UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Kaspar Papli

Morpheme-Aware Subword Segmentation for Neural Machine Translation

Bachelor's Thesis (9 ECTS)

Supervisor: Mark Fišel, PhD

Tartu 2017

Morpheme-Aware Subword Segmentation for Neural Machine Translation

Abstract:

Neural machine translation together with subword segmentation has recently produced state-of-the-art translation performance. The commonly used segmentation algorithm based on byte-pair encoding (BPE) does not consider the morphological structure of words. This occasionally causes misleading segmentation and incorrect translation of rare words. In this thesis we explore the use of morphological structure in subword segmentation and develop a novel segmentation algorithm that succeeds in preventing misleading BPE segmentations that occur due to its disregard for morphology. Analysis shows that the proposed algorithm decreases translation performance as measured by BLEU by 0.9 points while producing subjectively more intuitive segmentations and mildly better translations for sentences previously involving inaccurate baseline segmentation.

Keywords:

Machine translation, neural machine translation, artificial neural network, deep learning, word segmentation, subword segmentation, byte-pair encoding, morpheme, morphological structure

CERCS:

176 Artificial intelligence

Morfeemiteadlik sõnaosade segmenteerimine neuromasintõlke jaoks

Lühikokkuvõte:

Hiljuti kasutusele võetud neuromasintõlge koos sõnaosade segmenteerimisega on saavutanud masintõlke süsteemidest parima tõlkekvaliteedi. Tihti kasutatav bait-paar kodeeringul (BPK) põhinev segmenteerimisalgoritm ei arvesta sõnade morfoloogilist struktuuri, mis haruldaste sõnade puhul põhjustab aeg-ajalt eksitavat segmenteerimist ja ebakorrektset tõlget. Käesolevas töös esitatakse uus algoritm sõnaosade segmenteerimiseks, mis eemaldab BPK morfoloogilise struktuuri eiramise tõttu tekkinud segmenteerimisvead. Analüüs näitab, et esitatud algoritm vähendab BLEU poolt mõõdetud tõlkekvaliteeti 0.9 punkti võrra, kuid parandab eelnevalt ebatäpseid segmenteerimisi sisaldanud lausete segmenteerimist ja tõlget.

Võtmesõnad:

Masintõlge, neuromasintõlge, tehisnärvivõrkudega masintõlge, tehisnärvivõrk, neurovõrk, süvaõpe, sõnade segmenteerimine, sõnaosade segmenteerimine, bait-paar kodeering, morfeem, morfoloogiline struktuur

CERCS:

176 Tehisintellekt

Contents

1 Introduction

Machine translation is the task of producing output in one natural language while given input with equivalent meaning in another natural language. Until recently statistical methods such as (Koehn et al., 2003) have mainly been used in machine translation. However, recently specialized artificial neural network frameworks have been developed for machine translation that surpass the performance of state-of-the-art statistical machine translation systems (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2014).

One of the main limitations of neural machine translation (NMT) is vocabulary size—the number of distinct tokens the neural model recognizes. Typical neural translation models are restricted to 30 000–80 000 tokens (Bahdanau et al., 2014; Sutskever et al., 2014).

Proposed methods to address this problem include a back-off dictionary look-up (Luong et al., 2014; Jean et al., 2014) and segmentation of infrequent words into subword units (Sennrich et al., 2015). Subword segmentation has shown significantly improved performance compared to baseline models and models with a back-off dictionary.

However, current state-of-the-art segmentation algorithms do not consider the morphological structure of words, occasionally violate morpheme boundaries and produce misleading segmentations. The goal of this thesis is to develop an algorithm for subword segmentation that considers the morphological structure of words and to determine its applicability to NMT from Estonian compared to existing methods.

In Chapter 2 of this thesis we give an introduction into neural machine translation. We discuss subword segmentation in Chapter 3 and the current state-of-the-art segmentation algorithm for NMT in Chapter 3.2. We present a morpheme-aware segmentation algorithm as the main contribution of this thesis in Chapter 4. In Chapter 5 we analyze the performance of the proposed morpheme-aware segmentation algorithm and its effects on NMT.

2 Neural Machine Translation (NMT)

Machine translation (MT) is the task of converting text from one natural language into another natural language with the equivalent meaning.

Until recently statistical methods have been used for MT. Parameters for the underlying statistical models are derived from bilingual parallel corpora. Typically in statistical machine translation (SMT) separate statistical models are created for modelling language fluency and translation adequacy: the language model and the translation model (Koehn, 2009).

Neural machine translation (NMT) is a recently proposed approach to MT that employs an artificial neural network for end-to-end translation. NMT uses one single model for translation unlike SMT where several components are constructed, trained separately and combined to produce the translation.

2.1 Encoder-Decoder Framework

NMT neural models commonly belong to the encoder-decoder family of neural network models which is a subset of recurrent neural networks. For an introduction into neural networks and recurrent neural networks see for example Chapter 6 of *Deep Learning* by Ian Goodfellow, Yoshua Bengio and Aaron Courville (Goodfellow et al., 2016).

Fixed-length vocabularies are used both for the input and output languages that are automatically extracted from the training corpus based on frequency. If vocabulary size is n then the vocabulary usually consists of n most frequent tokens in the training corpus.

Figure 1 presents a high-level overview of a typical neural network used for translation. The encoder and decoder parts of the network are represented on the bottom and top parts of the figure respectively. An example translation from English into French is shown. The input sentence e and output sentence f are represented as sequences of tokens. Any space-delimited sequence of characters is considered a token.

The tokens are encoded using one-hot (1-of-K) encoding that represents each token as a vector of size $n \times 1$. These vectors consist of 0's except for one position in the vector that corresponds to the position of the word to be encoded in the vocabulary. This single position is filled with a 1.

One-hot vectors representing the words of the input sentence are transformed into continuous-space representations using an embedding layer in the network whose weights are learned during training.

Next the vectors are used to recurrently calculate the hidden state of the network. This state is represented as a fixed-length vector that is called a context vector.

The decoder uses this context vector to first generate a probability distribution over the output vocabulary and then choose the highest-scoring token for output. Each subsequent token is produced using the context vector and the previously generated token as input for the recurrent state. This process is repeated until a special end-of-word token

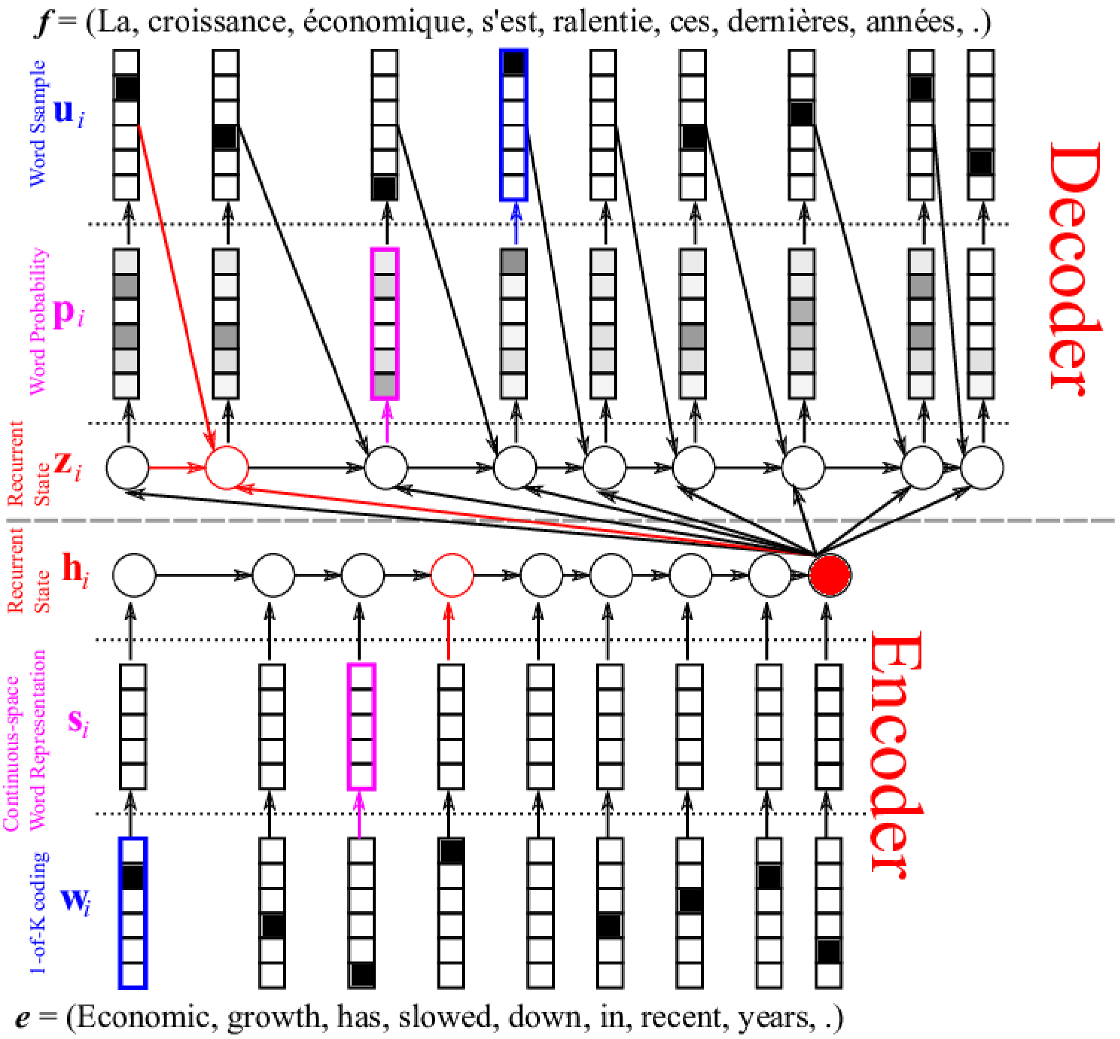
f = (La, croissance, économique, s'est, ralentie, ces, dernières, années, .)
Word Ssample
u_i
Word Probability
p_i
Recurrent State
z_i
Decoder
Recurrent State
h_i
Continuous-space Word Representation
s_i
1-of-K coding
w_i
Encoder
e = (Economic, growth, has, slowed, down, in, recent, years, .)

Figure 1. High-level view of a typical encoder-decoder framework for NMT (Cho, 2015a).

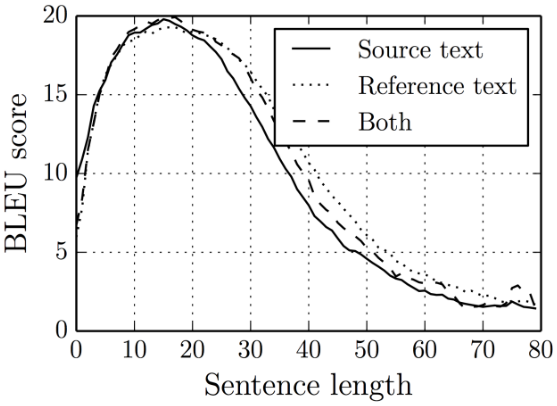

Figure 2. Translation performance of unseen sentences with respect to their length (Cho et al., 2014).

is produced by the network or a predefined maximum number of generated tokens in reached.

This approach presents a problem in the context of translation. Namely natural language sentences represented as sequences of tokens vary greatly in length. In practice sentences can vary from a few tokens in length to tens of tokens. While the input length is varied the internal representation of the sentence as a context vector always has a fixed length. This may introduce problems, especially for longer sentences (Bahdanau et al., 2014; Cho et al., 2014).

Figure 2 shows a correlation in translation performance as measured by BLEU (see Chapter 5.3) and sentence length. The continuous and dotted lines represent input and output sentence lengths respectively. The graph displays peak translation quality at a sentence length of approximately 15 tokens. A sudden drop in the BLEU score is observed for sentences longer than 20 tokens.

2.2 Attention Mechanism

To mitigate the problems that accompany encoding sentences into static fixed-length context vectors, a new approach was proposed that uses a bidirectional recurrent neural network (Schuster and Paliwal, 1997) to encode the sentence into a sequence of vectors called annotation vectors. The decoder uses a separate feed-forward neural network to find an alignment probability distribution over input tokens that shows how well each input token is aligned to the current output token to be generated (Bahdanau et al., 2014). This is called the attention mechanism since it represents the network's ability to concentrate its attention on some parts of the input sentence. The resulting context vector is calculated as a sum of annotation vectors weighted by alignment scores.

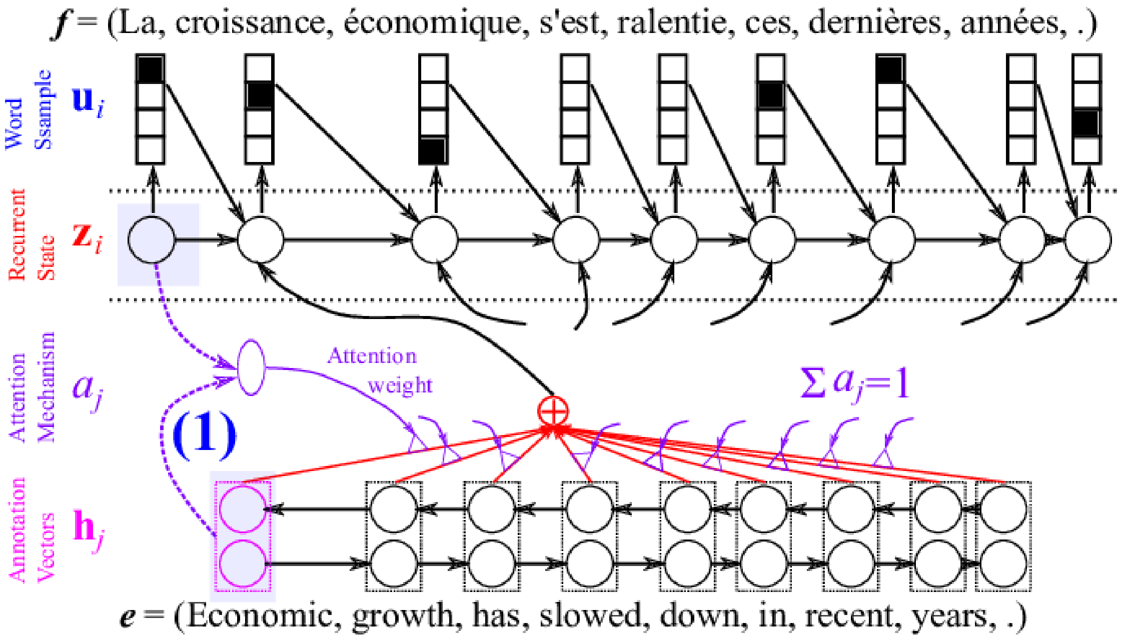
f = (La, croissance, économique, s'est, ralentie, ces, dernières, années, .)
Word Ssample
u_i
Recurrent State
z_i
Attention Mechanism
a_j
Attention weight
(1)
Σa_j=1
Annotation Vectors
h_j
e = (Economic, growth, has, slowed, down, in, recent, years, .)

Figure 3. Encoder-decoder framework for NMT with a bidirectional RNN encoder and attention mechanism (Cho, 2015b).

The attention mechanism allows the network to consider some input tokens more than others when generating different parts of the output sentence. For example when generating the first token of the output the network's attention mechanism can learn to consider tokens at the beginning of the source sentence more than tokens near the end. It can also learn to identify sentence structures where the alignment between tokens in the input and output sentences is not obvious.

Figure 3 demonstrates the use of a bidirectional recurrent neural network in the encoder and the attention mechanism assigning weights to annotation vectors while generating the second output token.

The bidirectional recurrent neural network (BiRNN) consists of two recurrent neural networks (RNN) that encode the sequence of tokens starting from opposite ends. The forward RNN reads input tokens as they are ordered and calculates a sequence of forward hidden states that correspond to each input word. The backward RNN reads tokens in the reverse order and produces a sequence of backward hidden states.

These states are concatenated to form an annotation vector for each input token. The attention mechanism is trained jointly with other parts of the framework to learn to predict alignment between parts of the input and output sentences. When generating an output token the attention mechanism calculates an attention weight for each input token taking as input the corresponding token's annotation vector and the decoder's state.

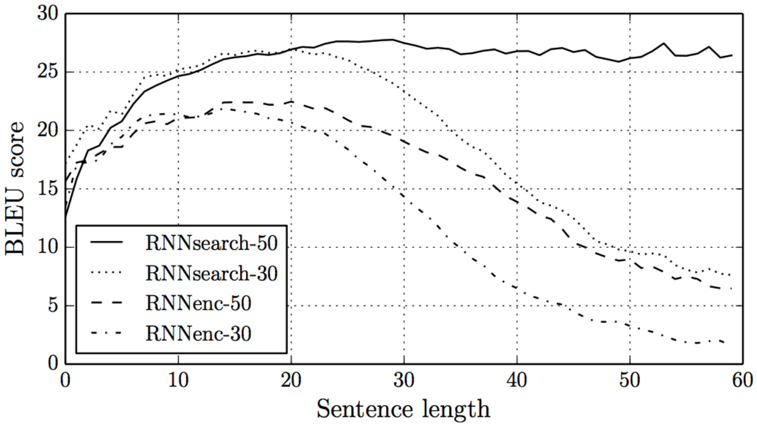

Figure 4. Translation performance of models with (RNNsearch) and without (RNNenc) the attention mechanism. Performance is measured on unseen sentences with respect to their length (Bahdanau et al., 2014).

After normalizing, the attention distribution over all input tokens is used to calculate the context vector as a weighted average of annotation vectors.

Since RNNs tend to represent recently processed inputs better the calculated context vector will also better represent tokens with a high predicted attention weight (Bahdanau et al., 2014).

This approach manages to address the problem of translating sentences with a large number of tokens. Figure 4 presents the translation performance with respect to sentence length of four models: a model without the attention mechanism (RNNenc) and a model with the attention mechanism (RNNsearch). Both models are trained twice: once with sentences up to length 50 and once with sentences up to length 30. It is clear that the model trained on long sentences with attention outperforms other models, especially on longer sentences.

The attention-equipped neural model's performance does not significantly degrade with longer sentences. In addition to translating long sentences this allows increasing the number of input tokens using subword segmentation techniques to mitigate a common problem in NMT—the limit on the size of the vocabulary.

2.3 Vocabulary Size Restriction

Recent NMT models typically outperform SMT systems but NMT has a major disadvantage compared to SMT: the complexity of training and using an NMT model is proportional to vocabulary size (Jean et al., 2014). Size of the vocabulary of an NMT model is typically restricted to 30 000–80 000 tokens (Bahdanau et al., 2014; Sutskever et al., 2014). Words that do not fit into the vocabulary are mapped to a special token UNK that represents an unknown word.

As the number of unknown words increases the translation performance of the model drops (Cho et al., 2014). This issue is even more prohibitive for languages with rich morphology and inflection (Jean et al., 2014) such as Estonian, Finnish and German.

Proposed solutions to this problem include a dictionary back-off (Luong et al., 2014; Jean et al., 2014), character-level translation (Costa-Jussà and Fonollosa, 2016; Chung et al., 2016; Lee et al., 2016) and segmentation into subword units (Sennrich et al., 2015).

NMT models have proven to produce state-of-the-art translation performance using subword segmentation as a text preprocessing technique (Sennrich et al., 2015). This method is discussed in the next chapter.

3 Subword Segmentation

Subword segmentation is a method of text preprocessing that segments words into subword units. Approaches can be commonly divided into unsupervised segmentation and linguistically-driven segmentation. The latter leverages morphological analysis to segment words according to predefined rules whereas the former aims to optimize an objective such as the probability of the segmented corpus (Fišel, 2011).

The goal of text segmentation in SMT is to reduce the number of unseen words i.e words that do not occur in the training data. This is often accomplished by separating compound nouns, stems, affixes and suffixes (Koehn and Knight, 2003; Popović and Ney, 2004; Popović et al., 2006).

In the context of MT it is beneficial to apply segmentation only to rare words that are not included in the vocabulary or are infrequent. Segmentation of frequent words should be discouraged since segmenting into smaller tokens typically increases ambiguity of the tokens and the distances between meaningful parts of the sentence.

3.1 Subword Segmentation in NMT

Unlike SMT, NMT is restricted to a small vocabulary size thus making it very important to limit the number of unique tokens in the text. A smaller number of tokens increases each token's ambiguity but recent NMT models employing the attention mechanism counteract this with high context-sensitivity similarly to earlier neural language models such as (Bengio et al., 2003).

One of the simplest methods of subword segmentation is using character n-gram models. This technique segments words into character sequences of at most length n with no word boundary crossing. This method decreases the number of unique tokens in the text but it segments all words regardless of their frequency and its segmentation has no meaningful reasoning (Sennrich et al., 2015).

Morfessor (Creutz and Lagus, 2005) is open-source software for unsupervised morphological analysis. It has widely been used in SMT, for example by (Virpioja et al., 2010), (de Gispert et al., 2009), (Kathol and Zheng, 2008) and (Virpioja et al., 2007). Morfessor attempts to segment words according to their morphological structure but similarly to n-gram models has no preference for infrequent words.

Current state-of-the-art NMT performance is achieved by a segmentation algorithm based on the byte-pair encoding scheme.

3.2 Byte-Pair Encoding (BPE) Segmentation

A segmentation algorithm was proposed by (Sennrich et al., 2015) for NMT that is inspired by the byte-pair encoding (BPE) technique. BPE is a data compression algorithm that iteratively replaces the most common pair of consecutive bytes with a single byte that does not appear in the data. The original data can be reconstructed using a table of the replacements (Gage, 1994).

This data compression algorithm can be adapted to subword segmentation. Every word is first viewed as a sequence of subword units of length 1 (characters). The most frequent pair of consecutive characters are then merged into a subword unit of length 2 consisting of those merged characters. The process is repeated for a predefined number of merges[1] (Sennrich et al., 2015).

If the number of merges is not limited then all subword units will be merged back together and the output will be equivalent to unsegmented text. After segmentation the number of unique tokens in the text is equal to the number of merge operations performed plus the number of unique characters in the text.

For example the 4-word corpus {'low', 'lowest', 'newer', 'wider'} with 2 allowed merges would be segmented into {'lo w', 'lo we s t', 'n e we r', 'w i d e r'} because the most frequent two pairs were ('w', 'e') and ('l', 'o').

The number of unique tokens can be controlled by the number of merge operations permitted. The number of merges can be set according to the NMT model's vocabulary size so that there will be no out-of-vocabulary words[2] while still utilizing the whole vocabulary.

BPE segmentation works well with NMT because it satisfies the requirement of only splitting rare words and leaving frequent words unsegmented.

However, the segmentation produced by BPE occasionally does not align with morphological structure and sometimes appears to be misleading.

[1] A Python implementation is available at https://github.com/rsennrich/subword-nmt.

[2] Unseen characters and subword units of which all occurrences have been merged into larger units are still out of vocabulary.

4 Morpheme-Aware Subword Segmentation

This chapter represents the main contribution of this thesis.

Due to the inaccuracies of BPE segmentation presented in Chapter 3.2 our goal was to modify the BPE segmentation algorithm to use morphological structure in segmentation decisions. The main idea was to restrict BPE from segmenting incomplete parts of different morphemes into one subword unit.

For example the word 'piimaalase' (meaning 'relating to milk') is segmented into 'pii', 'maal' and 'ase' by BPE since it is rare in our data while according to Morfessor the word consists of morphemes 'piima' and 'alase'. This BPE segmentation disagrees with morphological structure because the subword unit 'maal' consists of two incomplete parts of different morphemes: 'ma' from the first morpheme and 'al' from the second one. The segmentation of 'õppimisraskustele' into units 'õppimis' and 'raskustele' with identified morphemes as 'õppimis', 'raskus' and 'tele' agrees with morphological structure since no incomplete parts of different morphemes are merged into one subword unit.

Three additions to the original BPE segmentation algorithm were developed as part of this thesis. The most promising algorithm is an addition to BPE called temporary morpheme boundary restriction. The algorithm is described in Chapter 4.3, its performance in terms of segmentation is analyzed in Chapter 5.2 and its effects on translation are discussed in Chapter 5.3.

All algorithms require segmentation into morphemes as the first step, Morfessor is used for the provided examples and analysis. Python implementations of these algorithms are built on top of the original BPE Python implementation (Sennrich et al., 2015) and are freely available[3].

4.1 Morphemes as BPE Starting Tokens

BPE segmentation begins by considering each character a subword unit and starts merging the most frequent consecutive units. Our proposed addition uses morphemes as BPE starting subword units instead of characters. This prevents BPE from violating morpheme boundaries since subword units always consist of whole morphemes.

If no merges are permitted then the produced text is equivalent to segmentation into morphemes. If the number of merges is not limited then the produced output is equivalent to the original unsegmented text. This means that the full vocabulary size n will minimally be the number of unique morphemes and maximally the vocabulary size of the original text: $\#\ unique\ morphemes \leq n \leq \#\ unique\ words$.

[3] See https://github.com/kspar/subword-nmt.

4.2 Morpheme Boundary Restriction

Instead of considering morphemes as starting tokens (characters in normal BPE) they can also be considered ending tokens (words in normal BPE). In this algorithm words are first split into characters identically to normal BPE. Next the most frequent units are merged predefined number of times but no morpheme boundaries are allowed to be crossed while merging.

With no merges the output is identical to normal BPE with no merges. With an unlimited amount of merges the produced output is equivalent to the output of the algorithm described in Chapter 4.1 with no merges: segmentation into morphemes. Thus the full vocabulary size n will remain between the number of unique characters and number of unique morphemes: $\#\ unique\ characters \leq n \leq \#\ unique\ morphemes$.

4.3 Temporary Morpheme Boundary Restriction (TMBR)

The first method (Morphemes as BPE Starting Tokens) never allows segmentation to go below morpheme level. The second method (Morpheme Boundary Restriction) never allows segmentation to go above morpheme level. This means that resulting subword units are always subsequences of exactly one morpheme.

Both of these additions have problems when used with NMT. Firstly neither of these restrictions take full advantage of word and morpheme frequencies. If a word is rare then the first algorithm will be unable to segment it into smaller units than morphemes even when this is advantageous. The second algorithm is suboptimal for common words—it is unable to merge morphemes inside a word even if the word is very common.

Secondly neither of these algorithms allows as flexible control over the full vocabulary size as normal BPE does. BPE-segmented text has a full vocabulary size of $\#\ unique\ characters + \#\ merge\ operations$ while the number of merge operations can be freely set. The proposed restrictions confine the vocabulary size of the resulting text to more than the number of distinct morphemes (first method) or less than the number of distinct morphemes (second method).

Both of these problems are addressed in our proposed third method that starts out with characters as does normal BPE but dynamically decides for each word whether to allow merging separate morphemes. The algorithm keeps track of the words' morphological structure as identified by Morfessor and allows BPE to consider merging separate morphemes only if all morphemes in the word already consist of exactly one token. This means that only merging whole morphemes is allowed while merging of proper subsequences of different morphemes is forbidden.

This restriction disallows violating morphological structure while keeping merges dynamic enough to allow merging frequent tokens back into whole words and segmenting rare words into very small subword units. The resulting full vocabulary size is also not fixed and varies only with the number of allowed merges just as with normal BPE.

5 Evaluation

We aim to answer the following research questions:

- Does using the BPE segmentation algorithm with TMBR improve the segmentation of words whose segmentation by normal BPE disagrees with morphological structure and how does BPE with TMBR affect the overall segmentation quality?
- Does the translation of words whose segmentation by normal BPE disagrees with morphological structure improve compared to normal BPE and how does using BPE with TMBR affect the overall translation performance?

Chapter 5.1 presents a description of the experiment setup, chapters 5.2 and 5.3 contain the analysis of segmentation and translation respectively.

5.1 Experiment Setup

We performed experiments on the Estonian-English data from parallel corpora of Europarl (Koehn et al., 2002), EU Journal (Hajlaoui et al., 2014) and OPUS Open Subtitles (Tiedemann, 2012). The data consists of 14 million sentence pairs or approximately 125 million Estonian tokens and 160 million English tokens. Scripts from the open-source toolkit Moses (Koehn et al., 2007) are used for tokenization and truecasing. The data is split into testing (3000 sentences), development (3000 sentences) and training sets.

We analyze the segmentation and translation of two classes of sentences. Both classes are subsets of the testing set. The first class contains sentences that contain words whose segmentation by BPE violated morphological structure. The second class contains all sentences. We analyze randomly sampled sentences from both classes.

We train three separate models:

- a baseline model using normal BPE for segmentation (BPE)
- a model using Morfessor for segmentation (Morfessor)
- a model using BPE with TMBR for segmentation (BPE+TMBR)

All segmentation methods are applied both to the input and output texts. Training and development sets for BPE and BPE+TMBR models are segmented with 50 000 allowed merges. All models have a maximum vocabulary size of 51 000 tokens. Each model is trained for approximately 5 days an a machine equipped with an NVIDIA Tesla K80 GPU and 4 x 2.7 GHz CPUs. Translation experiments are performed with Neural Monkey (Helcl and Libovickȳ, 2017).

5.2 Segmentation Analysis

An analysis of BPE segmentation showed that out of 3000 sentences (41 001 words) in the Estonian test set the segmentation disagreed with morphological structure for 1906 words (4.6%) in 1096 sentences (36.5%).

We manually analyzed and subjectively rated the segmentation of 30 randomly selected BPE-violated sentences by BPE and BPE+TMBR based on adequacy and agreement with our intuition. Table 1 shows three sample sentences and their segmentations by BPE, Morfessor and BPE with TMBR. Segmentation is represented by the symbol ’· · · ’. Words with differing segmentations are shown in bold. BPE+TMBR was rated best in all three samples.

Morfessor	täpsus — hinnangute lähedus tundmatute· · · le õigete· · · le väärtuste· · · le ;
BPE	täpsus — hinnangute lähe· · · dus **tund**· · · **matutele** **õige**· · · **tele** **väärtustele** ;
BPE+TMBR	täpsus — hinnangute lähe· · · dus **tundmatu**· · · **te**· · · **le** **õigete**· · · **le** **väärtuste**· · · **le** ;
Morfessor	kujutle· · · ge , mille· · · le seda oleks võimalik kulu· · · tada seoses arenevate riikide haiglate , arstide ja õpetajate· · · ga .
BPE	kujutle· · · ge , millele seda oleks võimalik kulutada seoses **arene**· · · **vate** riikide **haig**· · · **late** , arstide ja **õpeta**· · · **jatega** .
BPE+TMBR	kujutle· · · ge , millele seda oleks võimalik kulutada seoses **areneva**· · · **te** riikide **haigla**· · · **te** , arstide ja **õpetajate**· · · **ga** .
Morfessor	tule nüüd . ära raiska oma aega selle jama seletamise· · · le .
BPE	tule nüüd . ära raiska oma aega selle jama **sele**· · · **tamisele** .
BPE+TMBR	tule nüüd . ära raiska oma aega selle jama **seleta**· · · **mise**· · · **le** .

Table 1. Segmentation of 3 randomly selected Estonian sentences where BPE segmentation violates morpheme boundaries. Segmentation boundaries are represented by ’· · · ’. Differing segmentations are shown in bold.

Out of 30 manually rated sentences normal BPE segmentation was rated best in 2 cases (6.7%), BPE with TMBR was rated best in 21 cases (70%) and no significant difference was observed in 7 cases (23.3%).

An identical analysis of BPE segmentation on the English test set revealed that out of 3000 sentences (54 146 words) the segmentation disagreed with morphological structure for 381 words (0.7%) in 308 sentences (10.3%).

Out of 30 manually rated sentences normal BPE segmentation was rated best in 3 cases (10%), BPE with TMBR was rated best in 16 cases (53.3%) and no significant difference was observed in 11 cases (36.7%). Table 2 shows 3 sample sentences and their segmentations by BPE, Morfessor and BPE with TMBR. Words with differing segmentations are shown in bold. In the first sample BPE+TMBR was rated best, no significant change was observed in the second sample and BPE segmentation was rated best in the third sample.

Morfessor	ac···curacy shall refer to the close···ness of estimate···s to the unknown true values ,
BPE	accuracy shall refer to the **clo···seness** of estimates to the unknown true values ,
BPE+TMBR	accuracy shall refer to the **close···ness** of estimates to the unknown true values ,
Morfessor	Mr Mon···net and Mr Schu···man were very far···-···sighted people .
BPE	Mr Monnet and Mr **Sch···uman** were very far-···sighted people .
BPE+TMBR	Mr Monnet and Mr **Schu···man** were very far···-···sighted people .
Morfessor	the climate is pre···valent···ly dry , with moderate wind and high temperature···s .
BPE	the climate is **preval···ently** dry , with moderate wind and high temperatures .
BPE+TMBR	the climate is **pre···val···ent···ly** dry , with moderate wind and high temperatures .

Table 2. Segmentation of 3 randomly selected English sentences where BPE segmentation violates morpheme boundaries. Segmentation boundaries are represented by '···'. Differing segmentations are shown in bold.

We also analyzed the segmentation of the whole test set. The Estonian segmentation of BPE and BPE with TMBR differed for 3431 words (8.37%) in 1569 sentences (52.3%) and the English segmentation differed for 880 words (1.63%) in 634 sentences (21.13%). Tables 3 and 4 show a sample segmentation of an Estonian and an English sentence where the segmentation of BPE and BPE with TMBR differed. Differing words are shown in bold.

On the whole Estonian test set out of 30 manually rated sentences normal BPE segmentation was rated best in 6 cases (20%), BPE with TMBR was rated best in 9 cases (30%) and no significant difference was observed in 15 cases (50%).

On the whole English test set normal BPE segmentation was rated best in 10 cases (33.3%), BPE with TMBR was rated best in 9 cases (30%) and no significant difference was observed in 11 cases (36.7%).

Morfessor	Liberaal· · · ide fraktsioon on esitanu· · · d taotluse pikenda· · · da tähtaega resolutsiooni ühisettepaneku· · · te ja muudatusettepaneku· · · te esitamise· · · ks Euroopa Parlamendi prioriteetide kohta seoses komisjoni õigusloome· · · - ja töö· · · programmiga homse· · · ni , teisipäeva· · · ni , 23.09· · · .2008 kella 10 : 00· · · -· · · ni .
BPE	Liberaalide fraktsioon on esitanud taotluse pikendada tähtaega resolutsiooni **ühise**· · · **ttepane**· · · **kute** ja muudatusettepanekute esitamiseks Euroopa Parlamendi prioriteetide kohta seoses komisjoni õigusloome- ja tööprogrammiga homseni , **teisi**· · · **päevani** , **23.**· · · **09.**· · · **2008** kella 10 : 00· · · -ni .
BPE+TMBR	Liberaalide fraktsioon on esitanud taotluse pikendada tähtaega resolutsiooni **ühisettepane**· · · **ku**· · · **te** ja muudatusettepanekute esitamiseks Euroopa Parlamendi prioriteetide kohta seoses komisjoni õigusloome- ja tööprogrammiga homseni , **teisipäeva**· · · **ni** , **23.**· · · **09**· · · **.2008** kella 10 : 00· · · -· · · ni .

Table 3. Segmentation of an Estonian sentence from the test set. Segmentation boundaries are represented by '· · · '. Differing segmentations are shown in bold.

Morfessor	led module (s) shall be tamper· · · proof .
BPE	led module (s) shall be **tam**· · · **per**· · · **proof** .
BPE+TMBR	led module (s) shall be **tamper**· · · **proof** .

Table 4. Segmentation of an English sentence from the test set. Segmentation boundaries are represented by '· · · '. The differing segmentation is shown in bold.

5.3 Effects on NMT

Translation performance is measured by BLEU (Papineni et al., 2002) and TER (Snover et al., 2006). Quantitative results are shown in Table 5.

Model	BLEU	TER
Morfessor	30.0	54.8
BPE	31.0	54.0
BPE+TMBR	30.1	54.6

Table 5. Translation performance of trained models as measured by BLEU and TER.

Morfessor denotes the model trained on data segmented into morphemes by Morfessor. BPE is the baseline model trained on data that is segmented by normal BPE. BPE+TMBR is the model trained on data that is segmented with the proposed algorithm.

We manually evaluate 30 translations of sentences where BPE disagrees with morphological structure and 30 random sentences from the test set. Each sentence is translated by the baseline BPE model and new the BPE+TMBR model. Samples of these sentences are shown in Table 6 and Table 7. The input text that was translated is denoted by Source.

Source	täpsus — hinnangute lähedus tundmatutele õigetele väärtustele ;
BPE	accuracy — evaluations of estimates for unknown right values ;
BPE+TMBR	accuracy — estimates of the estimated values to be unknown ;
Source	kujutlege , millele seda oleks võimalik kulutada seoses arenevate riikide haiglate , arstide ja õpetajatega .
BPE	imagine what it would be possible to spend with the hospitals , doctors and teachers .
BPE+TMBR	imagine what it could be possible to spend in relation to the hospital of the countries , doctors and teachers .
Source	tule nüüd . ära raiska oma aega selle jama seletamisele .
BPE	don &apos;t waste your time to explain this shit .
BPE+TMBR	don &apos;t waste your time to explain this shit .

Table 6. Translations of sentences where BPE segmentation disagrees with morphology.

For BPE-violated sentences the translation by the BPE model is rated better in 9 cases (30%) and BPE+TMBR in 13 cases (43.3%), the translations are rated equal in 8 cases (26.7%). For all sentences the translation by the BPE model is rated better in 13

Source	Liberaalide fraktsioon on esitanud taotluse pikendada tähtaega resolutsiooni ühisettepanekute ja muudatusettepanekute esitamiseks Euroopa Parlamendi prioriteetide kohta seoses komisjoni õigusloome- ja tööprogrammiga homseni , teisipäevani , 23.09.2008 kella 10 : 00-ni .
BPE	the applicant Group has submitted a request for a resolution on the priorities of the motion for a resolution on the priorities of the European Parliament &apos;s priorities for the Commission &apos;s legislative and work programme tomorrow , Tuesday , 23.09.2008 : 00 .
BPE+TMBR	the applicant Group has submitted a request for the submission of the motion for a resolution on the priorities and amendments to the European Parliament priorities in the Commission &apos;s legislative and work work work tomorrow , Tuesday , 23.09.2008 : 00 : 00-i .
Source	see on suurepärane mõte ja mul on selle üle üksnes hea meel .
BPE	it &apos;s a great idea , and I &apos;m just glad to be .
BPE+TMBR	that &apos;s a great idea , and I &apos;m just glad to do it .
Source	tunnen ennast juba nagu rohtlajänes .
BPE	I feel like a rock rabbit .
BPE+TMBR	I feel like a ro-ass rabbit .

Table 7. Translations of random sentences from the test set.

cases (43.3%) and BPE+TMBR in 12 cases (40%), the translations are rated equal in 5 cases (16.7%).

5.4 Discussion

The first research question was whether using the proposed TMBR restriction with BPE improves the segmentation of words whose segmentation by normal BPE disagrees with morphological structure and how does BPE with TMBR affect the overall segmentation quality.

Manual analysis of segmentation by BPE and BPE+TMBR showed that BPE+TMBR produced subjectively more adequate and intuitive segmentations for words whose segmentation by BPE did not agree with morphological structure. For Estonian sentences BPE+TMBR algorithm's segmentation was rated best in 70% of cases while in 23.3% of cases there was no significant difference in segmentation. For English sentences BPE+TMBR was best in 53.3% of cases and 36.7% of segmentations did not differ significantly.

Analysis of segmentations of the whole test set showed that BPE+TMBR was rated mildly better than normal BPE for Estonian sentences (30% vs 20%) and mildly worse for English sentences (30% vs 33.3%). We conclude that BPE+TMBR improves the segmentation of words whose segmentation by normal BPE disagrees with morphology and produces no significant difference in adequacy for other segmentations compared to BPE.

The second research question was whether the translation of words whose segmentation by normal BPE disagrees with morphology improves when using BPE+TMBR compared to normal BPE and how does using BPE+TMBR affect the overall translation performance?

Quantitative analysis revealed that using BPE+TMBR decreased the translation performance as measured by BLEU by 0.9 points. A manual analysis of translations showed that BPE+TMBR produced more adequate translations for sentences that contained words whose segmentation by normal BPE disagrees with morphology in 43.3% of cases while BPE was considered better in 30% of these cases. On samples from the whole test set the translation by BPE+TMBR was considered better in 40% of cases while BPE was considered better in 43.3% of cases.

We conclude that BPE+TMBR produced moderately better translations for sentences that contained words whose segmentation by BPE disagreed with morphology.

6 Conclusion

Current state-of-the-art translation performance in NMT is achieved using the byte-pair encoding (BPE) subword segmentation algorithm as an input preprocessing step. Due to only considering frequencies of word parts this algorithm occasionally fails to respect morphological structure of the words and leads to incorrect translations.

A new algorithm for subword segmentation was developed that refrains from violating morpheme boundaries if this leads to a disagreement with morphology. The algorithm allows only subsets of one morpheme or the concatenation of whole morphemes to be considered as one subword unit. This approach succeeds in preventing misleading segmentations that occurred due to the disregard for morphology in the baseline BPE system.

Using the proposed algorithm for segmentation of source and target texts improved the segmentation and translation of sentences from Estonian into English that contained words whose segmentation by BPE did not agree with morphological structure. Overall translation performance as measured by BLEU decreased by 0.9 points compared the baseline system.

Future research in this field includes determining the impact of incorrect BPE segmentation on the task of post-editing and further research into the refinement of BPE segmentation to prevent incorrect translations without using morphological structure. Other approaches such as linguistic morphological analysis could also be used for segmentation into morphemes instead of unsupervised methods.

References

Bahdanau, D., Cho, K., and Bengio, Y.: 2014, *Neural Machine Translation by Jointly Learning to Align and Translate*, CoRR abs/1409.0473

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C.: 2003, *A neural probabilistic language model*, Journal of machine learning research **3(Feb)**, 1137

Cho, K.: 2015a, *Introduction to Neural Machine Translation with GPUs (Part 2)*, `https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-2/`

Cho, K.: 2015b, *Introduction to Neural Machine Translation with GPUs (part 3)*, `https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-3/`

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y.: 2014, *On the Properties of Neural Machine Translation: Encoder-Decoder Approaches*, CoRR abs/1409.1259

Chung, J., Cho, K., and Bengio, Y.: 2016, *A Character-level Decoder without Explicit Segmentation for Neural Machine Translation*, CoRR abs/1603.06147

Costa-Jussà, M. R. and Fonollosa, J. A. R.: 2016, *Character-based Neural Machine Translation*, CoRR abs/1603.00810

Creutz, M. and Lagus, K.: 2005, *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*, Helsinki University of Technology

de Gispert, A., Virpioja, S., Kurimo, M., and Byrne, W.: 2009, *Minimum Bayes Risk Combination of Translation Hypotheses from Alternative Morphological Decompositions*, in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, NAACL-Short '09, pp 73–76, Association for Computational Linguistics, Stroudsburg, PA, USA

Fišel, M.: 2011, *Optimizing Statistical Machine Translation via Input Modification*, *Ph.D. thesis*, University of Tartu

Gage, P.: 1994, *A new algorithm for data compression*, The C Users Journal **12(2)**, 23

Goodfellow, I., Bengio, Y., and Courville, A.: 2016, *Deep Learning*, MIT Press, `http://www.deeplearningbook.org`

Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., and Varga, D.: 2014, *DCEP-Digital Corpus of the European Parliament.*, in LREC, pp 3164–3171, Citeseer

Helcl, J. and Libovickỳ, J.: 2017, *Neural Monkey: An Open-source Tool for Sequence Learning*, The Prague Bulletin of Mathematical Linguistics **107(1)**, 5

Jean, S., Cho, K., Memisevic, R., and Bengio, Y.: 2014, *On Using Very Large Target Vocabulary for Neural Machine Translation*, CoRR abs/1412.2007

Kalchbrenner, N. and Blunsom, P.: 2013, *Recurrent Continuous Translation Models.*, in EMNLP, Vol. 3, p. 413

Kathol, A. and Zheng, J.: 2008, *Strategies for building a Farsi-English SMT system from limited resources.*, in Interspeech, pp 2731–2734

Koehn, P.: 2009, *Statistical Machine Translation*, Cambridge University Press

Koehn, P. et al.: 2002, *Europarl: A multilingual corpus for evaluation of machine translation*

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 ACL on interactive poster and demonstration sessions, pp 177–180, Association for Computational Linguistics

Koehn, P. and Knight, K.: 2003, *Empirical Methods for Compound Splitting*, in Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1, EACL '03, pp 187–193, Association for Computational Linguistics, Stroudsburg, PA, USA

Koehn, P., Och, F. J., and Marcu, D.: 2003, *Statistical Phrase-based Translation*, in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL '03, pp 48–54, Association for Computational Linguistics, Stroudsburg, PA, USA

Lee, J., Cho, K., and Hofmann, T.: 2016, *Fully Character-Level Neural Machine Translation without Explicit Segmentation*, CoRR abs/1610.03017

Luong, T., Sutskever, I., Le, Q. V., Vinyals, O., and Zaremba, W.: 2014, *Addressing the Rare Word Problem in Neural Machine Translation*, CoRR abs/1410.8206

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 on association for computational linguistics, pp 311–318, Association for Computational Linguistics

Popović, M. and Ney, H.: 2004, *Towards the Use of Word Stems and Suffixes for Statistical Machine Translation.*, in LREC

Popović, M., Stein, D., and Ney, H.: 2006, in *Advances in Natural Language Processing*, pp 616–624, Springer Berlin Heidelberg

Schuster, M. and Paliwal, K.: 1997, *Bidirectional recurrent neural networks*, IEEE Transactions on Signal Processing **45(11)**, 2673

Sennrich, R., Haddow, B., and Birch, A.: 2015, *Neural Machine Translation of Rare Words with Subword Units*, CoRR abs/1508.07909

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J.: 2006, *A study of translation edit rate with targeted human annotation*, in Proceedings of association for machine translation in the Americas, Vol. 200

Sutskever, I., Vinyals, O., and Le, Q. V.: 2014, *Sequence to sequence learning with neural networks*, in Advances in neural information processing systems, pp 3104–

3112

Tiedemann, J.: 2012, *Parallel Data, Tools and Interfaces in OPUS*, in N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), Istanbul, Turkey

Virpioja, S., Väyrynen, J., Mansikkaniemi, A., and Kurimo, M.: 2010, *Applying Morphological Decomposition to Statistical Machine Translation*, in Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, WMT '10, pp 195–200, Association for Computational Linguistics, Stroudsburg, PA, USA

Virpioja, S., Väyrynen, J. J., Creutz, M., and Sadeniemi, M.: 2007, *Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner*, Machine Translation Summit XI **2007**, 491

Acknowledgements

I express my gratitude to Mark Fišel for strict supervision and relentless motivational speeches.

The research presented in this thesis was supported by IT Academy and the Estonian Students Fund in USA.

Appendix

I. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Kaspar Papli**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

   1.1 reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

   1.2 make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

   of my thesis

   Morpheme-Aware Subword Segmentation for Neural Machine Translation

   supervised by Mark Fišel

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 11.05.2017