TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer[*]

Zdeněk Žabokrtský, Jan Ptáček, Petr Pajas
Institute of Formal and Applied Linguistics
Charles University, Prague, Czech Republic
{zabokrtsky,ptacek,pajas}@ufal.mff.cuni.cz

Abstract

We present a new English→Czech machine translation system combining linguistically motivated layers of language description (as defined in the Prague Dependency Treebank annotation scenario) with statistical NLP approaches.

1 Introduction

We describe a new MT system (called TectoMT) based on the conventional analysis-transfer-synthesis architecture. We use the layers of language description defined in the Prague Dependency Treebank 2.0 (PDT for short, (Hajič and others, 2006)), namely (1) *word layer* – raw text, no linguistic annotation, (2) *morphological layer* – sequence of tagged and lemmatized tokens, (3) *analytical layer* – each sentence represented as a surface-syntactic dependency tree, and (4) *tectogrammatical layer* – each sentence represented as a deep-syntactic dependency tree in which only autosemantic words do have nodes of their own; prefixes w-, m-, a-, or t- will be used for denoting these layers.[1]

We use 'Praguian' tectogrammatics (introduced in (Sgall, 1967)) as the transfer layer because we believe that, first, it largely abstracts from language-specific (inflection, agglutination, functional words…) means of expressing non-lexical meanings, second, it allows for a natural transfer factorization, and third, local tree contexts in t-trees carry more information (esp. for lexical choice) than local linear contexts in the original sentences.

In order to facilitate separating the transfer of lexicalization from the transfer of syntactization, we introduce the concept of *formeme*. Each t-node's has a formeme attribute capturing which morphosyntactic form has been (in the case of analysis) or will be (synthesis) used for the t-node in the surface sentence shape. Here are some examples of formemes we use for English: n:subj (semantic noun (sn) in subject position), n:for+X (sn with preposition *for*), n:X+ago (sn with postposition *ago*), n:poss (possessive form of sn), v:because+fin (semantic verb (sv) as a subordinating finite clause introduced by *because*), v:without+ger (sv as a gerund after *without*), adj:attr (semantic adjective (sa) in attributive position), adj:compl (sa in complement position).

The presented system intensively uses the PDT technology (data formats, software tools). Special attention is paid to modularity: the translation is implemented (in Perl) as a long sequence of processing modules (called blocks) with relatively tiny, well-defined tasks, so that each module is independently testable, improvable, or substitutable. TectoMT allows to easily combine blocks based on different approaches, from blocks using complex probabilistic solutions (e.g., B2, B6, B35, see the next section), through blocks applying simpler Machine Learning techniques (e.g., B69) or empirically based heuristics (e.g., B7, B25, B36, B71), to blocks implementing 'crisp' linguistic rules (e.g., B48-B51, B59). There are also blocks for trivial technical tasks (e.g., B33, B72).

[*] The research reported in this paper is financially supported by grants GAAV ČR 1ET101120503, MSM0021620838, GAUK 52408/2008 and 7643/2007.

[1] In addition, we use also p-layer (phrase structures) as an a-layer alternative, the only reason for which is that we do not have a working a-layer parser for English at this moment.

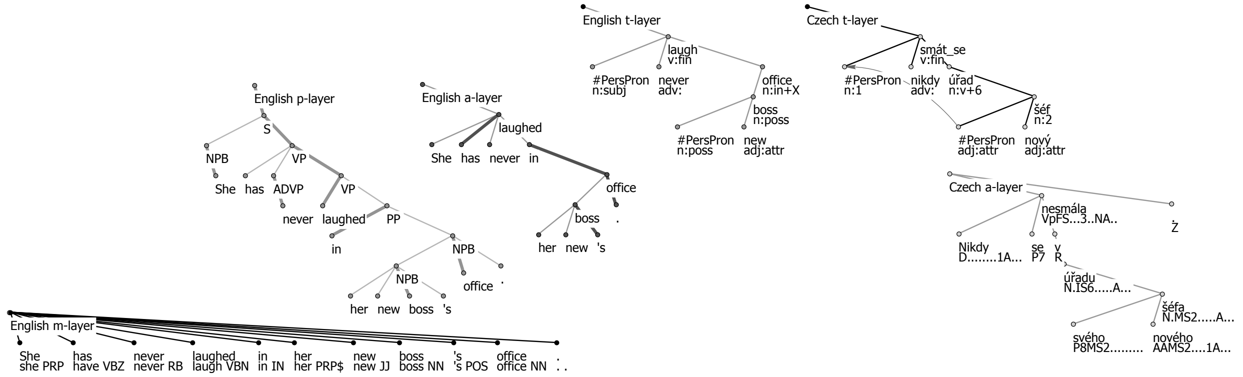

Figure 1: MT 'pyramid' as implemented in TectoMT. All the representations are rooted with artificial nodes, serving only as labels. Virtually, the pyramid is bottomed with the input sentence on the source side (*She has never laughed in her new boss's office.*) and its automatic translation on the target side (*Nikdy se nesmála v úřadu svého nového šéfa.*).

2 Translation Procedure

The structure of this section directly renders the sequence of blocks currently used for English-Czech translation in TectoMT. The intermediate stages of the translation process are illustrated in Figure 1; identifiers of the blocks affecting on the translation of the sample sentence are typeset in bold.

2.1 From English w-layer to English m-layer

B1: Segment the source English text into sentences. **B2**: Split the sentences into sequences of tokens, roughly according to Penn Treebank (PTB for short; (Marcus et al., 1994)) conventions. **B3**: Tag the tokens with PTB-style POS tags using a tagger (Brants, 2000). B4: Fix some tagging errors systematically made by the tagger using a rule-based corrector. **B5**: Lemmatize the tokens using `morpha`, (Minnen et al., 2000).

2.2 From English m-layer to English p-layer

B6: Build PTB-style phrase-structure tree for each sentence using a parser (Collins, 1999).

2.3 From English p-layer to English a-layer

B7: In each phrase, mark the head node (using a set of heuristic rules). **B8**: Convert phrase-structure trees to a-trees. B9: Apply some heuristic rules to fix apposition constructions. B10: Apply another heuristic rules for reattaching incorrectly positioned nodes. B11: Unify the way in which multiword prepositions (such as *because of*) and subordinating conjunctions (such as *provided that*) are treated. B12: Assign analytical functions (only if necessary for a correct treatment of coordination/apposition constructions).

2.4 From English a-layer to English t-layer

B13: Mark a-nodes which are auxiliary (such as prepositions, subordinating conjunctions, auxiliary verbs, selected types of particles, etc.) **B14**: Mark *not* as an auxiliary node too (but only if it is connected to a verb form). **B15**: Build t-trees. Each a-node cluster formed by an autosemantic node and possibly several associated auxiliary nodes is 'collapsed' into a single t-node. T-tree dependency edges are derived from a-tree edges connecting the a-node clusters. B16: Explicitely distinguish t-nodes that are members of coordination (conjuncts) from shared modifiers. It is necessary as they all are attached below the coordination conjunction t-node. **B17**: Modify t-lemmas in specific cases. E.g., all kinds of personal pronouns are represented by the 'artificial' t-lemma #PersPron. B18: Assign functors that are necessary for proper treatment of coordination and apposition constructions. B19: Distribute shared auxiliary words in coordination constructions. **B20**: Mark t-nodes that are roots of t-subtrees corresponding to finite verb clauses. B21: Mark passive verb forms. B22: Assign (a subset of) functors. B23: Mark t-nodes corresponding to infinitive verbs. B24: Mark t-nodes which are roots of t-subtrees corresponding to relative clauses. B25: Identify coreference links between relative pronouns (or other relative pronominal word) and their nominal antecedents. B26: Mark t-nodes that are the roots of t-subtrees corresponding to direct speeches. B27: Mark t-nodes that are the roots of t-subtrees corresponding to parenthesized expressions. **B28**: Fill the nodetype attribute (rough classification of t-nodes). **B29**: Fill the sempos attribute (fine-grained classification of t-nodes). **B30**: Fill the grammateme attributes (semantically indispensable morphological categories, such as number for nouns, tense for verbs). **B31**: Determine the formeme of each t-node. B32: Mark personal names, distinguish male and female first names if possible.

2.5 From English t-layer to Czech t-layer

B33: Initiate the target-side t-trees, simply by cloning the source-side t-trees. **B34**: In each t-node, translate its formeme.[2] **B35**: Translate t-lemma in each t-node as its most probable target-language counterpart (which is compliant with the previously chosen formeme), according to a probabilistic dictionary.[3] B36: Apply manual rules for fixing the formeme and lexeme choices, which are otherwise systematically wrong and are reasonably frequent. **B37**: Fill the gender grammateme in t-nodes corresponding to denotative nouns (it follows from the chosen t-lemma).[4] **B38**: Fill the aspect grammateme in t-nodes corresponding to verbs. Information about aspect (perfective/imperfective) is necessary for making decisions about forming complex future tense in Czech. B39: Apply rule-based correction of translated date/time expressions (several templates such as *1970's*, *July 1*, etc.). B40: Fix grammateme values in places where the English-Czech grammateme correspondence is not trivial (e.g., if an English gerund expression is translated using Czech subordinating clause, the tense grammateme has to be filled). **B41**: Negate verb forms where some arguments of the verbs bear negative meaning (double negation in Czech). B42: Verb t-nodes in active voice that have transitive t-lemma and no accusative object, are turned to reflexives. **B43**: The t-nodes with genitive formeme or prepositional-group formeme, whose counterpart English t-nodes are located in pre-modification position, are moved to post-modification position. B44: Reverse the dependency orientation between numeric expressions and counted nouns, if the value of the numeric expression is greater than four and the noun without the numeral would be expressed in nominative or accusative case. **B45**: Find coreference links from personal pronouns to their antecedents, if the latter are in subject position (needed later for reflexivization).

2.6 From Czech t-layer to Czech a-layer

B46: Create initial a-trees by cloning t-trees. **B47**: Fill the surface morphological categories (gender, number, case, negation, etc.) with values derived from values of grammatemes, formeme, semantic part of speech etc. B48: Propagate the values of gender and number of relative pronouns from their antecedents (along the coreference links). **B49**: Propagate the values of gender, number and person according to the subject-predicate agreement (i.e., from subjects to the finite verbs). **B50**: Resolve agreement of adjectivals in attributive positions (copying gender/number/case from their governing nouns). B51: Resolve complement agreement (copying gender/number from subject to adjectival complement). **B52**: Apply pro-drop – deletion of personal pronouns in subject positions. **B53**: Add preposition a-nodes (if implied by the t-node's formeme). B54: Add a-nodes for subordinating conjunction (if implied by the t-node's formeme). **B55**: Add a-nodes corresponding to reflexive particles for reflexiva tantum verbs. B56: Add an a-node representing the auxiliary verb *být* (to be) in the case of compound passive verb forms. B57: Add a-nodes representing modal verbs, accordingly to the deontic modality grammateme. B58: Add the auxiliary verb *být* in imperfective future-tense complex verb forms. B59: Add verb forms such as *by*/*bys*/*bychom* expressing conditional verb modality. B60: Add auxiliary verb forms such as *jsem*/*jste* in past-tense complex verb forms. B61:

[2] The translation mapping from English formemes to Czech formemes was obtained as follows: we analyzed 10,000 sentence pairs from the WMT'08 training data up to the t-layer (using a tagger shipped with the PDT and parser (McDonald et al., 2005) for Czech), added formemes to t-trees on both sides, aligned the t-trees (using a set of weighted heuristic rules, similarly to (Menezes and Richardson, 2001)), and from the aligned t-node pairs extracted for each English formeme its most frequent Czech counterpart.

[3] The dictionary was created by merging the translation dictionary from PCEDT ((Cuřín and others, 2004)) and a translation dictionary extracted from a part of the parallel corpus Czeng ((Bojar and Žabokrtský, 2006)) aligned at word-level by Giza++ ((Och and Ney, 2003)).

[4] Czech nouns have grammatical gender which is (among others) important for resolving grammatical agreement.

Partition a-trees into finite clauses (a-nodes belonging to the same clause are coindexed). **B62**: In each clause, a-nodes which represent clitics are moved to the so called second position in the clause (according to Wackernagel's law). **B63**: Add a-nodes corresponding to sentence-final punctuation mark. B64: Add a-nodes corresponding to commas on boundaries between governing and subordinated clauses. B65: Add a-nodes corresponding to commas in front of conjunction *ale* and also commas in multiple coordinations. B66: Add pairs of parenthesis a-nodes. **B67**: Choose morphological lemmas in a-nodes corresponding to personal pronouns. **B68**: Generate the resulting word forms (derived from lemmas and tags) using Czech word form generator described in (Hajič, 2004). B69: Vocalize prepositions *k*, *s*, *v*, and *z* (accordingly to the prefix of the following word). **B70**: Capitalize the first word in each sentence as well as in each direct speech.

2.7 From Czech a-layer to Czech w-layer

B71: Create the resulting sentences by flattening the a-trees. Heuristic rules for proper spacing around punctuation marks are used. B72: Create the resulting text by concatenating the resulting sentences.

3 Final remarks

We believe that the potential contribution of tectogrammatical layer of language representation for MT is the following: it abstracts from many language-specific phenomena (which could reduce the notorious data-sparsity problem) and offers a natural factorization of the translation task (which could be useful for formulating independence assumptions when building probabilistic models). Of course, the question naturally arises whether these properties can ever outbalance the disadvantages, especially cumulation and interference of errors made on different layers, considerable technical complexity, and the need for detailed linguistic insight. In our opinion, this question still remains open. On one hand, the translation quality offered now by TectoMT is below the state-of-the-art system according to the preliminary evaluation of the WMT08 Shared Task. But on the other hand, the potential of tectogrammatics has not been used fully, and moreover there are still many components with only pilot heuristic implementation which increase the number of translation errors and which can be relatively easily substituted by corpus-based solutions. In the near future, we plan to focus especially on the transfer blocks, which are currently based on the naive assumption of isomorphism of the source and target t-trees and which do not make use of the target language model, so far.

References

Ondřej Bojar and Zdeněk Žabokrtský. 2006. CzEng: Czech-English Parallel Corpus, Release version 0.5. *Prague Bulletin of Mathematical Linguistics*, 86:59–62.

Thorsten Brants. 2000. TnT - A Statistical Part-of-Speech Tagger . pages 224–231, Seattle.

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

Jan Cuřín et al. 2004. Prague Czech - English Dependency Treebank, Version 1.0. CD-ROM, Linguistics Data Consortium, LDC Catalog No.: LDC2004T25, Philadelphia.

Jan Hajič et al. 2006. Prague Dependency Treebank 2.0. CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia.

Jan Hajič. 2004. *Disambiguation of Rich Inflection – Computational Morphology of Czech*. Charles University – The Karolinum Press, Prague.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of HTL/EMNLP*, pages 523–530, Vancouver, Canada.

Arul Menezes and Stephen D. Richardson. 2001. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the workshop on Data-driven methods in machine translation*, volume 14, pages 1–8.

Guido Minnen, John Carroll, and Darren Pearce. 2000. Robust Applied Morphological Generation. In *Proceedings of the 1st International Natural Language Generation Conference*, pages 201–208, Israel.

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

Petr Sgall. 1967. *Generativní popis jazyka a česká deklinace*. Academia, Prague.