SMT and Hybrid systems of the QTLeap project in the WMT16 IT-task

Rosa Del Gaudio
Higher Functions Sistemas Inteligentes, Lisbon, Portugal
rosa.gaudio@pcmedic.pt

Gorka Labaka, Eneko Agirre
University of the Basque Country, UPV/EHU, San Sebastian, Spain
{gorka.labaka,e.agirre}@ehu.eus

Petya Osenova, Kiril Simov
Linguistic Modelling Department, IICT-BAS, Sofia, Bulgaria
{petya,kivs}@bultreebank.org

Martin Popel
Charles University in Prague, Faculty of Mathematics and Physics, ÚFAL, Czechia
popel@ufal.mff.cuni.cz

Dieke Oele, Gertjan van Noord
Rijksuniversiteit Groningen, Groningen, The Netherlands
{d.oele,g.j.van.noord}@rug.nl

Luís Gomes, João Rodrigues, Steven Neale, João Silva,
Andreia Querido, Nuno Rendeiro, António Branco
Universidade de Lisboa, Departamento de Informática, Faculdade de Ciências
luisgomes@gmail.com,{joao.rodrigues,steven.neale,jsilva,
andreia.querido,nuno.rendeiro,antonio.branco}@di.fc.ul.pt

Abstract

This paper presents the description of 12 systems submitted to the WMT16 IT-task, covering six different languages, namely Basque, Bulgarian, Dutch, Czech, Portuguese and Spanish. All these systems were developed under the scope of the QTLeap project, presenting a common strategy. For each language two different systems were submitted, namely a phrase-based MT system built using Moses, and a system exploiting deep language engineering approaches, that in all the languages but Bulgarian was implemented using TectoMT. For 4 of the 6 languages, the TectoMT-based system performs better than the Moses-based one.

1 Introduction

The QTLeap[1] project focuses on the development of an articulated methodology for machine translation that explores deep language engineering approaches and sophisticated semantic datasets. The underling hypothesis is that the deeper the level of representation, the better the translation becomes since deeper representations abstract away from surface aspects that are specific to a given language. At the limit, the representation of the meaning of a sentence, and of all its paraphrases, would be shared among all languages.

This purpose is supported by recent advances in terms of lexical processing. These advances have been made possible by enhanced techniques for referential and conceptual ambiguity resolution, and supported also by new types of datasets recently developed as linked open data.

The overall goal of the project is to produce quality translation between English (EN) and another language *X* by using deep linguistic information. All language pairs follow the same processing pipeline of analysis, transfer and synthesis (generation) and adopt the same hybrid MT approach of using both statistical as well as rule-based components in a tightly integrated way for the best possible results.

In this paper, we present the systems developed by the University of Basque Country for Basque

[1] http://www.qtleap.eu

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 435–441,
Berlin, Germany, August 11-12, 2016.

and Spanish, Charles University in Prague for Czech, by University of Groningen for Dutch, by University of Lisbon for Portuguese and by IICT-BAS of the Bulgarian Academy of Sciences for Bulgarian.

For each language two different systems were submitted, corresponding to different phases of the project, namely a phrase-based MT system built using Moses (Koehn et al., 2007), and a system exploiting deep language engineering approaches, that in all the languages but Bulgarian was implemented using TectoMT (Žabokrtský and Popel, 2009). For Bulgarian, its second MT system is not based on TectoMT, but on exploiting deep factors in Moses. All 12 systems are constrained, that is trained only on the data provided by the WMT16 IT-task organizers.

We present briefly the Moses common setting and the TectoMT structure and then more detailed information for each language system are provided. In the last Section, results based on BLEU and TrueSkill are given and discussed.

2 Moses

All the systems submitted that were based on Moses have been trained on a phrase-based model by Giza++ or mGiza with "grow-diag-final-and" symmetrization and "msd-bidirectional-fe" reordering (Koehn et al., 2003). For the language pairs where big quantities of domain-specific monolingual data were available along with the generic domain data, separate language models (domain-specific and generic) were interpolated against our ICT domain-specific development set. For LM training and interpolation, the SRILM toolkit (Stolcke, 2002) was used. The method of truecasing has been adopted for several language pairs where it proved useful.

3 TectoMT

The deep translation is based on the TectoMT system, an open-source MT system based on the Treex platform for general natural-language processing. TectoMT uses a combination of rule-based and statistical (trained) modules (blocks in Treex terminology), with a statistical transfer based on HMTM (Hidden Markov Tree Model) at the level of a deep, so-called tectogrammatical, representation of sentence structure. The general TectoMT pipeline is language independent, and consists of analysis, deep transfer, and synthesis steps.

The design of TectoMT is highly modular and consists of a language-universal core and language-specific additions and distinguishes two levels of syntactic description:

- Surface dependency syntax (a-layer) – surface dependency trees containing all the tokens in the sentence.
- Deep syntax (t-layer) – dependency trees that contain only content words (nouns, main verbs, adjectives, adverbs) as nodes. Each node has a deep lemma (t-lemma), a semantic function label (functor), a morpho-syntactic form label (formeme), and various grammatical attributes (grammatemes), such as number, gender, tense, or modality.

T-layer representations of the same sentence in different languages are closer to each other than the surface texts; in many cases, there is a 1:1 node correspondence among the t-layer trees. TectoMTs transfer exploits this by translating the tree isomorphically, i.e., node-by-node and assuming that the shape will not change in most cases (apart from a few exceptions handled by specific rules).

The translation is further factorized: t-lemmas, formemes, and grammatemes are translated using separate Translation Models (TM). The t-lemma and formeme TMs are an interpolation of maximum entropy discriminative models (MaxEnt) (Mareček et al., 2010) and simple conditional probability models. The MaxEnt models are in fact an ensemble of models, one for each individual source t-lemma/formeme. The combined translation models provide several translation options for each node along with their estimated probability. The best options are then selected using a Hidden Markov Tree Model (HMTM) with a target-language tree model (Žabokrtský and Popel, 2009).

For this specific task, where we need to work on a specific domain, an extended version of TectoMT was used allowing interpolation of multiple TMs (Rosa et al., 2015).

4 Basque

Both English-Basque submissions are trained on the same training corpora. That is, the PaCO2-eneu corpus for translation and language modeling, and the in-domain Batch1 corpus for domain

adaptation and MERT training. Batch2 domain corpus was used for testing during development.

The Moses system, `EU-Moses`, uses factored models to allow lemma-based word-alignment. After word alignment, the rest of the training process is based on lowercased word-forms and standard parameters: Stanford CoreNLP (Manning et al., 2014) and Eustagger (Alegria et al., 2002) tools are used for tokenization and lemmatization, MGIZA for word alignment with the "grow-diag-final-and" symmetrization heuristic, a maximum length of 75 tokens per sentence and 5 tokens per phrase, translation probabilities in both directions, lexical weightings in both directions, a phrase length penalty, a "phrase-mslr-fe" lexicalized reordering model and a target language model. As for the language model, a 5-gram model was trained. The weights for the different components were adjusted to optimize BLEU using MERT tuning over the Batch1 development set, with an n-best list of size 100.

For the TectoMT system, `EU-Treex` existing tools were used in order to get the a-layer. Eustagger is a robust and wide coverage morphological analyzer and POS tagger. The dependency parser is based on the MATE-tools (Bjrkelund et al., 2010). Basque models have been trained using the Basque Dependency Treebank (BDT) corpus (Aduriz et al., 2003). Transformation from the a-level analysis into t-level is partially performed with language-independent blocks thanks to the support of Interset (Zeman, 2008).

The English-to-Basque TectoMT system uses the PaCo2 and the Batch1 corpora to train two separate translation models, and they are used to create an interpolated list of translation candidates. In addition to that, the terminological equivalences extracted from the localization PO files (VLC, LO and KDE) as well as the domain terms extracted from Wikipedia are used to identify domain terms before syntactical analysis and to ensure domain translation on transfer. Finally, an extra module to treat non linguistic elements (URLs, shell commands, ...) has been used, to identify the elements that should be maintained untranslated on the output.

5 Bulgarian

Bulgarian team participated with two systems implemented using Moses: `BG-Moses` — a system that is based on standard factored Moses with factors retrieved from POS tagged, lemmatized parallel corpora; and `BG-DeepMoses` — a system that also is based on standard factored Moses but the translation is done in two steps: (1) semantics-based translation of the source language text to a mixed source-target language text which is then (2) translated to the target language via Moses. The latter system builds on Simov et al. (2015).

As training data for both systems the following corpora were used: 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[2] pipeline for the English part and the BTB pipeline for the Bulgarian. The analyses include POS tagging, lemmatization and WSD, using the UKB system,[3] which provides graph-based methods for Word Sense Disambiguation and lexical similarity measurements.

For the `BG-Moses` system, the following factors have been constructed: WordForm|Lemma|POStag.

For the `BG-DeepMoses` system, we exploited also the information from word sense annotation in order to predict some translations from English to Bulgarian based on the WordNet synsets and their mappings to the Bulgarian WordNet. Thus, we replaced the English word form with a *representative lemma* in Bulgarian. The motivation for using representative lemmas in Bulgarian is as follows: we aim at unifying the various synsets with similar translations in the Bulgarian language. After the creation of this intermediate English/Bulgarian text, we trained Moses with the following factors: ENWordForm-BGLemma|Lemma|BGPOStag, where ENWordForm-BGLemma is an English word form when there is no appropriate Bulgarian one, or the Bulgarian lemma; BGPOStag is the appropriate Bulgarian tag representing grammatical features like number, tense, etc.

6 Czech

The Czech Moses system follows the CU-Bojar system (Bojar et al., 2013). A factored phrase-based model was trained based on truecased forms translated directly to the pair <truecased form, morphological tag>. There were three LMs for Czech:

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

[3] http://ixa2.si.ehu.es/ukb/

- 8grams of morphological tags from the monolingual part of news and political corpora,
- 6grams of forms from the monolingual part of news and political corpora and
- 6grams from the Czech side of a bilingual Czech-English corpus CzEng.

The pre-processing of this SMT system has been harmonized with the pre-existing version of Tecto-MT: Tokenization and lemmatization is handled by Treex followed by further tokenization at any letter-digit-punctuation boundary. Additionally, casing is handled by a Czech-specific supervised truecasing method. The output of the lemmatizer is used, as names have lemmas capitalized, the casing of the lemma is cast to the token (lowercasing non-names at sentence beginnings, lowercasing also ALL CAPS if correctly lemmatized). Finally, the translation is done using case-sensitive tokens and finally the first letter in every sentence is only capitalized.

The TectoMT analysis pipeline is based on the annotation pipeline of the CzEng 1.0 corpus (Bojar et al., 2012) starting with a rule-based tokenizer and a statistical part-of-speech tagger (Straková et al., 2014) and dependency parser (McDonald et al., 2005; Novák and Žabokrtský, 2007). These steps result in a-layer trees, which are then converted to t-layer using a rule-based process.

The English-to-Czech transfer uses a combination of translation models and tree model reranking. The Czech synthesis pipeline has remained basically unchanged since the original TectoMT system (Žabokrtský et al., 2008).

7 Dutch

The Moses system for Dutch was trained on the third version of the Europarl corpus (Koehn, 2005) and the in-domain KDE4 Localization data (Tiedemann, 2012). Words are aligned with GIZA++ and tuning was done with MERT. The applied heuristics for the Dutch baselines were set to "grow-diag-final-and" alignment and "msd-bidirectional-fe" reordering. For the creation of the language models, IRSTLM was used to train a 5-gram language model with Kneser-Ney smoothing on the monolingual part of the training corpora.

For the TectoMT system, the analysis of Dutch input uses the Alpino system (Noord, 2006), a stochastic attribute value grammar. The transfer uses discriminative (context-sensitive) and dictionary translation models. In addition, a few rule-based modules are employed that handle changes in t-tree topology and Dutch grammatical gender.

The Dutch synthesis pipeline includes morphology initialization and agreements (subject-predicate and attribute-noun), insertion of prepositions and conjunctions based on formemes, and insertion of punctuation, possessive pronouns and Dutch pronominal adverbs. The t-tree resulting from the transfer phase is first converted into an Abstract Dependency Tree (ADT) using rule-based modules implemented in Treex. The ADT is then passed to the Alpino generator (de Kok and Noord, 2010), which handles the creation of the actual sentence including inflected word forms.

8 Spanish

The Moses system developed for the translation from English to Spanish, `ES-Moses`, uses standard parameters: tokenization and truecasing using tools available in Moses toolkit, MGIZA for word alignment with the "grow-diag-final-and" symmetrization heuristic, a maximum length of 80 tokens per sentence and 5 tokens per phrase, translation probabilities in both directions with Good–Turing discounting, lexical weightings in both directions, a phrase length penalty, an "msd-bidirectional-fe" lexicalized reordering model and a 5-gram target language model. The weights for the different components were adjusted to optimize BLEU using MERT tuning over the Batch1 development set, with an n-best list of size 100.

The English-to-Spanish TectoMT, `ES-Treex`, system uses the Europarl and the Batch1 corpora to train two separate translation models, and these were used to create an interpolated list of translation candidates. In addition to that, the terminological equivalences extracted from the localization PO files (VLC, LO and KDE) as well as the domain terms extracted from Wikipedia are used to identify domain terms before syntactic analysis and to ensure domain translation on transfer. Finally, an extra module to treat non linguistic elements (URLs, shell commands, ...) has been used to identify the elements that should be maintained untranslated on the output.

Both systems were trained using the same training corpora: the 7th version of the Europarl corpus was used for both translation and language mod-

eling, and the in-domain batch1 corpus was used for domain adaptation and MERT training. The Batch2 domain-specific corpus was used for testing during development. We have not used all the available parallel corpora, because of the computational restrictions in analyzing all those corpora at the tectogrammatical level of the TectoMT system.

9 Portuguese

The Moses system for the translation from English to Portuguese, `PT-Moses`, was obtained by using the default parameters and tools regarding the training of a phrase-based model. For the pre-processing, a sentence length of 80 words was used and the tokenization was performed by the Moses tokenizer. No lemmatization or compound splitting was used and the casing was obtained with the Moses truecaser. For the training, a phrase-based model was used with a language model order of 5, with Kneser-Ney smoothing, which was interpolated using the SRILM tool. The word alignment was done with Giza++ on full forms and the final tuning was done using MERT. The Europarl corpus was used for the training data, both as monolingual data for training language models and as parallel data for training the phrase-table.

Regarding the English-to-Portuguese TectoMT system (Silva et al., 2015)(Rodrigues et al., 2016a), `PT-Treex`, in order to get the a-layer the Portuguese system resorted to LX-Suite (Branco and Silva, 2006), a set of pre-existing shallow processing tools for Portuguese that include a sentence segmenter, a tokenizer, a POS tagger, a morphological analyser and a dependency parser, all with state-of-the-art performance. Treex blocks were created to be called and interfaced with these tools.

After running the shallow processing tools, the dependency output of the parser is converted into Universal Dependencies (UD) (de Marneffe et al., 2014). These dependencies are then converted into the a-layer tree (a-tree) in a second step. Both steps are implemented as rule-based Treex blocks. Converting the a-tree into a t-layer tree (t-tree) is done through rule-based Treex blocks that manipulate the tree structure.

The transfer phase is handled by a tree-to-tree maximum entropy translation model (Mareček et al., 2010) working at the deep syntactic level of tectogrammatical trees. Two separate models were trained and interpolated, the first model with over 1.9 million sentences from Europarl (Koehn, 2005) and the second model composed of the Batch1, the Microsoft Terminology Collection and the LibreOffice localization data (Štajner et al., 2016). Each pair of parallel sentences, one in English and one in Portuguese, are analyzed by Treex up to the t-layer level, where each pair of trees are fed into the model.

The TectoMT synthesis (Rodrigues et al., 2016b) included other two lexical-semantics-related modules, the HideIT and gazetteers. The HideIT module handles entities that do not require translation such as URLs and shell commands. The gazetteers are specialized lexicons that handle the translation of named entities from the IT-domain such as menu items and button names.

Finally, synset IDs were used as additional contextual features in the lemma-to-lemma Discriminative Translation Models (Neale et al., 2016).

10 Results

Table 1 presents the results of automatic and manual evaluation, based on BLEU and TrueSkill[4] scores respectively. For 4 of the 6 languages, the TectoMT-based system performs better than the Moses-based one when considering both BLEU and TrueSkill scores. For Bulgarian, the `BG-DeepFMoses` performs worse than the `BG-FMoses` on both scores. For Dutch, the Moses system outperforms the TectoMT only when considering the BLUE score, but not the TrueSkill score.

Regarding Bulgarian, although `BG-Deep-FMoses` system performed worse than `BG-Moses`, the automatic conversion of the source text into near-target language text represents a promising direction for further improvement of the English-to-Bulgarian MT system. We assume that the current drop might be overcome by improving the WordNet information for Bulgarian, its mapping to the English WordNet as well as the processing pipelines. Also, we plan to train this system on more data and to exploit other bilingual dictionaries.

For the English→Dutch translation direction, the Moses system outperforms TectoMT in terms of BLEU score. The results of the manual evaluation, however, are in favor of the TectoMT sys-

[4]For details, see the overview paper in these proceedings.

	Moses		TectoMT		Deep-Moses	
Language	BLEU	TrueSkill	BLEU	TrueSkill	BLEU	TrueSkill
Basque	8.3	−1.570	10.3	1.570		
Bulgarian	16.6	5.262	-	-	15.3	−5.262
Czech	20.8	−0.616	21.5	0.130		
Dutch	21.9	−2.462	19.0	0.154		
Spanish	16.0	−1.926	24.2	−0.809		
Portuguese	13.7	−2.276	15.2	−1.063		

Table 1: Automatic and manual evaluation results.

tem. This difference may in part be caused by the fact that BLEU only scores exact word or phrase matches and the TectoMT output shows more lexical flexibility as compared to Moses. We get better results, in terms of BLEU-score, in the opposite translation direction which indicates that more effort should be put into this translation direction. Our focus here lies on the Dutch synthesis pipeline where we still need to fix some basic errors. Also we intend to implement more modules that are based on lexical semantics.

We also presented at the IT-task a third system for Czech, Dutch, Spanish and Portuguese, called Chimera that combines Moses and TectoMT (Rosa et al., 2016).

Acknowledgments

This work has been supported by the 7th Framework Programme of the EU grant QTLeap (No. 610516).

References

Itzair Aduriz, Mara Jess Aranzabe, Jose Mari Arriola, Aitziber Atutxa, Arantza Daz de Ilarraza, Aitzpea Garmendia, and Maite Oronoz. 2003. Construction of a Basque dependency treebank. In *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories*, pages 201–204.

Iñaki Alegria, Maria Jesus Aranzabe, Aitzol Ezeiza, Nerea Ezeiza, and Ruben Urizar. 2002. Robustness and customisation in an analyser/lemmatiser for basque. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC). Customizing knowledge in NLP applications Workshop*.

Anders Bjrkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstrations*, pages 33–36, Beijing, China, August. Coling 2010 Organizing Committee.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2012. The joy of parallelism with czeng 1.0. In Nicoletta Calzolari, editor, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Ondřej Bojar, Rudolf Rosa, and Aleš Tamchyna. 2013. Chimera – three heads for English-to-Czech translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 92–98, Sofia, Bulgaria, August. Association for Computational Linguistics.

António Branco and João Silva. 2006. A Suite of Shallow Processing Tools for Portuguese: LX-Suite. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Daniel de Kok and Gertjan Van Noord. 2010. A sentence generator for dutch. In *LOT Occasional Series*, pages 75–90.

Marie-Catherine de Marneffe, Natalia Silveira, Timothy Dozat, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the 9th Language Resources and Evaluation Conference*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 48–54. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*. The Association for Computer Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, pages 79–86.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

David Mareček, Martin Popel, and Zdeněk Žabokrtský. 2010. Maximum entropy translation model in dependency-based MT framework. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 201–206.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Haji. 2005. Non-projective dependency parsing using spanning tree algorithms. In *In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530.

Steven Neale, Luis Gomes, Eneko Agirre, Oier Lopez de Lacalle, and António Branco. 2016. Word sense-aware machine translation: Including senses as contextual features for improved translation models. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC), European Language Resources Association*, Portorož, Slovenia. To appear.

Gertjan Van Noord. 2006. At last parsing is now operational. In *In TALN 2006, Verbum Ex Machina, Actes De La 13e Conference sur Le Traitement Automatique des Langues naturelles*, pages 20–42.

Václav Novák and Zdeněk Žabokrtský. 2007. Feature engineering in maximum spanning tree dependency parser. In Václav Matoušek and Pavel Mautner, editors, *Text, Speech and Dialogue: 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 3-7, 2007. Proceedings*, pages 92–98, Berlin, Heidelberg. Springer Berlin Heidelberg.

João Rodrigues, Luıs Gomes, Steve Neale, Andreia Querido, Nuno Rendeiro, Sanja Štajner, João Silva, and António Branco. 2016a. Domain-specific hybrid machine translation from english to portuguese. In *Lecture Notes in Artificial Intelligence*. Springer. To appear.

João Rodrigues, Nuno Rendeiro, Andreia Querido, Sanja Štajner, and António Branco. 2016b. Bootstrapping a hybrid mt system to a new language pair. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC), European Language Resources Association*, Portorož, Slovenia. To appear.

Rudolf Rosa, Ondřej Dušek, Michal Novák, and Martin Popel. 2015. Translation model interpolation for domain adaptation in tectomt. In *Proceedings of the 1st Deep Machine Translation Workshop*, pages 89–96, Praha, Czechia. ÚFAL MFF UK.

Rudolf Rosa, Roman Sudarikov, Michal Nov4́k, Martin Popel, and Ondřej Bojar. 2016. Dictionary-based domain adaptation of mt systems without retraining. In *Proceedings of the 1sh Conference of Machine Translation, WMT2016*, Berlin, Germany. To appear.

João Silva, João Rodrigues, Luıs Gomes, and António Branco. 2015. Bootstrapping a hybrid deep mt system. In *Proceedings of the ACL2015 Fourth Workshop on Hybrid Approaches to Translation*, pages 1–5.

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

Andreas Stolcke. 2002. SRILM - an Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 901–904.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-source tools for morphology, lemmatization, pos tagging and named entity recognition. In *In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, page 1318, Baltimore, Maryland. Association for Computational Linguistics.

Jrg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association.

Sanja Štajner, Andreia Querido, Nuno Rendeiro, João Rodrigues, and António Branco. 2016. Use of domain-specific language resources in machine translation. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC), European Language Resources Association*, Portorož, Slovenia. To appear.

Zdeněk Žabokrtský and Martin Popel. 2009. Hidden Markov Tree Model in Dependency-based Machine Translation. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 145–148.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. Tectomt: Highly modular mt system with tectogrammatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 167–170, Stroudsburg, PA, USA. Association for Computational Linguistics.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In Nicoletta Calzolari, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.