Morphosyntactic Tags in Statistical Machine Translation of Highly Inflectional Language

Mirjam Sepesy Maučec, Gregor Donaj
University of Maribor
Maribor, SLOVENIA
{mirjam.sepesy, gregor.donaj}@um.si

***Abstract*—In this paper, we investigate the usefulness of morphosyntactic information in statistical machine translation between English and Slovenian, which is a highly inflectional language. Translation in both directions is explored, while translation to inflectional language is a more challenging task. Morphosyntactic tags are attached to words by two different taggers (TreeTagger and Obelix) and utilized during translation in three different ways: for N-best list re-scoring, factored translation and OSM modeling. We investigate the usefulness of a complete set of morphosyntactic tags and a reduced set, containing only the most relevant morphosyntactic information. The results show that morphosyntactic information is an important factor in translation. When used in factored translation with OSM models it improves the BLEU score by almost 10% relative if translating from English to Slovenian, and by 2% if translating from Slovenian to English.**

I. Introduction

Statistical machine translation (SMT) has made significant progress in the last two decades, however, the results for morphologically rich languages are still weak. Translation into languages with rich morphology is a special challenge, as rich morphology causes data sparsity problems. Larger corpora are needed for estimating the translation model parameters. For many languages, the parallel corpora are of limited size. The research has shown that integrating linguistic knowledge could reduce data sparsity and improve the results. Morphological factors used in multi-factored phrase-based translation improved translation accuracy [2], [11]. Recently defined operation sequence n-gram models over morphological factors further outperformed state-of-the-art systems [7]. Morphological analysis was also used in two phases decoding, where the appropriate target inflection was selected in the second phase [3], [4].

In this paper, we report on experiments with English-Slovenian translation pair. Slovenian is a highly inflectional South Slavic language. It shares many linguistic characteristics with other Slavic languages, like Russian, Czech, Serbian etc. We examine translations in both directions. There is a big difference between translation from a morphologically-rich language and translation to a morphologically-rich language. If we are translating from morphologically-rich languages, the idea is to reduce the sparsity caused by the rich morphology of the source language through some form of morphology reduction. If we are translating to morphologically rich languages, some sort of morphology generation is needed.

In this paper, we are using morphosyntactic information in three different ways. First, we examine re-ranking the N-best list by external language model of morphosyntactic tags. In the second set of experiments different factored translation models are built. Surface forms, lemmas, and morphosyntactic tags are used as factors, that make various combinations in translation model possible. Finally, factored word representations are used in operation sequence models (OSM) that are added as additional features to the final factored translation models.

II. Corpora and baseline system

We investigate the models using the Slovenian-English parallel corpus from the Europarl corpus v7 (see http://www.statmt.org/europarl/). The corpus contains 623,490 sentences (14M Slovenian tokens, 16M English tokens). The corpus was split into training, development (2,000 sentences) and testing (2,000 sentences) sets. All words that appear in the training set were added to the vocabularies. Slovenian vocabulary contains 144,671 words and English vocabulary 66,604 words. The corpus was true-cased and tokenized before the SMT systems training took place. Standard Moses phrase-based SMT was used as the baseline system. All conditions, not only baseline, use word alignments produced by sequential iterations of IBM model 1, HMM, and IBM models 3 and 4 in GIZA++, followed by "grow-diag-final-and" symmetrization [12]. A 3-gram language model with modified Kneser-Ney discounting was built on the training corpus by the SRILM toolkit [15]. Singletons were excluded. The perplexity of Slovenian language model was 109, and of English 62. For all the setups we perform standard MERT training on the defined development set. In all experiments, the resulting translations were detruecased and detokenized before evaluated. We use standard BLEU [17] as a metric for evaluation. It is in no way tolerant to errors of inflected languages. Some modifications were proposed in literature [18]. Using them remains for our future work.

III. Morphosyntactic information

In sparse data conditions, it is reasonable to use a more generalized representation of words. The representation can reflect morphosyntactic features of words. The most basic morphosyntactic information is the information about part of speech. Part of speech (POS) tags assign words the corresponding grammatical categories: verb, noun, adjective, pronoun, etc. There are approximately 10 basic POS tags. Morphosyntactic tags or morphosyntactic descriptions (MSD) are tags in which additional subcategories are included, such as gender and case for nouns or tense and person for verbs. There is a great

ISBN 978-952-68397-8-3 (PDF)

TABLE I. SLOVENIAN POSS WITH FULL ATTRIBUTES IN MSD TAGS. IN BRACKETS IS THE NUMBER OF DIFFERENT VALUES FOR EACH ATTRIBUTE

POS	Attributes
Noun (N)	type (2), gender (3), number(3), case (6), animate (2)
Verb (V)	type (2), aspect (3), verb form (7), person (3), number (3), gender (3), negative (2)
Adjective (A)	type (3), degree (3), gender (3), number(3), case (6), definiteness (2)
Adverb (R)	type (2), degree (3)
Pronoun (P)	type (9), person (3), gender (3), number (3), case (6), owner number (3), owner gender (3), clitic (2)
Numeral (M)	form (3), type (4), gender (3), number (3), case (6), definiteness (2)
Adposition (S)	case (6)
Conjunction (C)	type (2)
Particle (Q)	-
Interjection (I)	-
Abbreviation (Y)	-
Residual (X)	type (7)
Punctioation (Z)	-

diversity between the number of different tags defined per language. Penn Treebank defines 36 tags for English [13].

Due to the rich morphosyntactic complexity of highly inflectional languages, there are for example 3922 plausible MSD tags defined for Czech (although only 1571 unique tags actually appear in most corpora) [16]. In the MULTEXT-East project standardized MSD tagsets for six Central and Eastern European languages were developed [8]. The latest release (Version 5 (http://nl.ijs.si/ME/V5/msd/html/)) for example defines: 135 tags for English, 1425 for Czech, 17279 tags for Hungarian, and 1903 MSD tags for Slovenian language. Table I lists Slovenian POS tags with attributes. In our experiments 58 tags were used for English (see https://www.sketchengine.co.uk/penn-treebank-tagset/) and 1903 for Slovenian (see http://nl.ijs.si/ME/V5/msd/html/msd-sl.html). Fig. 1 shows an example of annotated part of a sentence in English and in Slovenian.

agenda|NN for|IN next|JJ sitting|NN

dnevni|Agpmsny red|Ncmsn naslednje|Agpfsg seje|Ncfsg

Fig. 1. Part of an English sentence and Slovenian translation with MSD tags

The corpus was annotated by TreeTagger [14]. Slovenian part of the corpus was additionally tagged with Obelix tagger [9] and the results were compared. In literature the accuracy of 96.32% was reported for TreeTagger when annotating English, and the accuracy of 92.49% for the Obelix when annotating Slovenian. To the best of our knowledge the accuracy for TreeTagger when annotating Slovenian has not been published yet. After annotating our experimental corpus, the analysis showed that in the English part 54 different tags were used and in Slovenian part 977 tags.

A. Reduced tags

Highly inflectional languages face the problem of data sparsity. Using the extended set of MSD tags does not reduce it to a great extend. For translation complete morphosyntactic information of inflectional language is not needed, especially when paired with English. We reduced the tags to include only the most important attributes. The decision on attributes was experiential. Attributed that were kept are given in table II. For all other categories just POS tag is kept, with no additional attributes. For example a full attributed tag for an adjective "*zadnji*" (eng. last) is Agpmsay , and we reduced it to A--msa- . It could be that there are combinations of attributes that are more effective in translation. The presented schema was just a method of trial. Some experiments were performed using both the full and the restricted tagging scheme.

TABLE II. REDUCED SETS OF ATTRIBUTES IN MSD TAGS

POS	Attributes
Noun	gender, number, case
Verb	person, gender, number
Adjective	gender, number, case
Pronoun	person, gender, number, case
Numeral	gender, number, case

TABLE III. THE BLEU RESULTS OF TRANSLATION FROM ENGLISH TO SLOVENIAN WITHOUT/WITH RE-RANKING N-BEST LISTS

	English-Slovenian
Baseline	30.45
re-ranking (1000-best)	30.48

IV. RE-RANKING N-BEST LISTS

In the first set of experiments morphosyntactic information is used in two-pass decoding. In the first pass the baseline system was used to produce the output containing 1000-best hypotheses for each input sentence. The output underwent the second pass decoding, where the MSD language model was used to re-score the N-best list of translations. The hypotheses were annotated with MSD tags and MSD language model score was added as a new feature weight. The hypotheses were re-ranked according to a new cumulative scoring function:

$$score(e, f) = \lambda_1 \cdot score_b(e, f) + \lambda_2 \cdot score_{MSD}(f). \quad (1)$$

By e we denote the sentence in original language and by f the translation hypotheses. $score_b(e, f)$ is the original score given by the baseline system, and $score_{MSD}(f)$ is the score given by MSD language models used in re-scoring. λ_1 and λ_2 are the weights for both scores. We applied re-scoring only for English to Slovenian translation. The results are given in table III. In the first row the results obtained with the baseline systems are given. Only a small difference in BLEU score can be observed. We think that reasonable improvements could be obtained with re-scoring methods only if we have high-quality baseline system that generates good hypotheses for re-scoring.

V. FACTORED TRANSLATION MODELS

In experiments with re-scoring MSD information was used in the second pass of decoding. MSD tags can be used earlier, i.e. in the first pass, as part of the translation models training. Factored translation models were proved to be a useful framework for that [11], [2], [10]. Factored translation models are based on factored representation of words. We use the representation of words with three factors: surface form (S), lemma (L), and MSD tag (M). Different scenarios can be defined, based on factors. We investigated the following scenarios for translations in both directions(we follow the taxonomy in [1]):

- tSaM-SaM: translation of surface form and MSD tag in the source language to surface form and MSD tag in the target language,
- tSaMaL-SaMaL: translation of surface form, MSD tag, and lemma in the source language to surface form, MSD tag and lemma in the target language.

For translation from English to Slovenian, we added the following scenarios:

- tS-SaM: translation of surface form in the source language to surface form and MSD tag in the target language,
- tS-SaMaL: translation of surface form in the source language to surface form, MSD tag, and lemma in the target language,

In these scenarios target language was morphologically more complex, therefore the decomposition into factors was used on the target side.

For translation from Slovenian to English, we added the following scenarios:

- tL-S: translation of lemma in the source language to surface form in the target language,
- tLaM-SaM: translation of lemma and MSD tag in the source language to surface form and MSD tag in the target language.

In these scenarios target language was from a morphological point of view less complex, therefore morphological information on source side was not seen to be important for translation.

Other linguistically motivated scenarios were defined in literature (like tL-L+tM-M+gLaM-S). Due to the increased complexity of the setups, we were not able to train them. Factored configurations can include language models of different types. In our experiments, language models of MSD tags and lemmas were used in addition to surface form language models. Perplexities of language models are given in table IV.

A particular language model was always used, if the scenarios made the use of it possible. For example, if the scenario includes all three factors on the target side, all three types of language models were used: surface LM, MSD LM , and lemma LM.

The results of factored translations are given in tables V and VI In the first row, the results of the baseline systems are given for comparison. Each factored configuration was run twice, once with factors generated by TreeTagger and once with Obelix factors on Slovenian side.

TABLE IV. PERPLEXITIES OF LANGUAGE MODELS

	English	**Slovenian**
Surface LM	54	90
Lemma LM (TreeTagger)	45	52
Lemma LM (Obelix)	-	42
MSD LM (TreeTagger)	11	32
MSD LM (Obelix)	-	15

For English to Slovenian translation better results were obtained by using TreeTagger factors. Comparing the results based on TreeTagger factors, translation scenario tS-SaM brought the best BLEU score, whereas using Obelix factors, the scenario tSaMaL-SaMaL, where all factors were used, was the best one. Translation example of scenario tS-SaM based on TreeTagger factors is given in Fig. 2. We can notice wrong translation of negation.

Input: however , it is still not clear
Output: vendar|Cc pa|Cc je|Va-r3s-n še|Q vedno|Rgp ni|Va-r3s-y jasno|Agpnsn

Fig. 2. The translation example of scenario tS-SaM.

Comparing the results for opposite translation (Slovenian to English) only the scenario tSaM-SaM slightly improved the baseline results. All other factored scenarios caused a loss in terms of BLEU score. The TreeTagger factors brought again better BLEU scores than Obelix factors. If we eliminated surface forms on source side, the results got considerably worse. It seems that only adding MSD tag on both sides is reasonable for that translation direction.

MSD information in factored translation brought more improvements when translating from English to Slovenian than in opposite translation direction. For this translation direction the scenario tSaMaL-SaMaL was rerun using the reduced set of MSD tags. We observed only the slight drop in BLEU score.

VI. OSM MODELS

Recently operation sequence model (OSM) was defined and integrated into the phrase-based SMT [5], [6], [7]. It is a joint model for the translation and long distance reordering. OSM models translation by a linear sequence of operations. Operations generate the aligned sentence pair. An operation either generates source and target words or it performs reordering by inserting gaps and jumping forward and backward [6]. Operation sequences can be learned over words or over any other generalized representations [7]. In our experiments, OSM models were learned over surface forms, lemmas, and MSD tags. OSM models over surface forms were added to the baseline systems. Results are in the second row in tables VII and VIII. OSM models over surface forms and MSD tags were added to the factored system tSaM-SaM. Results are in the third row in both tables. Finally, for English to Slovenian translation, OSM models were additionally learned over lemmas and added to the factored system tSaMaL-SaMaL. Results are given in the last row in table VII. We can see that OSM models improved the results of all configurations. Overall the

TABLE V. THE BLEU RESULTS OF FACTORED TRANSLATION FROM ENGLISH TO SLOVENIAN

	English-Slovenian
Baseline (tS-S)	30.45
tSaM-SaM (TreeTagger)	32.19
tSaM-SaM (Obelix)	31.66
tS-SaM (TreeTagger)	32.21
tS-SaM (Obelix)	31.93
tS-SaMaL (TreeTagger)	31.31
tS-SaMaL (Obelix)	31.90
tSaMaL-SaMaL (TreeTagger)	32.08
tSaMaL-SaMaL (Obelix)	31.99
tSaMaL-SaMaL (Obelix-reduced)	31.55

TABLE VI. THE BLEU RESULTS OF FACTORED TRANSLATION FROM SLOVENIAN TO ENGLISH

	Slovenian-English
Baseline (tS-S)	40.54
tSaM-SaM (TreeTagger)	40.77
tSaM-SaM (Obelix)	40.56
tL-S (TreeTagger)	34.45
tL-S (Obelix)	31.42
tLaM-SaM TreeTagger)	37.28
tLaM-SaM (Obelix)	36.40
tSaMaL-SaMaL (TreeTagger)	39.96
tSaMaL-SaMaL (Obelix)	39.89

best results are obtained with factored system tSaM-SaM and OSM models over surface forms and MSD tags. OSM model over lemmas did not contribute to the improvement of results.

VII. CONCLUSION

We have investigated three different ways of using morphosyntactic information in phrase-based SMT. We did not gain any improvements with re-ranking N-best lists by MSD language models. In contrast, factored translation combined with OSM models outperformed the baseline system significantly. Larger improvements were obtained for English to Slovenian translation. We can conclude that morphosyntactic information improves the translation results in both directions if it is used appropriately. Similar results could reasonably be expected for other highly inflected languages when combined with English. In the future, we will analyze MSD tags and annotation schemas in more details, as there is high diversity among them. We believe that specific attributes in MSD tags play the major role in the search for the right translation. In the current research, the reduced tag set was an intuitive guess. In the future, we will try to find MSD attributes by a data-driven approach, that will base on an optimization metric.

TABLE VII. THE BLEU RESULTS OF FACTORED TRANSLATION WITH OSM FROM ENGLISH TO SLOVENIAN

	English-Slovenian
Baseline (tS-S)	30.45
Baseline (OSM: 0-0)	31.00
tSaM-SaM (OSM: 0-0, 1-1)	33.41
tSaMaL-SaMaL (OSM: 0-0, 1-1, 2-2)	32.77

TABLE VIII. THE BLEU RESULTS OF FACTORED TRANSLATION WITH OSM FROM SLOVENIAN TO ENGLISH

	Slovenian-English
Baseline (tS-S)	40.54
Baseline (OSM: 0-0)	41.49
tSaM-SaM (OSM: 0-0, 1-1)	41.81

ACKNOWLEDGMENT

This research work was partially funded by the Slovenian Research Agency ARRS under the contract number P2-0069.

REFERENCES

[1] O. Bojar, B. Jawaid, and A. Kamran, "Probes in a Taxonomy of Factored Phrase-Based Models", *In Proc. of the 7th Workshop on Statistical Machine Translation*, June 2012, pp. 253–260.

[2] O. Bojar, "English-to-Czech Factored Machine Translation", *In Proc. of the Second Workshop on Statistical Machine Translation*, June 2007, pp. 232–239.

[3] V. Chahuneau, E. Schlinger, N.A. Smith, and C. Dyer, "Translating into Morphologically Rich Languages with Synthetic Phrases", *In Proc. of Conference on Empirical Methods in Natural Language Processing*, Oct. 2013, pp. 1677–1687.

[4] J. Daiber and K. Sima'an, "Machine Translation with Source-Predicted Target Morphology", *In Proc. of the MT Summit XV*, 2015, pp. 283–296.

[5] N. Durrani, H. Schmid and A. Fraser A, "A Joint Sequence Translation Model with Integrated Reordering", *In Proc. of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-HLT)*, 2011, pp. 1045–1054.

[6] N. Durrani, A. Fraser, H. Schmid, H. Hoang and Koehn P, "Can Markov Models Over Minimal Translation Units Help Phrase-Based SMT?", *In Proc. of the 51st Annual Conference of the Association for Computational Linguistics (ACL)*, 2013, pp. 399–405.

[7] N. Durrani, P. Koehn, H. Schmid, and A. Fraser, "Investigating the Usefulness of Generalized Word Representations in SMT", *In Proc. of the 25th Annual Conference on Computational Linguistics (COLING)*, Aug. 2014, pp. 421–432.

[8] T. Erjavec, "MULTEXT-East: morphosyntactic resources for Central and Eastern European languages", *Language Resources & Evaluation*, vol. 46, 2012, pp. 131–142.

[9] M. Grčar, S. Krek, and K. Dobrovoljc, "Obeliks: statistini oblikoskladenjski oznaevalnik in lematizator za slovenski jezik", *In Proc. of the Language Technologies Conference*, 2012, pp. 82–88.

[10] S. Huet, E. Manishina, F. Lefevre, "Factored Machine Translation Systems for Russian-English", *In Proc. of the Eighth Workshop on Statistical Machine Translation*, 2013, pp. 154–157.

[11] P. Koehn and H. Hoang, "Factored Translation Models", *In Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 868–876.

[12] P. Koehn, F.J. Och and D. Marcu, "Statistical Phrase-Based Translation", *In Proc. of the Human Language Technology Conference*, 2003, pp. 48–54.

[13] M. Marcus, B. Santorini and M.A. Marcinkiewicz, "Building a Large Annotated Corpus of English: The Penn Treebank", 1993, Technical Report MS-CIS-93-87, University of Pennsylvania, Computer and Information Science Department.

[14] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees", *In Proc. of the International Conference on New Methods in Language Processing*, 1994, pp. 44–49.

[15] A. Stolcke, "SRILM an Extensible Language Modeling Toolkit", *In Proc. of the International Conference on Spoken Language Processing*, 2002, pp. 257–286.

[16] J. Straková, M. Straka, and J. Hajič, " Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition", *In Proc. of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 13–18.

[17] K. Papineni, S. Roukos, T. Ward, W.J. Zhu, "BLEU: a method for automatic evaluation of machine translation", 2004, Technical Report RC22176(W0109-022), IBM Research Report, IBM.

[18] J. Libovický and P. Pecina, "Tolerant BLEU: a Submission to the WMT14 Metrics Task", Proceedings of the Ninth Workshop on Statistical Machine Translation, 2015, pp. 409–413.