Treebanks in Machine Translation

Martin Čmejrek, Jan Cuřín, Jiří Havelka

Charles University in Prague
Center for Computational Linguistics
{cmejrek,curin,havelka}@ckl.mff.cuni.cz

1 Introduction

We present an approach using treebanks in machine translation. Our experiment in Czech-English machine translation is an attempt to develop a full machine translation system based on dependency trees (Dependency Based Machine Translation, DBMT). We use the following resources: Prague Dependency Treebank, a newly created Czech-English parallel corpus of Penn Treebank, English monolingual corpus, and translation lexicons. The fully automatic process includes analysis of the Czech input into tectogrammatical (semantic) representation, lexical and structural transfer, a simple rule-based system for generation to English surface realization, and an n-gram language model for scoring and choosing from translation hypotheses. The results are evaluated quantitatively with BLEU score.

2 Data Resources

Prague Dependency Treebank [7] is a project aiming at a complex annotation of a corpus containing about 1.5M word occurrences (about 100k sentences) in Czech. The annotation, which is based on dependency syntax, is carried out in three steps: morphological, analytical, and tectogrammatical. The first two steps have been finished so far, presently, there are about 50k sentences tectogrammatically annotated. Dependency trees represent the sentence structure as concentrated around the verb and its valency; we use tectogrammatical dependency trees devised to capture the linguistic meaning of the sentence. In a tectogrammatical dependency tree, only autosemantic (lexical) words are represented as nodes, dependencies (edges) are labeled by tectogrammatical functors denoting semantic roles, the information conveyed by auxiliary words is stored as special attributes of the nodes.

Czech-English Penn Treebank is a human translation of a considerable part of Penn Treebank [8]. The translators were asked to translate each English sentence as a single Czech sentence and also to stick to the original sentence construction if possible. In the experiment, we have used about 11k sentence pairs. For both training and evaluation measured by BLEU metrics, 490 sentences were retranslated back from Czech into English by 4 different translators. The Czech sentences were also manually annotated on tectogrammatical level.

The transfer module of the DBMT system uses a translation dictionary. There were three different sources of Czech-English manual dictionaries available, two of them were downloaded form the Web, and one was extracted from Czech and English EuroWordNets. To each Czech entry, POS tag and lemma were assigned. We selected a few relevant translations for each entry taking into account the reliability of the source dictionary, the frequencies of the translations in the English monolingual corpus, and the correspondence of the Czech and English POS tags. To make the dictionary more sensitive to the specific domain of financial news, the translations were weighted. By running GIZA++ (Och and Ney [9]) translation model training on the training part of the Czech-English parallel corpus (7,412 sentences) extended by the manual dictionaries, we obtained a probabilistic Czech-English dictionary.

3 DBMT System Overview

The DBMT has the vintage analysis–transfer–generation architecture.

The Czech sentence is automatically tokenized, morphologically tagged, and each word form is assigned a lemma, i.e. a basic form, by Hajič and Hladká [5] tagging tools. A statistical dependency parser (either Collins parser [4], or Charniak parser [2]) is used to obtain the analytical representation. Then the analytical structure is converted into tectogrammatical representation using linguistic rules (Böhmová [1]), and tectogrammatical functors are assigned by a C4.5 classifier (Žabokrtský [11]).

In transfer, tectogrammatical base-form attribute of autosemantic nodes is replaced by its English equivalent found in the Czech-English probabilistic dictionary. We use contextual boundness for the reordering of constituents in the English counterpart of the Czech tectogrammatical structure and for determining the definiteness of noun phrases; here we make use of the fact that Czech is a language with a relatively high degree of word order freedom and in its written form it uses mainly left to right ordering to express the information structure.

When generating the surface form, an appropriate verb form is chosen (the active or passive voice, tense, mood, person are determined according to or taken

MT system	*BLEU – devtest*	*BLEU – evaltest*
DBMT with Collins parser	0.1857	0.1634
DBMT with Charniak parser	0.1916	0.1705
DBMT on manually annotated trees	0.1974	0.1704
GIZA++ & ReWrite – word forms	0.0971	0.0590
GIZA++ & ReWrite – lemmatized	0.2222	0.2017
MAGENTA WS'02	0.0640	0.0420
Avg. BLEU score of human retranslations	—	0.5560

Table 1: BLEU score of different MT systems

over from the semantic representation of the sentence). A packed-tree format is used to represent multiple variants of insertions of prepositions and articles. Rules for possible insertions are both manually written and automatically extracted from the data. The rules describing surface realization by preposition take into consideration the tectogrammatical functor, the original Czech preposition, and the English translation. The main criteria for inserting articles are contextual boundness and syntactico-semantic properties of a noun phrase. A simple module combining statistical and rule-based approaches generates the surface word form from the lemma and morphological tag. Finally, the packed-tree representation is unwrapped into an n-best list, scored by an n-gram language model for English, and the translation with the highest score is selected as the result.

4 Evaluation of Results and Conclusion

We carried out two experiments: an experiment implementing fully automatic translation from Czech plain text, and another experiment skipping the analysis of Czech into tectogrammatical representation with manually annotated tectogrammatical trees as input. We evaluated our translations with IBM's BLEU evaluation metric (Papineni et al. [10]) on 490 sentences and their four different reference human retranslations. For comparison, we also evaluated, on the same test set and with the same metric, results of GIZA++/ISI ReWrite Decoder (Germann et al. [3]) and the MAGENTA system (Hajič et al. [6]). The final results are presented in Table 1. Both our experiments show a considerable improvement over MAGENTA's performance, and they also score better than GIZA++/ReWrite trained on word forms, but we were still outperformed by GIZA++/ReWrite trained on lemmas. In order to further improve our results, we plan to integrate the language model into the translation process as soon as the transfer and generation steps.

5 Acknowledgements

This research was supported by the following grants: MŠMT ČR Grant LN00A063, MŠMT ČR Grant Kontakt ME642, and NSF Grant IIS-0121285.

References

[1] Alena Böhmová. Automatic procedures in tectogrammatical tagging. *The Prague Bulletin of Mathematical Linguistics*, 76, 2001.

[2] Eugene Charniak. A maximum-entropy-inspired parser. Technical Report CS-99-12, 1999.

[3] Ulrich Germann, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada. Fast decoding and optimal decoding for machine translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 228–235, 2001.

[4] Jan Hajič, Eric Brill, Michael Collins, Barbora Hladká, Douglas Jones, Cynthia Kuo, Lance Ramshaw, Oren Schwartz, Christopher Tillmann, and Daniel Zeman. Core Natural Language Processing Technology Applicable to Multiple Languages. Technical Report Research Note 37, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, 1998.

[5] Jan Hajič and Barbora Hladká. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In *Proceedings of COLING-ACL Conference*, pages 483–490, Montreal, Canada, 1998.

[6] Jan Hajič, Martin Čmejrek, Bonnie Dorr, Yuan Ding, Jason Eisner, Daniel Gildea, Terry Koo, Kristen Parton, Gerald Penn, Dragomir Radev, and Owen Rambow. Natural language generation in the context of machine translation. Technical report, 2002. NLP WS'02 Final Report.

[7] Linguistic Data Consortium. Penn Treebank 3. LDC99T42, 1999.

[8] Linguistic Data Consortium. Prague Dependency Treebank 1. LDC2001T10, 2001.

[9] F. J. Och and H. Ney. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hongkong, China, October 2000.

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176, IBM, 2001.

[11] Zdeněk Žabokrtský, Petr Sgall, and Džeroski Sašo. Machine learning approach to automatic functor assignment in the prague dependency treebank. In *Proceedings of LREC 2002 (Third International Conference on Language Resources and Evaluation)*, volume V, pages 1513–1520, Las Palmas de Gran Canaria, Spain, 2002.