Asistent – A Machine Translation System for Slovene, Serbian and Croatian

Mihael Arčan* **Maja Popović†** **Paul Buitelaar***

*Insight Centre for Data Analytics, NUI Galway, Ireland
[firstname.lastname]@insight-centre.org

†Humboldt University of Berlin, Germany
maja.popovic@hu-berlin.de

Abstract

The META-NET research on language technologies in 2012 showed a weak support on tools for crossing the language barrier for many European languages, including the south Slavic languages. Therefore, we describe a statistical machine translation system, called *Asistent*, which enables automatic translations between English, Slovene, Croatian and Serbian. In addition to make this system publicly accessible, we focus on parallel data preparation as well as on using multiple pivot languages for translation quality improvement of the targeted Slavic languages. A comparison of translations generated by the *Asistent* translation system shows a significant improvement of translation quality over *Google Translate*.

1. Introduction

The statistical machine translation (SMT), in particular phrase-based SMT (Koehn et al., 2003), has become widely used to cross the language barrier in the last years. Nowadays, open source tools such as Moses (Koehn et al., 2007) have made it possible to build translation systems for many language pairs, domains or text types within days. Despite the fact that for certain language pairs, e.g. English-French, high quality SMT systems have been developed, a large number of languages and language pairs still suffer from underdeveloped resources. The largest study about European languages in the Digital Age, the META-NET Language White Paper Series[1] in year 2012 showed that only English has good machine translation support, followed by moderately supported French and Spanish. More languages are only fragmentary supported (such as German), whereby the majority of languages are weakly supported. Many of those languages are also morphologically rich, which makes the SMT task even more challenging, especially if translations are performed into the morphologically rich languages. A large part of the weakly supported languages consists of Slavic languages, where Slovene, Serbian and Croatian belong (Krek, 2012). Therefore, we describe *Asistent*,[2] an SMT system, which enables automatic translations between English, Slovene, Croatian and Serbian language. Despite the limited amount of resources and domain variations, specifically among the Slavic languages, we collected existing data and developed a system aimed at supporting human translators and enabling cross-lingual language technology tasks.

2. Related Work

One of the first results with automatic translations for Slovene was shown in the *Presis* System (Romih and Holozan, 2002). The rule-based translation system annotates each source sentence with grammatical features and uses built-in rules for converting annotated source sentences into the target language.

First publications dealing with SMT systems for Serbian-English (Popović et al., 2005) and Slovene-English (Maučec et al., 2006) are reporting results using small bilingual corpora. Using morpho-syntactic knowledge for the Slovene-English language pair was shown to be useful for both translation directions in Žganec Gros and Gruden (2007). However, no analysis of results has been carried out in terms of what actual problems were caused by the rich morphology and which of those were solved by the morphological preprocessing. Recent work in SMT also deals with the Croatian language, which is very closely related to Serbian. First results for Croatian-English are reported in Ljubešić et al. (2010) on a small weather forecast corpus, and an SMT system for the tourist domain is presented in Toral et al. (2014). Furthermore, SMT systems for both Serbian and Croatian are described in Popović and Ljubešić (2014) and more recently in Antonio Toral and Ramírez-Sánchez (2016) and Sánchez-Cartagena et al. (2016). Work on rule based machine translation between Croatian and Serbian was shown in Klubička et al. (2016).

Different SMT systems for subtitles were developed in the framework of the SUMAT project, including Serbian and Slovene (Etchegoyhen et al., 2014). First effort in the direction of collecting a larger amount of existing parallel data sets for Serbian and Slovene was carried out in Popović and Arcan (2015). The authors built several SMT systems in order to identify the most important language related issues which may help to build better translation systems. However, all the translation systems described were built and used only locally, mainly only on one particular genre and/or domain. In this proposed work, we are building a publicly available mixed-domain SMT system built on existing parallel corpora, which we believe will be useful for the given under-resourced language pairs.

[1] http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison

[2] http://server1.nlp.insight-centre.org/asistent/

3. Experimental Setup

The proposed system, called *Asistent*, is a freely accessible translation system, based on the widely used phrase-based SMT framework. It supports translations from English into Slovene, Croatian and Serbian and vice versa. In addition to that, translations between the Slavic languages can be obtained.

3.1. Statistical Machine Translation

Our approach is based on SMT, where we wish to find the best translation $\mathbf{e}$, of a source string $\mathbf{f}$, given by a log-linear model combining a set of features. The translation that maximizes the score of the log-linear model is obtained by searching all possible translations candidates. The decoder or search procedure, respectively, provides the most probable translation based on statistical translation model learned from the training data.

For generating the translation models, we use the statistical translation toolkit Moses (Koehn et al., 2007). Word alignments were built with GIZA++ (Och and Ney, 2003) and a 5-gram language model was built with KenLM (Heafield, 2011).

3.2. Automatic Translation Evaluation

For the automatic evaluation of translations between the targeted languages we report results based on the BLEU (Papineni et al., 2002), Meteor (Denkowski and Lavie, 2014) and the chrF3 (Popović, 2015) metric. The approximate randomization approach in MultEval (Clark et al., 2011) is used to test whether differences among system performances are statistically significant.

BLEU is calculated for individual translated segments (n-grams) by comparing them with a data set of reference translations. BLEU scores, between 0 and 100 (perfect translation), are averaged over the whole evaluation data set to reach an estimate of the translation's overall quality.

Meteor is based on the harmonic mean of precision and recall, whereby recall is weighted higher than precision. Along with standard exact word (or phrase) matching it has additional features, i.e. stemming, paraphrasing and synonymy matching.

chrF3 is a tokenisation-independent metric, which has shown very good correlations, especially for morphologically rich(er) languages, with human judgements on the WMT2015 shared metric task (Stanojević et al., 2015), both on the system level as well as on the segment level.

In addition to the described evaluation metrics, we performed an automatic error classification with the help of *Hjerson* (Popović, 2011). The publicly available tool estimates the frequencies of five error types, i.e. morphological (inflectional) errors, word order errors, omissions, additions and lexical errors (mistranslations).

4. Parallel Data Sets

In this section we describe the parallel corpora used to build the translation models as well as the data preparation approach for better translation performance.

Corpus Name	En-Sl	En-Hr	En-Sr	Sl-Hr	Sl-Sr	Hr-Sr
DGT	1.8M	196K	/	/	/	/
ECB	79K	/	/	/	/	/
EMEA	253K	/	/	/	/	/
Europarl	599K	/	/	/	/	/
Gnome	998	5K	126	4K	600K	300K
hrenWaC	/	86K	/	/	/	/
JRC-Acquis	29K	38K	/	/	/	/
KDE	58K	/	32K	85K	49k	33.2k
LangCourse	/	/	3K	/	/	
PHP	1K	/	/	/	/	/
OpenSubtitles	10.1M	16.3M	20.5M	6.1M	13.3M	22.3M
SETimes	/	198K	209K	/	/	200K
Tatoeba	/	777	633	/	/	/
TED	13K	76K	1K	/	/	/
Ubuntu	/	8K	/	557	86K	51K
Training Data	**En-Sl**	**En-Hr**	**En-Sr**	**Sl-Hr**	**Sl-Sr**	**Hr-Sr**
L1 words:	161M	165M	194M	39M	90M	137M
L2 words:	133M	133M	159M	40M	94M	139M
unique L1 w.:	631K	626K	658K	468K	775K	1.22M
unique L2 w.:	1.00M	1.26M	1.37M	579K	966K	1.24M
Par. sentences:	13.1M	16.9M	20.7M	5.5M	12.6M	19.4M

Table 1: Statistics on parallel corpora used to build the translation models accessed by the *Asistent* system (explanation: En-Sl → L1=En, L2=Sl).

4.1. Data Sets Description

The parallel data used to train the SMT system were mostly obtained from the OPUS web site (Tiedemann, 2012), which contains various corpora of different sizes and domains. For the Serbian-English language pair, a small language course corpus of about 3,000 sentence pairs was added as well. Furthermore, a small phrase book with about 1,000 entries was added to the Slovene-Serbian training set.

Table 1 illustrates the various corpora used to train the *Asistent* system. The upper part of the table shows the original amount of parallel entries in each corpus, while the lower part shows details on the concatenated and preprocessed data set (cf. Subsection 4.3.) used to train the translation models. While corpora for the English-Slavic language pairs consist of different domains, e.g. legal, medical, financial, IT, parallel data between Slavic language pairs consist mostly out of the OpenSubtitles corpus (Lison and Tiedemann, 2016).[3]

4.2. Evaluation Data Set

The in-domain data set used for evaluating *Asistent*'s performance consists of around 2.000 sentences for each language pair of various domains.[4] When translating from or into English, sentences from different corpora[5] were

[3] http://www.opensubtitles.org/

[4] The evaluation set can be obtained under: http://server1.nlp.insight-centre.org/asistent/data/asisten_evaluation_set.tar.gz

[5] DGT, EMEA, Europarl, KDE and OpenSubtitles for English-Slovene; DGT, hrenWaC, KDE, OpenSubtitles and SETimes for English-Croatian; KDE, OpenSubtitles and SETimes for English-Serbian

	English → Slovene		Slovene → English	
	non-preprocessed	preprocessed	non-preprocessed	preprocessed
BLEU	49.56	**49.97**	61.37	**63.52**
parallel sentences in training data	15.4M	13.1M	15.4M	13.1M
entries in translation model	201M	230M	201M	230M
unique source words in translation model	553K	604K	893K	972K

Table 2: Results on translation quality based on BLEU and statistics on training data and translation models before and after data preparation.

added to the evaluation data set (isolated from the training data set). The data used for evaluating translations between the Slavic languages consist mostly out of the OpenSubtitles corpus, since this corpus builds the largest part (≈95%) of the data used to train the translation models.

4.3. Data Preparation

The parallel corpora used for the proposed SMT systems were obtained from the OPUS web site, which contains various corpora of different sizes and domains. However, some of the corpora are rather noisy and therefore certain preprocessing steps were performed.

First, for Serbian as a bi-alphabetical language (Cyrillic and Latin), segments containing letters from both alphabets were removed (such segments were very frequent in the OpenSubtitles corpus). Cyrillic and Latin parts were separated, whereby the Cyrillic parts were converted into Latin. The original Latin parts were removed from falsely encoded special characters. The same approach was performed on the Croatian and Slovene corpora as well. After that, for all languages, technical texts were cleaned from segments containing "#", "%" and "@" symbols. In OpenSubtitles, the hyphen signs appearing at the beginning of a sentence were removed in all texts in order to obtain better consistency. Apart from the described conversions, a large portion of Slovak text was removed from the Slovene part of the Tatoeba corpus.

The next step consisted of filtering of all corpora based on the sentence length proportions. The source/target and target/source sentence length proportions were calculated on the preprocessed texts, and the confidence intervals were extracted based on average proportions and standard deviations. Then, for the texts to be cleaned, only the sentence pairs with proportions within the confidence intervals were kept. The confidence intervals based on average proportions and standard deviations were calculated on the preprocessed texts, i.e. Europarl (Koehn, 2005) for Slovene-English and SETimes (Tyers and Alperen, 2010) for Serbian-English and Croatian-English. For all other corpora, all sentence pairs with proportions within the corresponding confidence interval were kept, and the rest was removed. The last step was removing repetitions, i.e. keeping only unique sentence pairs in all corpora.

After preprocessing the data, tuning and evaluation data sets were extracted for each language pair, containing mostly clean segments from a diverse set of domains. Additionally, these data sets were extracted following the findings of (Song et al., 2014). Namely, too short and too long sentences were not included into the data set, with an optimal average sentence length of 25 words (between 10 and 40 words). Due to the heterogeneity of the used data sets, we extracted such segments from the Europarl and SETimes corpora, and shorter segments (5 to 15 words) from OpenSubtitles and technical texts in the IT domain.

Evaluation on preprocessed data set Due to the noisiness of the parallel corpora, we evaluated first the translation quality of translations generated from a translation model, which was learned from the concatenated data obtained directly from OPUS. We compared these results with translations obtained from the translation model learned from the preprocessed data set, using cleaning steps explained in Subsection 4.3. As seen in Table 2, we gain minor improvements in term of BLEU when translating from English into Slovene, but larger improvements are shown when translating from Slovene into English. Although the non-preprocessed training data set contains more parallel sentences (15.4M vs. 13.1M), the amount of entries as well as the vocabulary stored in the translation models based on the preprocessed data set increases. This illustrates that with this data set, more bilingual alignments can be learned in comparison to a non-preprocessed data set.

5. Evaluation

To evaluate the translation quality of our proposed system, we perform an in-domain and out-of-domain evaluation. The first is done on the evaluation set, which is constructed and isolated from the aforementioned preprocessed parallel corpora. The out-of-domain evaluation is performed on a domain, which is not primary used in the training step. We support the evaluation by illustrating the most frequent n-gram mismatches as well as an analysis of error classes.

5.1. In-domain Translation Evaluation

In this section we present the translation evaluation based on the data set isolated from parallel corpora described in Section 4. Table 3 shows the performance of the *Asistent* system for translating text between English, Slovene, Serbian and Croatian. We compare the translation quality in terms of the BLEU, Meteor and chrF metric to the *Google Translate* system.[6] As seen, we significantly outperform *Google Translate* in most of the examined language pairs ($p<0.01$). Only when translating from

[6] `https://translate.google.com/`, translations performed on April 3rd, 2016

	English → Slovene			Slovene → English		
	BLEU	Meteor	chrF3	BLEU	Meteor	chrF3
Google	34.46	32.90	61.75	44.41	40.59	62.94
Asistent	**49.82**	**36.36**	**69.26**	**64.14**	**45.77**	**76.71**
	English → Croatian			Croatian → English		
	BLEU	Meteor	chrF3	BLEU	Meteor	chrF3
Google	29.27	26.87	57.19	46.08	39.35	67.61
Asistent	**42.15**	**33.02**	**64.95**	**48.07**	**40.05**	**68.18**
	English → Serbian			Serbian → English		
	BLEU	Meteor	chrF3	BLEU	Meteor	chrF3
Google	27.48	26.33	55.80	**46.05**	**39.35**	**67.53**
Asistent	**42.47**	**32.63**	**63.59**	42.35	39.11	64.96
	Slovene → Serbian			Serbian → Slovene		
	BLEU	Meteor	chrF3	BLEU	Meteor	chrF3
Google	12.27	16.83	34.27	14.05	18.32	37.16
Asistent	**23.46**	**23.23**	**43.23**	**29.23**	**25.33**	**47.42**
	Croatian → Serbian			Serbian → Croatian		
	BLEU	Meteor	chrF3	BLEU	Meteor	chrF3
Google	59.88	41.72	73.84	64.90	46.11	78.65
Asistent	**67.39**	**46.34**	**77.98**	**70.09**	**48.97**	**80.89**
	Slovene → Croatian			Croatian → Slovene		
	BLEU	Meteor	chrF3	BLEU	Meteor	chrF3
Google	13.47	17.81	37.36	16.07	19.60	39.54
Asistent	**34.63**	**27.02**	**50.98**	**38.64**	**29.78**	**54.98**

Table 3: Automatic translation evaluation based on BLEU, Meteor and chrF for the *Asistent* and *Google Translate* translation system.

Serbian into English, *Google Translate* performs better than the *Asistent* translation system.

Table 4 reports the frequencies of five *Hjerson* error classes, i.e. morphological-inflectional errors (infl), word order errors (order), omissions (miss), additions (add) and lexical errors (lex). The last column represents the overall sum of errors. It can be seen that there is a larger number of inflectional errors when translating from English into a Slavic language, indicating that one of the first steps towards improving the current version of the system should be dealing with morphological generation. Apart from this, a high percentage of mistranslations is present, which is typical for state-of-the-art SMT systems and can be overcome with enlarging the training data set.

In addition to the automatic translation evaluation, we performed a semi-automatic analysis of the most frequent errors based on unmatched words and word sequences. The automatic evaluation tool *rgbF* (Popović, 2012), based on word n-gram F-score, enables the annotation of unmatched n-grams in the automatically generated translations in regards to reference translation. A manual inspection of these n-grams revealed some frequent patterns, which are shown in Table 5. It can be seen that prepositions, conjunctions, relative pronouns and auxiliary verbs are problematic for

	infl	order	miss	add	lex	$\sum$
Slovene → English	1.4	6.4	5.8	5.5	13.8	33.0
English→ Slovene	7.3	6.2	5.9	5.6	16.5	41.4
Croatian → English	2.8	7.3	6.0	5.1	17.8	39.0
English → Croatian	8.7	5.4	5.3	5.3	20.0	44.7
Serbian → English	2.8	8.8	6.4	5.8	18.6	42.4
English → Serbian	8.9	7.9	6.0	6.7	23.5	53.0

Table 4: Identified translation error classes of the *Asistent* system by the *Hjerson* tool for the in-domain evaluation set.

all translation directions. For translations into English, articles and pronouns are frequently problematic since these two classes are non-existing or often omitted in the Slavic languages. For other translation directions, negation and reflexive pronouns represent frequent issues.

Pivot Language Evaluation Additionally to the direct translation (source language → target language) evaluation, we performed an experiment on pivot translation (source language → pivot language → target language). This approach can enable a bridge between languages, when existing parallel corpora are under-resourced (Babych et al., 2007). Due to the language coverage within the *Asistent* system, we could use two pivot languages for our additional translation experiment. Since we do not know in advance which pivot language can contribute most in pivot translation, we perform a mixed approach, where we select the best translation out of the set of translations coming from the different pivot languages. In our approach we translate first the source sentence into a pivot language and use the most probable translation to translate it further into the target language. In this last step we collect the best 100 translations for each source sentence. Since the language model of the target language is same regardless which pivot language is used, we identify out of the set of 200 translations, provided by two different pivot languages, the most accurate target sentence based on the language model probability.

As seen in the last column in Table 6, the pivot translation quality declines mostly for the English-Slovene and English-Croatian language pairs. Only for the English-Serbian translation direction, the mixed pivot approach provides better translation quality compared to the direct translation. Focusing on translations between Slavic languages only, the proposed approach frequently shows improvements over the direct translations for the less resourced Slavic language pairs.

5.2. Out-of-Domain Translation Evaluation

Besides the in-domain translation evaluation, we perform an evaluation on a data set, which differs from the parallel data used to build the translation models. Massive Open Online Courses (MOOCs) have been growing in impact and popularity in recent years. However, the materials are available mostly in English and the translation solutions provided so far have been fragmentary and human-based. Therefore, in addition to the in-domain evaluation campaign, the *Asistent* system has been tested on a set of

Class	English → Slovene	Slovene → English
Article	/	the, a, an$_{ref}$
Preposition	z, v, na, z, s, pri$_{ref}$, o$_{ref}$,	of, in, to, on, for, with, as
Conjunction	in, da, ki, kot, pa$_{ref}$, tudi$_{ref}$, tako$_{ref}$,	that, which
Pronoun	se (reflexive), to	it, I, you, that, this, we
Verb	je, so, bi, sem, bo, bilo	is, are, have, be
Negative particle	ne, ni,	/
Sequence	,+da ,+ki ,+da+je, to+je ,+in, da+bi, da+se,	of+the, in+the, to+the, ,+the
Class	English → Croatian	Croatian → English
Article	/	the, a, an$_{ref}$
Preposition	u, na, za, s, sa, iz,	of, in, to, on, for, with, at$_{ref}$, by$_{ref}$,
Conjunction	i, a, da, koji, koje, koja, kao, kako, te$_{ref}$,	that, and, which
Pronoun	se(reflexive), to, ti,	it, I, you, that
Verb	je, su, biti, bi, će	i, are, be, have, will, has$_{ref}$, was$_{ref}$
Negative particle	ne	/
Sequence	to+je, ,+i, ,+a$_{ref}$, da+se$_{ref}$, bi+se$_{ref}$,	of+the, in+the, to+the, ,+the, ,+and,
class	English → Serbian	Serbian → English
Article	/	the, a, an$_{ref}$
Preposition	u, na, za, s, sa, od, iz$_{ref}$, o$_{ref}$,	of, in, to, on, for, with, at$_{ref}$, by$_{ref}$, as$_{ref}$
Conjunction	i, a, da, koji, koje, kao, što$_{ref}$,	that, and
Pronoun	se(reflexive), to	it, I, you, that, its$_{ref}$
Verb	je, su, će, bi$_{ref}$	is, are, has, was, have, will, be$_{ref}$
Negative particle	ne	/
Sequence	da+se, da+je, to+je, ,+koji, koji+je, ,+a$_{ref}$, da+će$_{ref}$	of+the, in+the, to+the, ,+the$_{ref}$

Table 5: Examples of most frequent unmatched n-grams by the *Asistent* translation system.

Translation Direction	BLEU with Pivot language		Mixed
English → Slovene	21.55 (Hr)	14.60 (Sr)	20.44
Slovene → English	24.20 (Hr)	26.77 (Sr)	28.77*
English → Croatian	19.04 (Sl)	36.23 (Sr)	38.42*
Croatian → English	29.82 (Sl)	44.84 (Sr)	34.87
English → Serbian	19.19 (Sl)	**60.80** (Hr)	**59.91**
Serbian → English	21.64 (Sl)	**52.44** (Hr)	31.39
Slovene → Serbian	18.03 (En)	**25.45** (Hr)	19.68
Serbian → Slovene	**30.21** (En)	**31.97** (Hr)	**35.39***
Croatian → Serbian	24.23 (En)	30.79 (Sl)	25.89
Serbian → Croatian	41.29 (En)	34.49 (Sl)	37.96
Slovene → Croatian	**35.95** (En)	30.25 (Sr)	**40.44***
Croatian → Slovene	**39.81** (En)	**48.62** (Sr)	**52.09***

Table 6: Automatic translation evaluation based on BLEU using pivot language (in brackets; bold results = improved translation quality compared to direct translation; *-improvement over individual pivot translations).

	Croatian → English							
	BLEU	CHRF3	infl	order	miss	add	lex	∑
Asistent (d)	23.7	48.0	2.7	8.2	14.8	3.4	25.7	54.8
Asistent (p)	19.7	44.1	2.7	7.0	17.4	3.5	30.3	60.8
Google	**26.2**	**51.9**	2.7	6.7	13.7	3.1	25.1	51.3
	English → Croatian							
	BLEU	CHRF3	infl	order	miss	add	lex	∑
Asistent (d)	15.6	45.3	9.0	5.4	5.4	9.5	30.3	59.6
Asistent (p)	12.9	38.8	8.4	6.6	9.5	6.7	35.7	66.9
Google	**18.4**	**50.4**	7.9	5.5	2.6	12.2	27.9	56.1
	Serbian → English							
	BLEU	CHRF3	infl	order	miss	add	lex	∑
Asistent (d)	23.0	48.2	2.6	7.8	11.6	4.4	28.6	55.1
Asistent (p)	18.6	42.5	2.6	8.3	14.3	3.9	34.0	63.2
Google	**24.6**	**50.8**	2.7	8.2	10.7	4.2	28.0	53.7
	English → Serbian							
	BLEU	CHRF3	infl	order	miss	add	lex	∑
Asistent (d)	12.8	38.9	8.3	7.0	7.7	6.1	36.4	65.5
Asistent (p)	10.0	33.8	7.6	6.5	11.7	5.6	40.4	70.8
Google	**17.0**	**46.4**	7.9	6.6	4.7	8.9	30.8	59.0

Table 7: Identified translation error classes of *Asistent* by the *Hjerson* tool for the out-of-domain test set (d=direct translation; p=pivot translation).

out-of-domain texts, originating from educational domain, i.e. lecture subtitles from Coursera. However, it should be noted that these data were not available for the Slovene-English language pair.

The results for Serbian-English and Croatian-English are shown in Table 7 in the form of BLEU, chrF3 as well as the aforementioned five *Hjerson* error rates. Additionally we perform the same evaluation for translations generated by *Google Translate*. It can be seen that although the results for *Google* are better for these texts, they are rather close. Differently to the in-domain evaluation, the pivot translation could not improve the translations over the direct translation approach. As for detailed error rates, the main advantage of *Google* is the smaller amount of omis-

sions (miss) and lexical errors (lex), which is usually the case when larger data sets are used.

6. Translation System as a Web Service

The generic translation models built with the default Moses settings are in general very large, and cannot be use in an online scenario. Therefore, to provide a user translations as good and as fast as possible, we limit the length of the source and target translation candidates in the translation models to five-grams.[7] Additionally we filter out those translation candidates, which are below the direct phrase probability p(e|f) of 1.0E-4.[8] With these strategies we exclude more than 80 million entries for the English-Slovene language pair without to significantly decrease the translation quality. At last, we compared the performance between the OnDisk binarization of the translation model (Zens and Ney, 2007) against the Compact implementation (Junczys-Dowmunt, 2012), where the compressed translation model relies on a perfect minimum hash for look-up.

We evaluated the results of an unfiltered translation models (binarized and non-binarized) against translation models, filtered on aforementioned thresholds. Compared to the default setting we observed insignificant differences in terms of BLEU (≈49.6) using thresholds between 1.0E-5 and 1.0E-3. Only when the threshold is set to 1.0E-2 or above, the performance declines in translation quality in terms of the BLEU score. Additionally, we did not detect any significant quality difference between the OnDisk and Compact implementation.

Optimization evaluation Considering the online scenario, we compress the translation models for all language pairs based on the 1.0E-4 threshold (direct translation probability) and binarize it with OnDisk implementation.[9] Table 8 shows the performance of the *Asistent* translation system, comparing unfiltered translation models for each language pair with filtered and binarized ones. As seen, although we reduce the amount of possible translation candidates in the translation models, the BLEU score does not always decreases significantly. In fact, by using the filtered model we observe improvements for the English → Serbian (+3.59 BLEU) and Slovene → Croatian (+1.3 BLEU) translation direction. This indicates that the filtering approach can exclude an extensive amount of misaligned translation candidates in the original models that may cause translation errors. On the other hand, a decrease in performance for English → Croatian (-1.44) has been observed. Nevertheless, the decrease of translation quality for other translation directions is moderate.

Webdemo API service The translation models, which are accessed through the *Asistent* web interface, can also be accessed by third-party tools.[10] When the *Asistent* service receives a translation request in form of a JSON object (upper part of Figure 1), the service queries the translation models for the best candidate translations. A ranked list based on log probabilities of candidate translations (accessible with JSON key $possible\_translations$, seen in the lower part of Figure 1) is generated from the web service and sent back to the user that can select either the best probable translation or a translation among the proposed translations.

Translation Direction	Asistent	org. models	δ
English → Slovene	49.82	**49.97**	-0.15
Slovene → English	**64.14**	63.52	+0.62
English → Serbian	**42.47**	38.88	+3.59
Serbian → English	42.35	**43.79**	-1.44
English → Croatian	42.15	**43.38**	-1.23
Croatian → English	48.07	**48.90**	-0.83
Slovene → Serbian	**23.46**	23.34	+0.12
Serbian → Slovene	**29.23**	28.97	+0.26
Slovene → Croatian	**34.63**	33.3	+1.33
Croatian → Slovene	38.64	**38.73**	-0.09
Serbian → Croatian	70.09	**70.29**	-0.20
Croatian → Serbian	67.39	**67.54**	-0.15

Table 8: Comparison of BLEU scores between default (original) translation models and *Asistent* accessed compressed translation models.

7. Conclusion

This paper presents a publicly accessible SMT system for translating between English, Slovene, Croatian and Serbian, called *Asistent*. Through the publicly accessible web interface and API request, the SMT system can support human translators and enable information access across languages. Based on the automatically extracted evaluation data set, *Asistent* outperformed *Google Translate* for the majority of the targeted translation directions. Furthermore, experiments on pivot translation show improvements in translation quality between closely related Slavic language pairs over a direct translation approach. Our ongoing work focuses on a better combination of the pivot translation and the comparison of feature based (linguistically annotated data sets) and hierarchical (synchronous context-free grammar rules) SMT for the Slavic languages.

8. Acknowledgements

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 and the TRAMOOC project (Translation for Massive Open Online Courses), partially funded by the European Commission under H2020-ICT-2014/H2020-ICT-2014-1 under grant agreement number 644333.

[7] Moses in its default setting aligns maximum seven source/target words.

[8] We tested different thresholds between 1.0E-5 and 1.0E-1, whereby 1.0E-4 showed best performance.

[9] In our experiments we observed that translating only one sentence at a time, the OnDisk implementation performs at fastest. On the other hand, Compact implementation of the translation model performs fastest when translating an entire document. This implementation also benefits more from parallelizing the translation approach.

[10] more on: `http://server1.nlp.insight-centre.org/asistent/rest_service.html`

```
{
  "nbest":"5",
  "translation_direction":"en_sl",
  "method":"phrase_based",
  "text2translate":[
     {
        "source":"Accusations of witchcraft are also common in other African countries."
     }
  ]
}
```

```
{
  "time":"6 wallclock secs ( 0.02 usr  0.01 sys +  5.16 cusr  0.42 csys =  5.61 CPU)",
  "translation_direction":"en_sl",
  "nbest":"3",
  "method":"phrase_based",
  "text2translate":[
     {
        "source":"Accusations of witchcraft are also common in other African countries.",
        "possible_translations":{
           "obtožbe so pogosti tudi v čarovništva , druge afriške države . ":"-9.741",
           "obtožbe čarovništva so pogosti tudi v drugih afriških državah . ":"-9.644",
           "obtožbe o čarovništvu so pogosti tudi v drugih afriških državah . ":"-9.706"
        },
       "best":"obtožbe čarovništva so pogosti tudi v drugih afriških državah . "
     }
  ],
  "key":""
}
```


Figure 1: Illustration of JSON representations provided to and from the *Asistent* translation service.

9. References

Raphael Rubino Antonio Toral and Gema Ramírez-Sánchez. 2016. Re-assessing the Impact of SMT Techniques with Human Evaluation: a Case Study on English-Croatian. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT)*, volume 4, Riga, Latvia. Baltic Journal of Modern Computing.

Bogdan Babych, Anthony Hartley, and Serge Sharoff. 2007. Translating from under-resourced languages: comparing direct transfer against pivot translation. In *Proceedings of the MT Summit XI*, pages 412–418, Copenhagen.

Jonathan Clark, Chris Dyer, Alon Lavie, and Noah Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability . In *Proceedings of the Association for Computational Linguistics*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Thierry Etchegoyhen, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard Van Loenhout, Arantza Del Pozo, Mirjam Sepesy Maučec, Anja Turner, and Martin Volk. 2014. Machine Translation for Subtitling: A Large-Scale Evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC14)*, Reykjavik, Iceland.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom.

Marcin Junczys-Dowmunt. 2012. Phrasal rank-encoding: Exploiting phrase redundancy and translational relations for phrase table compression. *Prague Bull. Math. Linguistics*, 98:63–74.

Filip Klubička, Gema Ramírez-Sánchez, and Nikola Ljubešić. 2016. Collaborative development of a rule-based machine translator between Croatian and Serbian. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT)*, volume 4, Riga, Latvia. Baltic Journal of Modern Computing.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *HLT03*, pages 127–133, Edmonton, Canada.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Stroudsburg, PA, USA.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for

Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand.

Simon Krek. 2012. *Slovenski jezik v digitalni dobi – The Slovene Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Available online at `http://www.meta-net.eu/whitepapers`.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Nikola Ljubešić, Petra Bago, and Damir Boras. 2010. Statistical machine translation of Croatian weather forecast: How much data do we need? In Vesna Lužar-Stiffler, Iva Jarec, and Zoran Bekić, editors, *Proceedings of the ITI 2010 32nd International Conference on Information Technology Interfaces*, pages 91–96, Zagreb. SRCE University Computing Centre.

Mirjam Sepesy Maučec, Janez Brest, and Zdravko Kačič. 2006. Slovenian to English Machine Translation using Corpora of Different Sizes and Morpho-syntactic Information. In *Proceedings of the 5th Language Technologies Conference*, pages 222–225, Ljubljana, Slovenia.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318.

Maja Popović and Mihael Arcan. 2015. Identifying main obstacles for statistical machine translation of morphologically rich south slavic languages. In *18th Annual Conference of the European Association for Machine Translation (EAMT)*.

Maja Popović and Nikola Ljubešić. 2014. Exploring cross-language statistical machine translation for closely related South Slavic languages. In *Proceedings of the EMNLP14 Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 76–84, Doha, Qatar.

Maja Popović, David Vilar, Hermann Ney, Slobodan Jovičić, and Zoran Šarić. 2005. Augmenting a Small Parallel Text with Morpho-syntactic Language Resources for Serbian–English Statistical Machine Translation. In *ACL05-DDMT*, pages 41–48, Ann Arbor, MI.

Maja Popović. 2011. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 96:59–68.

Maja Popović. 2012. rgbf: An open source tool for n-gram based automatic evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics (PBML)*, 98:99–108.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Miro Romih and Peter Holozan. 2002. Slovensko-angleški prevajalni sistem (a slovene-english translation system). In *Proceedings of the 3rd Language Technologies Conference (in Slovenian)*, Ljubljana, Slovenia.

Xingyi Song, Lucia Specia, and Trevor Cohn. 2014. Data selection for discriminative training in statistical machine translation. In *17th Annual Conference of the European Association for Machine Translation*, EAMT, pages 45–53, Dubrovnik, Croatia.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 Metrics Shared Task. In *Proceedings of the 10th Workshop on Statistical Machine Translati on (WMT-15)*, Lisbon, Portugal.

Víctor M. Sánchez-Cartagena, Nikola Ljubešić, and Filip Klubička. 2016. Dealing with data sparseness in SMT with factored models and morphological expansion: a Case Study on Croatian. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT)*, volume 4, Riga, Latvia. Baltic Journal of Modern Computing.

Jörg Tiedemann. 2012. Character-based pivot translations for under-resourced languages and domains. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 141–151, Avignon, France.

Antonio Toral, Raphael Rubino, Miquel Esplà-Gomis, Tommi Pirinen, Andy Way, and Gema Ramirez-Sanchez. 2014. Extrinsic Evaluation of Web-Crawlers in Machine Translation: a Case Study on Croatian–English for the Tourism Domain. In *Proceedings of the 17th Conference of the European Association for Machine Translation (EAMT)*, pages 221–224, Dubrovnik, Croatia.

Francis M. Tyers and Murat Alperen. 2010. South-East European Times: A parallel corpus of the Balkan languages. In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pages 49–53, Valetta, Malta.

Jerneja Žganec Gros and Stanislav Gruden. 2007. The voiceTRAN machine translation system. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH 07)*, pages 1521–1524, Antwerp, Belgium. ISCA.

Richard Zens and Hermann Ney. 2007. Efficient phrase-table representation for machine translation with applications to online mt and speech translation. In *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting*, pages 492–499, Rochester, NY.