From characters to words to in between: Do we capture morphology?

Citation for published version:

Digital Object Identifier (DOI):
10.18653/v1/P17-1184

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Abstract

Words can be represented by composing the representations of subword units such as word segments, characters, and/or character n-grams. While such representations are effective and may capture the morphological regularities of words, they have not been systematically compared, and it is not understood how they interact with different morphological typologies. On a language modeling task, we present experiments that systematically vary (1) the basic unit of representation, (2) the composition of these representations, and (3) the morphological typology of the language modeled. Our results extend previous findings that character representations are effective across typologies, and we find that a previously unstudied combination of character trigram representations composed with bi-LSTMs outperforms most others. But we also find room for improvement: none of the character-level models match the predictive accuracy of a model with access to true morphological analyses, even when learned from an order of magnitude more data.

1 Introduction

Continuous representations of words learned by neural networks are central to many NLP tasks (Cho et al., 2014; Chen and Manning, 2014; Dyer et al., 2015). However, directly mapping a finite set of word types to a continuous representation has well-known limitations. First, it makes a closed vocabulary assumption, enabling only generic out-of-vocabulary handling. Second, it cannot exploit systematic functional relationships in learning. For example, cat and cats stand in the same relationship as dog and dogs. While this relationship might be discovered for these specific frequent words, it does not help us learn that the same relationship also holds for the much rarer words sloth and sloths.

These functional relationships reflect the fact that words are composed from smaller units of meaning, or morphemes. For instance, cats consists of two morphemes, cat and -s, with the latter shared by the words dogs and tarsiers. Modeling this effect is crucial for languages with rich morphology, where vocabulary sizes are larger, many more words are rare, and many more such functional relationships exist. Hence, some models produce word representations as a function of subword units obtained from morphological segmentation or analysis (Luong et al., 2013; Botha and Blunsom, 2014; Cotterell and Schütze, 2015). A downside of these models is that they depend on morphological segmenters or analyzers.

Morphemes typically have similar orthographic representations across words. For example, the morpheme -s is realized as -es in finches. Since this variation is limited, the general relationship between morphology and orthography can be exploited by composing the representations of characters (Ling et al., 2015; Kim et al., 2016), character n-grams (Sperr et al., 2013; Wieting et al., 2016; Bojanowski et al., 2016; Botha and Blunsom, 2014), bytes (Plank et al., 2016; Gillick et al., 2016), or combinations thereof (Santos and Zadrozny, 2014; Qiu et al., 2014). These models are compact, can represent rare and unknown words, and do not require morphological analyzers. They raise a provocative question: Does NLP benefit from models of morphology, or can they be replaced entirely by models of characters?

The relative merits of word, subword, and character-level models are not fully understood because each new model has been compared on
different tasks and datasets, and often compared against word-level models. A number of questions remain open:

1. How do representations based on morphemes compare with those based on characters?

2. What is the best way to compose subword representations?

3. Do character-level models capture morphology in terms of predictive utility?

4. How do different representations interact with languages of different morphological typologies?

The last question is raised by Bender (2013): languages are typologically diverse, and the behavior of a model on one language may not generalize to others. Character-level models implicitly assume concatenative morphology, but many widely-spoken languages feature non-concatenative morphology, and it is unclear how such models will behave on these languages.

To answer these questions, we performed a systematic comparison across different models for the simple and ubiquitous task of language modeling. We present experiments that vary (1) the type of subword unit; (2) the composition function; and (3) morphological typology. To understand the extent to which character-level models capture true morphological regularities, we present oracle experiments using human morphological annotations instead of automatic morphological segments. Our results show that:

1. For most languages, character-level representations outperform the standard word representations. Most interestingly, a previously unstudied combination of character trigrams composed with bi-LSTMs performs best on the majority of languages.

2. Bi-LSTMs and CNNs are more effective composition functions than addition.

3. Character-level models learn functional relationships between orthographically similar words, but don’t (yet) match the predictive accuracy of models with access to true morphological analyses.

4. Character-level models are effective across a range of morphological typologies, but orthography influences their effectiveness.

<table>
<thead>
<tr>
<th>word</th>
<th>tries</th>
</tr>
</thead>
<tbody>
<tr>
<td>morphemes</td>
<td>try+s</td>
</tr>
<tr>
<td>morphs</td>
<td>tri+es</td>
</tr>
<tr>
<td>morph. analysis</td>
<td>try+VB+3rd+SG+Pres</td>
</tr>
</tbody>
</table>

Table 1: The morphemes, morphs, and morphological analysis of tries.

## 2 Morphological Typology

A morpheme is the smallest unit of meaning in a word. Some morphemes express core meaning (roots), while others express one or more dependent features of the core meaning, such as person, gender, or aspect. A morphological analysis identifies the lemma and features of a word. A morph is the surface realization of a morpheme (Morley, 2000), which may vary from word to word. These distinctions are shown in Table 1.

Morphological typing classifies languages based on the processes by which morphemes are composed to form words. While most languages will exhibit a variety of such processes, for any given language, some processes are much more frequent than others, and we will broadly identify our experimental languages with these processes.

When morphemes are combined sequentially, the morphology is concatenative. However, morphemes can also be composed by non-concatenative processes. We consider four broad categories of both concatenative and non-concatenative processes in our experiments.

**Fusional languages** realize multiple features in a single concatenated morpheme. For example, English verbs can express number, person, and tense in a single morpheme:

```
<table>
<thead>
<tr>
<th>word</th>
<th>tries</th>
</tr>
</thead>
<tbody>
<tr>
<td>morphemes</td>
<td>try+s</td>
</tr>
<tr>
<td>morphs</td>
<td>tri+es</td>
</tr>
<tr>
<td>morph. analysis</td>
<td>try+VB+3rd+SG+Pres</td>
</tr>
</tbody>
</table>
```

Fusional languages realize multiple features in a single concatenated morpheme. For example, English verbs can express number, person, and tense in a single morpheme:

```
<table>
<thead>
<tr>
<th>word</th>
<th>tries</th>
</tr>
</thead>
<tbody>
<tr>
<td>morphemes</td>
<td>try+s</td>
</tr>
<tr>
<td>morphs</td>
<td>tri+es</td>
</tr>
<tr>
<td>morph. analysis</td>
<td>try+VB+3rd+SG+Pres</td>
</tr>
</tbody>
</table>
```

For example (Haspelmath, 2010):

```
okursam (Turkish)
oku+r+sa+m
```

```
“read”+AOR+COND+1SG
```

Agglutinative languages assign one feature per morpheme. Morphemes are concatenated to form a word and the morpheme boundaries are clear. For example (Haspelmath, 2010):

```
okursam (Turkish)
oku+r+sa+m
```

```
“read”+AOR+COND+1SG
```

**Root and Pattern Morphology** forms words by inserting consonants and vowels of dependent morphemes into a consonantal root based on a given pattern. For example, the Arabic root *ktb* (“write”) produces (Roark and Sproat, 2007):

```
katab “wrote” (Arabic)
```
We use three variants of \( \sigma \) in Eq. 1, each returning a different type of subword unit: character, character trigram, or one of two types of morph. Morphs are obtained from Morfessor (Smit et al., 2014) or a word segmentation based on Byte Pair Encoding (BPE; Gage (1994)), which has been shown to be effective for handling rare words in neural machine translation (Sennrich et al., 2016). BPE works by iteratively replacing frequent pairs of characters with a single unused character. For Morfessor, we use default parameters while for BPE we set the number of merge operations to 10,000.\(^1\) When we segment into character trigrams, we consider all trigrams in the word, including those covering notional beginning and end of word characters, as in Sperr et al. (2013). Example output of \( \sigma \) is shown in Table 3.

3.2 Composition Functions

We use three variants of \( f \) in Eq. 1. The first constructs the representation \( w \) of word \( w \) by adding the representations of its subwords \( s_1, \ldots, s_n = \sigma(w) \), where the representation of \( s_i \) is vector \( s_i \).

\[
\begin{align*}
w &= f\left(W_s, \sigma(w)\right) \\
\text{where } \sigma \text{ is a deterministic function that returns a sequence of subword units; } W_s \text{ is a parameter matrix of representations for the vocabulary of subword units; and } f \text{ is a composition function which takes } \sigma(w) \text{ and } W_s \text{ as input and returns } w. \text{ All of the representations that we consider take this form, varying only in } f \text{ and } \sigma.
\end{align*}
\]

3.1 Subword Units

We consider four variants of \( \sigma \) in Equation 1, each returning a different type of subword unit: character, character trigram, or one of two types of morph. Morphs are obtained from Morfessor (Smit et al., 2014) or a word segmentation based on Byte Pair Encoding (BPE; Gage (1994)), which has been shown to be effective for handling rare words in neural machine translation (Sennrich et al., 2016). BPE works by iteratively replacing frequent pairs of characters with a single unused character. For Morfessor, we use default parameters while for BPE we set the number of merge operations to 10,000.\(^1\) When we segment into character trigrams, we consider all trigrams in the word, including those covering notional beginning and end of word characters, as in Sperr et al. (2013). Example output of \( \sigma \) is shown in Table 3.

3.2 Composition Functions

We use three variants of \( f \) in Eq. 1. The first constructs the representation \( w \) of word \( w \) by adding the representations of its subwords \( s_1, \ldots, s_n = \sigma(w) \), where the representation of \( s_i \) is vector \( s_i \).

\[
\begin{align*}
w &= f\left(W_s, \sigma(w)\right) \\
\text{where } \sigma \text{ is a deterministic function that returns a sequence of subword units; } W_s \text{ is a parameter matrix of representations for the vocabulary of subword units; and } f \text{ is a composition function which takes } \sigma(w) \text{ and } W_s \text{ as input and returns } w. \text{ All of the representations that we consider take this form, varying only in } f \text{ and } \sigma.
\end{align*}
\]

The only subword unit that we don’t compose by addition is characters, since this will produce the same representation for many different words.

Our second composition function is a bidirectional long-short-term memory (bi-LSTM), which we adapt based on its use in the character-level model of Ling et al. (2015) and its widespread use in NLP generally. Given \( s_i \) and the previous LSTM hidden state \( h_{i-1} \), an LSTM (Hochreiter and Schmidhuber, 1997) computes the following outputs for the subword at position \( i \):

\[
\begin{align*}
h_i &= LSTM\left(s_i, h_{i-1}\right) \\
\hat{s}_{i+1} &= g(V^T \cdot h_i)
\end{align*}
\]

where \( \hat{s}_{i+1} \) is the predicted target subword, \( g \) is the softmax function and \( V \) is a weight matrix.

A bi-LSTM (Graves et al., 2005) combines the final state of an LSTM over the input sequence with one over the reversed input sequence. Given the hidden state produced from the final input of the forward LSTM, \( h_n^{fw} \) and the hidden state produced from the final input of the backward LSTM \( h_0^{bw} \), we compute the word representation as:

\[
\begin{align*}
w_t &= W_f \cdot h_n^{fw} + W_b \cdot h_0^{bw} + b
\end{align*}
\]

\( W_f, W_b, \) and \( b \) are parameter matrices and \( h_n^{fw} \) and \( h_0^{bw} \) are forward and backward LSTM states, respectively.

The third composition function is a convolutional neural network (CNN) with highway layers, as in Kim et al. (2016). Let \( c_1, \ldots, c_k \) be the sequence of characters of word \( w \). The character embedding matrix is \( C \in \mathbb{R}^{d \times k} \), where the \( i \)-th column corresponds to the embeddings of \( c_i \). We first apply a narrow convolution between \( C \) and a filter \( F \in \mathbb{R}^{d \times n} \) of width \( n \) to obtain a feature map \( f \in \mathbb{R}^{k-n+1} \). In particular, the computation of the \( j \)-th element of \( f \) is defined as:

\[
\begin{align*}
f[j] &= tanh(\langle C[s, j : j + n - 1], F \rangle + b)
\end{align*}
\]

where \( \langle A, B \rangle = \text{Tr}(AB^T) \) is the Frobenius inner product and \( b \) is a bias. The CNN model applies filters of varying width, representing features
Models | Subword Unit(s) | Composition Function
--- | --- | ---
Sperr et al. (2013) | words, character n-grams | addition
Luong et al. (2013) | morphs (Morfessor) | recursive NN
Botha and Blunsom (2014) | words, morphs (Morfessor) | addition
Qiu et al. (2014) | words, morphs (Morfessor) | addition
Santos and Zadrozny (2014) | words, characters | CNN
Coțterell and Schütze (2015) | words, morphological analyses | addition
Sennrich et al. (2016) | morphs (BPE) | none
Kim et al. (2016) | characters | CNN
Ling et al. (2015) | characters | bi-LSTM
Wieting et al. (2016) | character n-grams | addition
Bojanowski et al. (2016) | character n-grams | addition
Vylomova et al. (2016) | characters, morphs (Morfessor) | bi-LSTM, CNN
Miyamoto and Cho (2016) | words, characters | bi-LSTM
Rei et al. (2016) | words, characters | bi-LSTM
Lee et al. (2016) | characters | CNN
Kann and Schütze (2016) | characters, morphological analyses | none
Heigold et al. (2017) | words, characters | bi-LSTM, CNN

Table 2: Summary of previous work on representing words through compositions of subword units.

<table>
<thead>
<tr>
<th>Unit</th>
<th>Output of $\sigma(wants)$</th>
<th>$\sigma$(the)</th>
<th>$\sigma$(dogs)</th>
<th>$\sigma$(are)</th>
<th>subword units</th>
<th>compositional function</th>
<th>word embedding</th>
<th>softmax</th>
<th>P(playing, are, dogs, the, …)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morfessor</td>
<td>‘want, s$</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>BPE</td>
<td>‘w, ants$</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>char-trigram</td>
<td>‘wa, wan, ant, nts ts$</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>character</td>
<td>‘w, a, n, t, s, $</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>analysis</td>
<td>want+VB, +3rd, +SG, +Pres</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
</tbody>
</table>

Table 3: Input representations for wants.

Figure 1: Our LSTM-LM architecture.

of character n-grams. We then calculate the max-over-time of each feature map.

$$y_j = \max_j f[j]$$  \hspace{1cm} (7)$$

and concatenate them to derive the word representation $w_t = [y_1, \ldots, y_m]$, where $m$ is the number of filters applied. Highway layers allow some dimensions of $w_t$ to be carried or transformed. Since it can learn character n-grams directly, we only use the CNN with character input.

3.3 Language Model

We use language models (LM) because they are simple and fundamental to many NLP applications. Given a sequence of text $s = w_1, \ldots, w_T$, our LM computes the probability of $s$ as:

$$P(w_1, \ldots, w_T) = \prod_{t=1}^{T} P(y_t|w_1, \ldots, w_{t-1})$$  \hspace{1cm} (8)$$

where $y_t = w_t$ if $w_t$ is in the output vocabulary and $y_t = \text{UNK}$ otherwise.

Our language model is an LSTM variant of recurrent neural network language (RNN) LM (Mikolov et al., 2010). At time step $t$, it receives input $w_t$ and predicts $y_{t+1}$. Using Eq. 1, it first computes representation $w_t$ of $w_t$. Given this representation and previous state $h_{t-1}$, it produces a new state $h_t$ and predicts $y_{t+1}$:

$$h_t = \text{LSTM}(w_t, h_{t-1})$$  \hspace{1cm} (9)$$

$$\hat{y}_{t+1} = g(V^T \cdot h_t)$$  \hspace{1cm} (10)$$

where $g$ is a softmax function over the vocabulary yielding the probability in Equation 8. Note that this design means that we can predict only words
from a finite output vocabulary, so our models differ only in their representation of context words. This design makes it possible to compare language models using perplexity, since they have the same event space, though open vocabulary word prediction is an interesting direction for future work.

The complete architecture of our system is shown in Figure 1, showing segmentation function \( \sigma \) and composition function \( f \) from Equation 1.

## 4 Experiments

We perform experiments on ten languages (Table 4). We use datasets from Ling et al. (2015) for English and Turkish. For Czech and Russian we use Universal Dependencies (UD) v1.3 (Nivre et al., 2015). For other languages, we use preprocessed Wikipedia data (Al-Rfou et al., 2013). For each dataset, we use approximately 1.2M tokens to train, and approximately 150K tokens each for development and testing. Preprocessing involves lowercasing (except for character models) and removing hyperlinks.

To ensure that we compared models and not implementations, we reimplemented all models in a single framework using Tensorflow (Abadi et al., 2015). For each dataset, we use approximately 1.2M tokens to train, and approximately 150K tokens each for development and testing. Preprocessing involves lowercasing (except for character models) and removing hyperlinks.

To make our results comparable to Ling et al. (2015), for each language we limit the output vocabulary to the most frequent 5,000 training words plus an unknown word token. To learn to predict unknown words, we follow Ling et al. (2015): in training, words that occur only once are stochastically replaced with the unknown token with probability 0.5. To evaluate the models, we compute perplexity on the test data.

<table>
<thead>
<tr>
<th>Typology</th>
<th>Languages</th>
<th>#tokens</th>
<th>#types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fusional</td>
<td>Czech</td>
<td>1.2M</td>
<td>125.4K</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>1.2M</td>
<td>81.1K</td>
</tr>
<tr>
<td></td>
<td>Russian</td>
<td>0.8M</td>
<td>103.5K</td>
</tr>
<tr>
<td>Agglutinative</td>
<td>Finnish</td>
<td>1.2M</td>
<td>188.4K</td>
</tr>
<tr>
<td></td>
<td>Japanese</td>
<td>1.2M</td>
<td>59.2K</td>
</tr>
<tr>
<td></td>
<td>Turkish</td>
<td>0.6M</td>
<td>126.2K</td>
</tr>
<tr>
<td>Root&amp;Pattern</td>
<td>Arabic</td>
<td>1.4M</td>
<td>137.5K</td>
</tr>
<tr>
<td></td>
<td>Hebrew</td>
<td>1.1M</td>
<td>104.9K</td>
</tr>
<tr>
<td>Reduplication</td>
<td>Indonesian</td>
<td>1.2M</td>
<td>76.5K</td>
</tr>
<tr>
<td></td>
<td>Malaysian</td>
<td>1.2M</td>
<td>77.7K</td>
</tr>
</tbody>
</table>

Table 4: Statistics of our datasets.

<table>
<thead>
<tr>
<th>Typology</th>
<th>Languages</th>
<th>#tokens</th>
<th>#types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fusional</td>
<td>Czech</td>
<td>1.2M</td>
<td>125.4K</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>1.2M</td>
<td>81.1K</td>
</tr>
<tr>
<td></td>
<td>Russian</td>
<td>0.8M</td>
<td>103.5K</td>
</tr>
<tr>
<td>Agglutinative</td>
<td>Finnish</td>
<td>1.2M</td>
<td>188.4K</td>
</tr>
<tr>
<td></td>
<td>Japanese</td>
<td>1.2M</td>
<td>59.2K</td>
</tr>
<tr>
<td></td>
<td>Turkish</td>
<td>0.6M</td>
<td>126.2K</td>
</tr>
<tr>
<td>Root&amp;Pattern</td>
<td>Arabic</td>
<td>1.4M</td>
<td>137.5K</td>
</tr>
<tr>
<td></td>
<td>Hebrew</td>
<td>1.1M</td>
<td>104.9K</td>
</tr>
<tr>
<td>Reduplication</td>
<td>Indonesian</td>
<td>1.2M</td>
<td>76.5K</td>
</tr>
<tr>
<td></td>
<td>Malaysian</td>
<td>1.2M</td>
<td>77.7K</td>
</tr>
</tbody>
</table>

from a finite output vocabulary, so our models differ only in their representation of context words. This design makes it possible to compare language models using perplexity, since they have the same event space, though open vocabulary word prediction is an interesting direction for future work.

The complete architecture of our system is shown in Figure 1, showing segmentation function \( \sigma \) and composition function \( f \) from Equation 1.

4 Experiments

We perform experiments on ten languages (Table 4). We use datasets from Ling et al. (2015) for English and Turkish. For Czech and Russian we use Universal Dependencies (UD) v1.3 (Nivre et al., 2015). For other languages, we use preprocessed Wikipedia data (Al-Rfou et al., 2013). For each dataset, we use approximately 1.2M tokens to train, and approximately 150K tokens each for development and testing. Preprocessing involves lowercasing (except for character models) and removing hyperlinks.

To ensure that we compared models and not implementations, we reimplemented all models in a single framework using Tensorflow (Abadi et al., 2015). For each dataset, we use approximately 1.2M tokens to train, and approximately 150K tokens each for development and testing. Preprocessing involves lowercasing (except for character models) and removing hyperlinks.

To make our results comparable to Ling et al. (2015), for each language we limit the output vocabulary to the most frequent 5,000 training words plus an unknown word token. To learn to predict unknown words, we follow Ling et al. (2015): in training, words that occur only once are stochastically replaced with the unknown token with probability 0.5. To evaluate the models, we compute perplexity on the test data.

5 Results and Analysis

Table 5 presents our main results. In six of ten languages, character-trigram representations composed with bi-LSTMs achieve the lowest perplexities. As far as we know, this particular model has not been tested before, though it is similar
<table>
<thead>
<tr>
<th>Language</th>
<th>word</th>
<th>character</th>
<th>char trigrams</th>
<th>BPE</th>
<th>Morfessor</th>
<th>%imp</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>bi-lstm</td>
<td>CNN add</td>
<td>bi-lstm add</td>
<td>bi-lstm add</td>
<td></td>
</tr>
<tr>
<td>Czech</td>
<td>41.46</td>
<td>34.25</td>
<td>36.60</td>
<td>42.73</td>
<td>33.59</td>
<td>49.96</td>
</tr>
<tr>
<td>English</td>
<td>46.40</td>
<td>43.53</td>
<td>44.67</td>
<td>45.41</td>
<td>42.97</td>
<td>47.51</td>
</tr>
<tr>
<td>Russian</td>
<td>34.93</td>
<td>28.44</td>
<td>29.47</td>
<td>35.15</td>
<td>27.72</td>
<td>40.10</td>
</tr>
<tr>
<td>Finnish</td>
<td>24.21</td>
<td>20.05</td>
<td>20.29</td>
<td>24.89</td>
<td>18.62</td>
<td>26.77</td>
</tr>
<tr>
<td>Japanese</td>
<td>98.14</td>
<td>98.14</td>
<td>91.63</td>
<td>101.99</td>
<td>101.09</td>
<td>126.53</td>
</tr>
<tr>
<td>Turkish</td>
<td>66.97</td>
<td>54.46</td>
<td>55.07</td>
<td>50.07</td>
<td>39.87</td>
<td>50.85</td>
</tr>
<tr>
<td>Arabic</td>
<td>48.20</td>
<td>42.02</td>
<td>43.17</td>
<td>50.85</td>
<td>30.40</td>
<td>44.15</td>
</tr>
<tr>
<td>Hebrew</td>
<td>38.23</td>
<td>31.63</td>
<td>33.19</td>
<td>39.67</td>
<td>59.49</td>
<td>59.49</td>
</tr>
<tr>
<td>Indonesian</td>
<td>46.07</td>
<td>45.47</td>
<td>46.60</td>
<td>58.51</td>
<td>43.37</td>
<td>59.33</td>
</tr>
<tr>
<td>Malay</td>
<td>54.67</td>
<td>53.01</td>
<td>50.56</td>
<td>68.51</td>
<td>50.74</td>
<td>68.20</td>
</tr>
</tbody>
</table>

Table 5: Language model perplexities on test. The best model for each language is highlighted in **bold** and the improvement of this model over the word-level model is shown in the final column.

to (but more general than) the model of Sperr et al. (2013). We can see that the performance of character, character trigrams, and BPE are very competitive. Composition by bi-LSTMs or CNN is more effective than addition, except for Turkish. We also observe that BPE always outperforms Morfessor, even for the agglutinative languages. We now turn to a more detailed analysis by morphological typology.

**Fusional languages.** For these languages, character trigrams composed with bi-LSTMs outperformed all other models, particularly for Czech and Russian (up to 20%), which is unsurprising since both are morphologically richer than English.

**Agglutinative languages.** We observe different results for each language. For Finnish, character trigrams composed with bi-LSTMs achieves the best perplexity. Surprisingly, for Turkish character trigrams composed via addition is best and addition also performs quite well for other representations, potentially useful since the addition function is simpler and faster than bi-LSTMs. We suspect that this is due to the fact that Turkish morphemes are reasonably short, hence well-approximated by character trigrams. For Japanese, we improvements from character models are more modest than in other languages.

**Root and Pattern.** For these languages, character trigrams composed with bi-LSTMs also achieve the best perplexity. We had wondered whether CNNs would be more effective for root-and-pattern morphology, but since these data are unvocalized, it is more likely that non-concatenative effects are minimized, though we do still find morphological variants with consonantal inflections that behave more like concatenation. For example, maktab (root:ktb) is written as mktb. We suspect this makes character trigrams quite effective since they match the tri-consonantal root patterns among words which share the same root.

**Reduplication.** For Indonesian, BPE morphs composed with bi-LSTMs model obtain the best perplexity. For Malay, the character CNN outperforms other models. However, these improvements are small compared to other languages. This likely reflects that Indonesian and Malay are only moderately inflected, where inflection involves both concatenative and non-concatenative processes.

### 5.1 Effects of Morphological Analysis

In the experiments above, we used unsupervised morphological segmentation as a proxy for morphological analysis (Table 3). However, as discussed in Section 2, this is quite approximate, so it is natural to wonder what would happen if we had the true morphological analysis. If character-level models are powerful enough to capture the effects of morphology, then they should have the predictive accuracy of a model with access to this analysis. To find out, we conducted an oracle experiment using the human-annotated morphological analyses provided in the UD datasets for Czech and Russian, the only languages in our set for which these analyses were available. In these experiments we treat the lemma and each morphological feature as a subword unit.

The results (Table 6) show that bi-LSTM composition of these representations outperforms all...
Table 6: Perplexity results using hand-annotated morphological analyses (cf. Table 5).

<table>
<thead>
<tr>
<th>Languages</th>
<th>Addition</th>
<th>bi-LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>51.8</td>
<td>30.07</td>
</tr>
<tr>
<td>Russian</td>
<td>41.82</td>
<td>26.44</td>
</tr>
</tbody>
</table>

Other models for both languages. These results demonstrate that neither character representations nor unsupervised segmentation is a perfect replacement for manual morphological analysis, at least in terms of predictive accuracy. In light of character-level results, they imply that current unsupervised morphological analyzers are poor substitutes for real morphological analysis.

However, we can obtain much more unannotated than annotated data, and we might guess that the character-level models would outperform those based on morphological analyses if trained on larger data. To test this, we ran experiments that varied the training data size on three representation models: word, character-trigram bi-LSTM, and character CNN. Since we want to see how much training data is needed to reach perplexity obtained using annotated data, we use the same output vocabulary derived from the original training. While this makes it possible to compare perplexities across models, it is unfavorable to the models trained on larger data, which may focus on other words. This is a limitation of our experimental setup, but does allow us to draw some tentative conclusions. As shown in Table 7, a character-level model trained on an order of magnitude more data still does not match the predictive accuracy of a model with access to morphological analysis.

5.2 Automatic Morphological Analysis

The oracle experiments show promising results if we have annotated data. But these annotations are expensive, so we also investigated the use of automatic morphological analysis. We obtained analyses for Arabic with the MADAMIRA (Pasha et al., 2014).4 As in the experiment using annotations, we treated each morphological feature as a subword unit. The resulting perplexities of 71.94 and 42.85 for addition and bi-LSTMs, respectively, are worse than those obtained with character trigrams (39.87), though it approaches the best perplexities.

Table 7: Perplexity results on the Czech development data, varying training data size. Perplexity using ~1M tokens annotated data is 28.83.

<table>
<thead>
<tr>
<th>#tokens</th>
<th>word</th>
<th>char trigram</th>
<th>char bi-LSTM</th>
<th>char CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1M</td>
<td>39.69</td>
<td>32.34</td>
<td>35.15</td>
<td></td>
</tr>
<tr>
<td>2M</td>
<td>37.59</td>
<td>36.44</td>
<td>35.58</td>
<td></td>
</tr>
<tr>
<td>3M</td>
<td>36.71</td>
<td>35.60</td>
<td>35.75</td>
<td></td>
</tr>
<tr>
<td>4M</td>
<td>35.89</td>
<td>32.68</td>
<td>35.93</td>
<td></td>
</tr>
<tr>
<td>5M</td>
<td>35.20</td>
<td>34.80</td>
<td>37.02</td>
<td></td>
</tr>
<tr>
<td>10M</td>
<td>35.60</td>
<td>35.82</td>
<td>39.09</td>
<td></td>
</tr>
</tbody>
</table>

5.3 Targeted Perplexity Results

A difficulty in interpreting the results of Table 5 with respect to specific morphological processes is that perplexity is measured for all words. But these processes do not apply to all words, so it may be that the effects of specific morphological processes are washed out. To get a clearer picture, we measured perplexity for only specific subsets of words in our test data: specifically, given target word $w_i$, we measure perplexity of word $w_{i+1}$. In other words, we analyze the perplexities when the inflected words of interest are in the most recent history, exploiting the recency bias of our LSTM-LM. This is the perplexity most likely to be strongly affected by different representations, since we do not vary representations of the predicted word itself.

We look at several cases: nouns and verbs in Czech and Russian, where word classes can be identified from annotations, and reduplication in Indonesian, which we can identify mostly automatically. For each analysis, we also distinguish between frequent cases, where the inflected word occurs more than ten times in the training data, and rare cases, where it occurs fewer than ten times. We compare only bi-LSTM models.

For Czech and Russian, we again use the UD annotation to identify words of interest. The results (Table 8), show that manual morphological analysis uniformly outperforms other subword models, with an especially strong effect for Czech nouns, suggesting that other models do not capture useful predictive properties of a morphological analysis. We do however note that character trigrams achieve low perplexities in most cases, similar to overall results (Table 5). We also observe that the subword models are more effective for rare words.

---

4We only experimented with Arabic since MADAMIRA disambiguates words in contexts; most other analyzers we found did not do this, and would require additional work to add disambiguation.
Table 8: Average perplexities of words that occur after nouns and verbs. Frequent words occur more than ten times in the training data; rare words occur fewer times than this. The best perplexity is in bold while the second best is underlined.

<table>
<thead>
<tr>
<th>Language</th>
<th>type-level (%)</th>
<th>token-level (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indonesian</td>
<td>1.10</td>
<td>2.60</td>
</tr>
<tr>
<td>Malay</td>
<td>1.29</td>
<td>2.89</td>
</tr>
</tbody>
</table>

Table 9: Percentage of full reduplication on the type and token level.

<table>
<thead>
<tr>
<th>Language</th>
<th>type-level (%)</th>
<th>token-level (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>word</td>
<td>91.71</td>
<td>156.98</td>
</tr>
<tr>
<td>characters</td>
<td>91.35</td>
<td>137.42</td>
</tr>
</tbody>
</table>

Table 10: Average perplexities of words that occur after reduplicated words in the test set.

Table 11 presents nearest neighbors under cosine similarity for in-vocabulary, rare, and out-of-vocabulary (OOV) words. For frequent words, standard word embeddings are clearly superior for lexical meaning. Character and morph representations tend to find words that are orthographically similar, suggesting that they are better at modeling dependent than root morphemes. The same pattern holds for rare and OOV words. We suspect that the subword models outperform words on language modeling because they exploit affixes to signal word class. We also noticed similar patterns in Japanese.

We analyze reduplication by querying reduplicated words to find their nearest neighbors using the BPE bi-LSTM model. If the model were sensitive to reduplication, we would expect to see morphological variants of the query word among its nearest neighbors. However, from Table 12, this is not so. With the partially reduplicated query berlembah-lembah, we do not find the lemma lembah.

6 Conclusion

We presented a systematic comparison of word representation models with different levels of morphological awareness, across languages with different morphological typologies. Our results confirm previous findings that character-level models are effective for many languages, but these models do not match the predictive accuracy of models with explicit knowledge of morphology, even after we increase the training data size by ten times. Moreover, our qualitative analysis suggests that they learn orthographic similarity of affixes, and lose the meaning of root morphemes.

Although morphological analyses are available...

---

5 We use Indonesian affixes listed in Larasati et al. (2011)

6 https://radimrehurek.com/gensim/
in limited quantities, our results suggest that there might be utility in semi-supervised learning from partially annotated data. Across languages with different typologies, our experiments show that the subword unit models are most effective on agglutinative languages. However, these results do not generalize to all languages, since factors such as morphology and orthography affect the utility of these representations. We plan to explore these effects in future work.

### Acknowledgments

Clara Vania is supported by the Indonesian Endowment Fund for Education (LPDP), the Centre for Doctoral Training in Data Science, funded by the UK EPSRC (grant EP/L016427/1), and the University of Edinburgh. We thank Sameer Bansal, Toms Bergmanis, Marco Damonte, Federico Fancellu, Sorcha Gilroy, Sharon Goldwater, Frank Keller, Mirella Lapata, Felicia Liu, Jonathan Mallinson, Joana Ribeiro, Naomi Saphra, Ida Szubert, and the anonymous reviewers for helpful discussion of this work and comments on previous drafts of the paper.

### References


