ITC 1/47
Journal of Information Technology and Control
Vol. 47 / No. 1 / 2018
pp. 63-74
DOI 10.5755/j01.itc.47.1.17887

Morphology in Statistical Machine Translation from English to a Highly Inflectional Language

Received 2017/05/25 | Accepted after revision 2018/01/12

http://dx.doi.org/10.5755/j01.itc.47.1.17887

Morphology in Statistical Machine Translation from English to a Highly Inflectional Language

Mirjam S. Maučec, Gregor Donaj
University of Maribor, Faculty of Electrical Engineering and Computer Science
Koroška c. 46, SI-2000 Maribor, Slovenia

Corresponding author: mirjam.sepesy@um.si

In this paper, we investigate the role of morphology in phrase-based statistical machine translation (SMT) from English to the highly inflectional Slovenian language. Translation to an inflectional language is a challenging task because of its morphological complexity. Rich morphology increases data sparsity and worsens the quality of statistical machine translation. The idea of the paper is to find the SMT configuration, based on morpho-syntactic information, with the best translation results, when translating from English to the highly inflectional Slovenian language. To address this issue, we added the morphological information in terms of morpho-syntactic description (MSD) tags that were attached to words. A MSD tag includes all morpho-syntactic information in position-dependent attributes. Tags were attached to words by TreeTagger. Several experiments were performed using MSD tags to improve the translation results. First, factored translation was studied, and different configurations were tested. They show that factored translation improves modeling of short distance collocations. To capture long-distance dependencies in languages, operation sequence models (OSM) were added in the second set of experiments. An additional improvement was obtained. The overall results show that the morpho-syntactic information of inflectional language is an important factor in translation. Factored translation with OSM models brought 9% relative improvement. The most successful configuration was tSaMaL-SaMaL (OSM: 0-0, 1-1, 2-2). The conclusions of our work can be applied to other Slavic languages, as they to some extent share the same morphological characteristics.

KEYWORDS: natural language processing, statistical machine translation, inflectional language, morphology.

1. Introduction

In the last decades, quite considerable progress has been made in natural language processing, not only for the English language, but also for other, more complex languages [28]. It is important to develop a

technology for a wide range of languages, even if they are spoken by smaller populations. Multilingualism is an important aspect of world's cultural diversity and provides for freedom of speech and expression. Making text or speech available in a language other than the source language involves translation. Depending on the type of content, translation can be quite expensive. For example, the Translation Directorate General of the European Commission reports 330 million euro per year for translation costs being spent. In 2012, it translated a total of 1,760,615 pages [8]. All 24 official languages of the EU enjoy equal status. With the candidate countries, the number of official languages may one day reach 29. As the European Parliament made cuts to translation budget, we are witnessing the intensive search for the solution of the problem of high translation costs in Europe. Not only Europe, the whole world is faced with the problem of translation, as the amount of multimedia material in very different languages available to the public is increasing at enormous speed. The use of machine translation may significantly reduce the costs and increase the translation speed.

Machine translation (MT) is the application of computers to the task of translating texts from one natural language to another. Natural language is enormously complex, and translation between languages is far from being an easy task. The challenge has been approached from various points of view, including linguistics, statistics, and computational linguistics. In this paper, we focus on statistical approaches that use linguistic information in terms of morpho-syntactic features. In statistical machine translation (SMT), the process of translating is modeled as a statistical decision process.

The translation task is more difficult if we are translating from a weakly inflected source language, like English, to a target language with richer grammatical features, such as gender, case, and number. The translation output commonly contains many errors. There are intuitive reasons for that. Take the example of translating noun phrases from English to Slovenian. The same English noun phrase can be translated into over 50 different patterns in Slovenian. A purely lexical mapping of English noun phrases to Slovenian noun phrases suffers from the lack of information about grammatical features. The same observations can be drawn for other Slavic languages.

Morphological variations of highly inflectional languages amplify the effective vocabulary size (i.e. the number of different words) and, consequently, a corpus of increased size is needed to estimate statistics reliably. For many highly inflectional languages, large corpora are not available, and we have to face data sparsity problem. In this research, we try to reduce the data sparsity problem by using the information contained in morpho-syntactic tags. Factored phrase-based SMT models are trained for that purpose.

The context in which a word is used influences its translation. The phrase-based SMT approach tackles this phenomenon by learning the translation of whole phrases instead of single words [15]. Here, "phrase" is not used in the linguistic sense but simply refers to a sequence of words, determined by a data-driven approach. As the phrase-based approach translates phrases independently, words outside the phrase are not considered for its translation. To overcome these deficiencies, the operation sequence model (OSM), proposed in [12], will be considered in this paper.

There are many different possibilities to combine factored phrase-based SMT models and OSM models. The novelty of our work is the definition of most promising combination for translation from English to Slovenian language. This translation direction is the most difficult one, as we are translating from language with poor morphology to language with rich morphology. To our knowledge, such research has not been done in any other previous work.

2. Related Work

There have been numerous efforts to study the effect of applying morphological processing or using morphological information on SMT quality. In one of the first efforts to enrich the source in word-based SMT, part-of-speech tags were used [30]. The approach improved single-word-based SMT that has been afterwards solved by adopting a phrase-based model [15].

There is a big difference between machine translation from a morphologically-rich language and translation to a morphologically-rich language or even between two morphologically-rich languages [20]. If we are translating from morphologically-rich languages, the idea is to reduce data sparsity caused by the rich morphology of the source language through some form of

morphology reduction. Some sort of morphology generation is needed if we are translating to morphologically rich languages. Translation between two morphologically-rich languages was studied in one of our previous works [19]. In this work, we are interested in translation from a language with poor morphology (i.e. English) to a language with rich morphology (i.e. Slovenian).

In general, there are two groups of approaches dealing with morphological processing today. The first group of approaches consists of factored SMTs, where morphological features are modeled jointly as factors in core translation process [17]. During training, the mapping from source factors to target factors is learned. Factors can be used in different mapping combinations. The translation process can be broken up into a sequence of mapping steps that either translate input factors into output factors or generate additional output factors from existing output factors. Different morphological factors and varied translation scenarios were tested in literature [2, 3]. One of the main drawbacks of these approaches is the combinatorial expansion of the number of translation options. One needs to be very careful when defining the topology of the translation system. Another group of approaches models translation and morphology in a sequential manner. These approaches are called "translate-and-inflect" models. The translation is decomposed in two steps. First, a meaning-bearing stem is chosen and then an appropriate inflection is selected. Some approaches use an independent morphological prediction component in the second step [21, 29]. They use maximum entropy Markov model as the learning framework. In [14], conditional random fields framework was implemented to combine the prediction of linguistic features with the prediction of surface forms. In [5], a feature-rich discriminative model was defined, which conditions on the source context of the word being translated to find the right inflection of the translation. The discriminative model was also used to create additional sentence specific word- and phrase-level translations that are added to a standard translation model as "synthetic" phrases. Target morphology was also explored as a source-side prediction task [7]. After word aligning was performed, source sentences were enriched with the useful target morphological information.

Our work in this paper is based on factored phrase-based statistical machine translation models. We focus on preprocessing the source data in terms of MSD tagging to acquire the needed information, and then use it within the factored models. As there are many different configurations, the most successful one is selected. For modeling long distance dependencies, OSM models were added.

We carried out experiments on English to Slovenian translation, a language pair that exemplifies the problems of translating from a morphologically poor to a morphologically rich language. The same experiments with very probably similar results can be performed on other inflected languages, like Slavic languages, paired with English, if MSD tagger and parallel corpus are available.

3. Phrase-Based SMT

The most widely used SMT systems are based on phrases. They succeeded SMT systems based on words. The idea of phrase-based SMT was originally proposed in [15] and since then much research has been done using this approach [19]. A few years ago, neural machine translation occurred with the potential to be the next step in the evolution in MT, but its disadvantage, namely, computational costs, discourage us from its use in the current work.

The phrase-based SMT model can be formulated based on the noisy-channel model:

$$\boldsymbol{e} = \arg\max_{\boldsymbol{e}} \boldsymbol{P}(\boldsymbol{e}|\boldsymbol{f}) = \arg\max_{\boldsymbol{e}} \boldsymbol{P}(\boldsymbol{e})\boldsymbol{P}(\boldsymbol{f}|\boldsymbol{e}), \quad (1)$$

where $\boldsymbol{f}$ denotes a sentence in the source language and $\boldsymbol{e}$ a sentence in the target language. $\boldsymbol{f}_j$ denotes a word in a source sentence, and $\boldsymbol{e}_i$ a word in a target sentence, respectively. Words, denoted with $\boldsymbol{f}_j$, belong to $\boldsymbol{F}$ (vocabulary in source language) and the words, denoted with $\boldsymbol{e}_i$, to $\boldsymbol{E}$ (vocabulary in target language). In the phrase-based model, the source sentence is broken down into I phrases (denoted with $\bar{\boldsymbol{f}}_i$) and the translation takes place based on the phrases. Each source phrase is translated into a target phrase, denoted with $\bar{\boldsymbol{e}}_i$.

Standard phrase-based SMT models consist of three main components:

1 The most important component is the translation

model of phrases (denoted as $\phi(\bar{f}|\bar{e})$. The conditioning of translation probabilities is inverted. In practice, both translation directions are used: $\phi(\bar{f}|\bar{e})$ and $\phi(\bar{e}|\bar{f})$. The translation probabilities are multiplied by a proper weight λ_ϕ.

2 Words in the source and target languages are in a different order. Reordering is modeled by reordering model. This model is based on certain distance metric. The reordering distance is computed as $start_i - end_{i-1} - 1$, where $start_i$ is the position of the first word in th phrase, end_{i-1} is the position of the last word of $(i-1)$th phrase, and d is the probability distribution of reordering.

3 The language model (denoted as $p_{LM}(e)$ makes the output a fluent sequence of words in the target language. This model assigns the probability to each sentence $\boldsymbol{e}$. These probabilities are trained on monolingual corpus from the target language. The most commonly used is an n-gram language model. In this work, 3-gram language model is used:

$$\boldsymbol{p_{LM}(e)} = \prod_{i=1}^{N} \boldsymbol{p_{LM}(e_i|e_{i-2}, e_{i-1})}. \quad (2)$$

N denotes the length of the sentence.

In the machine learning community, a well-known model structure is the log-linear model. It has the following form:

$$\boldsymbol{p(x)} = \boldsymbol{\exp} \prod_{i=1}^{n} \boldsymbol{\lambda_i h_i(x)}, \quad (3)$$

where $h_i(x)$ are feature functions and λ_i are the weights that scale the contribution of each of the feature functions. $\boldsymbol{n}$ denotes the number of feature functions. Log-linear models were adopted to the phrase-based SMT as well. They have the following form:

$$\begin{aligned} &\boldsymbol{p(e, a|f)} \\ &= \boldsymbol{\exp}[\boldsymbol{\lambda_\phi} \sum_{i=1}^{I} \boldsymbol{\log \phi(\bar{f}_i|\bar{e}_i)} \\ &+ \boldsymbol{\lambda_d} \sum_{i=1}^{I} \boldsymbol{\log d(start_i - end_{i-1} - 1)} \\ &+ \boldsymbol{\lambda_{LM}} \sum_{i=1}^{N} \boldsymbol{\log p_{LM}(e_i|e_1 \ldots e_{i-1})}], \end{aligned} \quad (4)$$

Where $\boldsymbol{a}$ is an alignment between source and target sentences. This structure allows us to extend the machine translation knowledge by including additional model components in the form of feature functions.

4. Morpho-Syntactic Information

In sparse data conditions, it is reasonable to use a more generalized representation of words. The representation can reflect morpho-syntactic features of words. The most basic morpho-syntactic information is the information about the part of speech. Part of speech (POS) tags assign words the corresponding grammatical categories, for example verb, noun, adjective, and pronoun. There are approximately 10 basic POS tags. Morpho-syntactic tags, or morpho-syntactic 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 diversity between the numbers of different tags defined per language.

Penn Treebank defines 36 tags for English [18]. Due to rich morpho-syntactic complexity of highly inflectional languages, there are, for example, 3,922 plausible MSD tags defined for Czech (although only 1,571 unique tags actually appear in most corpora) [27]. In the MULTEXT-East project, standardized MSD tag sets for six Central and Eastern European languages were developed [13]. The latest release (Version 5[1]), for example, defines 135 tags for English, 1,425 for Czech, 17,279 for Hungarian, and 1,903 for Slovenian language. Table 1 lists Slovenian POS tags with attributes. In our experiments, 58 tags were used for English[2] and 1,903 for Slovenian[3]. Figure 1 shows an example of annotated part of a sentence in English and in Slovenian.

The parameter file for English was trained on Penn Treebank and for Slovenian on ssj500k 1.3 corpus[4].

The corpus was annotated by TreeTagger [24] on both sides, English and Slovenian. In literature, the accuracy of 96.32% was reported for TreeTagger when annotating English. The accuracy of TreeTagger for Slovenian is not known. It is slightly lower than for

1 http://nl.ijs.si/ME/V5/msd/html/

2 https://www.sketchengine.co.uk/penn-treebank-tagset/

3 http://nl.ijs.si/ME/V5/msd/html/msd-sl.html

4 http://www.slovenscina.eu/tehnologije/ucni-korpus

Table 1
Slovenian POSs with full attributes in MSD tags. The number of different values for each attribute is given in parentheses

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)	-

English, as there are more tags for Slovenian, but generally it is not lower than 90%. After annotating our experimental corpus (it is described in Section 7.1), the analysis shows that in English part 54 different tags were used and 977 in Slovenian part.

Figure 1
Part of an English sentence and Slovenian translation, annotated with MSD tags

Agenda|NN for|IN next|JJ sitting|NN

Dnevni|Agpmsny red|Ncmsn
naslednje|Agpf sg seje|Ncf sg

4.1 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 extent. For translation, complete morpho-syntactic information of inflectional language is not needed, especially when paired with English. We reduced the tags to include only the most important attributes. Attributes that were kept are given in Table 2. 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-`. Experiments were performed using both, the full and the restricted tagging scheme.

Table 2
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

5. SMT Based on Generalized Word Representations

A generalized representation of words enables the extension of translation models and language models used in Eq. (4). Let us suppose that each source word f_j is represented as a vector of factors $(f_{jsurface}, f_{jlemma}, f_{jMSD})$ and target word e_i as a vector of factors $(e_{jsurface}, e_{jlemma}, e_{jMSD})$.

5.1. Factored Translation

Phrase-based translation model $\phi(\bar{f}_i|\bar{e}_i)$ is built from word-aligned corpus by extracting phrases that are in line with word alignments. Having a vector of factors instead of just a word, different mapping steps can be extracted to form a phrase translation model. For example, we can use only a surface form in source vector $(f_{j_{surface}}, -, -)$, but a surface form and MSD tag in target vector $\left(e_{j_{surface}}, -, e_{j_{MSD}}\right)$. Following the taxonomy defined in [4], this type of factored translation model is called tS-SaM, as it translates surface forms to surface

forms and MSD tags. The baseline SMT system, described in Section 3, is of type tS-S, as it translates surface forms to surface forms.

We investigated the following scenarios of factored translation:

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

Having additional factors on target side makes it possible to use more language models in addition to a language model based on surface forms. For example, in tS-SaMaL, we can add a language model based on MSD tags and a language model based on lemmas. For MSD tags, the 6-gram language model is commonly used:

$$p_{LM}(e_{MSD}) = \prod_{i=1}^{N} p_{LM}(e_{i_{MSD}} | e_{i-5_{MSD}} \dots e_{i-2_{MSD}}, e_{i-1_{MSD}}), \quad (5)$$

and for lemmas, 3-gram language model (similar to that defined in Eq. (2)).

5.2 The Operation Sequence Model

Recently, operation sequence model (OSM) was defined and integrated into the phrase-based SMT [9-11]. It is a joint model for the translation and long distance reordering. OSM models translation by a linear sequence of operations. For source sentence $\boldsymbol{f}$, target sentence $\boldsymbol{e}$ (being a translation of $\boldsymbol{f}$) and their alignment $\boldsymbol{a}$, the OSM model is defined as:

$$p_{OSM}(f, e, a) = \prod_{i=1}^{J} p(o_i | o_{i-4}, o_{i-3}, o_{i-2}, o_{i-1}). \quad (6)$$

oi denotes the ith operation, and *J* the number of operations. Operations work on one or more words. We talk about cerpts. *J* denotes the number of cerpts. A cerpt is a group of words in one language translated to another language as a minimal unit in one specific context.

We get a unique operation sequence for every sentence pair given the alignment. When viewed from the opposite perspective, we can say that operations generate the aligned sentence pair. An operation either [10]:

- generates source and target words (for example: _TRANS_obravnavati_TO_address, _TRANS_vzroke_TO_the_causes),
- performs reordering by inserting gaps (denoted as _INS_GAP_),
- performs reordering by jumping forward (denoted as _JMP_FWD_),
- performs reordering by jumping backward (for example, _JMP_BCK_1 jumps backward for two words) or
- continues with the operations at the current position (denoted as _CONT_CEPT

In Table 3, an excerpt from the OSM for an aligned sentence (taken from our experimental training corpus) is given.

Table 3

Training a sentence in OSM model. Before alignment, the corpus was tokenized and truecased

Type	Sentence
English sentence	we must also address the causes .
Slovenian sentence	obravnavati moramo tudi vzroke .
OSM	INS_GAP_ _TRANS_obravnavati_TO_address _JMP_BCK_1 _TRANS_moramo_TO_we_must _CONT_CEPT_ _TRANS_tudi_TO_also _JMP_FWD_ _TRANS_vzroke_TO_the_causes _CONT_CEPT_ _TRANS_ _TO_

Operation sequences can be learned over words or over any other generalized representations [11]. In our research, OSM models are learned over words, MSD tags or lemmas. Finally, we got three different types of OSM models.

6. Evaluation

We used MultEval [6] for evaluation of translation results. Evaluation was based on three runs of the MERT optimizer [22] on the development set in each experimental set-up. Each MERT run provided a different set of weights for components of SMT system (in Eqs. (3) and (4)). MultEval uses three popular metric scores: BLEU, TER, METEOR. It computes standard deviations via bootstrap re-sampling and p-values via approximate randomization.

The BLEU metric [23] is the geometric mean of the test corpus' modified precision scores, based on *n*-grams of different length, multiplied by an exponential brevity penalty factor.

The TER (Translation Error Rate) metric [25] is defined as the minimum number of edits needed to change a hypothesis so that it matches the reference exactly. Possible edits include the insertion, deletion, and substitution of single words as well as shifts of word sequences.

The METEOR [1] is based on unigram precision and unigram recall of matching between translations and references.

7. Experiments

7.1. Corpora and Baseline System

We experimented with the described models using the Slovenian-English parallel corpus from the Europarl corpus v7[5]. The corpus contains 623,490 sentences (14 million Slovenian tokens and 16 million 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 [16] 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 [15]. A 3-gram language model with modified Kneser-Ney discounting was built on the training corpus by the SRILM toolkit [26]. Singletons were excluded. The perplexity of Slovenian language model was 131, and that of English was 62. In the experiments, only language model of Slovenian language was used, as Slovenian was our target language. Perplexity of English language model is reported only for comparison. We also built language models on lemmas and MSD tags, as evident from Table 4. For all the setups, we perform standard MERT training on the defined development set. In all experiments, the resulting translations were evaluated in truecased and tokenized forms (Tables 5 and 7), as well as being detruecased and detokenized (Tables 6 and 8), before evaluated.

Table 4
Vocabulary sizes and perplexities of language models on test set

	V	PP
Surface LM	144,671	131
Lemma LM (TreeTagger)	31,564	52
MSD LM (TreeTagger)	1,903	32

7.2. The Factored Systems

All scenarios of factored translation, described in Section 5.1, were experimentally tested. The results are given in Tables 5 and 6. In the first row, the result of the baseline system is given for comparison. Each factored configuration was run twice, once with full MSD tags and once with reduced tags on Slovenian side. The new factors were added in sequential manner, first only on the target side and then on both sides. We can see that adding MSD tags only on the target side improved the results by more than 1 BLEU point. The relative improvement of factored systems compared to baseline on tokenized and truecased corpus is 3.62%. The improvement on detokenized and detruecased corpus is 4.27%. The results were not further improved by followed configurations. MSD tags seem to be more important factor than lemmas.

5 http://www.statmt.org/europarl/

Table 5
The results of factored translation (tokenized and truecased). The best results are in bold

System configuration	BLEU	TER	METEOR
tS-S	35.9	45.2	31.1
tS-SaM	**37.2**	**44.2**	**31.7**
tS-SaM with reduced MSD tags	36.8	44.5	31.5
tSaM-SaM	**37.2**	**44.2**	**31.7**
tSaM-SaM with reduced MSD tags	37.0	44.4	31.5
tS-SaMaL	36.8	44.9	31.4
tS-SaMaL with reduced MSD tags	36.6	45.0	31.3
tSaMaL-SaMaL	**37.2**	**44.2**	**31.7**
tSaMaL-SaMaL with reduced MSD tags	36.8	44.7	31.4

Table 6
The results of factored translation after detruecasing and detokenization. The best results are in bold

System configuration	BLEU	TER	METEOR
tS-S	30.4	51.3	28.2
tS-SaM	**31.7**	**50.1**	**28.8**
tS-SaM with reduced MSD tags	31.3	50.4	28.7
tSaM-SaM	**31.7**	**50.2**	**28.9**
tSaM-SaM with reduced MSD tags	31.5	50.4	28.7
tS-SaMaL	31.3	50.6	28.6
tS-SaMaL with reduced MSD tags	31.0	50.7	28.5
tSaMaL-SaMaL	**31.7**	**50.2**	**28.8**
tSaMaL-SaMaL with reduced MSD tags	31.2	50.7	28.6

Table 7
The results of factored translation with OSM models (tokenized and truecased). The best results are in bold

System configuration	BLEU	TER	METEOR
Baseline (tS-S)	35.9	45.2	31.1
Baseline (OSM: 0-0)	36.7	44.5	31.6
tSaM-SaM (OSM: 0-0, 1-1)	38.5	43.1	32.4
tSaMaL-SaMaL (OSM: 0-0, 1-1, 2-2)	**38.8**	**42.7**	**32.5**

Table 8
The results of factored translation with OSM models after detruecasing and detokenization. The best results are in bold

System configuration	BLEU	TER	METEOR
Baseline (tS-S)	30.4	51.3	28.2
Baseline (OSM: 0-0)	31.3	50.5	28.8
tSaM-SaM (OSM: 0-0, 1-1)	32.9	48.9	29.2
tSaMaL-SaMaL (OSM: 0-0, 1-1, 2-2)	**33.2**	**48.5**	**29.6**

Table 9
Examples of SMT outputs by not using/using MSD tags. Grammatical errors are given in italics

Ex.	Configuration	Translation
1.	Source: Ref.: tS-S: tS-SaM:	...unless the Council decides against this by qualified majority. ...razen če se Svet odloči drugače na podlagi kvalificirane večine. ...razen če se Svet odloči proti *tej* s kvalificirano večino. ...razen če se Svet odloči proti temu s kvalificirano večino.
2.	Source: Ref.: tS-S: tS-SaM:	The European Union needs stricter European economic supervision Evropska unija potrebuje strožji evropski gospodarski razvoj Evropska unija potrebuje *strožje evropskega gospodarskega nadzora* Evropska unija potrebuje strožji evropski gospodarski nadzor
3.	Source: Ref.: tS-S: tS-SaM: tS-SaMr:	a good idea for Europe's long-term growth dobra ideja za dolgoročno rast Evrope dobra zamisel za *evropske dolgoročno rast* dobra zamisel za dolgoročno rast. dobra zamisel za Evropo dolgoročne rasti..
4.	Source: Ref.: tS-S: tS-SaM:	That is the objective of the economic governance package To je cilj svežnja ukrepov na področju gospodarskega upravljanja. To je cilj *sveženj* o gospodarskem upravljanju. To je glavni cilj svežnja ukrepov na področju gospodarskega upravljanja.
5.	Source: Ref.: tS-S: tS-SaM:	...their budgetary plan is founded on... ...njihov proračunski načrt temeljil na ... ... *svoje proračunske načrt* temelji na... ... da proračunski načrt temelji na ...

7.3. The Factored Systems with OSM Components

In our experiments with OSM models, all three following factors were used: surface forms, lemmas, and MSD tags. We separately taught OSM models over each factor. OSM models over surface forms were added to the baseline systems. The results are summarized in the second row in Tables 7 and 8. OSM models over surface forms and MSD tags were added to the factored system tSaM-SaM. Results are reported in the third row in both tables. Finally, OSM models over surface forms, MSD tags, and lemmas were added to the factored system tSaMaL-SaMaL. Results are given in the last row. We can see that OSM models improved the results of all factored configurations. Overall, the best results were obtained with factored system tSaMaL-SaMaL and OSM models over surface forms, MSD tags, and lemmas. The relative improvement over baseline system on tokenized

Table 10
Examples of SMT outputs by not using/using OSM component. Long-range grammatical errors are given in italics

Ex.	Configuration	Translation
1.	Source:	The European Union needs stricter European economic supervision and a reform of the stability and growth pact.
	Ref.:	Evropska unija potrebuje strožji evropski gospodarski razvoj in reformo pakta za stabilnost in rast.
	tS-S:	Evropska unija potrebuje strožji evropski gospodarski nadzor in reforma pakta za stabilnost in rast.
	tSaM-SaM (OSM: 0-0, 1-1):	Evropska unija potrebuje strožji evropski gospodarski nadzor in reformo Pakta za stabilnost in rast.
2.	Source:	This strategy designates new areas that must be focused on...
	Ref.:	Ta strategija določa nova področja, ki se morajo osredotočiti...
	tS-S:	Ta strategija določa *novih* področjih, ki mora biti...
	tSaM-SaM (OSM: 0-0, 1-1):	Ta strategija določa nova področja, ki jih je treba...
3.	Source:	The measures which have been proposed...
	Ref.:	Ukrepi, ki so bili predlagani...
	tS-S:	Ukrepi, ki so *bile predlagane* ...
	tSaM-SaM (OSM: 0-0, 1-1):	Ukrepi, ki so bili predlagani...
4.	Source:	This signal could act as a guarantee against the risk ...
	Ref.:	Ta signal bi lahko služil kot jamstvo, da ne bo propadlo...
	tS-S:	To sporočilo bi lahko *deloval kot jamstva proti tveganje*...
	tSaMaL-SaMaL (OSM: 0-0, 1-1, 2-2):	Ta signal lahko deluje kot jamstvo proti tveganju...

and truecased corpus is 8.08%. On detokenized and detruecased corpus, it is 9.21%. Comparing successive configurations, we can notice that OSM models over MSD tags brought the greatest change of results. Adding OSM models over lemmas contributed to only minor relative improvement of results.

From the automatic evaluation of translation results, the system tSaMaL-SaMaL with OSM models over surface forms, MSD tags, and lemmas was selected as the best performing one.

7.4. A Qualitative Comparison of Different SMT Configurations

In addition to quantitative evaluation by automatic metrics, we qualitatively analysed and compared the system outputs where results were the most different compared to the baseline output. Table 9 gives some examples, where inflection in word forms was improved by using MSD tags in factored translation (tS-SaM). The same examples were also taken from SMT output of tS-SaM configuration with reduced MSD tags. We noticed some cases, in which the use of reduced MSD tags improved the translation (e.g. example 3 in Table 9), but there were many examples with no improvements. This observation indicates that further investigation into MSD tags reduction would be needed in the future. We have also analysed the outputs of systems that use lemma as additional factor. No improvements in lemma selection were recognized. It could be so because the system is trained and tested on the same narrow domain text, where the problem of wrong lemma selection is not as evident as in general domain translation.

In the second qualitative comparison, we were interested in improvements brought by OSM models. Table 10 gives some examples. Improvements in long range dependencies are evident. It is interesting that the best performing system did not bring the best translation in all cases. Different configurations also selected different lemmas in some contexts (e.g. example 4 in Table 10).

Conclusion

In this paper, we have investigated the use of morpho-syntactic information in SMT in two different ways, by factored translation and by OSM models.

Both approaches were proved to be beneficial in translation to a highly inflected language. The results obtained in the experiments showed a significant improvement in terms of automatic metrics by adding morpho-syntactic information to the translation process. The configuration tSaMaL-SaMaL (OSM: 0-0, 1-1, 2-2), which uses surface forms, MSD tags and lemmas on both sides and all three types of OSM models, brought the best translation results. A manual inspection of the translation examples showed improved inflections of translations in many cases, as well as improved ordering of words in translation; however, there are still many morphological and syntactical mismatches, among other error types. To improve grammatical fluency, could be one direction for future work. In cooperation with the linguists, we plan to collect rules for grammatical agreement and apply them to the machine translation output. The definition of some rules seems straightforward, like adjective-noun agreement, but others are more complicated, like subject-verb agreement, as Slovenian is a null-subject language, like many other Slavic languages.

Acknowledgment

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding P2-0069).

References

1. Banerjee, S., Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, 2005, 65–72.
2. Bojar, O. English-to-Czech Factored Machine Translation. Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, 2007, 232–239. https://doi.org/10.3115/1626355.1626390
3. Bojar, O., Kos, K. 2010 Failures in English-Czech Phrase-Based MT. Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics (MATR), Uppsala, Sweden, 2010, 60–66.
4. Bojar, O., Jawaid, B., Kamran, A. Probes in a Taxonomy of Factored Phrase-Based Models. Proceedings of the 7th Workshop on Statistical Machine Translation, Montréal, Canada, 2012, 253–260.
5. Chahuneau, V., Schlinger, E., Smith, N. A., Dyer, C. Translating into Morphologically Rich Languages with Synthetic Phrases. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA, 2013, 1677– 1687.
6. Clark, J. H., Dyer, C., Lavie, A., Smith, N. A. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, HLT'11, Portland, Oregon, 2011, 176–181.
7. Daiber, J., Sima'an, K. Machine Translation with Source-Predicted Target Morphology. Proceedings of the MT Summit XV, Miami, Florida, 2015, 283–296.
8. Davidel, L. Translation in the European Union Facts and Figures, 2013. [Online]. Available at: http://one-europe.info/translation-in-the-europeanunion-facts-and-figures.
9. Durrani, N., Schmid, H., Fraser, A. A Joint Sequence Translation Model with Integrated Reordering. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-HLT), Portland, Oregon, USA, 2011, 1045–1054.
10. Durrani, N., Fraser, A., Schmid, H., Hoang, H., Koehn, P. Can Markov Models Over Minimal Translation Units Help Phrase-Based SMT? Proceedings of the 51st Annual Conference of the Association for Computational Linguistics (ACL), Sofia, Bulgaria, 2013, 399–405.
11. Durrani, N., Koehn, P., Schmid, H., Fraser, A. Investigating the Usefulness of Generalized Word Representations in SMT. Proceedings of the 25th Annual Conference on Computational Linguistics (COLING), Dublin, Ireland, 2014, 421–432.
12. Durrani, N., Schmid, H., Fraser, A., Koehn, P., Schütze, H. The Operation Sequence Model – Combining N-Gram-Based and Phrase-Based Statistical Machine Translation. Computational Linguistics, 2015, 41(2), 185–214. https://doi.org/10.1162/COLI_a_00218
13. Erjavec, T. MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. Language Resources & Evaluation, 2012, 46, 131–142. https://doi.org/10.1007/s10579-011-9174-8
14. Fraser, A., Weller, M., Cahill, A., Cap, F. Modeling Inflection and Word-Formation in SMT. Proceedings of the

13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, 2012, 664–674.

15. Koehn, P., Och, F. J., Marcu, D. Statistical Phrase-Based Translation. Proceedings of the Human Language Technology Conference, Stroudsburg, PA, USA, 2003, 48–54. https://doi.org/10.21236/ADA461156
16. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ..., Dyer, C. Moses: Open Source Toolkit for Statistical Machine Translation. Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, 2007, 177–180. https://doi.org/10.3115/1557769.1557821
17. Koehn, P., Hoang, H. Factored Translation Models. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, Scotland, 2007, 868–876.
18. Marcus, M., Santorini, B., Marcinkiewicz, M. A. Building a Large Annotated Corpus of English: The Penn Treebank. Technical Report MS-CIS-93-87, University of Pennsylvania, Computer and Information Science Department, 1993.
19. Maučec, M. S., Kačič, Z., Verdonik, D. Statistical Machine Translation of Subtitles for Highly Inflected Language Pair. Pattern Recognition Letters, 2014, 46, 96–103. http://dx.doi.org/10.1016/j.patrec.2014.05.01
20. Maučec, M. S., Brest, J. Slavic Languages in Phrase-based Statistical Machine Translation: A Survey. Artificial Intelligence Review, 2017. https://doi.org/10.1007/s10462-017-9558-2
21. Minkov, E., Toutanova, K., Suzuki, H. Generating Complex Morphology for Machine Translation. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 2007, 128-135.
22. Och, F. J. Minimum Error Rate Training in Statistical Machine Translation. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, 2003, 160-167. https://doi.org/10.3115/1075096.1075117
23. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. BLEU: A Method for Automatic Evaluation of Machine Translation. Technical Report RC22176(W0109-022), IBM Research Report, IBM, 2004.
24. Schmid, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK, 1994, 44-49.
25. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J. A Study of Translation Edit Rate with Targeted Human Annotation. 5th Conference of the Association for Machine Translation in the Americas (AMTA), Boston, Massachusetts, 2006, 223-231.
26. Stolcke, A. SRILM – an Extensible Language Modeling Toolkit. Proceedings of the International Conference on Spoken Language Processing, 2002, 257-286.
27. Straková, J., Straka, M., Hajič J. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, Maryland USA, 2014, 13-18. https://doi.org/10.3115/v1/P14-5003
28. Šveikauskienė, D., Telksnys, L. Accuracy of the Parsing of Lithuanian Simple Sentences. Information Technology and Control, 2014, 43(4), 402-413. https://doi.org/10.5755/j01.itc.43.4.6700
29. Toutanova, K., Suzuki, H., Ruopp, A. Applying Morphology Generation Models to Machine Translation. Proceedings of ACL, Columbus, Ohio, USA, 2008, 514-522.
30. Ueffing, N., Ney, H. Using POS Information for SMT into Morphologically Rich Languages. EACL 2003, 10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, Hungary, 2003, 347.