Using TectoMT as a Preprocessing Tool for Phrase-Based Statistical Machine Translation

Daniel Zeman

Univerzita Karlova v Praze, ÚFAL,
Malostranské náměstí 25, 11800 Praha, Czechia
zeman@ufal.mff.cuni.cz
http://ufal.mff.cuni.cz/~zeman/

Abstract. We present a systematic comparison of preprocessing techniques for two language pairs: English-Czech and English-Hindi. The two target languages, although both belonging to the Indo-European language family, show significant differences in morphology, syntax and word order. We describe how TectoMT, a successful framework for analysis and generation of language, can be used as preprocessor for a phrase-based MT system. We compare the two language pairs and the optimal sets of source-language transformations applied to them. The following transformations are examples of possible preprocessing steps: lemmatization; retokenization, compound splitting; removing/adding words lacking counterparts in the other language; phrase reordering to resemble the target word order; marking syntactic functions. TectoMT, as well as all other tools and data sets we use, are freely available on the Web.

Key words: phrase-based translation, preprocessing, reordering

1 Introduction

It is widely accepted that linguistically informed preprocessing of training data can improve quality of statistical machine translation. The general goal is, in most cases, to make the source and the target texts grammatically more similar and thus easier to learn for a statistical machine translation system. Both source and target languages can be preprocessed. The task is easier if we restrict preprocessing to the source language. In this case, the source part of the training parallel data is preprocessed in the hope that the resulting string can be better aligned with the target string and thus better phrase translation model can be learned. During the decoding phase (i.e. applying the model to new unseen data), the source test corpus is preprocessed exactly the same way, then the model is applied.

If we choose to preprocess the target side of the training data, we need to be able to reverse the transformation in a postprocessing step after the decoding phase. The assumption is that the model trained on the preprocessed data will produce output similar to the preprocessed data. However, the required output of the MT system is natural, unpreprocessed target language string. On the

other hand, we cannot rely on any expected structure of the output, as the MT system can (and will) make errors.

There are several reasons for considering preprocessing of parallel corpora:

- Richer morphology on one side results in sparse data. For instance, the English word *woman* may appear in singular or in plural *(women)*. In contrast, its Czech translation *žena* is marked for number and case, resulting in 10 distinct forms *(žena, ženy, ženě, ženu, ženo, ženou, žen, ženám, ženách, ženami)*. It is not realistic to expect that each of these forms will occur frequently enough in the training data, with every possible English translation. By separating morphology from the lexical information the data sparseness can be reduced. Similar effect can be achieved by separating compound words (e.g. in German) or separating morphemes that would be standalone words in the target language (e.g. in Arabic-to-English translation).
- If the target language is the morphologically richer of the two, generating source pseudowords bearing necessary information such as syntactic functions (subject, object etc.) can help to figure out the correct target word forms. For other target languages, pseudowords can help generate target words that normally do not have direct source counterparts. For instance, pro-drop languages such as Czech or Spanish do not require that personal pronouns are present when otherwise there would be no subject. However, verbs in such languages are often marked for person, which can help with generating the correct personal pronoun in the non-pro-drop target language. Thus, in Czech-to-English MT, we have to learn *jdu* $\rightarrow$ *I go*, *jdeš* $\rightarrow$ *you go* etc. The English personal pronouns can be more easily generated if we augment the Czech source with explicit person+number information, supplied by morphological analysis of the Czech verbs.
- Significant differences in word order between the two languages. Preprocessing includes syntactic parsing of the source language, then the phrases are reordered according to some rules. The availability of a parser is crucial here because whole phrases have to be moved along, not just words.

The following transformations are examples of possible preprocessing steps: lemmatization; retokenization, compound splitting; removing/adding function words that systematically lack counterparts in the other language (articles, personal pronouns etc.); reordering of phrases in parsed source sentence in order to make the word order closer to that of the target language; adding pseudo-tokens for syntactic functions such as subject, predicate, object.

There have been numerous publications on various aspects of preprocessing for several language pairs. In this paper, we present a systematic comparison of preprocessing techniques for two language pairs: English-Czech and English-Hindi. Due to the reasons mentioned above, we restrict ourselves to preprocessing of the source language. The two target languages, although both belonging to the Indo-European language family, show significant differences in morphology, syntax and word order. We describe how TectoMT, a successful framework developed originally for deep-syntax-based machine translation, can be used as preprocessor for a phrase-based MT system, such as Moses or Joshua. We compare

the two language pairs and the optimal sets of source-language transformations applied to them.

The rest of the paper is organized as follows: in Section 2 we summarize the related work, in Section 3 we introduce TectoMT and other software, in Section 4 we describe the transformations used for each language. Then we describe the data sets (Section 5) and discuss preliminary results (Section 6).

2 Related Work

There exists a body of previous work that is related to ours in one or more aspects. We discuss a selection of related publications in this section.

Nießen and Ney [1] describe a German-to-English MT system that integrates morphology-based preprocessing of German. They split German compound words, join separable verb prefixes with verbs and augment German words with morphological information. They observe that while many German morphological features (such as the distinction between the nominative and the accusative) are not reflected in English, sometimes more morphological information is present in the English word than in its German counterpart: *das Zimmer* → *the room* vs. *die Zimmer* → *the rooms*.

Collins, Koehn and Kučerová [2] also experiment with German-to-English SMT. They use a syntactic parser to obtain an analysis of the source language string, then they apply a series of transformations to the parse tree, effectively reordering the source string. The goal of this step is to recover an underlying word order that is closer to the target language word order than the original string. They report a statistically significant improvement of the BLEU score on the Europarl corpus.

Popović et al. [3] present results on a very small Serbian-English corpus for both translation directions, sr-en and en-sr. For each direction, they preprocess the source side of the corpus. English preprocessing is limited to the removal of articles. Serbian preprocessing consists of two steps: lemmatization and special treatment of verbs (person verb feature is used to generate missing personal pronoun).

Goldwater and McClosky [4] discuss the Czech-to-English task on the Prague Czech-English Dependency Treebank. To reduce the data sparseness problem, they first lemmatize the source Czech text, than attempt to partially restore the lost information by introducing pseudowords or separated morphemes.

Different issues are encountered in Arabic-to-English translation (Habash and Sadat [5], El Isbihani et al. [6]). Here the preprocessing mostly involves English-like retokenization of Arabic (comparable to the compound splitting in German), i.e. separating conjunctions, prepositions and articles that are normally written jointly with the noun.

Prokopová [7] investigates various ways of enriching Czech input in Czech-to-English translation. Besides word reordering (to get the fixed English subject-verb-object word order), she also inserts into the Czech string frequent English

words that may not have any counterpart in Czech: articles, personal pronouns, the infinitival marker *to*, prepositions *of* and *by*.

Avramidis and Koehn [8] use parse trees of the source English text not to reorder it but rather to acquire information about syntactic functions of the English words. That information can then be made explicit and help generate the correct case marker in the target language within an English-to-Greek MT experiment. Reduction of errors in verb conjugation and in noun case agreement is reported.

Axelrod et al. [9] present another experiment with German stemming and compound splitting but this time for a German-to-Spanish MT system.

Popović et al. [10] apply part-of-speech-based (i.e., no parsing) reorderings of the source language to the German-, French- and Spanish-to-English tasks. Again, German compound splitting is found helpful, too.

Finally, Ramanathan et al. [11] address the large word-order discrepancy in English-to-Hindi MT, along with richer morphology of the target language. They use preprocessing to figure out the English syntactic functions and to get the target SOV word order; they also use postprocessing to generate Hindi case markers and suffixes.

In general, former work focused more on translation to English (which usually meant into the morphologically poorer language) than on translation into a morphologically rich language; however, the interest in the latter has been increasing recently.

3 TectoMT and Related Tools

TectoMT [12] is a highly modular NLP framework implemented in Perl under Linux. It was originally developed to facilitate machine translation within the classical analysis-transfer-synthesis paradigm. It is composed of numerous reusable processing modules (called "blocks"), which are equipped with uniform object-oriented interfaces. Some of the blocks wrap large NLP applications such as taggers and parsers (together with pre-trained models), others are designed to perform tiny specialized operations: for instance, operating on output of a particular parser, a block can apply some heuristics to correct treatment of coordination. Unified application programming interface allows for rapid development of such language transformations without having to care about the file format, task parallelization etc. Because of the unique modular environment, the usefulness of TectoMT extends beyond machine translation to virtually any natural language processing task.

We use TectoMT to analyze the English side of the parallel corpora. We do not use the transfer- and generation blocks of TectoMT; instead, we train a phrase-based SMT system on the preprocessed corpora. Two blocks wrapped in TectoMT deserve being mentioned separately: the morphosyntactic tagger Morče [13] and the MST (maximum spanning tree) dependency parser[14]. Besides and around these two, we reuse nearly 40 other blocks that the TectoMT developers

designed and routinely use to improve the analysis of English texts. On top of it, our reordering block takes care for the transformations described below.

As the phrase-based SMT component, we use and Joshua [15].

4 Overview of Transformations

4.1 English to Czech

Articles. There are no definite or indefinite articles in Czech. The SMT systems waste energy to align them to Czech, sometimes it makes the data unnecessarily sparse: e.g., Czech *pražskou* has two English counterparts, *the Prague* and *Prague the*. Solution: All occurrences of the words *a, an, the* tagged `DT` are removed.

Target case selection. English almost completely lacks the notion of grammatical case (except for the direct and oblique cases of pronouns). In Czech, there are 7 cases. In general, it is not easy to select the correct case (see also *Target agreement* below), however, the subject is typically in nominative. Hence appending `/Sb` to the root word of the English subject (provided we have parsed the English input) can help to generate the nominative on the Czech side.

Target agreement. It is difficult to generate target phrases that agree in gender, number and case as required by Czech grammar. For instance, English *trading day* can be translated as nominative *obchodní den*, genitive *obchodního dne*, dative *obchodním dni* etc. If the SMT system does not learn the Czech phrase in all the cases, it will attempt to translate each word separately, in which case however it will lose the agreement feature. Thus, incorrect translations such as **obchodním dne* are frequently seen in the output. The solution is more tricky in this case. We could separate lemma from the morphological features in the Czech text; however, this would mean preprocessing of the target text, which we prefer to avoid. We leave this problem open for further research.

Verbal groups. English has many analytical tenses of verbs and is richer than Czech in that respect. To make it easier for phrase-based SMT systems to get the correct tense, we move all auxiliaries, modal verbs etc. as close to the main verb as possible. Example: *will only make matters worse* → *only will make worse matters*.

Personal pronouns. English personal pronouns functioning as subjects are joined with their verbs. Word alignment tends to align them to Czech verbs anyway; however, there is room for mis-alignments and data sparseness is unnecessarily increased.

4.2 English to Hindi

Articles. Similarly to Czech, there are no definite articles in Hindi. However, indefinite articles are sometimes translated using the numeral एक *(eka)* ("one"). Solution: Remove occurrences of *the* from the English text.

Postpositions. English prepositions are usually translated as postpositions in Hindi. Sometimes the postpositions are compound and they require concrete case ending for the preceding noun or pronoun. Examples: *in the house* → घर में *(ghara meṁ)* ("house in"), *my teacher's book* → मेरे अध्यापक की किताब *(mere adhyāpaka kī kitāba)* ("my-`oblique` teacher of-`fem` book"), *towards Ram* → राम की तरफ़ *(rāma kī tarafa)* ("Ram of-`fem` direction"). Solution: Convert English prepositions to postpositions, i.e. move them after the noun phrase they govern.

Subject-object-verb order. English is an SVO language while Hindi is an SOV language, i.e. Hindi verbs occur mostly at the end of the clause, as in:

I'm doing some work with a friend.
एक मित्र के साथ कुछ काम कर रहा हूँ ।
(eka mitra ke sātha kucha kāma kara rahā hūṁ .)
"one friend of-`masc` with some work do -ing-`masc` I-am ."

Solution: reorder clauses so that the main finite verb goes to the end.

To have. There is no direct equivalent in Hindi for the English verb *to have*. Instead, various indirect constructions are used to convey the sense of having. Example:

We have time.
हमारे पास समय है ।
(hamāre pāsa samaya hai.)
"our-`oblique` at time is."

Solution: Make *to have* an exception to the verb reordering rule introduced above. Keep it with its subject and let the SMT learn translations like *we have* → हमारे पास *(hamāre pāsa)*, *X has* → X के पास *(X ke pāsa)* etc.

5 Data

All test sets mentioned in this section have only one reference translation per sentence.

5.1 English to Czech

We use the News Commentary 10 corpus (94,697 sentence pairs) for training, the WMT 2008 test set (2,051 sentence pairs) for development and the WMT 2009 test set (2,525 sentence pairs) for testing. All these corpora are freely available at `http://www.statmt.org/wmt10/translation-task.html`.

5.2 English to Hindi

We use a cleaned version of the IIIT Tides corpus. This dataset was originally collected for the DARPA-TIDES surprise-language contest in 2002, later refined at IIIT Hyderabad and provided for the NLP Tools Contest at ICON 2008 [16]. The corpus is a general domain dataset with news articles forming the greatest proportion. It is aligned on sentence level, and tokenized to some extent. There are 50K sentence pairs for training, 1K pairs for development and 1K for testing.

Table 1. Translation from preprocessed English to Czech/Hindi.

Method	BLEU
English-Czech, Baseline	0.0863
English-Czech, Preprocessed	0.0905
English-Hindi, Baseline	0.1006
English-Hindi, Preprocessed	0.1029

6 Results

We evaluate the impact of the preprocessing transformations in two ways. A manual evaluation focuses at the phenomena described in 4. Human inspection of 50 sentence pairs selected randomly from the test data revealed that the case selection in Czech, and the alignment in both Czech and Hindi (with the reordered English) improved.

The newly aligned corpora were then used to train new translation models for the Joshua decoder and BLEU score has been used to evaluate the accuracy of the new models on test data. Unfortunately, all quantitative improvements so far are statistically insignificant. Table 1 presents the BLEU scores; as the impact of the transformations is rather low, we do not present detailed figures for isolated transformations.

Space limitations of this paper do not allow to describe all details of error analysis; however, translations generated by the model for the test data seem to suggest that morphology generation on the target side is a more important source of errors than the alignment problems addressed by our transformations.

7 Conclusion

We proposed a number of source-side transformations of English text in order to make it grammatically more similar to the target language, namely Czech and Hindi. We gave a comprehensive overview of related work and argued that TectoMT, a modular NLP framework, is a tool highly suitable for the preprocessing task. Different sets of transformations were proposed w.r.t. the given target language. The preprocessed corpora led to better word alignment but not to significant improvement of translation quality in terms of BLEU score. Future research will focus on morphology of the target language, which seems to be a more important source of errors.

Acknowledgments. The work on this project has been supported by the grant MSM0021620838 by the Czech Ministry of Education.

References

1. Nießen, S., Ney, H.: Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information. Computational Linguistics 30(2):181–204 (2004)

2. Collins, M., Koehn, P., Kučerová, I.: Clause Restructuring for Statistical Machine Translation. In: Proceedings of the 43rd Annual Meeting of the ACL, pp. 531–540. ACL, Ann Arbor, Michigan, USA (2005)
3. Popović, M., Vilar, D., Ney, H., Jovičić, S., Šarić, Z.: Augmenting a small parallel text with morpho-syntactic language. In: Proceedings of the ACL Workshop on Building and Using Parallel Texts, pp. 41–48. ACL, Ann Arbor, Michigan, USA (2005)
4. Goldwater, S., McClosky, D.: Improving Statistical MT through Morphological Analysis. In: Proceedings of HLT-EMNLP, pp. 676–683. ACL, Vancouver, British Columbia, Canada (2005)
5. Habash, N., Sadat, F.: Arabic Preprocessing Schemes for Statistical Machine Translation. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pp. 49–52. ACL, New York, USA (2006)
6. El Isbihani, A., Khadivi, S., Bender, O., Ney, H.: Morpho-syntactic Arabic Preprocessing for Arabic-to-English Statistical Machine Translation. In: Proceedings of the Workshop on Statistical Machine Translation, pp. 15–22. ACL, New York, USA (2006)
7. Prokopová, M.: Automatic Simplification of Texts for Translation. Master's thesis, Univerzita Karlova v Praze, Praha, Czechia (2007)
8. Avramidis, E., Koehn, P.: Enriching Morphologically Poor Languages for Statistical Machine Translation. In: Proceedings of ACL-08: HLT, pp. 763–770. ACL, Columbus, Ohio, USA (2008)
9. Axelrod, A., Yang, M., Duh, K., Kirchhoff, K.: The University of Washington Machine Translation System for ACL WMT 2008. In: Proceedings of the Third Workshop on Statistical Machine Translation, pp. 123–126. ACL, Columbus, Ohio, USA (2008)
10. Popović, M., Vilar, D., Stein, D., Matusov, E., Ney, H.: The RWTH Machine Translation System for WMT 2009. In: Proceedings of the Fourth Workshop on Statistical Machine Translation, pp. 66–69. ACL, Athîna, Greece (2009)
11. Ramanathan, A., Choudhary, H., Ghosh, A., Bhattacharyya, P.: Case markers and Morphology: Addressing the crux of the fluency problem in English-Hindi SMT. In: Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pp. 800–808. ACL and AFNLP, Suntec, Singapore (2009)
12. Žabokrtský, Z., Ptáček, J., Pajas, P.: TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In: Proceedings of the Third Workshop on Statistical Machine Translation, pp. 167–170. ACL, Columbus, Ohio, USA (2008)
13. Votrubec, J.: Selecting an optimal set of features for the morphological tagging of Czech. Master thesis, Univerzita Karlova v Praze, Praha, Czechia (2005)
14. McDonald, R., Pereira, F., Ribarov, K., Hajič, J.: Non-projective Dependency Parsing using Spanning Tree Algorithms. In: Proceedings of the Human Language Technology / Empirical Methods in Natural Language Processing conference (HLT-EMNLP), pp. 523–530. ACL, Vancouver, British Columbia, Canada.
15. Li, Z., Callison-Burch, C., Khudanpur, S., Thornton, W.: Decoding in Joshua: Open Source, Parsing-Based Machine Translation. In: The Prague Bulletin of Mathematical Linguistics, vol. 91, pp. 47–56.
16. Venkatapathy, S.: NLP Tools Contest – 2008: Summary. In: Proceedings of ICON 2008 NLP Tools Contest. Pune, India (2008)