Linguistically-Enriched Models for Bulgarian-to-English Machine Translation

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-enriched Bulgarian-to-English statistical machine translation model, which takes a statistical machine translation (SMT) system as backbone various linguistic features as factors. The motivation is to take advantages of both the robustness of the SMT system and the rich linguistic knowledge from morphological analysis as well as the hand-crafted grammar resources. The automatic evaluation has shown promising results and our extensive manual analysis confirms the high quality of the translation the system delivers. The whole framework is also extensible for incorporating information provided by different sources.

1 Introduction

Incorporating linguistic knowledge into statistical models is an everlasting topic in natural language processing. The same story happens in the machine translation community. Along with the success of statistical machine translation (SMT) models (summarized by Koehn (2010)), various approaches have been proposed to include linguistic information, ranging from early work by Wu (1997) to recent work by Chiang (2010), from deep transfer-based models (Graham and van Genabith, 2008) to mapping rules at the syntactic level (Galley et al., 2004; Liu et al., 2006; Zhang et al., 2008). Although the purely data-driven approaches achieve significant results as shown in the evaluation campaigns (Callison-Burch et al., 2011), according to the human evaluation, the final outputs of the SMT systems are still far from satisfactory.

Koehn and Hoang (2007) proposed a factored SMT model as an extension of the traditional phrase-based SMT model, which opens up an easy way to incorporate linguistic knowledge at the token level. Birch et al. (2007) and Hassan et al. (2007) have shown the effectiveness of adding supertags on the target side, and Avramidis and Koehn (2008) have focused on the source side, translating a morphologically-poor language (English) to a morphologically-rich language (Greek). However, all of them attempt to enrich the English part of the language pairs being translated. For the language pairs like Bulgarian-English, there has not been much study on it, mainly due to the lack of resources, including corpora, preprocessors, etc, on the Bulgarian part. 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 popular domains like news. They reported a very high BLEU score (Papineni et al., 2002) on the Bulgarian-English translation direction (61.3).

Apart from being morphologically-rich, Bulgarian has a number of challenging linguistic phenomena to consider, including free word order, long distance dependency, coreference relations, clitic doubling, etc. For instance, the following two sentences:

(1) Momcheto j go dava buketa na momicheto.
Boy-the her-dat it-acc gives bouquet-the to girl-the.
The boy gives the bouquet to the girl.

(2) Momcheto j go dava.
Boy-the her-dat it-acc gives.
The boy gives it to her.

Proceedings of SSST-6, Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 10–19,
Jeju, Republic of Korea, 12 July 2012.

are difficult for the traditional phrase-based SMT system, because the clitic in the first sentence must not be translated, while in the second case it is obligatory. Via the semantic analysis (e.g., Minimal Recursion Semantics), the clitic information will be incorporated in the representation of the corresponding arguments.

In this work, we rely on the linguistic processing to cope with some of these phenomena and improve the correspondences between the two languages: 1) The lemmatization factors out the difference between word forms and ensures better coverage of the Bulgarian-English lexicon. 2) The dependency parsing helps to identify the grammatical functions such as subject, object in sentences with a non-standard word order. 3) The semantic analysis provides a further abstraction which hides some of the language specific features. Example of the last is the case of clitic doubling.

As for the Bulgarian-to-English translation model, we basically 'annotate' the SMT baseline with various linguistic features derived from the preprocessing and hand-crafted grammars. There are three contributions of this work:

- The models trained on a decent amount of parallel corpora output **surprisingly good results**, in terms of automatic evaluation metrics.
- The enriched models give us more space for experimenting with **different linguistic features** without losing the 'basic' robustness.
- According to our **extensive manual analyses**, the approach has shown promising results for future integration of more knowledge from the continued advances of the deep grammars.

The rest of the paper will be organized as follows: Section 2 briefly introduces some background of the hand-crafted grammar resources we use and also some previous related work on transfer-based MT. Section 3 describes the linguistic analyses we perform on the Bulgarian text, whose output is used in the factored SMT model. We show our experiments in Section 4 as well as both automatic and detailed manual evaluation of the results. We summarize this paper in Section 5 and point out several directions for future work.

2 Machine Translation with Deep Grammars

Our work is also enlightened by another line of research, transfer-based MT models using deep linguistic knowledge, which are seemingly different but actually very related. In this section, before we describe our model of incorporating linguistic knowledge from the hand-crafted grammars, we firstly introduce the background of such resources as well as some previous work on MT using them.

Our usage of Minimal Recursion Semantic (MRS) 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). Although being a semantic representation, MRS 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)[1], 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) used their generator to generate text fragments instead of full sentences, in order to increase the robustness.

In our approach, 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 particular, we take a robust SMT system as our 'backbone' and then we augment it with deep linguistic knowledge. In general, what we are doing

[1] Although their grammars are automatically induced from treebanks, the formalism supports rich linguistic information.

is still along the lines of previous work utilizing deep grammars, but we build a more 'light-weighted' but yet extensible *statistical transfer* model.

3 Factor-based SMT Model

Our translation model is built on top of the factored 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, or other linguistic features, if they can be (somehow) represented as annotations to each token.

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. Instead, 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 within and out of our project.

Firstly, the data was processed by the NLP pipe for Bulgarian (Savkov et al., 2012) including a morphological tagger, GTagger (Georgiev et al., 2012), a lemmatizer and a dependency parser[2]. Then 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, 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.
- DEPREL is the dependency relation between the current word and the parent node.
- HLEMMA is the lemma of the parent node.
- HPOS is the POS tag of the 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 are presented in Table 1.

As for the deep linguistic knowledge, we also extract features from the semantic analysis — Minimal Recursion Semantics (MRS). 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 predicate contains exactly four components: 1) a handle which is the label of

[2]We have trained the MaltParser[3] (Nivre et al., 2007) on the dependency version of BulTreeBank: `http://www.bultreebank.org/dpbtb/`. The trained model achieves 85.6% labeled parsing accuracy.

No	WF	Lemma	POS	Ling	DepRel	HLemma	HPOS
1	spored	spored	R	_	adjunct	zloupotrebyavam	VP
2	odita	odit	Nc	npd	prepcomp	spored	R
3	v	v	R	_	mod	odit	Nc
4	elektricheskite	elektricheski	A	pd	mod	kompaniya	Nc
5	kompanii	kompaniya	Nc	fpi	prepcomp	v	R
6	politicite	politik	Nc	mpd	subj	zloupotrebyavam	Vp
7	zloupotrebyavat	zloupotrebyavam	Vp	tir3p	root	-	-
8	s	s	R	_	indobj	zloupotrebyavam	Vp
9	dyrzhavnite	dyrzhaven	A	pd	mod	predpriyatie	Nc
10	predpriyatiya	predpriyatie	Nc	npi	prepcomp	s	R

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

No	EP	EoV	EP_1/POS_1	EP_2/POS_2	EP_3/POS_3
1	spored_r	e	zloupotrebyavam_v/Vp	odit_n/Nc	-
2	odit_n	v	-	-	-
3	v_r	e	odit_n/Nc	kompaniya_n/Nc	-
4	elekticheski_a	e	kompaniya_n/Nc	-	-
5	kompaniya_n	v	-	-	-
6	politik_n	v	-	-	-
7	zloupotrebyavam_v	e	politik_n/Nc	-	s_r/R
8	s_r	e	zloupotrebyavam_v/Vp	predpriyatie_n/Nc	-
9	dyrzhaven_a	e	predpriyatie_n/Nc	-	-
10	predpriyatie_n	v	-	-	-

Table 2: Representation of MRS factors for each wordform in the sentence.

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).

Robust MRS (RMRS) is introduced as a modification of MRS which captures 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 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 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 parts of speech. 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 RMRS structures and how they are constructed.

As for the factors, 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.
- ARG_nEP indicates the elementary predicate of the argument which belongs to the predicate. n is usually from 1 to 3.
- ARG_nPOS indicates the POS tag of the argument which belongs to the predicate.

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. Those kinds of information is not straightforward to be represented in such 'tagging'-style models, which will be tackled in the future.

The extra information for the example sentence is represented in Table 2. All these factors encoded

within the corpus provide us with a rich selection of features for different experiments.

4 Experiments

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 SETIMES parallel corpus has been used, which is part of the OPUS parallel corpus[5]. As for the choice of the datasets, the language is more diverse in the news articles, compared with other corpora in more controlled settings, e.g., the JRC-Acquis corpus[6] used by Koehn et al. (2009).

We split the corpus into the training set and the test set by 150,000 and 1,000 sentence pairs respectively[7]. Both datasets are preprocessed with the tokenizer and lowercase converter provided by Moses. Then the procedure is quite standard: We run GIZA++ (Och and Ney, 2003) for bi-directional 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). For the rest of the parameters we use the default setting provided by Moses.

Notice that, since on the target language side (i.e., English) we do not have any other factors than the word form, the factor-based models we use here only differentiate from each other in the translation phase, i.e., there is no 'generation' models involved.

4.1 Automatic Evaluation Metrics

The baseline results (non-factored model) under the standard evaluation metrics are shown in the first row of Table 3 in terms of BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2011). We then design various configurations to test the effectiveness of different linguistic annotations described in Section 3. The detailed configurations we considered are shown in the first column of Table 3.

The first impression is that the BLEU scores in general are high. These models can be roughly grouped into six categories (separated by double lines): 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); EP features without word forms; and EP features with MRS ARG$_n$ features.

In terms of the resulting scores, POS and Lemma seem to be effective features, as Model 2 has the highest BLEU score and Model 4 the best METEOR 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. Incorporating HLEMMA feature largely decreases the results due to the vastly increasing vocabulary, i.e., aligning and translating bi-grams instead of tokens. Therefore, we did not include the results in the table. After replacing the HLEMMA with HPOS, the result is close to the others (Model 8). Model 9 may also indicate that increasing the number of factors does not guarantee performance enhancement. The experiments with predicate features (EP and EOV) from the MRS analyses (Model 10-12) show improvements over the baseline consistently and using only the MRS features (Model 13-14) also delivers descent results. Concerning the MRS ARG$_n$ features, the models with ARG$_n$EP again suffer from the sparseness problem as the dependency HLEMMA features, but the models with ARG$_n$POS (Model 15-16) achieve better performance than those with dependency HPOS features. This is mainly because the dependency information is encoded together with the (syntactically) dependent word, while the MRS arguments are grouped around the semantic heads.

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 automatic evaluation scores already indicate that the system can deliver decent translation quality consistently.

4.2 Manual Evaluation

We manually validated the output for all the models mentioned in Table 3. The guideline includes two

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

[5] OPUS — an open source parallel corpus, http://opus.lingfil.uu.se/

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

[7] We did not preform MERT (Och, 2003), as it is quite computationally heavy for such various configurations.

ID	Model	BLEU	1-gram	2-gram	3-gram	4-gram	METEOR
1	WF (Baseline)	38.61	**69.9**	44.6	31.5	22.7	0.3816
2	WF, POS	**38.85**	**69.9**	**44.8**	**31.7**	**23.0**	0.3812
3	WF, LEMMA, POS, LING	38.84	**69.9**	44.7	**31.7**	**23.0**	0.3803
4	LEMMA	37.22	68.8	43.0	30.1	21.5	**0.3817**
5	LEMMA, POS	37.49	68.9	43.2	30.4	21.8	0.3812
6	LEMMA, POS, LING	38.70	69.7	44.6	31.6	22.8	0.3800
7	WF, DEPREL	36.87	68.4	42.8	29.9	21.1	0.3627
8	WF, DEPREL, HPOS	36.21	67.6	42.1	29.3	20.7	0.3524
9	WF, LEMMA, POS, LING, DEPREL	36.97	68.2	42.9	30.0	21.3	0.3610
10	WF, POS, EP	38.74	69.8	44.6	31.6	22.9	0.3807
11	WF, EP, EOV	38.74	69.8	44.6	31.6	22.9	0.3807
12	WF, POS, LING, EP, EOV	38.76	69.8	44.6	**31.7**	22.9	0.3802
13	EP, EOV	37.22	68.5	42.9	30.2	21.6	0.3711
14	EP, EOV, LING	38.38	69.3	44.2	31.3	22.7	0.3691
15	EP, EOV, ARG$_n$POS	36.21	67.4	41.9	29.2	20.9	0.3577
16	WF, EP, EOV, ARG$_n$POS	37.37	68.4	43.2	30.3	21.8	0.3641

Table 3: Results of the factor-based model (Bulgarian-English, SETIMES 150,000/1,000)

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 for one Cyrillic token and score 1 for more than one tokens in the output translation. We have two annotators achieving the inter-annotator agreement according to Cohen's Kappa (Cohen, 1960) $\kappa = 0.73$ for grammaticality and $\kappa = 0.75$ for content, both of which are *substantial* agreement. For the conflict cases, we take the average value of both annotators and rounded the final score up or down in order to have an integer.

The current results from the manual validation are on the basis of randomly sampled 150 sentence pairs. The numbers shown in Table 4 are the number of sentences given the corresponding scores. The 'Sum' column shows the average score of all the output sentences by each model and the 'Final' column shows the average of the two 'Sum' scores.

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, although incorporating the MRS ARG$_n$ features shows similar performance drops as dependency features. Including more factors in general reduces the results because of the sparseness effect over the dataset, which is consistent with the automatic evaluation. The last two rows are shown

ID	Model	Grammaticality						Content						Final
		1	2	3	4	5	Sum	1	2	3	4	5	Sum	
1	WF (Baseline)	20	47	5	32	**46**	3.25	20	46	5	23	56	3.33	3.29
2	WF, POS	20	48	5	37	40	3.19	20	48	5	24	53	3.28	3.24
3	WF, LEMMA, POS, LING	20	47	6	34	43	3.22	20	47	1	24	58	3.35	3.29
4	LEMMA	**15**	34	11	46	44	**3.47**	15	32	5	**33**	**65**	**3.67**	**3.57**
5	LEMMA, POS	**15**	38	12	**51**	34	**3.34**	15	35	9	32	59	**3.57**	**3.45**
6	LEMMA, POS, LING	20	48	5	34	43	3.21	20	48	5	22	55	3.29	3.25
7	WF, DEPREL	32	48	3	29	38	2.95	32	49	4	14	51	3.02	2.99
8	WF, DEPREL, HPOS	45	41	7	23	34	2.73	45	41	2	21	41	2.81	2.77
9	WF, LEMMA, POS, LING, DEPREL	34	47	5	30	34	2.89	34	48	3	20	45	2.96	2.92
10	WF, POS, EP	19	49	4	34	44	3.23	19	49	3	20	59	3.34	3.29
11	WF, EP, EOV	20	49	2	41	38	3.19	19	50	4	16	61	3.33	3.26
12	WF, POS, LING, EP, EOV	19	49	5	37	40	3.20	19	50	3	24	54	3.29	3.25
13	EP, EOV	**15**	41	10	44	40	**3.35**	**14**	38	7	31	60	**3.57**	**3.46**
14	EP, EOV, LING	20	49	7	38	36	3.14	19	49	7	20	55	3.29	3.21
15	EP, EOV, ARG$_n$POS	23	49	9	34	35	3.06	23	47	8	**33**	39	3.12	3.09
16	WF, EP, EOV, ARG$_n$POS	34	47	10	30	29	2.82	34	47	10	20	39	2.89	2.85
*	GOOGLE	0	2	20	52	76	4.35	1	0	9	42	98	4.57	4.46
*	REFERENCE	0	0	5	51	94	4.59	1	0	5	37	107	4.66	4.63

Table 4: Manual evaluation of the grammaticality and the content

for reference. 'Google' shows the results of using the online translation service provided by `http://translate.google.com/` on 06.02.2012. 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.

Concerning the impact from the linguistic processing pipeline to the final translation results, Lemma and MRS elementary predicates help at the level of rich morphology. For example, the baseline model correctly translates the adjective 'Egyptian' in 'Egyptian Scientists' (plural), but not in 'Egyptian Government, as in the second phrase the adjective has a neutral gender. Model 4 and Model 13 are correct for both.

Generally speaking, if we roughly divide the linguistic processing pipeline in two categories: statistical processing (POS tagger and dependency parser) and rule-based processing (lemmatizer and MRS construction), the latter category (almost perfect) highly relies on the former one. For example, the lemma depends on the word form and the tag, and the result is unambiguous in more than 98% of the morphological lexicon and in text this is almost 100% (because the ambiguous cases are very rare). The errors come mainly from new words and errors in the tagger. Similarly, the RMRS rules are good when the parser is correct. Here, the main problems are duplications of the ROOT elements and the subject elements, which we plan to fix using heuristics in the future.

4.3 Question-Based Evaluation

Although the reported manual evaluation in the previous section demonstrates that linguistic knowledge improves the translation, we notice that the evaluators tend to give marks at the two ends of scale, and less in the middle. Generally, this is because the measurement is done on the basis of the content that the evaluators extract from the Bulgarian sentence using there own cognitive capacity. Then they start to overestimate or underestimate the translation, knowing in advance what has to be translated. In order to avoid this subjectivity, we design a different manual evaluation in which the evaluator does not know the original Bulgarian sentences. Then the evaluation is based only on the content represented within the English translation.

In order to do this, we represent the content of the Bulgarian sentences as a set of questions that have a list of possible answers, assigned to them. During the judgement of the content transfer, the evaluators

need to answer these questions. As the list of answers also contains false answers, the evaluators are forced to select the right answer which can be inferred from the English translation.

The actual questions are created semi-automatically from the dependency analysis of the sentences. We defined a set of rules for generation of the questions on the basis of the dependency relations. For example, if a sentence has only a subject relation presented within the analysis, the question will be about who is doing the event. If the analysis presents subject and direct object, the question will be about who is doing something with what/whom. These automatically generated questions are manually investigated and, if necessary, edited. Also, additional answers are formulated on the basis of general language knowledge. The main idea is that the possible answers are conceptually close to each other, but not in a hypernymy relation. Always there is an answer "none".

Then the questions are divided into small groups and distributed to be answered by three evaluators in such a way that each question is answered by two evaluators, but no evaluator answers the whole set of questions for a given sentence. In this way, we try to minimize the influence of one question to the answers of the next questions. The answers are compared to the true answers of the questions for each given sentence. We evaluated 192 questions for each model and sum up the scores (correctly answered questions) in Table 5.

This evaluation is more expensive, but we expect them to be more objective. As for a related work, (Yuret et al., 2010) used textual entailment to evaluate different parser outputs. The way they constructed the *hypotheses* is similar to our creation of questions (based on dependency relations). However, they focused on the automatic evaluation and we adopt it for the manual evaluation.

ID	Model	Score
1	WF (Baseline)	127
2	WF, POS	126
3	WF, LEMMA, POS, LING	131
4	LEMMA	**133**
5	LEMMA, POS	**133**
6	LEMMA, POS, LING	128
7	WF, DEPREL	131
8	WF, DEPREL, HPOS	120
9	WF, LEMMA, POS, LING, DEPREL	124
10	WF, POS, EP	125
11	WF, EP, EOV	126
12	WF, POS, LING, EP, EOV	128
13	EP, EOV	**138**
14	EP, EOV, LING	122
15	EP, EOV, ARG$_n$POS	130
16	WF, EP, EOV, ARG$_n$POS	121

Table 5: Question-based evaluation

5 Conclusion and Future Work

In this paper, we report our work on building a linguistically-enriched statistical machine translation model from Bulgarian to English. Based on our observations of the previous approaches on transfer-based MT models, we decide to build a factored model by feeding an SMT system with deep linguistic features. We perform various experiments 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 various manual analyses confirm the consistency of the system.

There are various aspects of the current approach we can improve: 1) The MRSes are not fully explored yet, although we have considered the most important predicate and argument features. 2) We would like to add factors on the target language side (English) as well to fulfill a 'complete' transfer. 3) Incorporating reordering rules on the Bulgarian side may help the alignment and larger language models on the English side should also help improving the translation results. 4) Due to the morphological complexity of the Bulgarian language, the other translation direction, from Bulgarian to English, is also worth investigation in this framework.

Acknowledgements

This work was partially supported by the EuroMatrixPlus project (IST-231720) funded by the European Community's Seventh Framework Programme. The authors would like to thank Laska Laskova, Stanislava Kancheva and Ivaylo Radev for doing the human evaluation of the data.

References

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proceedings of ACL*.

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

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*.

David Chiang. 2010. Learning to translate with source and target syntax. In *Proceedings of ACL*, pages 1443–1452.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

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.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of HLT-NAACL*, Boston, Massachusetts, USA, May.

G. Georgiev, V. Zhikov, P. Osenova, K. Simov, and P. Nakov. 2012. Feature-rich part-of-speech tagging for morphologically complex languages: Application to bulgarian. In *EACL 2012*.

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.

Y. Graham, A. Bryl, and J. 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)*.

P. Koehn, A. Birch, and R. Steinberger. 2009. 462 machine translation systems for europe. In *Proceedings of MT Summit XII*.

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

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of COLING-ACL*, pages 609–616.

Joakim Nivre, Jens Nilsson, Johan Hall, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. Maltparser: a language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(1):1–41.

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.

K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Aleksandar Savkov, Laska Laskova, Stanislava Kancheva, Petya Osenova, and Kiril Simov. 2012. Linguistic processing pipeline for bulgarian. In *Proceedings of LREC*, Istanbul, Turkey.

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.

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

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404, September.

Deniz Yuret, Aydın Han, and Zehra Turgut. 2010. Semeval-2010 task 12: Parser evaluation using textual entailments. In *Proceedings of the SemEval-2010 Evaluation Exercises on Semantic Evaluation*.

Min Zhang, Hongfei Jiang, Aiti Aw, Haizhou Li, Chew Lim Tan, and Sheng Li. 2008. A tree sequence alignment-based tree-to-tree translation model. In *Proceedings of ACL-HLT*, pages 559–567.