An Evaluation of Two Vocabulary Reduction Methods for Neural Machine Translation

Duygu Ataman ataman@fbk.eu
Fondazione Bruno Kessler, Trento, Italy
University of Trento, Trento, Italy

Marcello Federico federico@fbk.eu
MMT Srl, Trento, Italy
Fondazione Bruno Kessler, Trento, Italy

Abstract

Neural machine translation (NMT) models are conventionally trained with fixed-size vocabularies to control the computational complexity and the quality of the learned word representations. This, however, limits the accuracy and the generalization capability of the models, especially for morphologically-rich languages, which usually have very sparse vocabularies containing rare inflected or derivated word forms. Some studies tried to overcome this problem by segmenting words into subword level representations and modeling translation at this level. However, recent findings have shown that if these methods interrupt the word structure during segmentation, they might cause semantic or syntactic losses and lead to generating inaccurate translations. In order to investigate this phenomenon, we present an extensive evaluation of two unsupervised vocabulary reduction methods in NMT. The first is the well-known byte-pair-encoding (BPE), a statistical subword segmentation method, whereas the second is linguistically-motivated vocabulary reduction (LMVR), a segmentation method which also considers morphological properties of subwords. We compare both approaches on ten translation directions involving English and five other languages (Arabic, Czech, German, Italian and Turkish), each representing a distinct language family and morphological typology. LMVR obtains significantly better performance in most languages, showing gains proportional to the sparseness of the vocabulary and the morphological complexity of the tested language.

1 Introduction

Neural machine translation (NMT) has provided significant improvements to the state-of-the-art in machine translation (Bentivogli et al., 2016). However, it has also brought quite a few practical issues. A very important one of these is the low accuracy in translating rare words, caused by two of the main properties of the model. The first is related to the requirement of observing many examples of a word until its internal representation becomes accurate, and the second is due to the difficulty of handling large vocabularies, as this has an impact on the computational complexity of the model. Current implementations of NMT models require long training time and large memory space due to the high number of parameters to optimize. Hence, even with the most advanced machinery, deploying networks that can learn reliable representations for all words observed in the training corpus becomes practically impossible. In order to control the model complexity and the quality of the word representations, a straightforward approach is to fix the vocabularies to a maximum size, *e.g.* 100,000 lexical units, prior to training.

Clearly, a word can only be translated if an exact match of it is found in the vocabulary. This requirement leads to critical restrictions in translating morphologically-rich languages, where the word vocabulary tends to be very large and sparse. For example, in our case study, despite the relatively small size of our training corpora, the size of the source vocabulary found in the Turkish-English training corpus is around 170,000, *i.e.* much larger than the maximum size that is generally used.

Some studies have tried to overcome this problem by redefining the model vocabulary in terms of interior orthographic units compounding the words. These units could be individual characters (Ling et al., 2015; Lee et al., 2017), hybrid word/character units (Luong and Manning, 2016), or subwords (Sennrich et al., 2016), *i.e.* character sequences segmented according to their frequency in the training corpus. The prominent approach used today is to treat these subwords as individual lexical units. Hence, NMT is learned as a bilingual mapping between subword units of two languages. In addition to providing a new perspective to modeling translation at the sublexical level, these approaches have alleviated the out-of-vocabulary problem in NMT.

The sore point of these methods, however, is that they disregard any linguistic notion during segmentation. Many studies have shown that using subword segmentation methods which do not preserve the morpheme boundaries inside words may lead to loss of information related to the semantic or syntactic properties of words and generate inaccurate translations (Niehues et al., 2016; Ataman et al., 2017; Pinnis et al., 2017; Huck et al., 2017; Tamchyna et al., 2017). A more linguistically motivated solution was recently proposed by Ataman et al. (2017), which segments words into subwords by estimating their likeliness of being morphemes and their morphological categories. This approach provided significant improvements for translation of Turkish, an agglutinative language with a very sparse vocabulary.

In this paper, we present a comparative study on two unsupervised word segmentation methods: Byte-Pair Encoding (BPE) (Sennrich et al., 2016) and the Linguistically-Motivated Vocabulary Reduction (LMVR) method by Ataman et al. (2017) for NMT. Our analysis aims at understanding the important factors related to the statistical and formal characteristics of lexical units, mainly induced by morphology, and how they affect the translation quality. For this purpose, we set up an evaluation benchmark pairing English with five inflected languages: Arabic, Czech, German, Italian and Turkish, where each language represents a language family with distinct morphological characteristics.

The experimental results show that the translation quality obtained using LMVR (Ataman et al., 2017) in three of the languages (Arabic, Czech and Turkish) is significantly better than that with BPE (Sennrich et al., 2016). All of these languages share the common feature of having a high level of sparseness or a morphology with agglutinating or concatenating properties. For the remaining two languages with fusional characteristics and lower sparseness: German and Italian, the two segmentation methods yield comparable performance. In general, both word segmentation methods outperform the simple frequency-based vocabulary reduction method proposed by (Luong et al., 2015). Our study suggests that considering the morphological characteristics of the chosen language pair is essential in order to choose the most appropriate subword segmentation approach in NMT.

2 Neural Machine Translation (NMT)

In this paper, we use the NMT model described in (Bahdanau et al., 2014). The model essentially estimates the conditional probability of translating a source text *x*, represented by the input word sequence $x = (x_1, x_2, \dots x_m)$ of length m, into a target text *y*, represented as the target word sequence $y = (y_1, y_2, \dots y_j \dots y_l)$ of length l. The conditional probability is decomposed as follows:

$$p(y|x;\theta) = \prod_{i=1}^{l} p(y_i|y_{i-1}, ..., y_0, x_m, ..., x_1; \theta) \tag{1}$$

where θ represents the model parameters. The model is trained by maximizing the log-likelihood of a parallel training set *D*:

$$L(D, \theta) = \sum_{x,y \in D} \log p(y|x;\theta) \tag{2}$$

The inputs of the network are one-hot vectors – *i.e.* binary vectors that have a single bit set to 1 to identify a specific word in the vocabulary. Each word vector is then mapped to an embedding, a continuous representation in a lower dimension but more dense space. Hence, a distributed representation of the source words is learned using a bi-directional recurrent neural network, the *encoder*, which encodes *x* into *m* dense sentence vectors, corresponding to its hidden states.

Next, a unidirectional recurrent neural network, the *decoder*, predicts the target sequence y word by word using the information provided by the encoder. Each target word y_j is predicted by sampling from a word distribution computed from the previous target word y_{j-1}, the previous hidden state of the decoder, and a so-called *context vector*. The context vector is a linear combination of the encoder hidden states, whose weights are dynamically computed by a feed-forward neural network called the *attention* model. The attention model predicts each weight on the basis of the previous target word, the previous decoder hidden state and the corresponding encoder hidden state.

The overall network is trained to minimize the cost function in Equation 2 via Stochastic Gradient Descent (SGD) (Bottou, 2010) and the Back Propagation Through Time algorithm (Werbos, 1990). During training, the learning algorithm iteratively updates the parameters of the network, including the weights of the hidden units in each layer and the word embeddings, until the value of the cost function calculated in the training corpus is optimized, or a maximum number of iterations is reached. In practice, this process is computationally very expensive due to the many parameters to adjust and the fact that the probability of generating each target word y_j is normalized via a softmax function, as shown below:

$$p(y_i = e_j|x;\theta) = \frac{e^{e_j^T o_i}}{\sum_{k=1}^{K} e^{e_k^T o_i}} \tag{3}$$

where e_j is the j^{th} one-hot vector of the target vocabulary of size K, and o_i is the decoder output vector for the i^{th} target word y_i.

From equation (3) we see that the computational cost of predicting each word scales linearly with the target vocabulary size K. In general, larger source and target vocabulary sizes imply higher levels of data sparseness, longer training and inference time and a larger dynamic memory usage. Bahdanau et al. (2014) suggested using a fixed-size vocabulary of size k containing only the top k frequent words in the corpus in order to control the size of the source and target vocabularies. Nevertheless, this prevents translating any out-of-vocabulary words that might be encountered in new sentences. Luong et al. (2015) extended this approach to integrate a word alignment model as a post-processing step to the NMT system, where the words that do not fit in the vocabulary are marked as an unknown word token (*i.e. 'UNK'*) and the sentence is translated disregarding these words. After translation, the unknown tokens on the target side can be replaced with the original words on the source side or simply left as is. This approach is useful for translating rare words like numbers or named entities that are not found in the vocabulary, however, it does not provide a complete solution as rare words can be of different nature in each language. For instance, a large portion of the vocabularies in a synthetic language (see Section

4) can contain infrequent words that are derivated or inflected word forms, which often carry important information related to the syntax and semantics of the rest of the sentence.

3 Unsupervised Word Segmentation for NMT

A conventional solution to limit the vocabulary size in NMT is to segment words into smaller units and perform translation at the sublexical level. In this paper, we discuss two such methods: BPE and LMVR.

3.1 Byte-Pair Encoding (BPE)

BPE is the prominent method of subword segmentation for NMT that has been applied to many languages (Bojar et al., 2017). It is originally a data compression algorithm that minimizes the length of sequences of bytes by finding the most frequent consecutive byte pairs and encoding them using unused byte values (Gage, 1994). It was recently modified by Sennrich et al. (2016) for vocabulary reduction, where the most frequent character sequences are iteratively merged to find the optimal description of the corpus vocabulary. This purely statistical method is based on the assumption that many types of words can be translated when segmented into smaller units, such as named entities and loanwords. Nevertheless, in cases of common morphological paradigms such as derivational or inflectional transformations which are typically observed in morphologically-rich languages, the method lacks a linguistic notion that could allow it to better generalize syntactic patterns in the data and use the vocabulary space more effectively (Ataman et al., 2017; Huck et al., 2017; Tamchyna et al., 2017). Moreover, by disregarding morpheme boundaries during splitting, it can lead to semantically ambiguous subwords which would be translated inaccurately (Niehues et al., 2016; Ataman et al., 2017; Pinnis et al., 2017).

3.2 Linguistically-Motivated Vocabulary Reduction (LMVR)

Similar to BPE, LMVR constitutes a pre-processing step to NMT. The method is an extension of *Morfessor FlatCat* (Grönroos et al., 2014), an unsupervised morphology learning algorithm based on a Hidden Markov Model (HMM), which models the composition of a word based on the transitions between different morphemes and their categories (*i.e.* prefix, stem or suffix). The category-based HMM is essential for a linguistically motivated segmentation, as words are split considering the possible categories of the generated subwords and not only their frequencies. Ataman et al. (2017) has recently modified this method in order to optimize the complexity of the model with a constraint on the number of morphemes to be found in the corpus after segmentation, *i.e.* the lexicon size, which eventually allows it to be deployed as a stand-alone vocabulary reduction technique for NMT.

Similar to the two-level morphology model of Koskenniemi (1983), the model (M) consists of mainly two parts, a *lexicon* that contains the list of morphemes and a *grammar* which defines a set of rules that combine different morphemes together to generate new words. The model is estimated via Maximum A-Posteriori (MAP) optimization in order to avoid overfitting, by finding a balance between model accuracy and complexity. The MAP estimate of the overall system is given as:

$$M^* = \arg\max_M \Pr(D|M)\Pr(M) \tag{4}$$

where the two factors respectively represent the likelihood of the training corpus *D* and the prior probability of the model *M*.

While the former is computed on the data by a HMM, the latter is modeled by considering individual properties of the generated lexicon[1] of morphemes:

$$\Pr(M = \{\mu_1, \ldots, \mu_m\}) \approx m!\, P(usage(\mu_1, \ldots, \mu_m))P(form(\mu_1, \ldots, \mu_m)) \tag{5}$$

[1] The grammar is assumed as a fixed component of the model and is thus disregarded from the prior.

where *m* is the number of distinct morphemes (μ_i) in the lexicon (Creutz and Lagus, 2007). The *usage* of morphemes are modeled by their frequencies, lengths, and their left and rightwards perplexities. The *form* of morphemes considers instead the probability of their internal structure, composed either of other morphemic categories or a sequence of characters.

Using the a-posteriori probability, one can train a segmentation model considering both the model complexity and the likelihood of the corpus, without any control on the size of the output lexicon. In order to achieve a desired rate of vocabulary reduction for NMT, Ataman et al. (2017) inserts a regularization weight over the lexicon prior and thus force the optimization to give more importance to reducing the model complexity. The general formula for optimization then becomes:

$$L(D, M) = \log P(D|M) + \alpha \log P(M)$$

where $\alpha > 1$ would force the optimization algorithm to find a smaller lexicon size and a finer segmentation. Ataman et al. (2017) empirically sets α equal to $\frac{m_1}{m_2}$, where m_1 is the initial vocabulary size of the corpus, and m_2 is the target vocabulary size.

4 Morphology and Language Families

In NMT, translation is conventionally modeled at the lexical level. Thus, the statistical distribution of the words observed in training data has a crucial role to guide the NMT models. A high level of variance in the lexical distribution implies a high level of sparseness and a low expectation to observe each individual word. This increases the difficulty to learn translations, especially of the infrequent words, and limits the accuracy of the model. An important factor that affects the sparseness in a corpus is the morphological properties of a language. In order to illustrate this aspect, we hereby introduce basic concepts of morphology and how it is formed in different languages.

The smallest units inside a word that carry meaning are called *morphemes* (O'Grady et al., 1997). They can typically have one of two main functions: aiding the grammatical role or the meaning of the word in which they occur. The main component required to form a word is the root morpheme, or the *base*, which has the most crucial role of defining the meaning and contains one of several categories (*i.e.* noun, verb, adjective, or preposition). Other components may include *affixes*, which do not belong to a lexical category and are attached to the base to form new words. An affix that is attached to the front of the base is called a *prefix*, and an affix that is attached to the end of the base is called a *suffix*. In Italian both prefixes and suffixes can be observed, whereas in Turkish words expand only through attachment of suffixes. In very few languages like Arabic, it is also possible to observe *infixes*, types of affixes that are attached to the root within a base (O'Grady et al., 1997). In some languages, independent words from different lexical categories can be combined to create a larger word with a new meaning. This common morphological process is called *compounding*. In such a case, the same word may be expected to contain multiple bases and affixes. In German, compounding is frequently observed. From a functional perspective, morphemes can be combined to produce words mainly in two ways. *Derivational* morphemes are added to a root to change its category or function. On the other hand, *inflectional* morphemes carry grammatical meaning and do not change the category of the root. Both ensure the transformation of the root in a correct surface form so that the sentence is grammatically acceptable.

Depending on the language, a word may contain a limited number of morphemes. For instance, *analytic* languages, such as Mandarin Chinese or Vietnamese, usually preserve a one-to-one correspondence between a word and a morpheme (Shopen, 1985). On the other hand, in *synthetic* languages, a word can contain several morphemes. Synthetic languages are generally grouped into two morphology families. *Fusional* languages are characterized by their

Language	**Family**	**Morphological Complexity**	**Morphological Typology**
Arabic	Semitic	High	Concatenative, Templatic
Czech	Slavic	High	Mostly Fusional, Partially Agglutinating
German	Germanic	Medium	Fusional
Italian	Italic	Low	Fusional
Turkish	Turkic	High	Agglutinating

Table 1: Families and morphological characteristics of languages we translate from/to English.

tendency to use a single inflectional morpheme to denote multiple grammatical, syntactic, or semantic features. On the other hand, in *agglutinative* languages, each morpheme in a word remains in every aspect unchanged after their composition, allowing a direct identification of the morpheme boundaries. In fusional and agglutinating typologies, morphemes are generally composed continuously to construct new word forms. On the other hand, it is also possible to observe *templatic* typologies, for instance in Arabic, where morphemes are inserted in certain templates in a discontinuous fashion to achieve certain derivative or inflective transformations.

Most languages do not belong exclusively to one category of morphological typology. In fact, there are many languages where different morphological phenomena are observed together. Based on how much such phenomena are typical in a language, it is expected to observe increased sparseness in the lexical surface forms. Consequently, the morphological characteristics of a language would be directly influential on the statistical distribution obtained from a textual corpus in the given language. In order to enlighten this aspect, we have chosen five languages which have been commonly studied in official machine translation evaluation campaigns. Each of them represents a different language family and falls into a distinct combination of morphological typology. The selected languages consist of Arabic (*Semitic*), Czech (*Slavic*), German (*Germanic*), Italian (*Italic*) and Turkish (*Altaic*). As a bridge between all the languages, we choose English from the Germanic family, which is a moderately analytic language but contains some different morphological features compared to German, the other Germanic language in our study. A summary of the main linguistic and morphological features of the listed languages can be seen in Table 1.

5 Evaluation

In order to evaluate different subword segmentation methods, we set up a common benchmark to observe the effect of each method on languages with different statistical properties. Our benchmark couples English (either as source or target) with five languages: Arabic, Czech, German, Italian and Turkish. Thus, each language pair represents different statistical properties reflected by the level of agglutination or fusion observed in their formal morphology. We perform NMT by keeping the segmentation on the English side fixed and applying different segmentation approaches to the other languages. This aids us in avoiding a combinatorial explosion in the number of experiments, while ensuring the results between each language are comparable. We later vary the segmentation method applied to the English side to investigate its effects on both sides of translation. We limit these experiments only to the English-Italian and English-Turkish pairs, as they represent two extreme cases in our setting, *i.e.* from low to high morphological complexity. All experiments consider each tested language both at the source and the target side of translations. The two subword segmentation methods, *LMVR* and *BPE*, are also compared with the frequency-based vocabulary pruning method suggested by Luong et al. (2015), and described at the end of Section 2, which is henceforth referred to as *Word* method.

Language	# sentences	# tokens	# types
Arabic-English	238,511	3,9M(AR) - 4,9M(EN)	220K(AR) - 120K(EN)
Czech-English	117,966	2M(CS) - 2,3M(EN)	118K(CS) - 50K(EN)
German-English	212,151	4M(DE) - 4,3M(EN)	144K(DE) - 69K(EN)
Italian-English	184,642	3,5M(IT) - 3,8M(EN)	95K(IT) - 63K(EN)
Turkish-English	135,734	2,7M(TR) - 2M(EN)	171K(TR) - 53K(EN)

Language	Data sets		# sentences	# tokens
Arabic-English	Development	dev2010, test2010, test2011, test2012	5,835	89K(AR) - 114K(EN)
	Testing	test2013, test2014	4,121	66K(AR) - 83K(EN)
Czech-English	Development	dev2010, test2010, test2011	3,112	52K(CS) - 61K(EN)
	Testing	test2012, test2013	2,836	47K(CS) - 55K(EN)
German-English	Development	dev2010, test2010, test2011, test2012	5,777	108K(DE) - 113K(EN)
	Testing	test2013, test2014, test2015	3,543	67K(DE) - 70K(EN)
Italian-English	Development	dev2010, test2010,	3,517	74K(IT) - 79K(EN)
	Testing	test2011, test2012	3,230	55K(IT) - 60K(EN)
Turkish-English	Development	dev2010, test2010	2,433	34K(TR) - 47K(EN)
	Testing	test2011, test2012	2,720	39K(TR) - 53K(EN)

Table 2: *Above:* Training sets. *Below:* Development and Testing Sets. All data set are official evaluation sets from IWSLT. (*M*: Million, *K*: Thousand.)

5.1 Data

We train our NMT models using the TED Talks corpora (Cettolo et al., 2012) and test them on official data sets of the IWSLT[2] evaluation campaign from 2010 to 2015. This aids us in having a variety of languages with different morphological typology within the same benchmark. We select multiple development and testing sets from different years to obtain more reliable results. All data sets are tokenized and truecased using the Moses scripts[3] (Koehn et al., 2007), except

[2] The International Workshop on Spoken Language Translation with shared tasks organized between 2003-2017.

[3] www.statmt.org/moses

for Arabic, which is normalized and tokenized using the QCRI normalization tool[4] (Sajjad et al., 2013). The details of the statistical characteristics of each data set used in our experiments and the chosen development and testing sets are given in Table 2.

5.2 Segmentation Models

The two subword segmentation methods that we compare in our experiments, *BPE* and *LMVR*, as well as the baseline vocabulary reduction method, *Word*, are applied to fit the same dictionary sizes (30,000) set in the NMT models. Since our training sets are small, choosing a small vocabulary size allows to illustrate large vocabulary reduction rates encountered in practical NMT tasks. We learn the merge rules of BPE at an equal size to the dictionary. Similarly, the LMVR models are trained with an *output lexicon size* of the same size. The rest of the settings are kept as default except for the perplexity threshold, for which we keep the default value of 10 for five languages, while for Arabic we use the value 70 as suggested by Al-Mannai et al. (2014). The translated sentences are desegmented based on the splitting characters ("@@" for BPE, "+" for LMVR) before measuring the translation quality.

5.3 NMT Models

The NMT models used in the evaluation are based on the Nematus toolkit (Sennrich et al., 2017). They have a hidden layer and embedding dimension of 1024 and a dictionary size of 30,000 for both source and target languages. We train the models using the Adagrad (Duchi et al., 2011) optimizer with a mini-batch size of 100, a learning rate of 0.01, and a dropout rate of 0.1 (in the input and output layers) and 0.2 (in the embeddings and hidden layers). In order to prevent over-fitting, we stop training if the perplexity on the validation has not decreased for 5 epochs or the maximum number of epochs are reached. After 50 epochs, we choose the model with the highest performance on the development set for translating the test set. In order to present a comprehensive evaluation, we evaluated the accuracy of each model output using both BLEU (Papineni et al., 2002) and chrF3 (Popovic, 2015) metrics. Significance tests are computed only for BLEU with Multeval (Clark et al., 2011).

6 Results

The findings of the experiment, presented in Table 3, illustrates the translation qualities using different approaches and how these qualities vary among different languages. The results of the experiments where the English side is segmented with BPE show that LMVR generally achieves the best results by outperforming BPE with a significant improvement in three out of four morphologically-rich languages. The improvements are **1.55** BLEU points in Turkish-to-English, **1.08** BLEU points in Arabic-English and **0.99** BLEU points in Czech-to-English. In the German-to-English translation task, the difference between the performance of two subword segmentation methods is not statistically significant. In the Italian-to-English direction, BPE produces better translations with an accuracy of 0.29 BLEU points higher than LMVR. The improvements are in general more evident in the chrF3 score, where we observe an improvement of **0.017** points in Turkish-to-English, and around **0.015** points in Arabic-to-English and Czech-to-English. In the German-to-English direction, LMVR provides slightly higher accuracy in terms of the chrF3 score. The improvements over the weak baseline of word-based translation are also significant, ranging from **5.16** BLEU points in Turkish-to-English to **1.63** BLEU points in German-to-English and **0.62** BLEU points in Italian-to-English.

The experiments conducted in the opposite translation directions show that the performance characteristics of LMVR are consistent in either directions. Translating into a morphologically-rich language is a more challenging task and the output quality is generally

[4] alt.qcri.org/tools/arabic-normalizer

Language	Segmentation		BLEU	chrF3
Direction	(Src)	(Tgt)		
Arabic-English	Word	BPE	26.76	0.4793
	BPE	BPE	29.59	0.5102
	LMVR	BPE	**30.67**$^{\blacktriangle}$	**0.5248**
Czech-English	Word	BPE	26.82	0.4689
	BPE	BPE	28.21	0.494
	LMVR	BPE	**29.2**$^{\blacktriangle}$	**0.5091**
German-English	Word	BPE	30.71	0.5109
	BPE	BPE	**32.57**	0.5432
	LMVR	BPE	32.53	**0.5437**
Italian-English	Word	BPE	31.41	0.5237
	BPE	BPE	**32.50**	0.5322
	LMVR	BPE	32.21	0.5302
	BPE	LMVR	32.16	0.5416
	LMVR	LMVR	**32.50**	**0.5446**
Turkish-English	Word	BPE	17.58	0.3859
	BPE	BPE	21.28	0.4335
	LMVR	BPE	22.83	0.4501
	BPE	LMVR	20.99	0.4390
	LMVR	LMVR	**23.13**$^{\blacktriangle}$	**0.4599**

chrF3	BLEU	Segmentation		Language
		(Src)	(Tgt)	Direction
0.3460	15.20	BPE	Word	English-Arabic
0.4490	17.91	BPE	BPE	
0.4610	**18.95**$^{\blacktriangle}$	BPE	LMVR	
0.3731	18.46	BPE	Word	English-Czech
0.4378	19.09	BPE	BPE	
0.4483	**19.98**$^{\blacktriangle}$	BPE	LMVR	
0.4927	26.35	BPE	Word	English-German
0.5431	27.24	BPE	BPE	
0.5485	**27.38**	BPE	LMVR	
0.5120	27.77	BPE	Word	English-Italian
0.5415	28.28	BPE	BPE	
0.5451	**28.30**	BPE	LMVR	
0.5412	27.99	LMVR	BPE	
0.5432	28.24	LMVR	LMVR	
0.2968	10.05	BPE	Word	English-Turkish
0.4183	11.31	BPE	BPE	
0.4410	12.53	BPE	LMVR	
0.4378	11.13	LMVR	BPE	
0.4435	**12.63**$^{\blacktriangle}$	LMVR	LMVR	

Table 3: Experiment results. Best scores for each translation direction are in bold font. Those marked with $^{\blacktriangle}$ are also statistically significantly better (p-value < 0.05) than the baseline.

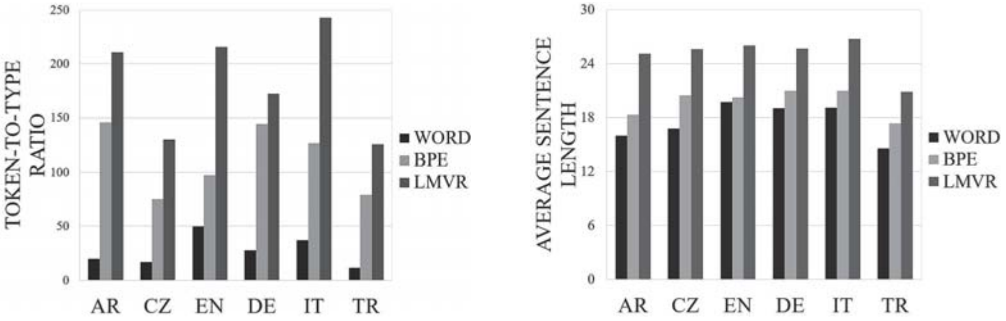
lower. In these directions, however, the effect of the chosen segmentation methods can be visualized more clearly, as the experiments show that the improvements of the two vocabulary reduction methods over the weak baseline of word-based translation is quite significant for the most sparse languages, whereas as sparseness in the languages decreases we observe very comparable performances. For instance, in English-to-Italian translation, the word vocabulary is naturally not very sparse, and segmentation by either method provides around **0.5** BLEU and **0.033** chrF3 points, *i.e.* 1,8% improvement. On the other hand, for English-to-Arabic translation, segmenting words generally provides an improvement of at least **3.75** BLEU and **0.103** chrF3 points. Similarly, in English-to-Turkish translation, we observe very large improvements over the baselines, **2.49** BLEU and **0.14** chrF3 points above the weak baseline, reaching an overall of 11% improvement. The findings also suggest that choosing LMVR for vocabulary reduction for NMT in the task of generating translations in the morphologically-rich language provides better translation quality than BPE. The highest improvement is observed again in the English-to-Turkish direction, where LMVR outperforms BPE by **1.32** BLEU and **0.025** chrF3 points. The improvements follow the same trend for English-to-Arabic, with **1.04** BLEU and **0.029** chrF3 points and in English-to-Czech translation, where LMVR achieves an accuracy of around 4.7% above the strong baseline. For translation directions involving German and Italian, the performances of two segmentation methods are generally comparable.

Using LMVR on both sides of the parallel data aids in obtaining full advantage from the method, especially in the morphologically-rich language setting. In Turkish-to-English direction, using LMVR also on the English side improves the performance by **0.3** BLEU and **0.0099** chrF3 points over the approach of using BPE on the English and LMVR on the target side, whereas in English-to-Turkish direction it provides an improvement of **0.1** BLEU and **0.0025** chrF3 points. In Italian-to-English direction, the performance is increased by **0.29** BLEU and **0.0124** chrF3 points, reaching the best performance among all vocabulary reduction methods. In English-to-Italian direction, all methods are comparable. A comparison between the approaches of applying either LMVR or BPE on both sides of the corpus does not yield a significantly different translation accuracy in English-to-Italian and Italian-to-English directions.

7 Discussion

A first glance at the findings of our experiments confirms the benefit of subword segmentation as a vocabulary reduction approach. This is mainly due to the higher reduction of vocabulary sparseness that is achieved with respect to filtering out infrequent words (*Word* method). When rare words that do not fit into the limited NMT model vocabulary are segmented into sequences of subwords, a new vocabulary with a lower sparseness is obtained. The lower data sparseness obtained by BPE and LMVR versus Word is evident from Figure 1a, which plots the corresponding type-to-token ratios of each training corpus. After segmentation, the new vocabulary of subwords has a less sparse frequency distribution and each subword is observed in more types of context. This allows to learn better representations for each subword and increases the translation accuracy. The significant difference in output quality observed with two different segmentation approaches tells, however, that their impact highly depends on the nature of the splitting process and the characteristics of the language they are applied on.

An interesting outcome of our experiments is that as the complexity of morphology (*i.e.* sparseness in the lexical vocabulary) and the level of agglutination observed in the language increases, the more beneficial it becomes to use LMVR for vocabulary reduction. Arabic, Czech and Turkish all have a high level of lexical sparseness, and the higher translation quality obtained using LMVR proves that a segmentation method that preserves morphological information contained at the subword level can generate better translations. As can be seen in Figures 1a and 1b, LMVR can reduce the sparseness, *i.e.* increase the token-to-type ratio, in the corpus



(a) Token-to-type ratios

(b) Average sentence lengths

Figure 1: Token-to-type ratios and average sentence lengths after vocabulary reduction.

at higher degrees by applying a more homogeneous segmentation among the corpus, indicated in the levels of increase in the average sentence lengths. Sentence lengths generally remain close to their original lengths with BPE, on the other hand, with LMVR, the sentences can become much longer. One can also see that the improvements in each language are related to the formal characteristics of subwords. All three languages have inflections or derivations that are formed in a rather concatenating fashion, compared to German and Italian, where the affixes cannot be observed independently on the surface level. This explains the success of LMVR in learning subwords that are more consistent with the morphological boundaries. In Turkish, an agglutinative language where morpheme boundaries are transparent, it is possible to achieve a complete segmentation of the morphemes inside a word. However, this is not the case for the others. Arabic morphemes are discontinuous, and in Czech, there is less transparency in terms of the morpheme boundaries. Therefore, when a clear segmentation at the surface level is not possible, LMVR can fail to predict the correct morphemes inside a word.

In German and Italian, languages with highly fusional morphology, different approaches in vocabulary reduction do not yield large differences in the output quality. This is mainly due the formal properties of fusional morphology, where typographic changes at the input may not yield sufficient information for the model to learn significantly better translations. In addition to fusional transformations, German is also rich in compounding, which can be defined as an agglutinating transformation. However, the small difference in the performance of either segmentation methods suggests that both methods can handle this phenomena similarly, leading to comparable performances. Another factor that affects the results is related to the statistical characteristics of the languages, which, as can be seen in the vocabulary sizes listed in Table 2, do not hold a large amount of sparseness. The quantity of rare words in the vocabulary that could better be translated by different approaches could be an important indicator in the overall output accuracy. Italian, unlike German, has a morphology of comparably lower complexity and the word vocabulary is quite compact, where rare words (singletons and less frequently observed words that are in the long tail of the frequency distribution) mostly consist of named entities or numeric expressions. This is in contrast to morphologically-rich languages, where the majority of rare words also include inflected or derivated word forms. Hence, in Italian, vocabulary reduction with either segmentation methods can provide similar performances. When BPE is used, most words in the corpus are translated without segmentation. Although rare inflected words can exist in the corpus, they are not observed many times, and the improvement in their translation through LMVR may not be significant enough to be observed at the output. English is also a language of this group, with a morphology of very low complexity, although most of the affixes are easily separable. Therefore, LMVR can be trained to segment words according to their morphological boundaries. The benefit of applying LMVR also on the English side can be seen in our experimental results, which show that the best translation accuracy is obtained when LMVR is applied on both sides of the training data. Nevertheless, these improvements are not as significant as in the morphologically-rich language settings, as in the cases of Arabic, Czech and Turkish.

8 Conclusion

NMT is a novel and successful approach to machine translation that can provide high quality translations in many languages. However, the limitation in the size of the model vocabulary prevents to take full advantage of the approach, especially in morphologically-rich languages. These languages usually have large and sparse vocabularies which contain rare inflected or derivated word forms that cannot be included in the model vocabulary, and consequently, translated. A conventional solution to this is to reduce the vocabulary of the training corpus by segmenting words into subword level representations and perform their translations at this level. In this paper, we have compared two such methods, BPE, the prominently used approach which is a statistical method that disregards any linguistic notion during segmentation, and LMVR, a recently proposed method which also takes morphological coherence into consideration during prediction of the subwords. We have evaluated two methods in a common machine translation benchmark consisting of six languages with distinct morphological characteristics. Our findings showed that using LMVR provides better translation in NMT applied on morphologically-rich languages by trying to maintain a coherence between the generated subwords and the morphological boundaries. On the other hand, for fusional languages with low sparseness, using BPE and LMVR provided comparable translation quality. Our analysis supports that morphology is an important factor in determining the statistical characteristics of the language and should be taken into consideration for choosing the most appropriate vocabulary reduction method for NMT.

Acknowledgments

This work has been partially supported by the EC-funded H2020 project QT21 (grant no. 645452).

References

Al-Mannai, K., Sajjad, H., Khader, A., Al Obaidli, F., Nakov, P., and Vogel, S. (2014). Unsupervised Word Segmentation Improves Dialectal Arabic to English Machine Translation. In *Proceedings of the Workshop on Arabic Natural Language Processing (ANLP)*, pages 207–216.

Ataman, D., Negri, M., Turchi, M., and Federico, M. (2017). Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English. In *The Prague Bulletin of Mathematical Linguistics*, volume 108, pages 331–342.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.

Bentivogli, L., FBK, T., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 257–267.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., et al. (2017). Findings of the 2017 Conference on Machine Translation. In *Proceedings of the Second Conference on Machine Translation (WMT)*, pages 169–214.

Bottou, L. (2010). Large-Scale Machine Learning with Stochastic Gradient Descent. In *Proceedings of 19th International Conference on Computational Statistics (COMPSTAT)*, pages 177–186. Springer.

Cettolo, M., Girardi, C., and Federico, M. (2012). Wit3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268.

Clark, J. H., Dyer, C., Lavie, A., and Smith, N. A. (2011). Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 176–181.

Creutz, M. and Lagus, K. (2007). Unsupervised Models for Morpheme Segmentation and Morphology Learning. In *ACM Transactions on Speech and Language Processing (TSLP)*, volume 4, pages 1–34.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. In *Journal of Machine Learning Research*, volume 12, pages 2121–2159.

Gage, P. (1994). A New Algorithm for Data Compression. In *The C Users Journal*, volume 12, pages 23–38.

Grönroos, S.-A., Virpioja, S., Smit, P., Kurimo, M., et al. (2014). Morfessor FlatCat: An HMM-based Method for Unsupervised and Semi-supervised Learning of Morphology. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 1177–1185.

Huck, M., Riess, S., and Fraser, A. (2017). Target-Side Word Segmentation Strategies for Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation (WMT)*, pages 56–67.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 177–180.

Koskenniemi, K. (1983). Two-Level Model for Morphological Analysis. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, volume 83, pages 683–685.

Lee, J., Cho, K., and Hofmann, T. (2017). Fully character-Level Neural Machine Translation without Explicit Segmentation. In *Transactions of the Association for Computational Linguistics (TACL)*, volume 5, pages 365–378.

Ling, W., Trancoso, I., Dyer, C., and Black, A. W. (2015). Character-based Neural Machine Translation. *arXiv preprint arXiv:1511.04586*.

Luong, M.-T. and Manning, C. D. (2016). Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1054–1063.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421.

Niehues, J., Cho, E., Ha, T.-L., and Waibel, A. (2016). Pre-translation for Neural Machine Translation. In *Proceedings of The 26th International Conference on Computational Linguistics (COLING)*, pages 1828–1836.

O'Grady, W., Dobrovolsky, M., and Katamba, F. (1997). *Contemporary Linguistics*. St. Martin's.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Pinnis, M., Krišlauks, R., Deksne, D., and Miks, T. (2017). Neural Machine Translation for Morphologically Rich Languages with Improved Subword Units and Synthetic Data. In *Proceedings of the International Conference on Text, Speech, and Dialogue (TSD)*, pages 237–245.

Popovic, M. (2015). chrF: Character n-gram F-score for Automatic MT Evaluation. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT)*, pages 392–395.

Sajjad, H., Guzmán, F., Nakov, P., Abdelali, A., Murray, K., Al Obaidli, F., and Vogel, S. (2013). QCRI at IWSLT 2013: Experiments in Arabic-English and English-Arabic Spoken Language Translation. In *Proceedings of the 10th International Workshop on Spoken Language Translation (IWSLT)*, pages 75–82.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Barone, A. V. M., Mokry, J., et al. (2017). Nematus: a toolkit for Neural Machine Translation. In *Proceedings of the 15th Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 65–68.

Sennrich, R., Haddow, B., and 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 (ACL)*, pages 1715–1725.

Shopen, T. (1985). *Language Typology and Syntactic Description*, volume 3. Cambridge University Press.

Tamchyna, A., Marco, M. W.-D., and Fraser, A. (2017). Modeling Target-Side Inflection in Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation (WMT)*, pages 32–42.

Werbos, P. J. (1990). Backpropagation Through Time: What it does and How to do it. In *Proceedings of the Institute of Electrical and Electronics Engineers (IEEE)*, volume 78, pages 1550–1560.