Linguistically-Augmented Bulgarian-to-English Statistical Machine Translation Model

Rui Wang
Language Technology Lab
DFKI GmbH
Saarbrücken, Germany
ruiwang@dfki.de

Petya Osenova and Kiril Simov
Linguistic Modelling Department, IICT
Bulgarian Academy of Sciences
Sofia, Bulgaria
{petya,kivs}@bultreebank.org

Abstract

In this paper, we present our linguistically-augmented statistical machine translation model from Bulgarian to English, which combines a statistical machine translation (SMT) system (as backbone) with deep linguistic features (as factors). The motivation is to take advantages of the robustness of the SMT system and the linguistic knowledge of morphological analysis and the hand-crafted grammar through system combination approach. The preliminary evaluation has shown very promising results in terms of BLEU scores (38.85) and the manual analysis also confirms the high quality of the translation the system delivers.

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.

Fortunately, recent error analysis shows that the two trends of the MT approaches tend to be complementary to each other, in terms of the types of the errors they made (Thurmair, 2005; Chen et al., 2009). Roughly speaking, RBMT systems often have missing lexicon and thus lack of robustness, while handling linguistic phenomena requiring syntactic information better. SMT systems, on the contrary, are in general more robust, but sometimes output ungrammatical sentences.

In fact, instead of competing with each other, there is also a line of research trying to combine the advantages of the two sides using a hybrid framework. Although many systems can be put under the umbrella of "hybrid" systems, there are various ways to do the combination/integration. Thurmair (2009) summarized several different architectures of hybrid systems using SMT and RBMT systems. Some widely used ones are: 1) using an SMT to post-edit the outputs of an RBMT; 2) selecting the best translations from several hypotheses coming from different SMT/RBMT systems; and 3) selecting the best segments (phrases or words) from different hypotheses.

For the language pair Bulgarian-English, there has not been much study on it, 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), which inspired us to further investigate this direction.

In this paper, we focus on the Bulgarian-to-English translation and mainly explore the approach of annotating the SMT baseline with linguistic features derived from the preprocessing and hand-crafted grammars. There are three motivations behind our approach: 1) the SMT baseline trained on a decent amount of parallel corpora outputs surprisingly good results, in terms of both statistical evaluation metrics and preliminary manual evaluation; 2) the augmented model gives

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 119–128,
Avignon, France, April 23 - 27 2012. ©2012 Association for Computational Linguistics

us more space for experimenting with different linguistic features without losing the 'basic' robustness; 3) the MT system can profit from continued advances in the development of the deep grammars thereby opening up further integration possibilities.

The rest of the paper will be organized as follows: Section 2 presents our work on cleaning the corpora and Section 3 briefly describes the preprocessing of the data. Section 4 introduces our factor-based SMT model which allows us to incorporate various linguistic features into an SMT baseline, among which those features coming from the MRS are described in Section 5 in detail. We show our experiments in Section 6 as well as both automatic and manual evaluation of the results. Section 7 briefly mentions some related work and then we summarize this paper in Section 8.

2 Data Preparation

In our experiments we are using the SETIMES parallel corpus, which is part of the OPUS parallel corpus[1]. The data in the corpus was aligned automatically. Thus, we first checked the consistency of the automatic alignments. It turned out that more than 25% of the sentence alignments were not correct. Since SETIMES appeared to be a noisy dataset, our effort was directed into cleaning it as much as possible before the start of the experiments. We first corrected manually more than 25.000 sentence alignments. The the rest of the data set includes around 135,000 sentences. Altogether the data set is about 160,000 sentences, when the manually checked part is added. Thus, 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. Similarly, we have cleaned EMEA dataset.

The size of the resulting datasets are: 151,718 sentence pairs for the SETIMES dataset. Similar approach was undertaken for another dataset from OPUS corpus - EMEA. After the cleaning 704,631 sentence pairs were selected from the EMEA dataset. Thus, the size of the original datasets was decreased by 10%.

3 Linguistic Preprocessing

The data in SETIMES dataset was analysed on the following levels:

- **POS tagging.** POS tagging is performed by a pipe of several modules. First we apply SVM POS tagger which takes as an input a tokenised text and its output is a tagged text. The performance is near 91% accuracy. The SVM POS tagger is implemented using SVMTool (Gimnez and Mrquez, 2004). Then we apply a morphological lexicon and a set of rules. The lexicon add all the possible tags for the known words. The rules reduce the ambiguity for some of the sure cases. The result of this step is a tagged text with some ambiguities unresolved. The third step is application of the GTagger (Georgiev et al., 2012). It is trained on an ambiguous data and select the most appropriate tags from the suggested ones. The accuracy of the whole pipeline is 97.83%. In this pipeline SVM POS Tagger plays the role of guesser for the GTagger.

- **Lemmatization.** The lemmatization module is based on the same morphological lexicon. From the lexicon we extracted functions which convert each wordform into its basic form (as a representative of the lemma). The functions are defined via two operations on

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

wordforms: remove and concatenate. The rules have the following form:

if **tag = Tag** *then* {*remove* **OldEnd**; *concatenate* **NewEnd**}

where **Tag** is the tag of the wordform, **OldEnd** is the string which has to be removed from the end of the wordform and **NewEnd** is the string which has to be concatenated to the beginning of the word form in order to produce the lemma. The rules are for word forms in the lexicon. Less than 2% of the wordforms are ambiguous in the lexicon (but they are very rare in real texts). Similar rules are defined for unknown words. The accuracy of the lemmatizer is 95.23%.

- **Dependency parsing.** We have trained the MALT Parser on the dependency version of BulTreeBank[2]. We did this work together with Svetoslav Marinov who has experience in using the MALT Parser and Johan Hall who is involved in thedevelopment of Malt Parser. The trained model achieves 85.6% labeled parsing accuracy. It is integrated in a language pipe with the POS tagger and the lemmatizer.

After the application of the language pipeline, the result is represented in a table form following the CoNLL shared task format[3].

4 Factor-based SMT Model

Our approach is built on top of the factor-based SMT model proposed by Koehn and Hoang (2007), 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, part-of-speech (POS), or other linguistic features. Furthermore, this extension actually allows us to incorporate various kinds of features if they can be (somehow) represented as annotations to 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 primarily experiment with the source language side, Bulgarian. This potentially huge feature space provides us with various possibilities of using our linguistic resources developed in and out of our project.

In particular, we consider the following factors on the source language side (Bulgarian):

- WF - word form is just the original text token.
- LEMMA is the lexical invariant of the original word form. We use the lemmatizer described in Section 3, which operates on the output from the POS tagging. Thus, the 3rd person, plural, imperfect tense verb form 'varvyaha' ('walking-were', They were walking) is lemmatized as the 1st person, present tense verb 'varvya'.
- POS - part-of-speech of the word. We use the positional POS tag set of the BulTreeBank, where the first letter of the tag indicates the POS itself, while the next letters refer to semantic and/or morphosyntactic features, such as: Dm - where 'D' stands for 'adverb', and 'm' stand for 'modal'; Ncmsi - where 'N' stand for 'noun', 'c' means 'common', 'm' is 'masculine', 's' is 'singular',and 'i' is 'indefinite'.
- LING - other linguistic features derived from the POS tag in the BulTreeBank tagset (see above).

In addition to these, we can also incorporate syntactic structure of the sentence by breaking down the tree into dependency relations. For instance, a dependency tree can be represented as a set of triples in the form of <parent, relation, child>. <loves, subject, John> and <loves, object, Mary> will represent the sentence "John loves Mary". Consequently, three additional factors are included for both languages:

[2] http://www.bultreebank.org/dpbtb/

[3] http://ufal.mff.cuni.cz/conll2009-st/task-description.html

- DEPREL - is the dependency relation between the current word and the parent node.

- HLEMMA is the lemma of the current word's parent node.

- HPOS is the POS tag of the current word's parent node.

Here is an example of a processed sentence. The sentence is "spored odita v elektricheskite kompanii politicite zloupotrebyavat s dyrzhavnite predpriyatiya." The glosses for the words in the Bulgarian sentence are: spored (*according*) odita (*audit-the*) v (*in*) elektricheskite (*electrical-the*) kompanii (*companies*) politicite (*politicians-the*) zloupotrebyavat (*abuse*) s (*with*) dyrzhavnite (*state-the*) predpriyatiya (*enterprises*). The translation in the original source is : "electricity audits prove politicians abusing public companies." The result from the linguistic processing and the addition of information about head elements are presented in the first seven columns of Table 1.

We extend the grammatical features to have the same size. All the information is concatenated to the word forms in the text. In the next section we present how we extend this format to incorporate the MRS analysis. In the next section we will extend this example to incorporate the MRS analysis of the sentence.

5 MRS Supertagging

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). In this section we present first a short overview of MRS and RMRS. Then we discuss the new features added on the basis of the RMRS structures.

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 the formalism 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 we will present only basic definitions from (Copestake et al., 2005). For more details the cited publication should be consulted. 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 the following assumption is taken into account the shallow processor does not have access to a lexicon. Thus it does not have access to 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 wordform in the text and the main argument for the relation is specified. This main argument could be of two types: *referential index* for nouns and *event* for the other part of speeches.

Because in this work we are using 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 semantically.

- EOV indicates the current EP is either an event or a reference variable.

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, which will be tackled in the future.

This information for the example sentence is

WF	Lemma	POSex	Ling	DepRel	HLemma	HPOS	EP	EoV
spored	spored	R	_	adjunct	zloupotrebyavam	VP	spored_r	e
odita	odit	Nc	npd	prepcomp	spored	R	odit_n	v
v	v	R	_	mod	odit	Nc	v_r	e
elektricheskite	elektricheski	A	pd	mod	kompaniya	Nc	elekticheski_a	e
kompanii	kompaniya	Nc	fpi	prepcomp	v	R	kompaniya_n	v
politicite	politik	Nc	mpd	subj	zloupotrebyavam	Vp	politik_n	v
zloupotrebyavat	zloupotrebyavam	Vp	tir3p	root	-	-	zloupotrebyavam_v	e
s	s	R	_	indobj	zloupotrebyavam	Vp	s_r	e
dyrzhavnite	dyrzhaven	A	pd	mod	predpriyatie	Nc	dyrzhaven_a	e
predpriyatiya	predpriyatie	Nc	npi	prepcomp	s	R	predpriyatie_n	v

Table 1: The sentence analysis with added head information — HLemma and HPOS.

represented for each word form in the last two columns of Table 1.

All these factors encoded within the corpus provide us with a rich selection of factors for different experiments. Some of them are presented within the next section. The model of encoding MRS information in the corpus as additional features does not depend on the actual semantic analysis — MRS or RMRS, because both of them provide enough semantic information.

6 Experiments

6.1 Experiments with the Bulgarian raw corpus

To run the experiments, we use the phrase-based translation model provided by the open-source statistical machine translation system, Moses[4] (Koehn et al., 2007). For training the translation model, the parallel corpora (mentioned in Section 2) were preprocessed with the tokenizer and lowercase converter provided by Moses. Then the procedure is quite standard:

- We run GIZA++ (Och and Ney, 2003) for bidirectional word alignment, and then obtain the lexical translation table and phrase table.
- A tri-gram language model is estimated using the SRILM toolkit (Stolcke, 2002).
- Minimum error rate training (MERT) (Och, 2003) is applied to tune the weights for the set of feature weights that maximizes the official f-score evaluation metric on the development set.

The rest of the parameters we use the default setting provided by Moses.

We split the corpora into the training set, the development set and the test set. For SETIMES, the split is 100,000/500/1,000 and for EMEA, it is 700,000/500/1,000. For reference, we also run tests on the JRC-Acquis corpus[5]. The final results under the standard evaluation metrics are shown in the following table in terms of BLEU (Papineni et al., 2002):

Corpora	Test	Dev	Final	Drop
SETIMES → SETIMES	34.69	37.82	36.49	/
EMEA → EMEA	51.75	54.77	51.62	/
SETIMES → EMEA	13.37	/	/	61.5%
SETIMES → JRC-Acquis	7.19	/	/	79.3%
EMEA → SETIMES	7.37	/	/	85.8%
EMEA → JRC-Acquis	9.21	/	/	82.2%

Table 2: Results of the baseline SMT system (Bulgarian-English)

As we mentioned before, the EMEA corpus is mainly about the description of medicine usage, and the format is quite fixed. Therefore, it is not surprising to see high performance on the in-domain test (2nd row in Table 2). SETIMES, consisting of news articles, is in a less controlled setting. The BLEU score is lower[6]. The results on the out-of-domain tests are in general much lower with a drop of more than 60% in BLEU score (the last column). For the JRC-Acquis corpus, in contrast to the in-domain scores given by Koehn et al. (2009) (61.3), the low out-of-domain results shows a very similar situation as EMEA. A brief manual check of the results indicate that the out-of-domain tests suffer severely from the missing

[4] http://www.statmt.org/moses/

[5] http://optima.jrc.it/Acquis/

[6] Actually, the BLEU score itself is higher than for most of the other language pairs http://matrix.statmt.org/. As the datasets are different, the results are not directly comparable. Here, we just want to get a rough picture. Achieving better performance for Bulgarian-to-English translation than for other language pairs is not the focus of the paper.

lexicon, while the in-domain test for the news articles contains more interesting issues to look into. The better translation quality also makes the system outputs human readable.

6.2 Experiments with the Linguistically-Augmented Bulgarian Corpus

As we described the factor-based model in Section 4, we also perform experiments to test the effectiveness of different linguistic annotations. The different configurations we considered are shown in the first column of Table 3.

These models can be roughly grouped into five categories: word form with linguistic features; lemma with linguistic features; models with dependency features; MRS elementary predicates (EP) and the type of the main argument of the predicate (EOV); and MRS features without word forms. The setting of the system is mostly the same as the previous experiment, except for 1) increasing the training data from 100,000 to 150,000 sentence pairs; 2) specifying the factors during training and decoding; and 3) without doing MERT[7]. We perform the finer-grained model only on the SETIMES data, as the language is more diverse (compared to the other two corpora). The results are shown in Table 3.

The first model is served as the baseline here. We show all the n-gram scores besides the final BLEU, since the some of the differences are very small. In terms of the numbers, POS seems to be an effective factor, as Model 2 has the highest score. Model 3 indicates that linguistic features also improve the performance. Model 4-6 show the necessity of including the word form as one of the factors, in terms of BLEU scores. Model 10 shows significant decrease after incorporating HLEMMA feature. This may be due to the data sparsity, as we are actually aligning and translating bi-grams instead of tokens. This may also indicate that increasing the number of factors does not guarantee performance enhancement. After replacing the HLEMMA with HPOS, the result is close to the others (Model 8). The experiments with features from the MRS analyses (Model 11-16) show improvements over the baseline consistently and using only the MRS features (Model 17-18) also delivers descent results. In future experiments we will consider to include more feature from the MRS analyses.

So far, incorporating additional linguistic knowledge has not shown huge improvement in terms of statistical evaluation metrics. However, this does not mean that the translations delivered are the same. In order to fully evaluate the system, manual analysis is absolutely necessary. We are still far from drawing a conclusion at this point, but the preliminary scores calculated already indicate that the system can deliver decent translation quality consistently.

6.3 Manual Evaluation

We manually validated the output for all the models mentioned in Table 3. The guideline includes two aspects of the quality of the translation: *Grammaticality* and *Content*. *Grammaticality* can be evaluated solely on the system output and *Content* by comparison with the reference translation. We use a 1-5 score for each aspect as follows:

Grammaticality

1. The translation is not understandable.
2. The evaluator can somehow guess the meaning, but cannot fully understand the whole text.
3. The translation is understandable, but with some efforts.
4. The translation is quite fluent with some minor mistakes or re-ordering of the words.
5. The translation is perfectly readable and grammatical.

Content

1. The translation is totally different from the reference.
2. About 20% of the content is translated, missing the major content/topic.
3. About 50% of the content is translated, with some missing parts.
4. About 80% of the content is translated, missing only minor things.
5. All the content is translated.

For the missing lexicons or not-translated Cyrillic tokens, we ask the evaluators to score 2

[7] This is mainly due to the large amount of computation required. We will perform MERT on the better-performing configurations in the future.

ID	Model	BLEU	1-gram	2-gram	3-gram	4-gram
1	WF	38.61	**69.9**	44.6	31.5	22.7
2	WF, POS	**38.85**	**69.9**	**44.8**	**31.7**	**23.0**
3	WF, LEMMA, POS, LING	38.84	**69.9**	44.7	**31.7**	**23.0**
4	LEMMA	37.22	68.8	43.0	30.1	21.5
5	LEMMA, POS	37.49	68.9	43.2	30.4	21.8
6	LEMMA, POS, LING	38.70	69.7	44.6	31.6	22.8
7	WF, DEPREL	36.87	68.4	42.8	29.9	21.1
8	WF, DEPREL, HPOS	36.21	67.6	42.1	29.3	20.7
9	WF, LEMMA, POS, LING, DEPREL	36.97	68.2	42.9	30.0	21.3
10	WF, LEMMA, POS, LING, DEPREL, HLEMMA	29.57	60.8	34.9	23.0	15.7
11	WF, POS, EP	38.74	69.8	44.6	31.6	22.9
12	WF, POS, LING, EP	38.76	69.8	44.6	**31.7**	22.9
13	WF, EP, EOV	38.74	69.8	44.6	31.6	22.9
14	WF, POS, EP, EOV	38.74	69.8	44.6	31.6	22.9
15	WF, LING, EP, EOV	38.76	69.8	44.6	**31.7**	22.9
16	WF, POS, LING, EP, EOV	38.76	69.8	44.6	**31.7**	22.9
17	EP, EOV	37.22	68.5	42.9	30.2	21.6
18	EP, EOV, LING	38.38	69.3	44.2	31.3	22.7

Table 3: Results of the factor-based model (Bulgarian-English, SETIMES 150,000)

for one Cyrillic token and score 1 for more than one tokens in the output translation.

The results are shown in the following two tables, Table 4 and Table 5, respectively. The current results from the manual validation are on the basis of 150 sentence pairs. The numbers shown in the tables are the number of sentences given the corresponding scores. The 'Total' column sums up the scores of all the output sentences by each model.

The results show that linguistic and semantic analyses definitely improve the quality of the translation. Exploiting the linguistic processing on word level — LEMMA, POS and LING — produces the best result. However, the model with only EP and EOV features also delivers very good results, which indicates the effectiveness of the MRS features from the deep hand-crafted grammars. Including more factors (especially the information from the dependency parsing) drops the results because of the sparseness effect over the dataset, which is consistent with the automatic evaluation BLEU score. The last two rows are shown for reference. 'Google' shows the results of using the online translation service provided by `http://translate.google.com/`. The high score (very close to the reference translation) may be because our test data are not excluded from their training data. In future we plan to do the same evaluation with a larger dataset.

The problem with the untranslated Cyrillic tokens in our view could be solved in most of the cases by providing additional lexical information from a Bulgarian-English lexicon. Thus, we also evaluated the possible impact of such a lexicon if it had been available. In order to do this, we substituted each copied Cyrillic token with its translation when there was only one possible translation. We did such substitutions for 189 sentence pairs. Then we evaluated the result by classifying the translations as acceptable or unacceptable. The number of the acceptable translations are 140 in this case.

The manual evaluation of the translation models on a bigger scale is in progress. The current results are promising. Statistical evaluation metrics can give us a brief overview of the system performance, but the actual translation quality is much more interesting to us, as in many cases, the different surface translations can convey exactly the same meaning in the context.

7 Related Work

Our work is also enlightened by another line of research, transfer-based MT models, which are seemingly different but actually very close. In this section, before we mention some previous work in this research direction, we firstly introduce the background of the development of the deep HPSG grammars.

The MRSes are usually delivered together with the HPSG analyses of the text. There already

ID	Model	1	2	3	4	5	Total
1	WF	20	47	5	32	**46**	487
2	WF, POS	20	48	5	37	40	479
3	WF, Lemma, POS, Ling	20	47	6	34	43	483
4	Lemma	**15**	34	11	46	44	**520**
5	Lemma, POS	**15**	38	12	**51**	34	**501**
6	Lemma, POS, Ling	20	48	5	34	43	482
7	WF, DepRel	32	48	3	29	38	443
8	WF, DepRel, HPOS	45	41	7	23	34	410
9	WF, Lemma, POS, Ling, DepRel	34	47	5	30	34	433
10	WF, Lemma, POS, Ling, DepRel, HLemma	101	**32**	0	8	9	242
11	WF, POS, EP	19	49	4	34	44	485
12	WF, POS, Ling, EP	19	49	3	39	40	482
13	WF, EP, EoV	20	49	2	41	38	478
14	WF, POS, EP, EoV	19	50	3	31	47	487
15	WF, Ling, EP, EoV	19	48	5	37	41	483
16	WF, POS, Ling, EP, EoV	19	49	5	37	40	480
17	EP, EoV	**15**	41	10	44	40	**503**
18	EP, EoV, Ling	20	49	7	38	36	471
19	Google	0	2	20	52	76	652
20	Reference	0	0	5	51	94	689

Table 4: Manual evaluation of the grammaticality

exist quite extensive implemented formal HPSG grammars for English (Copestake and Flickinger, 2000), German (Müller and Kasper, 2000), and Japanese (Siegel, 2000; Siegel and Bender, 2002). HPSG is the underlying theory of the international initiative LinGO 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 first version of the Bulgarian Resource Grammar - BURGER. In the research reported here, we use the linguistic modeled knowledge from the existing English and Bulgarian grammars. Since the Bulgarian grammar has limited coverage on news data, dependency parsing has been performed instead. Then, mapping rules have been defined for the construction of RMRSes.

However, the MRS representation is still quite close to the syntactic level, which is not fully language independent. This requires a *transfer* at the MRS level, if we want to do translation from the source language to the target language. The transfer 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 included 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 involving 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. We want to make use of the grammar resources while keeping the robustness, therefore, we experiment with another way of transfer involving information derived from the grammars.

In our approach, we take an SMT system as our 'backbone' which robustly delivers some translation for any given input. Then, we 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.

8 Conclusion and Future Work

In this paper, we report our work on building a linguistically-augmented statistical machine translation model from Bulgarian to English.

ID	Model	1	2	3	4	5	Total
1	WF	20	46	5	23	56	499
2	WF, POS	20	48	5	24	53	492
3	WF, Lemma, POS, Ling	20	47	1	24	58	503
4	Lemma	15	**32**	5	**33**	**65**	**551**
5	Lemma, POS	15	35	9	32	59	**535**
6	Lemma, POS, Ling	20	48	5	22	55	494
7	WF, DepRel	32	49	4	14	51	453
8	WF, DepRel, HPOS	45	41	2	21	41	422
9	WF, Lemma, POS, Ling, DepRel	34	48	3	20	45	444
10	WF, Lemma, POS, Ling, DepRel, HLemma	101	**32**	0	6	11	244
11	WF, POS, EP	19	49	3	20	59	501
12	WF, POS, Ling, EP	19	50	2	20	59	500
13	WF, EP, EoV	19	50	4	16	61	500
14	WF, POS, EP, EoV	19	50	2	23	56	497
15	WF, Ling, EP, EoV	19	48	4	18	61	504
16	WF, POS, Ling, EP, EoV	19	50	3	24	54	494
17	EP, EoV	**14**	38	7	31	60	**535**
18	EP, EoV, Ling	19	49	7	20	55	493
19	Google	1	0	9	42	98	686
20	Reference	1	0	5	37	107	699

Table 5: Manual evaluation of the content

Based on our observations of the previous approaches on transfer-based MT models, we decide to build a hybrid system by combining an SMT system with deep linguistic resources. We perform a preliminary evaluation on several configurations of the system (with different linguistic knowledge). The high BLEU score shows the high quality of the translation delivered by the SMT baseline; and manual analysis confirms the consistency of the system.

There are various aspects we can improve the ongoing project: 1) The MRSes are not fully explored yet, since we have only considered the EP and EoV features. 2) We would like to add factors on the target language side (English) as well. 3) The guideline of the manual evaluation needs further refinement for considering the missing lexicons as well as how much of the content is *truly* conveyed (Farreús et al., 2011). 4) We also need more experiments to evaluate the robustness of our approach in terms of out-domain tests.

Acknowledgements

This work was supported by the EuroMatrix-Plus project (IST-231720) funded by the European Community under the Seventh Framework Programme for Research and Technological Development. The authors would like to thank Tania Avgustinova for fruitful discussions and her helpful linguistic analysis; and also to Laska Laskova, Stanislava Kancheva and Ivaylo Radev for doing the human evaluation of the data.

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. *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*.

Yu Chen, M. Jellinghaus, A. Eisele, Yi Zhang, S. Hunsicker, S. Theison, Ch. Federmann, and H. Uszkoreit. 2009.

Combining multi-engine translations with moses. In *Proceedings of the 4th 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.

Mireia Farreús, Marta R. Costa-jussà, and Maja Popović Morse. 2011. Study and correlation analysis of linguistic, perceptual and automatic machine translation evaluations. *Journal of the American Society for Information Sciences and Technology*, 63(1):174–184, October.

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.

Jess Gimnez and Llus Mrquez. 2004. Svmtool: A general pos tagger generator based on support vector machines. In *Proceedings of the 4th LREC*.

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. 2007. Factored translation models. In *Proceedings of EMNLP*.

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 *Proceedings of ACL (demo session)*.

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.

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.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1).

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*.

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.

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.

Andreas Stolcke. 2002. Srilm – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2.

Gregor Thurmair. 2005. Hybrid architectures for machine translation systems. *Language Resources and Evaluation*, 39(1).

Gregor Thurmair. 2009. Comparing different architectures of hybrid machine translation systems. In *Proceedings of MT Summit XII*.