approach was to substitute the word form or lemma in the source text with a WordNet ID as a representation of the concept encoded by the corresponding synset. Additionally, we selected an appropriate lemma in the target language for the synset. Thus we relied on the concept information returned by the WSD software for the source text in two ways: a) to use the synset ID directly as a factor, or b) choose a *representative lemma* in the target language for that synset and present this representative lemma as a factor (in addition to the source word-form factor). The motivation for using representative lemmas in the target language is as follows: we aim at unifying the various synsets with similar translations in the target language. For example, in the en→bg direction, the two concepts referred by *donor*: `wn30-10025730-n` ("person who makes a gift of property") and `wn30-10026058-n` ("a medical term denoting someone who gives blood or tissue or an organ to be used in another person") are very close to each other, but they have the same translation in Bulgarian: ***донор***. The representative word is selected on the basis of a frequency list of Bulgarian lemmas.

4 Experiments

Three experiments have been performed: using synset IDs returned by the WSD software (ExpA); using representative target language lemmas for the synsets returned by the WSD software (ExpB); using representative target lemmas where the WSD software is run on domain-adapted wordnets, extended with domain gazetteers and terms (ExpC). The experiments for en→bg were performed through these steps: (1) annotation of the English text with the IXA pipeline, including tokenization, sentence splitting, part-of-speech tagging and word sense annotation with UKB; (2) substitution of the English word form with the synset (in ExpA) or Bulgarian representative lemma (in ExpB and ExpC); and (3) creating a factored model in the Moses system. In the direction bg→en we performed similar processing. Additionally, we provided part-of-speech tags[3] (PoS) from the pipeline, as well as the source language lemma, as factors for Moses. The PoS factor is important for Bulgarian, since Bulgarian is a morphologically rich language. Here is an example for the procedure we performed with respect to the training, testing and tuning tasks:

English sentence:
This is real progress ...

English sentence with factors:
this|this|dt is|be|vbz реален|real|jj напредък|progress|nn .|.|.

Bulgarian sentence with factors:
това|това|pd е|съм|vx реален|реален|a напредък|напредък|nc .|.|pu

Bulgarian sentence:
Това е реален напредък.

In order to adapt the semantic processing, we incorporated a Linked Open Data resource (DBPedia) in the en↔bg experiments via a mapping of the DBpedia ontology to WordNet. Our goal was to use again the IXA pipeline for the WSD task, similarly to the lexical semantics experiments. Unfortunately, the DBpedia ontology contains very few relevant classes, like *software, website, database*. For that reason, we decided to use an additional ontology created in a previous European project, LT4eL,[4] which covers about 1500 classes in the domain of Information technology. This ontology is already aligned to OntoWordNet (Gangemi et al., 2003), which is the basis for the extension of the existing word-

[3] In our case PoS tags include some morphosyntactic features.

[4] http://www.lt4el.eu/

System	Source factors	Domain terms	en→bg BLEU	en→bg NIST	bg→en BLEU	bg→en NIST
baseline	form	no	17.72	–	22.56	–
ExpA	synset-id|form, lemma, PoS	no	16.23	4.81	16.72	4.99
ExpB	repres-lemma|form, lemma, PoS	no	17.23	4.95	20.05	5.61
ExpC	repres-lemma|form, lemma, PoS	yes	17.41	4.98	19.92	5.58

Table 1: BLEU and NIST results of en↔bg experiments with concepts on Batch3 (Batch3a for en→bg and Batch3q for bg→en). The baseline is Pilot 0, where no synsets are used. In ExpA, the synset ID is added (if it exists). In ExpB, *repres-lemma* is added, which is a representative target language lemma for the given (source language) synset. ExpC is the same as ExpB, but the WSD resource used was enriched with domain terms.

nets as used by the WSD system (UKB). Then we performed the substitution of synsets with selected representative target language lemmas and trained the Moses system with the following factors: SubstitutedWordform[5], Lemma, PoS tag.[6] Table 1 presents the results. All three en↔bg experiments with lexical semantics show a drop in the results with respect to the baseline. But the addition of substituted lemmas improves over the usage of the concept itself (synset-id). The addition of domain knowledge also increases the performance. This justifies further experiments in the direction of extending the rules for substitution of the source language chunks with target language chunks.

5 Discussion and Conclusions

Here we reported on the initial steps of implementation of hybrid semantic statistical MT systems between Bulgarian and English. Our next development goal is to improve on the initial translation steps from the source to the target language by learning transfer rules from corpora. The main points of improvement will be:

Selection of appropriate representative lemma for a given sense. The current mechanism for selection of a representative lemma for a given WordNet synset is not the most appropriate. Thus we envisage the following options. First, word embeddings are trained on large corpora (which are first lemmatized). Then, for the lemmas in the synset we select their vectors and calculate their centroid. Our hope is that the centroid will be the best vector representation of the synset in the vector space. Afterwards, as a representative lemma we will select the lemma with a vector that is closest to the centroid. If more than one lemmas are very close to the centroid, we will look for additional information like frequency counts in the corpus, or we will perform manual selection.

Processing of analytical verb forms. Both languages in the translation pair have rich analytical verb complexes. On the basis of aligned parallel corpora, we will collect information about the most frequent translations. Then, with the help of the statistics we will construct rules for the translation of verbs between languages. These rules will be integrated with the rules for selection of the representative lemmas for verbal synsets.

Transfer of relations including close-class words. By using sense-annotated parallel corpora, we plan to learn the most frequent translations of prepositions between words that belong to given senses in both wordnets.

Using such rules we expect to be able to translate bigger portions of the source language text to the target language, and in this way to improve the overall translation.

Acknowledgments

This research has received partial funding from the EC's FP7 under grant agreement number 610516: "QTLeap: Quality Translation by Deep Language Engineering Approaches".

[5]For some word forms like prepositions, conjunctions, etc., the original word-form is kept.

[6]The parameters for training the Moses system are: `--translation-factors 0,2-0,2+1,2-0,2 --decoding-steps t0:t1`

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 33–41, Athens, Greece, March. Association for Computational Linguistics.

Eneko Agirre, Ander Barrena, and Aitor Soroa. 2015. Studying the wikipedia hyperlink graph for relatedness and disambiguation. *arXiv preprint arXiv:1503.01655*.

Clara Cabezas and Philip Resnik. 2005. Using wsd techniques for lexical selection in statistical machine. Technical report, Translation Technical report CS-TR-4736.

Marine Carpuat and Dekai Wu. 2005. Word sense disambiguation vs. statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 387–394. Association for Computational Linguistics.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *EMNLP-CoNLL*, volume 7, pages 61–72.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Annual Meeting-Association for Computational Linguistics*, volume 45, page 33. Citeseer.

Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003. The ontowordnet project: Extension and axiomatization of conceptual relations in wordnet. In Robert Meersman, Zahir Tari, and DouglasC. Schmidt, editors, *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, volume 2888 of *Lecture Notes in Computer Science*, pages 820–838. Springer Berlin Heidelberg.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

Kiril Simov, Alexander Popov, and Petya Osenova. 2015a. Improving word sense disambiguation with linguistic knowledge from a sense annotated treebank. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 596–603.

Kiril Simov, Iliana Simova, Velislava Todorova, and Petya Osenova. 2015b. Factored Models for Deep Machine Translation. In *Proceedings of the 1st Deep Machine Translation Workshop (DMTW 2015)*, pages 97–105.

Gregor Thurmair. 2009. Comparing different architectures of hybrid machine translation systems. In *Proceedings of MT Summit XII*.

David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-sense disambiguation for machine translation. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 771–778. Association for Computational Linguistics.