English-Czech MT in 2008 *

Ondřej Bojar, David Mareček, Václav Novák, Martin Popel,
Jan Ptáček, Jan Rouš, Zdeněk Žabokrtský
Charles University, Institute of Formal and Applied Linguistics
Malostranské nám. 25, Praha 1, CZ-118 00, Czech Republic
{bojar,marecek,novak,ptacek,zabokrtsky}@ufal.mff.cuni.cz
{popel,jan.rous}@matfyz.cz

Abstract

We describe two systems for English-to-Czech machine translation that took part in the WMT09 translation task. One of the systems is a tuned phrase-based system and the other one is based on a linguistically motivated analysis-transfer-synthesis approach.

1 Introduction

We participated in WMT09 with two very different systems: (1) a phrase-based MT based on Moses (Koehn et al., 2007) and tuned for English→Czech translation, and (2) a complex system in the TectoMT platform (Žabokrtský et al., 2008).

2 Data

2.1 Monolingual Data

Our Czech monolingual data consist of (1) the Czech National Corpus (CNC, versions SYN200[056], 72.6%, Kocek et al. (2000)), (2) a collection of web pages downloaded by Pavel Pecina (Web, 17.1%), and (3) the Czech monolingual data provided by WMT09 organizers (10.3%). Table 1 lists sentence and token counts (see Section 2.3 for the explanation of a- and t-layer).

Sentences	52 M
with nonempty t-layer	51 M
a-nodes (i.e. tokens)	0.9 G
t-nodes	0.6 G

Table 1: Czech monolingual training data.

2.2 Parallel Data

As the source of parallel data we use an internal release of Czech-English parallel corpus CzEng (Bojar et al., 2008) extended with some additional texts. One of the added sections was gathered from two major websites containing Czech subtitles to movies and TV series[1]. The matching of the Czech and English movies is rather straightforward thanks to the naming conventions. However, we were unable to reliably determine the series number and the episode number from the file names. We employed a two-step procedure to automatically pair the TV series subtitle files. For every TV series:

1. We clustered the files on both sides to remove duplicates
2. We found the best matching using a provisional translation dictionary. This proved to be a successful technique on a small sample of manually paired test data. The process was facilitated by the fact that the correct pairs of episodes usually share some named entities which the human translator chose to keep in the original English form.

Table 2 lists parallel corpus sizes and the distribution of text domains.

	English	Czech
Sentences	6.91 M	
with nonempty t-layer	6.89 M	
a-nodes (i.e. tokens)	61 M	50 M
t-nodes	41 M	33 M

Distribution:	[%]		[%]
Subtitles	68.2	Novels	3.3
Software Docs	17.0	Commentaries/News	1.5
EU (Legal) Texts	9.5	Volunteer-supplied	0.4

Table 2: Czech-English data sizes and sources.

* The work on this project was supported by the grants MSM0021620838, 1ET201120505, 1ET101120503, GAUK 52408/2008, MŠMT ČR LC536 and FP6-IST-5-034291-STP (EuroMatrix).

[1] www.opensubtitles.org and titulky.com

Proceedings of the Fourth Workshop on Statistical Machine Translation , pages 125–129,
Athens, Greece, 30 March – 31 March 2009.

2.3 Data Preprocessing using TectoMT platform: Analysis and Alignment

As we believe that various kinds of linguistically relevant information might be helpful in MT, we performed automatic analysis of the data. The data were analyzed using the layered annotation scheme of the Prague Dependency Treebank 2.0 (PDT 2.0, Hajič and others (2006)), i.e. we used three layers of sentence representation: morphological layer, surface-syntax layer (called analytical (a-) layer), and deep-syntax layer (called tectogrammatical (t-) layer).

The analysis was implemented using TectoMT, (Žabokrtský et al., 2008). TectoMT is a highly modular software framework aimed at creating MT systems (focused, but by far not limited to translation using tectogrammatical transfer) and other NLP applications. Numerous existing NLP tools such as taggers, parsers, and named entity recognizers are already integrated in TectoMT, especially for (but again, not limited to) English and Czech.

During the analysis of the large Czech monolingual data, we used Jan Hajič's Czech tagger shipped with PDT 2.0, Maximum Spanning Tree parser (McDonald et al., 2005) with optimized set of features as described in Novák and Žabokrtský (2007), and a tool for assigning functors (semantic roles) from Klimeš (2006), and numerous other components of our own (e.g. for conversion of analytical trees into tectogrammatical ones).

In the parallel data, we analyzed the Czech side using more or less the same scenario as used for the monolingual data. English sentences were analyzed using (among other tools) Morce tagger Spoustová et al. (2007) and Maximum Spanning Tree parser.[2]

The resulting deep syntactic (tectogrammatical) Czech and English trees are then aligned using T-aligner—a feature based greedy algorithm implemented for this purpose (Mareček et al., 2008). T-aligner finds corresponding nodes between the two given trees and links them. For deciding whether to link two nodes or not, T-aligner makes use of a bilingual lexicon of tectogrammatical lemmas, morphosyntactic similarities between the two candidate nodes, their positions in the trees and other similarities between their parent/child nodes. It also uses word alignment generated from surface shapes of sentences by GIZA++ tool, Och and Ney (2003). We use acquired aligned tectogrammatical trees for training some models for the transfer.

As analysis of such amounts of data is obviously computationally very demanding, we run it in parallel using Sun Grid Engine[3] cluster of 40 4-CPU computers. For this purpose, we implemented a rather generic tool that submits any TectoMT pipeline to the cluster.

3 Factored Phrase-Based MT

We essentially repeat our experiments from last year (Bojar and Hajič, 2008): GIZA++ alignments[4] on a-layer lemmas (a-layer nodes correspond 1-1 to surface tokens), symmetrized using grow-diag-final (no -and) heuristic[5].

Probably due to the domain difference (the test set is news), including Subtitles in the parallel data and Web in the monolingual data did not bring any improvement that would justify the additional performance costs. For most of the phrase-based experiments, we thus used only 2.2M parallel sentences (27M Czech and 32M English tokens) and 43M Czech sentences (694 M tokens).

In Table 3 below, we report the scores for the following setups selected from about 50 experiments we ran in total:

Moses T is a simple phrase-based translation (T) with no additional factors. The translation is performed on truecased word forms (i.e. sentence capitalization removed unless the first word seems to be a name). The 4-gram language model is based on the 43M sentences.

Moses T+C is a factored setup with form-to-form translation (T) and target-side morphological coherence check following Bojar and Hajič (2008). The setup uses two language models: 4-grams of word forms and 7-grams of morphological tags.

Moses T+C+C&T+T+G 84k is a setup desirable from the linguistic point of view. Two independent translation paths are used: (1) form→form translation with two target-side checks (lemma and tag generated from the target-side form) as a fine-grained baseline

[2] In some previous experiments (e.g.Žabokrtský et al. (2008)), we used phrase-structure parser Collins (1999) with subsequent constituency-dependency conversion.

[3] `http://gridengine.sunsource.net/`

[4] Default settings, IBM models and iterations: $1^5 3^3 4^3$.

[5] Later, we found out that the grow-diag-final-and heuristic provides insignificantly superior results.

with the option to resort to (2) an independent translation of lemma→lemma and tag→tag finished by a generation step that combines target-side lemma and tag to produce the final target-side form.

We use three language models in this setup (3-grams of forms, 3-grams of lemmas, and 10-grams of tags).

Due to the increased complexity of the setup, we were able to train this model on 84k parallel sentences only (the Commentaries section) and we use the target-side of this small training data for language models, too.

For all the setups we perform standard MERT training on the provided development set.[6]

4 Translation Setup Based on Tectogrammatical Transfer

In this translation experiment, we follow the traditional analysis-transfer-synthesis approach, using the set of PDT 2.0 layers: we analyze the input English sentence up to the tectogrammatical layer (through the morphological and analytical ones), then perform the tectogrammatical transfer, and then synthesize the target Czech sentence from its tectogrammatical representation. The whole procedure consists of about 80 steps, so the following description is necessarily very high level.

4.1 Analysis

Each sentence is tokenized (roughly according to the Penn Treebank conventions), tagged by the English version of the Morce tagger Spoustová et al. (2007), and lemmatized by our lemmatizer. Then the dependency parser (McDonald et al., 2005) is applied. Then the analytical trees resulting from the parser are converted to the tectogrammatical ones (i.e. functional words are removed, only morphologically indispensable categories are left with the nodes using a sequence of heuristic procedures). Unlike in PDT 2.0, the information about the original syntactic form is stored with each t-node (values such as `v:inf` for an infinitive verb form, `v:since+fin` for the head of a subordinate clause of a certain type, `adj:attr` for an adjective in attribute position, `n:for+X` for a given prepositional group are distinguished).

[6] We used the full development set of 2k sentences for "Moses T" and a subset of 1k sentences for the other two setups due to time constraints.

One of the steps in the analysis of English is named entity recognition using Stanford Named Entity Recognizer (Finkel et al., 2005). The nodes in the English t-layer are grouped according to the detected named entities and they are assigned the type of entity (location, person, or organization). This information is preserved in the transfer of the deep English trees to the deep Czech trees to allow for the appropriate capitalization of the Czech translation.

4.2 Transfer

The transfer phase consists of the following steps:

- Initiate the target-side (Czech) t-trees simply by "cloning" the source-side (English) t-trees. Subsequent steps usually iterate over all t-nodes. In the following, we denote a source-side t-node as *S* and the corresponding target-side node as *T*.

- Translate formemes using two probabilistic dictionaries ($p(T.formeme|S.formeme, S.parent.lemma)$ and $p(T.formeme|S.formeme)$) and a few manual rules. The formeme translation probability estimates were extracted from a part of the parallel data mentioned above.

- Translate lemmas using a probabilistic dictionary ($p(T.lemma|S.lemma)$) and a few rules that ensure compatibility with the previously chosen formeme. Again, this probabilistic dictionary was obtained using the aligned tectogrammatical trees from the parallel corpus.

- Fill the grammatemes (deep-syntactic equivalent of morphological categories) *gender* (for denotative nouns) and *aspect* (for verbs) according to the chosen lemma. We also fix grammateme values where the English-Czech grammateme correspondence is non-trivial (e.g. if an English gerund expression is translated to Czech as a subordinating clause, the *tense* grammateme has to be filled). However, the transfer of grammatemes is definitely much easier task than the transfer of formemes and lemmas.

4.3 Synthesis

The transfer step yields an abstract deep syntactico-semantical tree structure. Firstly,

we derive surface morphological categories from their deep counterparts taking care of their agreement where appropriate and we also remove personal pronouns in subject positions (because Czech is a pro-drop language).

To arrive at the surface tree structure, auxiliary nodes of several types are added, including (1) reflexive particles, (2) prepositions, (3) subordinating conjunctions, (4) modal verbs, (5) verbal auxiliaries, and (6) punctuation nodes. Also, grammar-based node ordering changes (implemented by rules) are performed: e.g. if an English possessive attribute is translated using Czech genitive, it is shifted into post-modification position.

After finishing the inflection of nouns, verbs, adjectives and adverbs (according to the values of morphological categories derived from agreement etc.), prepositions may need to be vocalized: the vowel *-e* or *-u* is attached to the preposition if the pronunciation of prepositional group would be difficult otherwise.

After the capitalization of the beginning of each sentence (and each named entity instance), we obtain the final translation by flattening the surface tree.

4.4 Preliminary Error Analysis

According to our observations most errors happen during the transfer of lemmas and formemes. Usually, there are acceptable translations of lemma and formeme in respective n-best lists but we fail to choose the best one. The scenario described in Section 4.2 uses quite a primitive transfer algorithm where formemes and lemmas are translated separately in two steps. We hope that big improvements could be achieved with more sophisticated algorithms (optimizing the probability of the whole tree) and smoothed probabilistic models (such as $p(\textit{T.lemma}|\textit{S.lemma}, \textit{T.parent.lemma})$ and $p(\textit{T.formeme}|\textit{S.formeme}, \textit{T.lemma}, \textit{T.parent.lemma})$).

Other common errors include:

- Analysis: parsing (especially coordinations are problematic with McDonald's parser).
- Transfer: the translation of idioms and collocations, including named entities. In these cases, the classical transfer at the t-layer is not appropriate and utilization of some phrase-based MT would help.
- Synthesis: reflexive particles, word order.

5 Experimental Results and Discussion

Table 3 reports lowercase BLEU and NIST scores and preliminary manual ranks of our submissions in contrast with other systems participating in English→Czech translation, as evaluated on the official WMT09 unseen test set. Note that automatic metrics are known to correlate quite poorly with human judgements, see the best ranking but "lower scoring" PC Translator this year and also in Callison-Burch et al. (2008).

System	BLEU	NIST	Rank
Moses T	**14.24**	**5.175**	-3.02 (4)
Moses T+C	13.86	5.110	–
Google	13.59	4.964	-2.82 (3)
U. of Edinburgh	13.55	5.039	-3.24 (5)
Moses T+C+C&T+T+G 84k	10.01	4.360	-
Eurotran XP	09.51	4.381	-2.81 (2)
PC Translator	09.42	4.335	**-2.77 (1)**
TectoMT	07.29	4.173	-3.35 (6)

Table 3: Automatic scores and preliminary human rank for English→Czech translation. Systems in italics are provided for comparison only. Best results in bold.

Unfortunately, this preliminary evaluation suggests that simpler models perform better, partly because it is easier to tune them properly both from computational point of view (e.g. MERT not stable and prone to overfitting with more features[7]), as well as from software engineering point of view (debugging of complex pipelines of tools is demanding). Moreover, simpler models run faster: "Moses T" with 12 sents/minute is 4.6 times faster than "Moses T+C". (Note that we have not tuned either of the models for speed.)

While "Moses T" is probably nearly identical setup as Google and Univ. of Edinburgh use, the knowledge of correct language-dependent tokenization and the use of relatively high quality large language model data seems to bring moderate improvements.

6 Conclusion

We described our experiments with a complex linguistically motivated translation system and various (again linguistically-motivated) setups of factored phrase-based translation. An automatic evaluation seems to suggest that simpler is better, but we are well aware that a reliable judgement comes only from human annotators.

[7] For "Moses T+C+C&T+T+G", we observed BLEU scores on the test set varying by up to five points absolute for various weight settings yielding nearly identical dev set scores.

References

Ondřej Bojar and Jan Hajič. 2008. Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 143–146, Columbus, Ohio, June. Association for Computational Linguistics.

Ondřej Bojar, Miroslav Janíček, Zdeněk Žabokrtský, Pavel Češka, and Peter Beňa. 2008. CzEng 0.7: Parallel Corpus with Community-Supplied Translations. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. ELRA.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, June. Association for Computational Linguistics.

Michael Collins. 1999. *Head-driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, Morristown, NJ, USA. Association for Computational Linguistics.

Jan Hajič et al. 2006. Prague Dependency Treebank 2.0. CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T0 1, Philadelphia.

Václav Klimeš. 2006. *Analytical and Tectogrammatical Analysis of a Natural Language*. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University, Prague, Czech Rep.

Jan Kocek, Marie Kopřivová, and Karel Kučera, editors. 2000. *Český národní korpus - úvod a příručka uživatele*. FF UK - ÚČNK, Praha.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

David Mareček, Zdeněk Žabokrtský, and Václav Novák. 2008. Automatic Alignment of Czech and English Deep Syntactic Dependency Trees. In *Proceedings of European Machine Translation Conference (EAMT 08)*, pages 102–111, Hamburg, Germany.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver, British Columbia, Canada.

Václav Novák and Zdeněk Žabokrtský. 2007. Feature engineering in maximum spanning tree dependency parser. In Václav Matoušek and Pavel Mautner, editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 10th I nternational Conference on Text, Speech and Dialogue*, Lecture Notes in Computer Science, pages 92–98, Pilsen, Czech Republic. Springer Science+Business Media Deutschland GmbH.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. The best of two worlds: Cooperation of statistical and rule-based taggers for czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*, pages 67–74, Praha.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. TectoMT: Highly Modular Hybrid MT System with Tectogrammatics Used as Transfer Layer. In *Proc. of the ACL Workshop on Statistical Machine Translation*, pages 167–170, Columbus, Ohio, USA.