Factored Models for Deep Machine Translation

Kiril Simov, Iliana Simova, Velislava Todorova, Petya Osenova
Linguistic Modelling Department, IICT
Bulgarian Academy of Sciences
Sofia, Bulgaria
{kivs|iliana|slava|petya}@bultreebank.org

Abstract

In this paper, we present some preliminary results on Statistical Machine Translation from Bulgarian-to-English and English-to-Bulgarian. Linguistic knowledge has been added gradually as factors in the MOSES system. The tests were performed on the QTLeap corpus data in IT domain for Pilot 1. The training was done on news parallel data as well as on IT domain data. The BLEU scores show that the addition of linguistic knowledge improves the Machine Translation.

1 Introduction

In the recent years, machine translation (MT) has achieved significant improvement in terms of translation quality (Koehn, 2010). Both data-driven approaches (e.g., statistical MT (SMT)) and knowledge-based (e.g., rule-based MT (RBMT)) have achieved comparable results shown in the evaluation campaigns (Callison-Burch et al., 2011). However, according to the human evaluation, the final outputs of the MT systems are still far from satisfactory. For that reason, we explore an approach that incrementally incorporates linguistic knowledge into an SMT system.

There has not been much study on the language pair Bulgarian – English, mainly due to the lack of resources, including corpora, preprocessors, etc. There was a system published by Koehn et al. (2009), which was trained and tested on the European Union law data, but not on other domains like news. They reported a very high BLEU score (Papineni et al., 2002) on the Bulgarian – English translation direction (61.3). The direction from English to Bulgarian was even less explored.

In the QTLeap project[1] linguistic knowledge is gradually added to SMT systems with the aim to achieve better translation in both directions: EN-to-X language and X language-to-English. The incremental process is organized in several pilots. Pilot 0 sets the baseline, which means that no linguistic knowledge is added. Pilot 1 introduces some initial linguistic knowledge through the incorporation of some features such as part-of-speech, lemma, etc. In the setting that involved Bulgarian, we also added some general information on the ontological type of the word: referent or event. Pilots 2 and 3 will integrate further knowledge, such as lexicons, semantic annotations, etc.

In this paper, we focus on the Bulgarian-to-English and English-to-Bulgarian translation, and mainly explore the approach of building on the SMT baseline, which is already augmented with linguistic features. More precisely, we explore the impact of the bilingual morphological lexicons in the translation process.

These are the motivations behind our approach: 1) the SMT baseline trained on a decent amount of parallel corpora already proved to be a good direction to go. Thus, more knowledge has to be added

[1] http://qtleap.eu/

Proceedings of the 1st Deep Machine Translation Workshop (DMTW 2015), pages 97–105,
Praha, Czech Republic, 3–4 September 2015.

for further lines of improvement; 2) the MT system can profit from the incorporation of knowledge additional to the common linguistic factors. Such additions include lexicons, gazetteers, etc.

The rest of the paper is organized as follows: Section 2 mentions some related approaches. Section 3 presents information on preparation of the data and Section 4 describes the preprocessing of the data and introduces our factor-based SMT model which allows us to incorporate various linguistic features into an SMT baseline, including some semantic features. We show our experiments in Section 5 as well as some preliminary evaluation of the results. The conclusions and future work are presented in Section 6.

2 Related Work

Our work is closely connected to the transfer-based MT models. Ideally, given the availability of two deep grammars for some language pair, we would be able to translate through the transfer of the deep representations.

One such setting was developed in the framework of the Head-driven Phrase Structure Grammar (HPSG) within the DELPH-IN community[2]. The deep representation is delivered by the Minimal Recursion Semantics (MRS) analyses. They are usually delivered together with the syntactic analyses of the text. There already exist quite extensive implemented formal HPSG grammars for English (Copestake and Flickinger, 2000), Spanish (Marimon, 2010), German (Müller and Kasper, 2000), and Japanese (Siegel, 2000; Siegel and Bender, 2002). All grammars are harmonized with a Grammar Matrix (Bender et al., 2002). At the moment, precise and linguistically motivated grammars, customized on the base of the Grammar Matrix, have been or are being developed for Norwegian, French, Korean, Italian, Modern Greek, Spanish, Portuguese, Chinese, etc. There also exists a Bulgarian Resource Grammar – BURGER[3].

The transfer in this setting is usually implemented in the form of rewriting rules. For instance, in the Norwegian LOGON project (Oepen et al., 2004), the transfer rules were hand-written (Bond et al., 2005; Oepen et al., 2007), which involved a large amount of manual work. Graham and van Genabith (2008) and Graham et al. (2009) explored the automatic rule induction approach in a transfer-based MT setting two Lexical Functional Grammars (LFGs), which was still restricted by the performance of both – the parser and the generator. Lack of robustness for target side generation is one of the main issues, when various ill-formed or fragmented structures come out after transfer. Oepen et al. (2007) use their generator to generate text fragments instead of full sentences, in order to increase the robustness.

However, since a real large-scale grammar for Bulgarian is still not available, we take an SMT system as our 'backbone' which robustly delivers some translation for any given input. Then, we incrementally augment SMT with deep linguistic knowledge. In general, what we are doing is still along the lines of previous work utilizing deep grammars, but we build a more 'light-weighted' transfer model over dependency parses.

One of the MRS-related semantic formalisms is the Abstract Meaning Representation (AMR[4]), which also aims at achieving whole-sentence deep semantics instead of addressing various isolated holders of semantic information (such as NER, coreferences, temporal anchors, etc.). AMR also builds on the available syntactic trees, thus contributing to the efforts on sembanking.

Another stream of research is related to the TectoMT approach (Žabokrtský et al., 2008). The Prague Dependency Treebank (PDT)[5] is a Czech treebank, annotated in accordance to the linguistic theory of Functional Generative Description (P. Sgall and Panevova, 1986). The tectogrammatical layer[6] is the third layer of the PDT. It represents the syntactic-semantic interface, adding the functional dimension and collapsing the structural information, thus aiming at a more language-independent level of abstraction. The other two layers are the morphological and analytical ones. The morphological layer operates over tokens, assigning to them POS and lemma tags. The analytical layer reflects the surface sentence structure.

[2] http://www.delph-in.net/wiki/index.php/Home
[3] http://www.bultreebank.org/BURGER/index.html
[4] http://www.isi.edu/natural-language/amr/a.pdf
[5] https://ufal.mff.cuni.cz/pdt2.0/
[6] https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/ch01.html

The tectogrammatical annotation builds on the analytical level. It presents the deep semantic structure of the sentence. At the tectogrammatical level, each sentence has at least one representation unambiguously characterizing the meaning of the sentence. The tectogrammatical level representation contains all the information necessary for translating the tectogrammatical representation into the lower levels, as well as for its interpretation in the sense of intentional semantics.

In contrast to the analytical level, which follows the surface sentence structure and encodes analytical functions (in particular, grammatical relations like *Subject, Object, Predicate, Attribute*, etc.), while preserving the word order, the tectogrammatical level highlights the functional dimension (such as the semantic roles *Actor, Patient, Addressee*, etc.). Additionally, it abstracts away from the synsemantic (functional) parts-of-speech (prepositions, conjunctions, etc.) in the dependency trees, thus focusing on the autosemantic (content) words (nouns, verbs, etc.). The structural information is not lost, but just "collapsed" into the content words representations. In this way, a more abstract level of language representation is achieved, which then is used for the transfer step within the MT systems. The result on the tectogrammatical level heavily depends on the results from the processed analytical level.

In the future, we plan to have transfer architectures for Bulgarian and English in both directions in both approaches – MRS and TectoMT. However, since these endeavors require more work, for the moment we test our ideas in the already built-in setting of the factored-based MOSES system. Thus, we build on the previous language model translation experience described in (Wang et al., 2012a) and (Wang et al., 2012b). However, while in the above-mentioned publications only Bulgarian-to-English translation was explored, in this paper also the English-to-Bulgarian direction is presented.

3 Data Preparation

Two types of data are used in our experiments. The first type includes parallel news data. It is the training data. The second type includes parallel QTLeap data in the IT domain. It is the training and test data.

The parallel news data comprises the following sources:

1. SETIMES parallel corpus, which is part of the OPUS parallel corpus[7].
2. EuroParl parallel corpus[8].
3. LibreOffice Document Foundation.

The data in SETIMES corpus was aligned automatically. We first checked the consistency of the automatic alignments. It turned out that more than 25% of the sentence alignments were not correct. We corrected manually more than 25.000 sentence alignments. (The the rest of the data set includes around 135,000 sentences. The whole data set is about 160,000 sentences.) Then, two actions were taken:

1. **Improving the tokenization of the Bulgarian part**. The observations from the manual check of the set of 25,000 sentences showed systematic errors in the tokenized text. Hence, these cases have been detected and fixed semi-automatically.
2. **Correcting and removing the suspicious alignments**. Initially, the ratio of the lengths of the English and Bulgarian sentences was calculated in the set of the 25,000 manually annotated sentences. As a rule, the Bulgarian sentences are longer than the English ones. The ratio is 1.34. Then we calculated the ratio for each pair of sentences. After this, the optimal interval was manually determined, such that if the ratio for a given pair of sentences is within the interval, then we assume that the pair is a good one. The interval for these experiments is set to [0.7; 1.8]. All the pairs with ratio outside of the interval have been deleted.

The test dataset was the Bulgarian-English parallel part from the QTLeap multilingual corpus. The QTLeap corpus is composed of 4 000 pairs of questions and respective answers in the domain of ICT

[7]OPUS–an open source parallel corpus, `http://opus.lingfil.uu.se/`

[8]`http://www.statmt.org/europarl/`

troubleshooting for both hardware and software. This material was collected using a real-life commercial online support service via chat. The corpus is thus composed of naturally occurring utterances produced by users while interacting with that service. The support system, denominated PcWizard, aims to be the first point of contact for troubleshooting trying to offer a rapid reply and solution to not too complex questions from the users. For more information see[9].

4 Linguistic Preprocessing and Factor-based SMT Model

For the current experiments the data in the training datasets was analyzed at two levels – POS tagging and Lemmatization: **POS tagging:** POS tagging was performed by a pipe of several modules. First, we applied a morphological lexicon and a set of rules. The lexicon added all the possible tags for the known words. The rules reduced the ambiguity for some of the sure cases. The result of this step was a tagged text with some ambiguities unresolved. The next step was the application of the GTagger (see (Georgiev et al., 2012)). It was trained on ambiguous data and thus selected the most appropriate tags from the suggested ones. The accuracy of the whole pipeline is 97.83%. **Lemmatization:** The lemmatization module is based on the same morphological lexicon that was used in the tagger. From the lexicon we extracted functions which convert each word form into its lemma.

Then we built our approach on top of the factor-based SMT model proposed by Koehn and Hoang (2007a), as an extension of the traditional phrase-based SMT framework. Instead of using only the word form of the text, it allows the system to take a vector of factors to represent each token, both for the source and target languages. The vector of factors can be used for different levels of linguistic annotations, like lemma or part-of-speech. Furthermore, this extension actually allows us to incorporate various kinds of features if they can be (somehow) represented as annotations of the tokens.

The process is quite similar to supertagging (Bangalore and Joshi, 1999), which assigns "rich descriptions (supertags) that impose complex constraints in a local context". In our case, all the linguistic features (factors) associated with each token form a supertag to that token. Singh and Bandyopadhyay (2010) had a similar idea of incorporating linguistic features, while they worked on Manipuri – English bidirectional translation. Our approach is slightly different from (Birch et al., 2007) and (Hassan et al., 2007), who mainly used the supertags on the target language side, English. We experiment with both sides.

In particular, we consider the following morphosyntactic factors for both languages:

- WF - word form, which is the original text token.
- LEMMA is the lexical invariant of the original word form.
- POS - part-of-speech of the word.
- LING - other linguistic features derived from the POS tag in the BulTreeBank tagset.

In comparison to the experiments described in ((Wang et al., 2012a), (Wang et al., 2012b)) the number of the linguistic factors were reduced in comparison to the ones that contributed best to the improvement of the translation results. Thus, we have excluded all the factors based on dependency parsing of the data.

Our work on Minimal Recursion Semantic analysis of Bulgarian text is inspired by the work on MRS and RMRS (Robust Minimal Recursion Semantic) (see (Copestake, 2003) and (Copestake, 2007)) and the previous work on transfer of dependency analyses into RMRS structures described in (Spreyer and Frank, 2005) and (Jakob et al., 2010).

MRS is introduced as an underspecified semantic formalism (Copestake et al., 2005). It is used to support semantic analyses in the English HPSG grammar ERG (Copestake and Flickinger, 2000), but also in other grammar formalisms like LFG. The main idea is that it avoids spelling out the complete set of readings resulting from the interaction of scope bearing operators and quantifiers, instead providing a single underspecified representation from which the complete set of readings can be constructed. Here

[9] http://qtleap.eu/wp-content/uploads/2015/05/QTLEAP-2015-D2.51.pdf

we will present only basic definitions from (Copestake et al., 2005). An MRS structure is a tuple $\langle GT, R, C \rangle$, where GT is the top handle, R is a bag of EPs (elementary predicates) and C is a bag of handle constraints, such that there is no handle h that outscopes GT. Each elementary predication contains exactly four components: (1) a handle which is the label of the EP; (2) a relation; (3) a list of zero or more ordinary variable arguments of the relation; and (4) a list of zero or more handles corresponding to scopal arguments of the relation (i.e., holes). RMRS is introduced as a modification of MRS which to capture the semantics resulting from the shallow analysis. Here an assumption is made that the shallow processor does not have access to a lexicon. Thus it does not have access to the arity of the relations in EPs. Therefore, the representation has to be underspecified with respect to the number of arguments of the relations. The names of relations are constructed on the basis of the lemma for each word form in the text. This main argument could be of two types: *referential index* (v) for nouns and *event* (e) for the other parts of speech. In our implementation we extend the types of the main argument of the elementary predicates. Especially for the event arguments we introduce a new type[10] "ef" for adverbs and "ec" for subordinators, because they modify other events and thus they are special type of events. In Bulgarian some parts of speech can have main attribute of both types: "v" and "e". For them we introduce a new type "e-v".

Similarly to our previous experiments, here we use only the RMRS relation and the type of the main argument as features to the translation model. We will skip here the explanation of the full structure of RMRS structures and how they are constructed. Thus, we firstly do a match between the surface tokens and the MRS elementary predicates (EPs) and then extract the following features as extra factors:

- EP – the name of the elementary predicate, which usually indicates an event or an entity from a semantic point of view.
- EoV – indicates the current EP as either an event, a reference variable or their subtypes.

Notice that we do not take all the information provided by the MRS, e.g., we throw away the scopal information and the other arguments of the relations. This kind of information is not straightforward to be represented in such 'tagging'-style models, but it will be tackled in the future.

All these factors encoded within the corpus provide us with a rich selection of factors for various experiments.

5 Experiments

For our entry level deep machine translation system (Pilot 1) we make use of the Moses open source toolkit to build a factored SMT model (Koehn and Hoang, 2007b). As it was mentioned above in the analysis stage, we create a representation of the text which encodes various levels of linguistic information as factors. These include morphological, syntactic and semantic abstractions in the source and target language.

We have experimented with several combinations of factors derived from the preprocessing with the Bulgarian and English analysis pipelines, together with semantic factors based on Minimal Recursion Semantics (see Table 1 for a subset of the results).

The following are some examples of factors for this model: word form, lemma, and morphosyntactic tags, factors modeling the parent word (lemma of the parent word, part of speech of the parent word) as well as the type of dependency relation (syntactic factors), and MRS-based factors (elementary predicate and variable type).

We contributed mainly in two directions: better analysis with an improved pipeline for Bulgarian, and different more complex types of factored models to explore successful factor combinations. We have experimented with a number of combinations of the listed factors, language model types (word and POS), translation and generation steps. The best performing model featuring a semantic factor for the direction BG→EN includes four factors: word form, lemma, POS and variable type; a word and POS-based language model. In the transfer step, two alternative approaches are used. If possible a mapping

[10] In fact these types are subtypes of the basic ones.

Factors	LM	Translation	Generation	Decoding	BLEU	
					BG→EN	EN→BG
WF, EP, EoV	0	0,1,2-0	–	–	31.53	24.00
WF, POS, EoV	0	0,1,2-0	–	–	32.07	24.13
WF, LEMMA, EP, EoV	0	1-1+2-2+3-3	1,2,3-0	–	23.94	13.69
WF, LEMMA, POS	0,2	0-0,2+1-0,2	–	t0:t1	32.59	22.86
WF, LEMMA, POS, LING	0,2	1-1+3-2+0-0,2	1-2+1,2-0	t0,g0,t1,g1:t2	32.78	22.73
WF, LEMMA, POS, EoV	0,2	0,3-0,2+1,3-0,2	–	t0:t1	32.59	22.77

Table 1: A subset of the results from the factored experiments, evaluated on the second half of the QTLeap data set.

is performed between the source word form and the variable type and the target word form candidates and POS candidates. However, if the source word form has not been seen during the training phase, the source lemma together with the variable type is used instead.

For the translation direction EN→BG the model includes three factors: word form, part of speech, and variable type. In the translation step, the source word, POS, and variable type are translated into the target word form.

The automatic evaluation for both directions is described in D2.4 of the QTLeap project.

The BG-to-EN direction was evaluated on questions. Here are the numbers for Pilot 0 and Pilot 1 per metric:

1. BLEU Pilot 0 (29.7); Pilot 1 (27.7)
2. wordF Pilot 0 (22.8); Pilot 1 (22.4)
3. chartF Pilot 0 (46.7); Pilot 1 (**47.4**)

The EN-to-BG direction was evaluated on the answers:

1. BLEU Pilot 0 (25.3); Pilot 1 (24.5)
2. wordF Pilot 0 (25.6); Pilot 1 (25.0)
3. chartF Pilot 0 (46.7); Pilot 1 (46.6)

The results from the two pilots are comparable. More linguistic knowledge is needed for the translation improvement. The only small improvement was noted in BG-to-EN direction in chartF. Since up to now this translation direction was the focus, more effort is needed for improvement in the other direction as well.

5.1 Preliminary Experiments with a Parallel Morphological Lexicon

One of the main problems in the translation in both directions are the so-called out-of-training word forms. These are word form pairs of translations that do not appear in the parallel corpora used for the training. For example, in Bulgarian each adjective has 9 forms. For many adjectives many of these forms are not present in the parallel corpora. In order to solve this problem we decided to add a parallel Bulgarian-English morphological lexicon to the parallel corpora.

The lexicon was constructed by exploiting the following resources: BTB-Morphological lexicon containing all wordforms for more than 110 000 Bulgarian lemmas; BTB-bilingual Bulgarian-English lexicons (with about 8000 entries); English Wiktionary. From it the English wordforms were extracted for the English lemmas. Then we mapped the wordform lexicons for both languages to the corresponding part of the bilingual lexicon. Afterwards, the corresponding wordforms were aligned on the basis of their morphological features like *number* (singular, plural); *degree* (comparative, superlative); *definiteness* (definite, indefinite), etc.

Bulgarian	English
visok|visok|a	a|a|d high|high|g
visok|visok|a	high|high|g
visok|visok|a	a|a|d tall|tall|g
visok|visok|a	tall|tall|g
—	—
naj-visokata|visok|a	highest|highest|g
naj-visokata|visok|a	the|the|d highest|highest|g
naj-visokata|visok|a	tallest|tallest|g
naj-visokata|visok|a	the|the|d tallest|tallest|g

Table 2: Wordform aligned parallel lexicon. It shows the Bulgarian adjective "visok" with its two translations in English: "high" and "tall". The table represents the encoding of singular, masculine, indefinite forms and superlative, singular, feminine, definite forms. Each triple represents wf|lm|pos, where *wf* is the wordform, *lm* is the lemma and *pos* is the part-of-speech. For example, the triple a|a|d means: wordform "a"; lemma "a" and part-of-speech "determiner".

In this preliminary experiment we used only the noun and the adjective parts-of-speech from the wordform aligned bilingual lexicon. Bulgarian language encodes definiteness as an ending to the nouns and adjectives in contrast to English which encodes it as a separate determiner in front of the noun or adjective. For this reason we also encode the English definite and indefinite articles for the English wordforms. Since in some contexts the English articles are not obligatory, the English wordforms were encoded with or without them. In addition, we also represented factors for each wordform (in the example below we encode the lemma and POS). Tab. 2 shows an example from the resulting lexicon.

The lexicon represents more than 70 000 aligned wordforms. It was added to the training data. Each aligned pair of word forms is added as a pair of sentences with length one or two depending on determiners. We got the results presented in Tab. 3. They show a positive impact of the aligned wordform parallel lexicon on the translation in both directions. The table shows also that the addition of the definite forms for English does not change the result.

	without lexicon	with lexicon; with only indefinite forms	with lexicon; with all forms
BG→EN	32.59	33.02	32.88
EN→BG	22.86	23.91	22.97

Table 3: Preliminary experiments with parallel morphological lexicons.

Although the reported here experiments are only preliminary they demonstrate a possible direction of improving of the training corpus for solving the "out-of-training-wordforms" problem. There is still room for improvements which include the incorporation also of other parts-of-speech, compositional and multiword phrases, etc.

6 Conclusions and Future Work

In this paper, we reported our initial work towards building deep statistical machine translation models between Bulgarian and English in both directions. Based on previous experiments, in Pilot 1 we extended the semantic factors with new types of main arguments for MRS elementary predicates, which improved the results in English-to-Bulgarian direction and shows promising results for the Bulgarian-to-English direction. The paper also showed that the addition of a wordform aligned parallel lexicon improved the results in both translation directions.

In our future work we plan to incorporate more linguistic knowledge from the lexicon. Also we will aim at improving the incorporation of deep factors within the translation models.

Acknowledgements

This research has received support by the EC's FP7 (FP7/2007-2013) project under grant agreement number 610516: "QTLeap: Quality Translation by Deep Language Engineering Approaches."

We thank the two anonymous for their valuable comments on the initial version of the paper. All errors remain our own responsibility.

References

Srinivas Bangalore and Aravind K. Joshi. 1999. Supertagging: an approach to almost parsing supertagging: an approach to almost parsing supertagging: an approach to almost parsing supertagging: an approach to almost parsing. *Computational Linguistics*, 25(2), June.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The grammar Matrix. An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammar. In *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2007. Ccg supertags in factored statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 9–16, Prague, Czech Republic, June.

Francis Bond, Stephan Oepen, Melanie Siegel, Ann Copestake, and Dan Flickinger. 2005. Open source machine translation with DELPH-IN. In *Proceedings of the Open-Source Machine Translation Workshop at the 10th Machine Translation Summit*, pages 15 – 22, Phuket, Thailand, September.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F. Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the 6th Workshop on SMT*.

Ann Copestake and Dan Flickinger. 2000. An open source grammar development environment and broad-coverage english grammar using hpsg. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language & Computation*, 3(4):281–332.

Ann Copestake. 2003. Robust minimal recursion semantics (working paper).

Ann Copestake. 2007. Applying robust semantics. In *Proceedings of the 10th Conference of the Pacific Assocation for Computational Linguistics (PACLING)*, pages 1–12.

Georgi Georgiev, Valentin Zhikov, Petya Osenova, Kiril Simov, and Preslav Nakov. 2012. Feature-rich part-of-speech tagging for morphologically complex languages: Application to bulgarian. In *Proceedings of EACL 2012*. MIT Press, Cambridge, MA, USA.

Yvette Graham and Josef van Genabith. 2008. Packed rules for automatic transfer-rule induction. In *Proceedings of the European Association of Machine Translation Conference (EAMT 2008)*, pages 57–65, Hamburg, Germany, September.

Yvette Graham, Anton Bryl, and Josef van Genabith. 2009. F-structure transfer-based statistical machine translation. In *Proceedings of the Lexical Functional Grammar Conference*, pages 317–328, Cambridge, UK. CSLI Publications, Stanford University, USA.

Hany Hassan, Khalil Sima'an, and Andy Way. 2007. Supertagged phrase-based statistical machine translation. In *Proceedings of ACL*, Prague, Czech Republic, June.

Max Jakob, Markéta Lopatková, and Valia Kordoni. 2010. Mapping between dependency structures and compositional semantic representations. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2491–2497.

Philipp Koehn and Hieu Hoang. 2007a. Factored translation models. In *Proceedings of EMNLP*.

Philipp Koehn and Hieu Hoang. 2007b. Factored translation models. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 868–876.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 2009. 462 machine translation systems for europe. In *Proceedings of MT Summit XII*.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, January.

Montserrat Marimon. 2010. The spanish resource grammar. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*.

Stefan Müller and Walter Kasper. 2000. HPSG analysis of German. In Wolfgang Wahlster, editor, *Verbmobil. Foundations of Speech-to-Speech Translation*, pages 238 – 253. Springer, Berlin, Germany, artificial intelligence edition.

Stephan Oepen, Helge Dyvik, Jan Tore Lønning, Erik Velldal, Dorothee Beermann, John Carroll, Dan Flickinger, Lars Hellan, Janne Bondi Johannessen, Paul Meurer, Torbjørn Nordgård, , and Victoria Rosén. 2004. Som å kapp-ete med trollet? towards MRS-based norwegian to english machine translation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD.

Stephan Oepen, Erik Velldal, Jan Tore Lønning, Paul Meurer, Victoria Rosén, and Dan Flickinger. 2007. Towards hybrid quality-oriented machine translation — on linguistics and probabilities in MT. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, Skovde, Sweden.

E. Hajicova P. Sgall and J. Panevova. 1986. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects.* Dordrecht: Reidel Publishing Company and Prague: Academia.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Melanie Siegel and Emily M. Bender. 2002. Efficient deep processing of japanese. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

Melanie Siegel. 2000. HPSG analysis of Japanese. In Wolfgang Wahlster, editor, *Verbmobil. Foundations of Speech-to-Speech Translation*, pages 265 – 280. Springer, Berlin, Germany, artificial intelligence edition.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010. Manipuri-english bidirectional statistical machine translation systems using morphology and dependency relations. In *Proceedings of the Fourth Workshop on Syntax and Structure in Statistical Translation*, pages 83–91, Beijing, China, August.

Kathrin Spreyer and Anette Frank. 2005. Projecting RMRS from TIGER Dependencies. In *Proceedings of the HPSG 2005 Conference*, pages 354–363, Lisbon, Portugal.

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.

Rui Wang, Petya Osenova, and Kiril Simov. 2012a. Linguistically-augmented bulgarian-to- english statistical machine translation model. In *Proceedings of the Joint Workshop on Exploiting Synergies Between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra), EACL 2012*, pages 119–128.

Rui Wang, Petya Osenova, and Kiril Simov. 2012b. Linguistically-enriched models for bulgarian-to-english machine translation. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation, SSST-6, 2012*, pages 10–19.