Mach Translat (2018) 32:167–189
https://doi.org/10.1007/s10590-017-9203-5

Neural machine translation for low-resource languages without parallel corpora

Alina Karakanta[1] · **Jon Dehdari**[1,2] · **Josef van Genabith**[1,2]

Received: 31 May 2017 / Accepted: 2 October 2017 / Published online: 7 November 2017

Abstract The problem of a total absence of parallel data is present for a large number of language pairs and can severely detriment the quality of machine translation. We describe a language-independent method to enable machine translation between a low-resource language (LRL) and a third language, e.g. English. We deal with cases of LRLs for which there is no readily available parallel data between the low-resource language and any other language, but there is ample training data between a closely-related high-resource language (HRL) and the third language. We take advantage of the similarities between the HRL and the LRL in order to transform the HRL data into data similar to the LRL using transliteration. The transliteration models are trained on transliteration pairs extracted from Wikipedia article titles. Then, we automatically back-translate monolingual LRL data with the models trained on the transliterated HRL data and use the resulting parallel corpus to train our final models. Our method achieves significant improvements in translation quality, close to the results that can be achieved by a general purpose neural machine translation system trained on a significant amount of parallel data. Moreover, the method does not rely on the existence of any parallel data for training, but attempts to bootstrap already existing resources in a related language.

✉ Alina Karakanta
alinak@coli.uni-saarland.de

Jon Dehdari
jon.dehdari@dfki.de

Josef van Genabith
josef.van_genabith@dfki.de

1 Saarland University, Saarbrücken, Germany

2 German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany

Keywords Neural machine translation · Low-resource languages · Closely related languages · Transliteration

1 Introduction

Human languages contribute significantly to the cultural and linguistic heritage of humankind. Natural language processing (NLP) can play a significant role in the survival and further development of all languages by offering state-of-the-art tools and applications to the speakers. Machine translation (MT) is one such application. MT can contribute to the rapid spread of vital information, such as in a crisis or emergency. A notable example is the recent refugee crisis where information has to be transferred rapidly from a large number of Asian languages (Farsi, Dari, Pashto) into European Languages and vice versa. However, many of these languages and a large number of other languages of the world are considered low-resource languages (LRL), because they lack in linguistic resources, e.g. grammars, POS taggers, corpora. For MT, the problem is further exacerbated by the lack of large amounts of quality parallel resources for training MT systems.

A number of methods have been developed to address this challenge, such as pivot-based approaches and zero-shot translation, which are described in detail in Sect. 2. The purpose of this paper is to explore an alternative scenario where there is no substantial readily available parallel data between the LRL and any other language. This paper is based on two observations. First, many low-resource languages are closely-related to a language for which there is significant amounts of parallel data, i.e. a high-resource language (HRL). Second, even though a language might lack in parallel resources, it is possible that monolingual data is available or can be collected from online news and other accurate web resources. The goal of this paper is to develop a language independent method that enables MT for low-resource languages and can be easily applied, for example in an emergency situation. The method is divided in two steps:

Transliteration HRL → LRL Based on the observation that many closely-related languages have high vocabulary overlap and almost completely regular orthographic or morphological differences, we bootstrap existing parallel resources between the HRL and the third language and use transliteration to transform the related HRL side of the parallel data so that it resembles the LRL. Since our language scenario does not assume any readily available parallel text between the LRL and any other language, we automatically extract transliteration pairs from Wikipedia article titles and explore an optional small bilingual glossary of the most frequent terms.

Back-translation of monolingual LRL data Although a language might lack in parallel texts, it is often the case that accurate monolingual online resources, such as newspapers, and web data from official sources are available. Contrary to phrase-based statistical MT—where monolingual data forms the language model—in neural machine translation (NMT) monolingual data is not so easily integrated without changes to the network architecture. For this reason, we use models trained on the transliterated HRL data to back-translate the monolingual LRL data into the third language and train our final models with the resulting data. Our assumption

is that accurate data is more efficient if used on the target side, a question explored in our work.

The article is structured as follows: In Sect. 2 we describe the related work and in Sect. 3 we explain in detail our method and experimental setting. In Sect. 4 we present the results of our experiments and in Sect. 5 we conduct a qualitative and quantitative analysis of the method presented and explore further possibilities for improvement. The conclusions of our work are presented in Sect. 6.

2 Related work

Our work is motivated by concepts related to MT without explicit parallel texts between the source and target language, but for which resources are created with methods such as transliteration after cognate extraction and back-translation of monolingual corpora. The related work in these fields is discussed in the sections below.

2.1 Machine translation without direct parallel data

The problem of lack of parallel resources in a language pair is often addressed using a third language as a pivot (Tiedemann 2012). This entails translating from language *A* to an intermediate language *I* and then to the target language *B*—a process called triangulation (Singla et al. 2014). However, this method requires parallel data between languages *A* and *B*, and at least one more language *I*. Decipherment techniques (Dou and Knight 2013; Dou et al. 2014) have been used to induce translation lexicons from non-parallel data and improve translation not only for out-of-vocabulary (OOV) words but also for observed words. More recently, Zoph et al. (2016) use transfer learning to enhance MT for low-resource languages. They first train a high-resource language pair (the parent model) and then transfer the learned parameters to the low-resource pair (the child model) to initialize and constrain training, achieving large improvements in quality. Johnson et al. (2016) have proposed a method for translating between multiple languages with a single model, using a shared vocabulary and a special token in the source sentence to specify the target language. They achieve zero-shot translation between a language pair for which no explicit training data has been seen. Zero-shot approaches have shown results close to the state-of-the-art while sparing the time and resources to train separate models for each language pair, but nonetheless rely on substantial parallel data for a given language, paired with other languages.

Another approach that might prove beneficial for low-resource languages—although it does not directly involve MT—is SuperPivot (Asgari and Schütze 2017). Based on a superparallel corpus build from the Parallel Bible Corpus (PBC) in 1169 languages, SuperPivot is able to search for a linguistic feature in any language. Based on the alignment of a head language to other languages, it projects the head pivot to a set of pivots and from there to all languages. These linguistic features, if extracted, can be incorporated in a NMT system and lead to significant improvements in quality, as shown in Sennrich and Haddow (2016).

2.2 Transliteration with cognate extraction

Transliteration has long been a subcomponent of MT and consists of the task of transforming the orthography of a string from one language into another, usually preserving pronunciation (Karimi et al. 2011; Zhang et al. 2015). Transliteration has been successfully used as a back-off especially between languages with different writing systems, like English ↔ Arabic (Durrani and Koehn 2014). It is a useful technique for the cases where MT fails to deal with OOV words, based on the observation that OOVs quite often constitute named entities, which have similar pronunciation among many languages.

For closely-related languages, transliteration expands beyond the concept of dealing with OOVs. The similarities among these languages on lexical and syntactic level enable transliteration methods using cognate extraction to translate between closely-related languages, improve word alignments (Simard et al. 1993; Bergsma and Kondrak 2007) or expand the parallel resources for (related) low-resource languages (Mann and Yarowsky 2001; Hana et al. 2006; Ismail and Manandhar 2010; Fišer and Ljubešic 2011a, b).

Nakov and Ng (2012) train a phrase-based transliteration model after using string similarity measures to extract cognates, and concatenate the resulting bi-text with the bi-text of the low-resource language. Their method produces significant improvements of + 1-3 BLEU points and reduces the amount of LRL parallel data needed to achieve the same result. Similarly, Currey et al. (2016) apply rule-based transliteration to enhance SMT for low-resource languages by transforming data from related languages and adding it to the training data. These approaches combine the transformed data from the related HRL with a small parallel corpus from the LRL. Contrary to these previous approaches, our work aims at utilising only transformed data.

2.3 Monolingual data

Recently, more research has been focusing on the use of monolingual corpora for NMT. Previous work combines NMT models with separately trained language models (Gülçehre et al. 2015). Sennrich et al. (2015) show that target-side monolingual data can greatly enhance the decoder model. They do not propose any changes in the network architecture, but rather pair monolingual data with automatic back-translations and treat it as additional training data. Contrary to this, Zhang and Zong (2016) exploit source-side monolingual data by employing the neural network to generate the synthetic large-scale parallel data and multi-task learning to predict the translation and the reordered source-side monolingual sentences simultaneously. Similarly, Jing et al. (2016) adapt NMT to the very low-resource language pair Mongolian–Chinese by representing words as lexeme-based sub-words, using an attention mechanism, monolingual data and a NMT correction model. This model is trained separately from the main model, with machine translated data as source and correct data on the target side, and acts as a re-translation of the output, with a view to correcting the translation results into the 'right' target language.

3 Method

The following sections describe the data, the transliteration method employed for transforming the data from the HRL to the LRL, and the neural MT system workflow used in our experiments.

3.1 Choice of languages

For the purpose of our experiments, we chose Belarusian as the low-resource language. According to Lewis et al. (2016), Belarusian or Belorussian is spoken by more than 2 million people in Belarus, Ukraine, Lithuania and Poland. Even though the majority of Belarusians speak Russian as their first language (L1), there has been an increasing interest in using Belarusian over the past few years. It is morphologically rich and linguistically close to Russian and Ukrainian, with transitional dialects to both. Russian was chosen as the related, high-resource language. Both Russian and Belarusian belong to the East Slavic family of languages and they share a high degree of mutual intelligibility. Russian is a good candidate for our task, i.e. translation LRL $\leftrightarrow$ EN using data from a related HRL, since there is a large amount of training data between English–Russian.

Since testing the method directly on a low-resource language poses several challenges, such as absence of readily available test sets and language specific pre-processing tools, we additionally experiment with a non-low-resource language, as an extra validation process for our method. In this scenario, Spanish acts as the pseudo-low-resource language (PS-LRL) and Italian as the closely-related HRL. Spanish and Italian both belong to the Romance family of Indo-European languages. Spanish has strong lexical similarity with Italian (82%) (Lewis et al. 2016). Among major Romance languages, Spanish and Italian have been found to be the second-closest pair (following Spanish and Portuguese) in automatic corpus comparisons (Ciobanu and Dinu 2014) and in comprehension studies (Voigt and Gooskens 2014). One more reason for choosing Spanish and Italian is to compare our results with the previous work of Nakov and Ng (2012) and Currey et al. (2016).

3.2 Data

For the low-resource setting, the training data consists of Russian–English data from WMT2016,[1] and the Belarusian monolingual data is taken from Web to Corpus (W2C), built from texts available on the Web and Wikipedia and is available for a large number of languages (Majliš and Žabokrtský 2012). In order to provide a proper test scenario, the development and test sets were compiled from bilingual articles extracted from the Belarusian Telegraph Agency (BelTA).[2] 150 articles in Belarusian and English were

[1] http://www.statmt.org/wmt16/translation-task.html.

[2] http://eng.belta.by.

Table 1 Datasets used for the low-resource language experiments and their size

	Dataset	Words	Sentences
Train RU–EN	Yandex	18.5 M	1,196,245
	NewsCommentary v11	4.2 M	
Dev BE–EN	BelTA	37 k	1546
Test BE–EN	BelTA	22 k	1006
Test RU–EN	newstest2013	56 k	3000
Back-trans (BE monol.)	W2C	23 M	1,821,359

Train/dev/test bitext corpus sizes are given in words/sentences on the English side

Table 2 Datasets used for the pseudo-low-resource language experiments and their size

	Dataset	Words	Sentences
Train IT–EN	Europarl-v7	23 M	950,000
Dev ES–EN	newstest2009-src	68 k	3027
Test ES–IT–EN	test2008	54 k	2000
Back-trans (ES monol.)	Europarl-v7	23 M	950,000

Train/dev/test bitext corpus sizes are given in words/sentences on the English side

manually collected. Then the main text was extracted using the tool Justext,[3] dates were removed and the remaining sentences were written to files one sentence per line. The bilingual, tokenized text was aligned at sentence level using the Hunalign sentence aligner.[4] Since the highest possible quality was required, sentences with a score lower than 0.4 were excluded and only bisentences (one-to-one alignment segments) were preserved, for which both the preceding and the following segments were one-to-one. The number of sentences left in the test set is 1006. The remaining sentences (1547) were used to create the validation set. Statistics of the data are presented in Tables 1 and 2.

For the pseudo-low-resource scenario, the training data is taken from Europarl (Koehn 2005). The size of the training data was paired with the size of the training data for our low-resource language scenario, in order to achieve comparison of results between the two scenarios.

As a preprocessing step, all data was tokenized using the Moses tokeniser with the language setting for each of the languages we experiment with. Since there is no language specific tokeniser for Belarusian, the Belarusian data was tokenised with Russian language setting. Then, the data was truecased and sentences longer than 50 tokens were ignored. Byte-Pair Encoding (Sennrich et al. 2016) was applied for subword segmentation to achieve common subword units. The number of merge oper-

[3] http://corpus.tools/wiki/Justext.

[4] http://mokk.bme.hu/resources/hunalign/.

ations was set to 50,000. In the low-resource language scenario, BPE was trained separately for English and jointly for Russian, Belarusian and transliterated Russian, since the languages have different script. For training BPE jointly, the corpora of the respective languages were concatenated. The final vocabulary sizes for the WMT data were 49,292 for Russian, 24,175 for English and 30,850 for Russian transliterated into Belarusian (BE_{ru}), and for the W2C data 46,584 for monolingual Belarusian and 32,042 for back-translated English. For the pseudo-low-resource scenario, BPE was trained jointly for all languages, i.e. English, Spanish, Italian and transliterated Italian. The final vocabulary sizes were 25,988 for Italian, 20,079 for English, 29,179 for Italian transliterated into Spanish (ES_{it}), 27,040 for the monolingual Spanish and 15,754 for back-translated English.

3.3 Transliteration

In the following sections we present the creation of a bilingual glossary of the most frequent words in the HRL corpus and their LRL translations, the method of extracting transliteration pairs and the transliteration system used for transforming the HRL data into data more similar to the LRL.

3.3.1 Glossary

Although transliteration pairs from Wikipedia articles might be suitable training data for transliterating named entities, the goal of the presented method is to transliterate sentential data. After examining the extracted pairs, we noticed that very frequent words, which are usually function words, are not present in the pairs, which might deteriorate the transliteration output in the case of sentences. For this purpose, we created a bilingual glossary containing many of these words.

The 200 most frequent words were extracted from the Russian training corpus and manually translated into Belarusian by linguists. In cases where multiple translations were possible, the frequency of the two variants was checked in the Belarusian corpus. If the difference in frequency was large, the most frequent word was kept. When the words were equally or almost equally frequent, the more similar word to the Russian word was kept. Table 3 shows examples of bilingual glossary entries, their English translation and their frequency. Most of them represent function words. It is worth noting that most multiple translations were caused by the letter $ў$ or $у$, which is particular to pronunciation in Belarusian. The letter $ў$ is called the *non-syllabic u* because it does not form syllables. When a word beginning with $у$ follows a vowel, it is replaced by $ў$. Following the frequency principle above, the $у$ variants were chosen in the glossary. As a first step, the glossary was used to substitute the words in the Russian part of the data with their translations in Belarusian, similar to word-based translation, in order to obtain an additional baseline for our NMT experiments; later, it was used as part of the training data for the transliteration system.

The same procedure was followed for creating the glossary for the pseudo-low-resource language, IT → ES.

Table 3 Examples of glossary entries and their frequency in the Russian corpus

Russian	Belarusian	Translation	Frequency
и	і	and	771, 273
в	у	in(to)	761,383
на	на	on	323,262
с	з	off/from	237,650
не	не(ня)	not	213,694
что	што	what	210,355
по	па	by	146,415
как	як	as	113,104
В	у	at	103,823
или	ці	or	93,674
его	яго	him	57,503
также	таксама	also	56,544
может	можа	can/may	43,636
уже	ужо(ўжо)	already	20,973

The words in parenthesis are the alternative translations which were removed from the glossary

3.3.2 Cognate pair extraction from Wikipedia titles

One of the best-suited sources for obtaining bilingual resources when there is no explicit parallel text available is Wikipedia. Even 'small' languages can have strong Wikipedia presence. A notable example is Aragonese, with 300 articles per speaker. The most straightforward way to extract bilingual vocabularies from Wikipedia is extracting the titles of the articles. Wikipedia dumps offer per-page data, which includes article IDs and their titles in a particular language and the interlanguage link records, which include links to articles on the same entity in other languages. The data was downloaded and the titles extracted as described in the project `wikipedia-parallel-titles`.[5] Table 4 shows examples of extracted bilingual entries, their translation into English, and the similarity score based on the Longest Common Subsequence Ratio (LCSR).

It is clear from Table 4 above that the majority of the titles are Named Entities referring to the same object, person or abstract notion. In other words, they are words that have similar orthography and are mutual translations of each other, i.e. *cognates* (Melamed 1999; Mann and Yarowsky 2001; Bergsma and Kondrak 2007). Therefore, the bilingual entries are suitable data to train a transliteration model between Russian and Belarusian after using string similarity measures to extract transliteration pairs, as shown in Nakov and Ng (2012).

For RU → BE, 145,910 titles were extracted in total. First, the data had to be cleaned. Indexes preceding the entities, such as *Category:*, *Annex:* etc. were erased and then duplicate entries and entries with Latin, Greek or Asian scripts were removed (11,269 entries). For the Russian part of the extracted data, the format for names was in many cases *Surname, Name*, which did not correspond to the Belarusian format *Name*

[5] https://github.com/clab/wikipedia-parallel-titles.

Table 4 Examples of extracted bilingual entries from Wikipedia articles between Belarusian and Russian and their translation into English

Belarusian	Russian	Translation	LCSR
Аазіс Гадамес	Гадамес	Ghadames	0.5600
Авіяцыйныя дывізіі СССР	Авиационные дивизии СССР	Aviation division of the USSR	0.5870
Скуп**шч**ына	Скуп**щ**ина	Assembly	0.7222
Транснептунавы аб’ект	Транснептуновый об**ъ**ект	Trans-neptunian object	0.8605
В**і**к**і**педыя на **і**дышы	В**и**к**и**пед**и**я на **и**д**и**ше	Yiddish Wikipedia	0.6471
Заходняя Е**ў**ропа	Западная Европа	Western Europe	0.7586
Федэральны Урад Германіі	Федеральное правительство Германии	Federal Government of Germany	0.5455

Letters in bold represent some regular spelling differences between the languages. Only pairs with $LCSR > 0.58$ are retained for training the transliteration model

Surname. This would cause the entries to be left out when the similarity measure is applied. For this reason name and surname were switched. For the Belarusian part, some entries contained a description after the named entity, e.g. *Вікінг, фільм, 2016* (Viking, film, 2016) whose equivalent in Russian was simply *Викинг* (Viking). For the reason mentioned above, the description was removed too and only the first part of the entry was kept. As a similarity measure, we chose the *Longest Common Subsequence Ratio* (LCSR) (Melamed 1999), which is defined as follows:

$$LCSR(s_1, s_2) = \frac{|LCS(s_1, s_2)|}{\max(|s_1|, |s_2|)} \tag{1}$$

where LCS(s_1,s_2) is the *longest common subsequence* of strings s_1 and s_2 and |s| is the length of string s.

Following Kondrak et al. (2003), we retained only pairs of entities with LCSR > 0.58, a value they found to be useful for cognate extraction in many language pairs. We also retained pairs without spelling differences, i.e. with LCSR = 1. The remaining transliteration pairs were 80,608. This was further split into train (78,608 pairs), development (1000 pairs) and test (1000 pairs) set. The data was further tokenized.

It should be noted that the preprocessing method described above was followed to achieve the best possible quality of transliteration pairs while at the same time preserve as many pairs as possible for training. If LCSR is applied directly on the extracted titles, without any other preprocessing, only 70,911 titles in Cyrillic script are left for training, development and testing.

^ А д р и а н $ ^ С м и т $ - ^ А д р ы я н $ ^ С м і т $

Fig. 1 Example of a bilingual title pair split into characters

The same procedure was followed in the pseudo-low-resource scenario, for extracting pairs IT → ES, allowing us to obtain 409,582 pairs after applying the data cleaning techniques mentioned above. This was further split into train (407,582 pairs), development (1000 pairs) and test (1000 pairs) set.

The resulting bilingual titles were tokenised and split into characters. Special word-initial and word-final symbols (∧ and $) were introduced to mark word boundaries, as shown in Fig. 1.

Then, transliteration was applied to the sentences of the related HRL data, i.e. the Russian part of the EN–RU corpus, and the Italian part of the EN–IT corpus. In order to transliterate the sentences, the data had to be split into characters too. In addition, the text was split one word per line, since the system was trained on short titles and therefore unable to translate long sequences. A special end-of-sentence token (⟨end⟩) was introduced to restore the original sentence alignment with the English part of the parallel data.

3.3.3 Transliteration system

The system chosen for the purpose of the experiments is OpenNMT, an open-source neural MT system utilising the Torch toolkit (Klein et al. 2017). The RNNs consisted of two layers of 500-node LSTMs, with 500-node embeddings on the source and target sides. For optimisation, Adam was used for training with a minibatch size of 64 and a learning rate of 0.002. We trained the models for 10 epochs and conducted two experimental runs, one with a bidirectional encoder and one without. We evaluated the model transliterating the titles test set as well as a test set with full sentences. The results of the transliteration experiments are presented in Sect. 5.1.

3.4 Neural machine translation system

For the translation system, we used the same attention-based NMT system that we used for our transliteration experiments. We trained the sequence-to-sequence models with global attention and a 2-layer LSTM with 500 hidden units each on both the encoder/decoder with the same optimisation parameters as in the transliteration experiments. Drop-out was set to 0.3 and vocabulary size to 50,000. Models were trained for 7 epochs, after which we did not observe large improvements on the development set. We propose a method for integrating the monolingual data from the low-resource language without any changes in the system architecture, presented in Fig. 2.

Our final goal is to create a system for translating from the third language (English) into LRL language. Since there is no readily available parallel data for training such a system, we exploit the RU–EN data and train two systems, as seen in Fig. 2. The steps are the following:

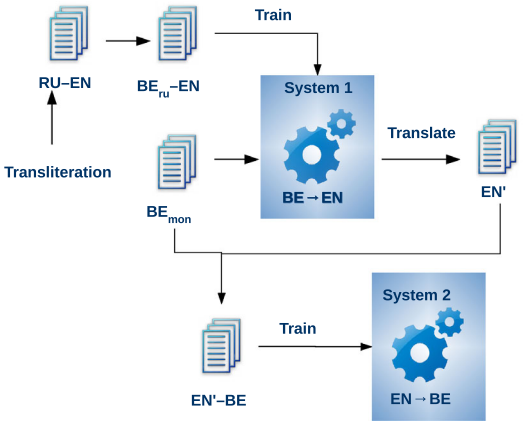

Fig. 2 Experimental work-flow for the NMT system

1. After training the transliteration system as described in Sect. 3.3, transform the HRL–EN (RU–EN) data to LRL–EN (BE$_{ru}$–EN) data.
2. With the transliterated data, train a BE → EN MT system (System 1).
3. Translate monolingual LRL data (BE$_{mon}$) into English, using System 1.
4. Train our final system (System 2) to translate from English into the LRL (EN → BE), using the parallel corpus generated from System 1; i.e. the monolingual LRL data and the machine-generated English (BE$_{mon}$–EN$'$).

System 2 can also be used to translate from the LRL into English, however, we expect that target-side monolingual data of high quality will be more beneficial, a question explored in Sect. 5.

For evaluation, BLEU (Papineni et al. 2002) was calculated on tokenised output. Because our method involves transliteration, which is applied at a character level, we found it also useful to evaluate the output with character-based metrics, which reward some translations even if the morphology is not completely correct. For this reason, we additionally report BEER (Stanojević and Sima'an 2014) and CHRF3 (Popović 2015) scores.

4 Experiments

In the following sections we present the experiments carried out. First, we describe the baseline system for the low-resource and the pseudo-low-resource language scenario. Then, we perform experiments to explore (1) the effect of our transliteration method applied on the training data and (2) the use of monolingual LRL data.

4.1 Baseline system

In the baseline, we used the following setups: First, we trained models on the related HRL data without any transformation. Then, we applied simple word substitution HRL → LRL using the glossary to establish whether a minimum transformation would

Table 5 Results for the baseline systems RU ↔ EN and BE_{ru} ↔EN after training for 7 epochs

Train	Test	BLEU	BEER	chrF3
RU → EN	RU→EN	17.36	0.495	0.447
RU → EN	BE→EN	1.19	0.285	0.129
BE_{ru} →EN (GLOS)	RU → EN	6.00	0.381	0.262
BE_{ru} →EN (GLOS)	BE → EN	1.36	0.320	0.209
EN → RU	EN→RU	15.68	0.476	0.416
EN → RU	EN→BE	1.23	0.318	0.202
EN → BE_{ru} (GLOS)	EN → RU	0.67	0.367	0.290
EN → BE_{ru} (GLOS)	EN → BE	0.35	0.315	0.210

GLOS refers to the minimum transformation of the HRL data by word substitution using only the bilingual glossary

affect the quality of the output and in which way. We evaluated both with an HRL–EN and an LRL–EN test set in order to determine whether the transformed data is more 'HRL-like' or 'LRL-like'. Unfortunately, the test sets between Russian and Belarusian are not parallel, which does not allow an accurate comparison, but this is explored in the pseudo-low-resource scenario.

4.1.1 Low-resource language EN ↔ BE

For the low-resource scenario, the results for the models are shown in Table 5. For the into-English direction, the model trained on non-transformed HRL data scores high when evaluated on RU–EN. However, the scores for evaluation with BE–EN are very low. This suggests that even though they have the same script, Russian and Belarusian still have important differences in spelling, syntax and vocabulary. When word substitution is performed using the glossary (GLOS), the scores for testing on BE–EN improve only slightly, while the quality detriments for testing on RU–EN. Despite this, the scores for RU–EN are still higher than for BE–EN. It seems that the minimum transformation with the glossary only inserts noise in the training data and therefore is not sufficient to achieve improvements in the translation output. This can also be attributed to the morphological complexity of Belarusian, which implies that even thought the translations in the glossary are correct, the morphology is not and therefore simple word substitution is too naive to achieve improvements.

The same tendency is observed in the opposite direction, i.e. EN → LRL. There is one difference; the score for the minimum transformation and testing on EN–RU drops and is almost equal to testing with EN–BE. While in the into-English direction the transformation on the source side affected the quality of the output, the effect was much larger for the target side. This shows the importance of quality target-side data and supports our assumption of using monolingual instead of transformed data on the target side.

Table 6 Results for the baseline systems IT → EN and ES_{it} →EN after training for 7 epochs

Train	Test	BLEU	Beer	chrF3
IT → EN	IT→EN	27.86	0.600	0.543
IT → EN	ES→EN	3.64	0.339	0.202
ES_{it} →EN (GLOS)	IT → EN	21.68	0.554	0.488
ES_{it} →EN (GLOS)	ES → EN	6.69	0.409	0.306
ES_{it} →EN (RB)	IT → EN	7.10	0.413	0.303
ES_{it} →EN (RB)	ES → EN	9.09	0.436	0.327
EN → IT	EN→IT	23.59	0.582	0.551
EN → IT	EN→ES	1.76	0.332	0.264
EN → ES_{it} (GLOS)	EN → IT	7.03	0.493	0.457
EN → ES_{it} (GLOS)	EN → ES	6.27	0.452	0.368
EN → ES_{it} (RB)	EN → IT	2.99	0.431	0.359
EN → ES_{it} (RB)	EN → ES	8.08	0.473	0.391

GLOS refers to the minimum transformation of the HRL data by word substitution using only the bilingual glossary. RB refers to the rule-based transliteration

4.1.2 Pseudo-low-resource language EN ↔ ES

In the pseudo-low-resource language scenario, we experimented with two types of transformations, in addition to the non-transformed baseline. First, we only applied word substitution using the glossary, as in the low-resource scenario. Then, we used the rule-based substitution proposed by Currey et al. (2016), which consists of 50 string substitution rules in combination with the glossary of the 200 most-frequent words. The results are presented in Table 6.

In this case, the test sets are parallel in all languages involved (HRL, LRL, third language), which allows us a more accurate comparison. As in the low-resource scenario, training with non-transformed data scores very low for EN–ES while results close to the state-of-the-art are achieved for evaluation with EN–IT. The rule-based transliteration method is more sophisticated than the simple word substitution with the glossary, which leads to further improvements.

4.2 Low-resource languages: Belarusian ↔ English using Russian

This section presents the results achieved for the experiments Belarusian ↔ English using Russian. It is divided in two parts; first, we explore the efficiency of the transliteration method in a NMT application (System 1), and second, we experiment with back-translating monolingual LRL data with System 1 and using the resulting parallel corpus to train our final models (System 2).

4.2.1 Belarusian ↔ English using transliterated Russian

The experiments described in this section consist of applying the transliteration models (see Sect. 3.3) to transform the data from the HRL into data similar to the LRL (LRL_{hrl}-EN). Results are shown in Table 7. Using this method, we achieve an improvement

of approximately 4 BLEU points for EN → BE and 9 points for BE → EN over the non-transformed baseline. Again, having non-transformed data on the target side is more beneficial for the models.

4.2.2 Belarusian ↔ English using back-translated monolingual Belarusian data

As a second step, the monolingual LRL data was translated into English using the models trained in Sect. 4.2.1, in order to generate parallel data between the LRL and English. Results are shown in Table 7. In this case, the English side of the data is machine-generated and therefore, its quality is not optimal. As a result, there is a decrease of 1 BLEU point between the BE → EN′ model using monolingual Belarusian data and machine-translated English (MONO), and the BE_{ru} →EN model where HRL data is transliterated and the target EN is original (TRANSLIT). However, the BE → EN′ model still scores higher than the EN′ →BE model, even though there is an improvement of 2 BLEU points for using monolingual BE data over transliterated BE_{ru}. It is also worth noticing the scores from the other two metrics apart from BLEU. Although there is a BLEU score improvement of 2 points for the EN′ →BE case, the scores for BEER and CHRF3 are almost the same. Contrary to this, in the BE → EN′ direction, the difference in these scores is larger, even though the difference in BLEU score is smaller. This can be attributed to the fact that BEER and CHRF3 take into account character *n*-grams, while BLEU does not. All scores show a similar tendency inside the same direction. However, BLEU does not agree with the other two metrics in the (MONO) experiment, where BLEU rewards the BE → EN′ direction while with BEER and CHRF3 the opposite direction scores higher.

All in all, the method presented in this paper led to significant improvements, but when the results for the low-resource scenario are compared with the results of the pseudo-low-resource languages, the former are lower. This can be attributed to a number of factors. First, the number of transliteration pairs extracted is small and might not have been sufficient to transform the HRL data accurately. Another factor is the quality of the monolingual data. Even though it is original Belarusian language data, the W2C corpus is built exclusively from web data and often contains missing words, multiple punctuation marks, etc. A corpus built from texts from official newspapers and websites could lead to better performance. It is also possible that the automatically-aligned test set contains some misaligned sentences, which would reduce the score

Table 7 Results for systems BE ↔ EN after training for 7 epochs

Train	Test	BLEU	BEER	chrF3
EN → BE_{ru} (TRANSLIT)	EN → BE	5.31	0.428	0.358
BE_{ru} →EN (TRANSLIT)	BE → EN	10.83	0.439	0.342
EN′ →BE (MONO)	EN → BE	7.73	0.429	0.360
BE → EN′ (MONO)	BE → EN	9.82	0.421	0.313

TRANSLIT refers to System 1, where the related HRL data is transliterated (BE_{ru}) and the model is trained with the resulting parallel data. MONO refers to System 2, where monolingual LRL data is back-translated and the model is trained with the resulting parallel data

even for correct translations. Lastly, the morphological complexity of the languages in this scenario poses an additional challenge for our models.

4.3 Pseudo-low-resource languages: Spanish ↔ English using Italian

The following sections report the results of the experiments performed with the pseudo-low-resource languages.

4.3.1 Spanish ↔ English using transliterated Italian

Even larger improvements are achieved for the pseudo-low-resource language scenario. Results are shown in Table 8. Here, our transliteration method leads to the transformation of the HRL data to a greater extent and into data that is more similar to the LRL. This claim is supported by the very low scores for testing with EN ↔ IT. The scores achieved here are higher compared to the scores of the low-resource language scenario. A possible reason for this could be the larger amount of training data for transliteration, since the number of transliteration pairs for IT → ES is four times larger than for RU → BE.

Using the transliteration method proposed in this work, we obtain an increase of 4 BLEU points for EN → ES_{it} and 3.5 points for ES_{it} → EN over transliterating with the rule-based method proposed by Currey et al. (2016). Additionally, we restrict the human effort to creating the glossary, instead of drafting rules separately for each language pair.

4.3.2 Spanish ↔ English using back-translated monolingual Spanish data

The results achieved for the models using back-translated monolingual ES for training are close to results that are possible to achieve with NMT systems using original, non-transformed data. The EN′ → ES direction improves by +4 BLEU points over the setting with the transliterated Spanish data (ES_{it}), showing that incorporating the non-synthetic monolingual data on the target enhances greatly the translation output.

Table 8 Results for systems ES ↔ EN after training for 7 epochs

Train	Test	BLEU	BEER	chrF3
EN → ES_{it} (TRANSLIT)	EN → ES	12.20	0.497	0.430
EN → ES_{it} (TRANSLIT)	EN → IT	2.50	0.414	0.335
ES_{it} → EN (TRANSLIT)	ES → EN	13.67	0.488	0.406
ES_{it} → EN (TRANSLIT)	IT → EN	5.70	0.391	0.272
ES → EN′ (MONO)	ES → EN	13.45	0.479	0.388
EN′ → ES (MONO)	EN → ES	18.59	0.529	0.463

TRANSLIT refers to System 1, where the related HRL data is transliterated (ES_{it}) and the model is trained with the resulting parallel data. MONO refers to System 2, where monolingual LRL data is back-translated and the model is trained with the resulting parallel data

For the ES → EN′ direction, there is a slight drop, similar to the low-resource language scenario.

The results obtained with the method described in this paper are lower than the ones mentioned in Nakov and Ng (2012) and Currey et al. (2016). However, the results are to a great extend incomparable due to a number of factors. Nakov and Ng (2012) (1) created their training/dev/test dataset out of Europarl v.3, while our training data is a part of Europarl v.7 and the trilingual test set comes from WMT; (2) they only perform experiments Spanish → English, while our main objective is to improve MT into the LRL; (3) we use 23 M words IT–EN for training instead of 44 M and (4) they use a small amount of parallel data ES–EN which they enhance with additional transliterated data.

Currey et al. (2016) use a vary small amount of training data; 30 k sentences of Spanish parallel data, aided by 30 k of transliterated Italian and 30 k of transliterated Portuguese, and a Spanish language model of 14 M sentences, which could have 'filtered' out some mistransliterations. Training/dev/test data was obtained from the same source as for Nakov and Ng (2012). Their experimental results show a slight improvement over baseline, which is significant only for enhancing the translation model with two related languages (+ 0.24 BLEU points), while when using only Italian data there is no improvement. When training our model with transliterated data using our proposed transliteration method (TRANSLIT) we achieve an increase of +4 BLEU points over using data transliterated with the rule-based transliteration method. A total increase of +10 BLEU points over their method was reached when back-translated monolingual data was used to train the model (MONO), which is a large improvement compared to the previous work.

It should be noted that both approaches rely heavily on the existence of some parallel data, starting with a minimum of 10 k sentences. For most low-resource languages, even this small amount of data is hard to obtain. Instead, our method deals with extreme cases of low-resource languages by developing parallel data from scratch. In the analysis of our work, we attempt to integrate very small amounts of parallel data, to determine the effect on the translation output (see Sect. 5).

5 Analysis

5.1 Transliteration

For the transliteration system we experimented with several settings. We found that the best results were achieved when training a sequence-to-sequence LSTM model with a bidirectional encoder and global attention after splitting the input into characters. We trained the system for 10 epochs, after which we did not observe any improvement on the development set. For RU → BE tuning BLEU was 94% and for IT → ES 95%, a value comparable to the one reported by Nakov and Ng (2012) (94.92%). The glossary of the 200 most frequent words gave further small improvements to the performance of the transliteration only in earlier epochs, but we show below that it could be valuable for transliterating sentences.

Original Russian Все люди рождаются свободными и равными в своем достоинстве и правах.
Transliteration Усе людзі Горад Намдаюця Увабоднымі і Раўнымі у своем дастоінствэ і правах.
Original Belarusian Усе людзі нараджаюцца свабоднымі і роўнымі ў сваёй годнасці і правах.
English Translation All human beings are born free and equal in dignity and rights.

Fig. 3 Example of a Russian sentence transliterated into Belarusian and the original Belarusian sentence

Original Italian Tutti gli esseri umani nascono liberi ed eguali in dignità e diritti.
Transliteration Tuttos los eserios humanos nascono libres e eguales en dignidad y derechos.
Rule-based transliteration Todos los eseri nines nascono liberi y eguales en diñidad y derechos.
Original Spanish Todos los seres humanos nacen libres e iguales en dignidad y derechos.

Fig. 4 Example of an Italian sentence transliterated into Spanish with our method and the method proposed by Currey et al. (2016) and the original Spanish sentence

In order to explore whether the approach described above can be applied to achieve our final goal in this part, i.e. transliterating sentential data from a related language, we present some quantitative results along with the automatic evaluation. The first sentence of Article 1 of the Universal Declaration of Human Rights is transliterated from the related HRL into the LRL with our best model and compared to the official translation in the target language. The sentences are presented in Figs. 3 and 4.

At first sight, the original Russian and Belarusian sentences have an exact string match of only one word. However, their structure is similar and they contain a large number of cognates. If we ignore the orthographic differences, the actual vocabulary overlap is almost 70%. After applying our transliteration, the vocabulary overlap at the individual word level increases from 9% to 45%. It is worth nothing the changes at a character-level. Character error rate (CER) is 53% and BLEU at a character level 41% even thought regular word-level BLEU score does not reward any similarities in the sentence. Moreover, the model was able to distinguish the cases when the Russian "*и*" – which is not present in the Belarusian alphabet – is transliterated as either "i" or "Зі". On the other hand, when transliterating the word свободными (free) the model inserted an error, the character "y" instead of preserving the "c", producing увабоднымі.

For the Italian → Spanish transliteration, we observe a 67% vocabulary overlap after applying our model, compared to a 0% vocabulary overlap for the untransformed sentences. It seems that some regular spelling differences, such as the adjective ending *-i* to *-os* or *-es*, are successfully modelled. On the other hand, the rule-based method

Table 9 Scores for different data selection options for transliteration IT → ES and in comparison with the rule-based transliteration by (Currey et al. 2016)

Train	Test	BLEU	WER	BEER	CHRF3
LCSR > 0.58 + 1xGlos	Newstest	12.14	0.744	0.489	0.432
LCSR > 0.58 + 10xGlos	Newstest	15.08	0.699	0.512	0.453
1 > LCSR > 0.58 + 1xGlos	Newstest	11.47	0.732	0.502	0.442
1 > LCSR > 0.58 + 10xGlos	Newstest	13.63	0.710	0.516	0.459
Rule-based	Newstest	10.38	0.727	0.495	0.422
LCSR > 0.58 + 1xGlos	Titles	74.12	0.107	0.841	0.919
LCSR > 0.58 + 10xGlos	Titles	73.91	0.109	0.839	0.916
1 > LCSR > 0.58 + 1xGlos	Titles	47.64	0.333	0.740	0.831
1 > LCSR > 0.58 + 10xGlos	Titles	48.97	0.320	0.739	0.830
Rule-based	Titles	36.70	0.331	0.719	0.757

All scores are calculated after post-processing the output into words

transforms some words correctly, but introduces several errors, such as the transformation of the word *umani* in *nines* instead of *humanos* or does not take some words into account, e.g. *liberi* is left untransliterated.

In the pseudo-low-resource language scenario, the trilingual (IT, ES, EN) test set available allowed us to test different data selection and concatenation options. We wish to establish whether adding more weight to the glossary and preserving transliteration pairs without spelling differences is beneficial to transliterating sentences. For this reason, we additionally experimented with (1) concatenating the glossary x10 with the training data and (2) excluding pairs without any spelling differences (140,278 remaining pairs).

Table 9 shows the results for transliteration with different data selection options. The glossary might not have yielded any improvement for transliterating the titles, but it proved useful in the case of sentences. The best scores for transliterating sentences according to the two character-based metrics were achieved when pairs without spelling differences are preserved and the glossary is concatenated x10 with the training data, even though BLEU and Word Error Rate (WER) score favour preserving only pairs with spelling differences. This effect could simply have been caused by the existence of significantly more training data. The results suggest that for closely-related languages, simple methods, such as transliteration, might be sufficient for achieving quality output. Additionally, we report results after applying the transliteration method proposed by Currey et al. (2016) on the same test sets. For both test sets, the rule-based method scores significantly lower for all metrics.

5.2 Effect of parallel data

We explore whether the presence of a very small amount of parallel data between the low-resource language and the third language can significantly enhance the quality of the output and what is the most effective method to exploit the parallel data. To

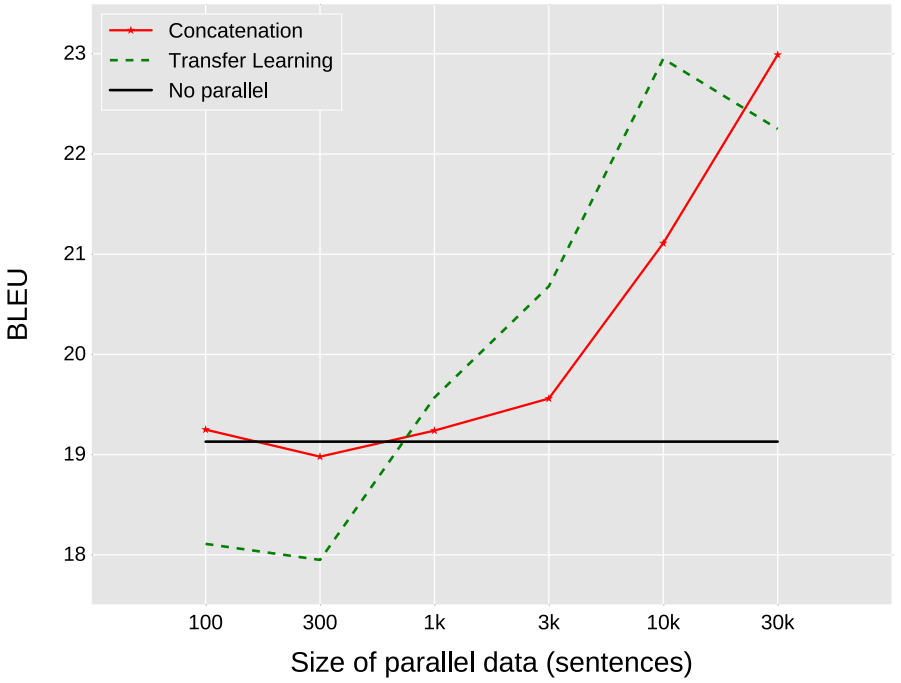

Fig. 5 BLEU scores for different parallel corpus (EN → ES) sizes and methods

determine the effect of such small amounts of parallel data, we used 100, 300, 1 k, 3 k, 10 k and 30 k sentences of parallel text, strictly distinct from the data used so far, in further experimental runs. The parallel data was combined with the data for training System 2, i.e. the monolingual LRL and the machine-translated English data. We experimented with two different ways of incorporating the parallel data into the models: concatenating the parallel corpus; and using transfer learning, i.e. training the existing models for further epochs with the parallel corpus. For transfer learning, the models were first trained with the EN′ →ES data for 7 epochs and then training was resumed with the small parallel EN → ES data.

The results for training with the concatenation and the transfer learning methods for 10 epochs are shown in Fig. 5 and compared to the setting without any parallel data. The concatenation did not yield any significant improvements on the BLEU score. The largest improvement was of +0.2 BLEU points for concatenating the 10 k corpus with the training data. On the contrary, the transfer learning method (TF) significantly improved the quality parallel corpus sizes above 1 k, but not with the very small sizes of 100 and 300 sentences. In addition, we observe that larger parallel corpus sizes require longer training with the transfer learning method, as can be seen with the 30 k corpus size. Figure 6 shows the effect of the each parallel data size for different numbers of epochs for transfer learning.

Another question related to the use of parallel data, which can also prove the effectiveness of the proposed method, is how much parallel data between the LRL and EN

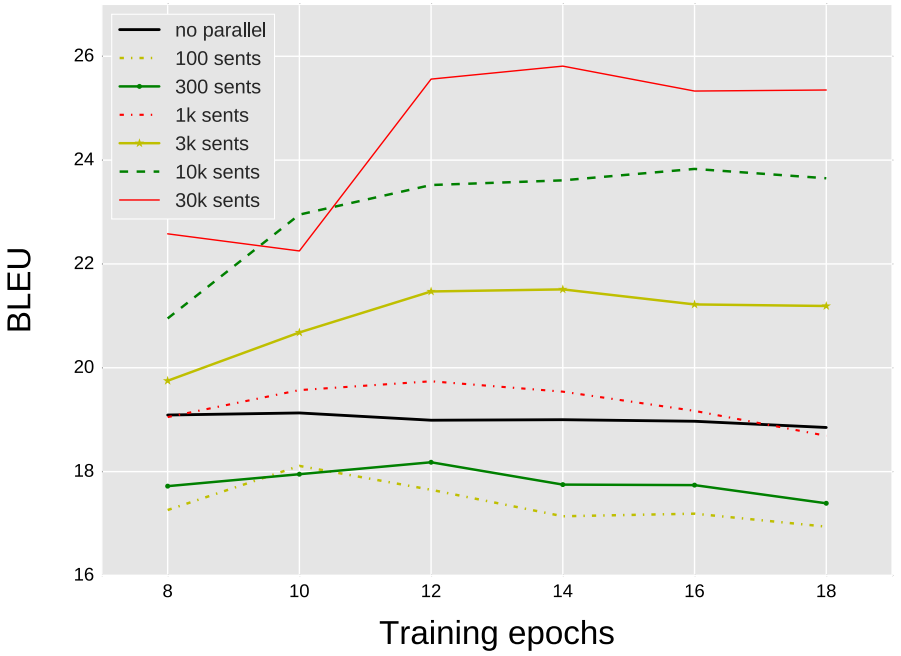

Fig. 6 BLEU scores for different parallel corpus (EN → ES) sizes and epochs for using transfer learning

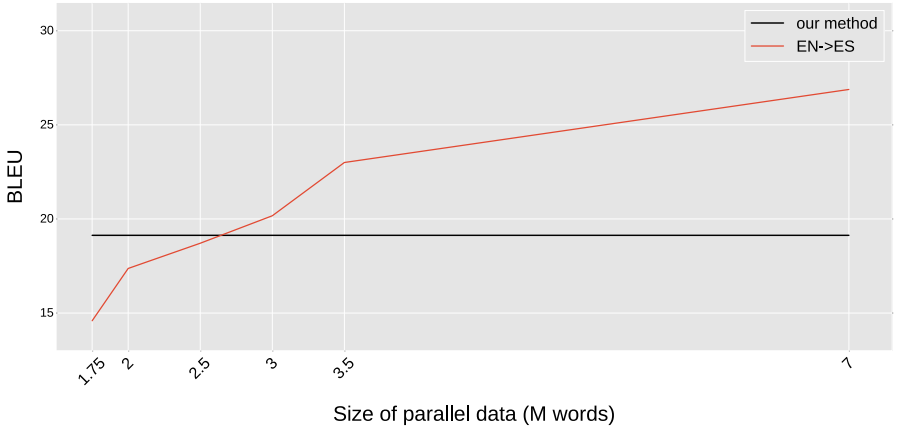

Fig. 7 Size of parallel data EN → ES required to achieve our best model's score

would be required to achieve the same results. One might argue that the method is complex and training NMT systems is too time-consuming, and that the same effort could have been used to collect or create parallel data. However, Fig. 7 shows that for EN → ES the same scores are achieved with around 2.7 M words of training data. This is a significant amount of data that cannot be collected in a short time, especially for LRLs. To give a measure of the amount of data required, we provide the following example: In order to translate this amount of data, a professional translator with

the average daily capacity of 2500 words/day would need 1080 days. This clearly shows that the proposed method can be efficient for the extreme cases of low-resource languages.

6 Conclusion

Enabling MT for low-resource languages poses several challenges due to the lack of parallel resources available for training. In this paper, we present a language-independent method that enables MT for low-resource languages for which no parallel data is available between the LRL and any other language. We take advantage of the similarities between a closely-related HRL and the LRL in order to transform the HRL data into data similar to the LRL. For this purpose, we train neural transliteration models with transliteration pairs extracted from Wikipedia article titles and a glossary of the 200 most frequent words in the HRL corpus. Then, we automatically back-translate monolingual LRL data with the models trained on the transliterated HRL data and use the resulting parallel corpus to train our final models. Our method achieves significant improvements in MT quality, especially when original quality data is used on the target side. Additional experiments in the pseudo-LRL scenario revealed that there is potential to achieve quality MT if sufficient transliteration pairs are available, as well as high quality monolingual corpora and accurate test sets for evaluation. Finally, we showed that the existence of parallel corpora can lead to further improvements for a size larger than 1000 sentences with the transfer learning method. In general, the proposed method could be used to contribute to spreading vital information in disaster and emergency scenarios.

Acknowledgements This work has been partially funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 645452 (QT21).

References

Asgari E, Schütze H (2017) Past, present, future: a computational investigation of the typology of tense in 1000 languages. ArXiv preprint

Bergsma S, Kondrak G (2007) Alignment-based discriminative string similarity. In: Proceedings of the 45th annual meeting of the Association of Computational Linguistics, Prague, Czech Republic. Association for Computational Linguistics, pp 656–663

Ciobanu AM, Dinu LP (2014) On the Romance languages mutual intelligibility. In: Proceedings of the 9th international conference on language resources and evaluation, LREC, pp 3313–3318

Currey A, Karakanta A, Dehdari J (2016)Using related languages to enhance statistical language models. In: Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics: student research workshop, San Diego, CA, USA. Association for Computational Linguistics, pp 24–31

Dou Q, Knight K (2013) Dependency-based decipherment for resource-limited machine translation. In: Proceedings of the 2013 conference on empirical methods in natural language processing, Seattle, Washington, USA. Association for Computational Linguistics, pp 1668–1676

Dou Q, Vaswani A, Knight K (2014) Beyond parallel data: joint word alignment and decipherment improves machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar. Association for Computational Linguistics, pp 557–565

Durrani N, Koehn P (2014) Improving machine translation via triangulation and transliteration. In: Proceedings of the 17th annual conference of the European Association for Machine Translation, Dubrovnik, Croatia. EAMT'14, pp 71–78

Fišer D, Ljubešic N (2011a) Bilingual lexicon extraction from comparable corpora for closely related languages

Fišer D, Ljubešic N (2011b) Bootstrapping bilingual lexicons from comparable corpora for closely related languages. Springer, Berlin, pp 91–98

Gülçehre Ç, Firat O, Xu K, Cho K, Barrault L, Lin H, Bougares F, Schwenk H, Bengio Y (2015) On using monolingual corpora in neural machine translation. CoRR, abs/1503.03535

Hana J, Feldman A, Brew C, Amaral L (2006) Tagging Portuguese with a Spanish tagger using cognates. In: Proceedings of the international workshop on cross-language knowledge induction, CrossLangInduction '06, Stroudsburg, PA, USA. Association for Computational Linguistics, pp 33–40

Ismail A, Manandhar S (2010) Bilingual lexicon extraction from comparable corpora using in-domain terms. In: Proceedings of the 23rd international conference on computational linguistics: posters, COLING '10, Stroudsburg, PA, USA. Association for Computational Linguistics, pp 481–489

Jing W, Hongxu H, Zhipeng S, Jian D, Jinting L(2016) Adapting attention-based neural network to low-resource Mongolian–Chinese machine translation. Natural language understanding and intelligent applications, p 201

Johnson M, Schuster M, Le QV, Krikun M, Wu Y, Chen Z, Thorat N, Viégas F, Wattenberg M, Corrado G, Hughes M, Dean J (2016) Google's multilingual neural machine translation system: enabling zero-shot translation. ArXiv Preprint

Karimi S, Scholer F, Turpin A (2011) Machine transliteration survey. ACM Comput Surv 43(3):1–17

Klein G, Kim Y, Deng Y, Senellart J, Rush AM (2017) OpenNMT: open-source toolkit for neural machine translation. ArXiv preprint

Koehn P (2005) Europarl: a parallel corpus for statistical machine translation. In: Conference proceedings: the tenth machine translation summit, Phuket, Thailand. AAMT, AAMT, pp 79–86

Kondrak G, Marcu D, Knight K (2003) Cognates can improve statistical translation models. In: Proceedings of the 2003 conference of the North American chapter of the Association for Computational Linguistics on human language technology: companion volume of the proceedings of HLT-NAACL 2003—short Papers—Volume 2, NAACL-Short '03, Stroudsburg, PA, USA. Association for Computational Linguistics, pp 46–48

Lewis MP, Gary FS, Charles DF (eds) (2016) Ethnologue: languages of the World, 19th edn. SIL International, Dallas, TX, USA

Majliš M, Žabokrtský Z (2012) Language richness of the web. In: Proceedings of the 8th international conference on language resources and evaluation (LREC 2012), İstanbul, Turkey. European Language Resources Association, pp 2927–2934

Mann GS, Yarowsky D (2001) Multipath translation lexicon induction via bridge languages. In: Proceedings of the second meeting of the North American chapter of the Association for Computational Linguistics on Language Technologies, NAACL '01, Stroudsburg, PA, USA. Association for Computational Linguistics, pp 1–8

Melamed ID (1999) Bitext maps and alignment via pattern recognition. Comput Linguist 25(1):107–130

Nakov P, Ng HT (2012) Improving statistical machine translation for a resource-poor languages using related resource-rich languages. J Artif Intell Res 44:179–222

Papineni K, Roukos S, Ward T, Zhu W-J (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on Association for Computational Linguistics, ACL '02, Stroudsburg, PA, USA. Association for Computational Linguistics, pp 311–318

Popović M (2015) chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the tenth workshop on statistical machine translation, Lisbon, Portugal. 2015 Association for Computational Linguistics, pp 392–395

Sennrich R, Haddow B (2016) Linguistic input features improve neural machine translation. In: Proceedings of the first conference on machine translation, Berlin, Germany. Association for Computational Linguistics, pp 83–91

Sennrich R, Haddow B, Birch A (2015) Improving neural machine translation models with monolingual data. CoRR, abs/1511.06709

Sennrich R, Haddow B, Birch A (2016) Neural machine translation of rare words with subword units. In: Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 1: long papers), Berlin, Germany. Association for Computational Linguistics, pp 1715–1725

Simard M, Foster GF, Isabelle P (1993) Using cognates to align sentences in bilingual corpora. In: Proceedings of the 1993 conference of the centre for advanced studies on collaborative research: distributed computing—volume 2, CASCON '93. IBM Press, pp 1071–1082

Singla K, Shastri N, Jhunjhunwala M, Singh A, Bangalore S, Sharma DM (2014). Exploring system combination approaches for Indo-Aryan MT systems. LT4CloseLang 2014

Stanojević M, Sima'an, K (2014). BEER: BEtter evaluation as ranking. In: Proceedings of the ninth workshop on statistical machine translation, Baltimore, Maryland, USA, pp 414–419

Tiedemann J (2012) Character-based pivot translation for under-resourced languages and domains. In: Proceedings of the 13th conference of the European chapter of the Association for Computational Linguistics, EACL '12, Stroudsburg, PA, USA. Association for Computational Linguistics, pp 141–151

Voigt S, Gooskens C (2014) Mutual intelligibility of closely related languages within the Romance language family. The State of the Art, Language Contact

Zhang J, Zong C (2016) Exploiting source-side monolingual data in neural machine translation. In: Proceedings of the 2016 conference on empirical methods in natural language processing, EMNLP 2016, Austin, Texas, USA, November 1–4, pp 1535–1545

Zhang M, Li H, Banchs RE, Kumaran A (2015) Whitepaper of NEWS 2015 shared task on machine transliteration. In: Proceedings of the fifth named entity workshop, Beijing, China, pp 1–9

Zoph B, Yuret D, May J, Knight K (2016). Transfer learning for low-resource neural machine translation. In: Proceedings of the 2016 conference on empirical methods in natural language processing. Association for Computational Linguistics, pp 1568–1575