On the Importance of Subword Information for Morphological Tasks in Truly Low-Resource Languages

Yi Zhu[1*] **Benjamin Heinzerling**[2,3*]**, Ivan Vulić**[1]**,**
Michael Strube[4]**, Roi Reichart**[5]**, Anna Korhonen**[1]
[1]Language Technology Lab, University of Cambridge
[2]RIKEN AIP, [3]Tohoku University
[4]Heidelberg Institute for Theoretical Studies, [5]Technion, IIT
`{yz568,iv250,alk23}@cam.ac.uk`, `benjamin.heinzerling@riken.jp`
`michael.strube@h-its.org`, `roiri@technion.ac.il`

Abstract

Recent work has validated the importance of subword information for word representation learning. Since subwords increase parameter sharing ability in neural models, their value should be even more pronounced in low-data regimes. In this work, we therefore provide a comprehensive analysis focused on the usefulness of subwords for word representation learning in truly low-resource scenarios and for three representative morphological tasks: fine-grained entity typing, morphological tagging, and named entity recognition. We conduct a systematic study that spans several dimensions of comparison: **1)** *type of data scarcity* which can stem from the lack of task-specific training data, or even from the lack of unannotated data required to train word embeddings, or both; **2)** *language type* by working with a sample of 16 typologically diverse languages including some truly low-resource ones (e.g. Rusyn, Buryat, and Zulu); **3)** the choice of the *subword-informed word representation method*. Our main results show that subword-informed models are universally useful across all language types, with large gains over subword-agnostic embeddings. They also suggest that the effective use of subwords largely depends on the language (type) and the task at hand, as well as on the amount of available data for training the embeddings and task-based models, where having sufficient in-task data is a more critical requirement.

1 Introduction and Motivation

Recent studies have confirmed the usefulness of leveraging subword-level information in learning word representations (Peters et al., 2018; Heinzerling and Strube, 2018; Grave et al., 2018; Zhu et al., 2019, *inter alia*), and in a range of tasks such as sequence tagging (Lample et al., 2016; Akbik

et al., 2018; Devlin et al., 2019), fine-grained entity typing (Zhu et al., 2019), neural machine translation (Sennrich et al., 2016; Luong and Manning, 2016; Lample et al., 2018; Durrani et al., 2019), or general and rare word similarity (Pilehvar et al., 2018; Zhu et al., 2019). The subword-informed word representation architectures leverage the internal structure of words and assume that a word's meaning can be inferred from the meaning of its constituent (i.e., subword) parts. Instead of treating each word as an atomic unit, subword-informed neural architectures reduce data sparsity by relying on parameterization at the level of subwords (Bojanowski et al., 2017; Pinter et al., 2017; Chaudhary et al., 2018; Kudo, 2018).

An increasing body of work focuses on various aspects of subword-informed representation learning such as segmentation of words into subwords and composing subword embeddings into word representations (Lazaridou et al., 2013; Cotterell and Schütze, 2015, 2018; Avraham and Goldberg, 2017; Vania and Lopez, 2017; Kim et al., 2018; Zhang et al., 2018; Zhao et al., 2018, *inter alia*).[1] The increased parameter sharing ability of such models is especially relevant for learning embeddings of rare and unseen words. Therefore, the importance of subword-level knowledge should be even more pronounced in *low-data regimes* for *truly low-resource languages*. Yet, a systematic study focusing exactly on the usefulness of subword information in such settings is currently missing in the literature. In this work, we fill this gap by providing a comprehensive analysis of subword-informed representation learning focused on low-resource setups.

Our study centers on the following axes of comparison, focusing on three representative tasks where subword-level information can guide learn-

---

[*]Equal contribution, work partly done while at HITS.

[1]An overview of a variety of subword-informed word representation architectures and different segmentation and composition strategies is provided by Zhu et al. (2019).

ing, namely fine-grained entity typing (FGET), morphological tagging (MTAG), and named entity recognition (NER): **1)** Since data scarcity can stem from unavailability of (i) task-specific training data or (ii) unannotated corpora to train the embeddings in the first place, or (iii) both, we analyse how different data regimes affect the final task performance. **2)** We experiment with 16 languages representing 4 diverse morphological types, with a focus on truly low-resource languages such as Zulu, Rusyn, Buryat, or Bambara. **3)** We experiment with a variety of subword-informed representation architectures, where the focus is on unsupervised, widely portable language-agnostic methods such as the ones based on character n-grams (Luong and Manning, 2016; Bojanowski et al., 2017), Byte Pair Encodings (BPE) (Sennrich et al., 2016; Heinzerling and Strube, 2018), Morfessor (Smit et al., 2014), or BERT-style pretraining and fine-tuning (Devlin et al., 2019) which relies on WordPieces (Wu et al., 2016). We demonstrate that by tuning subword-informed models in low-resource settings we can obtain substantial gains over subword-agnostic models such as skip-gram with negative sampling (Mikolov et al., 2013) across the board.

The main goal of this study is to identify viable and effective subword-informed approaches for truly low-resource languages and offer modeling guidance in relation to the target task, the language at hand, and the (un)availability of general and/or task-specific training data. As expected, our key results indicate that there is no straightforward *"one-size-fits-all"* solution, although certain approaches (e.g., BPE-based or character n-grams) emerge as more robust in general. The optimal subword-informed configurations are largely task-, language-, and resource-dependent: their performance hinges on a complex interplay of the multiple factors mentioned above. For instance, we show that fine-tuning pretrained multilingual BERT (Devlin et al., 2019; Wu and Dredze, 2019) is a viable strategy for "double" low-resource settings in the NER and MTAG tasks, but it fails for the FGET task in the same setting; furthermore, its performance can be matched or surpassed by other subword-informed methods in NER and MTAG as soon as they obtain sufficient embedding training data.

2 Methodology

In what follows, we further motivate our work by analyzing two different sources of data scarcity:

embedding training data (termed *WE data*) and task-specific training data (termed *task data*). Following that, we motivate our selection of test languages and outline the subword-informed representation methods compared in our evaluation.

Types of Data Scarcity. The majority of languages in the world still lack basic language technology, and progress in natural language processing is largely hindered by the lack of annotated *task data* that can guide machine learning models (Agić et al., 2016; Ponti et al., 2018). However, many languages face another challenge: the lack of *large unannotated text corpora* that can be used to induce useful general features such as word embeddings (Adams et al., 2017; Fang and Cohn, 2017; Ponti et al., 2018):[2] i.e. WE data.

The absence of data has over the recent years materialized the *proxy fallacy*. That is, methods tailored for low-resource languages are typically tested only by proxy, simulating low-data regimes exclusively on resource-rich languages (Agić et al., 2017). While this type of evaluation is useful for analyzing the main properties of the intended low-resource methods in controlled *in vitro* conditions, a complete evaluation should also provide results on true low-resource languages *in vivo*. In this paper we therefore conduct both types of evaluation. Note that in this work we still focus on low-resource languages that have at least some digital footprint (see the statistics later in Table 1), while handling zero-resource languages without any available data (Kornai, 2013; Ponti et al., 2018) is a challenge left for future work.

(Selection of) Languages. Both *sources of data scarcity* potentially manifest in degraded task performance for low-resource languages: our goal is to analyze the extent to which these factors affect downstream tasks across morphologically diverse language types that naturally come with varying data sizes to train their respective embeddings and task-based models. Our selection of test languages is therefore guided by the following goals: **a)** following recent initiatives (e.g. in language modeling) (Cotterell et al., 2018; Gerz et al., 2018), we aim to ensure coverage of different genealogical and typological properties; **b)** we aim to cover low-resource languages with varying amounts of available WE data and task-specific data.

We select 16 languages in total spanning 4 broad

---

[2]For instance, as of April 2019, Wikipedia is available only in 304 out of the estimated 7,000 existing languages.

	Agglutinative						Fusional						Introflexive		Isolating	
	BM	BXR	MYV	TE	TR	ZU	EN	FO	GA	GOT	MT	RUE	AM	HE	YO	ZH
EMB	40K	372K	207K	5M	5M	69K	5M	1.6M	4.4M	18K	1.5M	282K	659K	5M	542K	5M
FGET	29K	760	740	13K	60K	36K	60K	30K	56K	289	2.7K	1.5K	2.2K	60K	15K	60K
NER	345	2.4K	2.1K	9.9K	167K	425	8.9M	4.0K	7.6K	475	1.9K	1.6K	1.0K	107K	3.4K	–
MTAG	–	–	–	1.1K	3.7K	–	24K	–	–	3.4K	1.1K	–	–	5.2K	–	4.0K
BERT				✓	✓		✓		✓					✓	✓	✓

Table 1: Overview of test languages and data availability. EMB denotes the maximum number of tokens in corresponding Wikipedias used for training embeddings. Actual Wikipedia sizes are larger than 5M for (TE, TR, EN, HE, ZH), but were limited to 5M tokens in order to ensure comparable evaluation settings for data scarcity simulation experiments across different languages. FGET, NER, and MTAG rows show the number of instances for the three evaluation tasks (see §3): number of entity mentions for FGET, number of sentences for NER and MTAG. In MTAG, we omit languages for which UDv2.3 provides only a test set, but no training set. The BERT row shows the languages supported by multilingual BERT. Languages are identified by their ISO 639-1 code.

morphological types, listed in Table 1. Among these, we chose one (relatively) high-resource language for each type: Turkish (agglutinative), English (fusional), Hebrew (introflexive), and Chinese (isolating). We use these four languages to *simulate* data scarcity scenarios and run experiments where we *control* the degree of data scarcity related to both embedding training data and task-related data. The remaining 12 languages are treated as test languages with varying amounts of available data (see Table 1. For instance, relying on the Wikipedia data for embedding training, Gothic (GOT) is the language from our set that contains the fewest number of word tokens in its respective Wikipedia (18K, in terms of Wikipedia size this ranks it as 273th out of 304 Wikipedia languages); Irish Gaelic (GA) with 4.4M tokens is ranked 87/304.

Subword-Informed Word Representations. We mainly follow the framework of Zhu et al. (2019) for the construction of subword-informed word representations; the reader is encouraged to refer to the original paper for more details. In short, to compute the representation for a given word $w \in V$, where V is the word vocabulary, the framework is based on three main components: 1) *segmentation* of words into subwords, 2) *interaction* between subword and position embeddings, and 3) a *composition function* that yields the final word embedding from the constituent subwords. Zhu et al. (2019) explored a large space of possible subword-informed configurations. Based on their findings, we select a representative subset of model configurations. They can be obtained by varying the components listed in Table 2.

Concretely, w is first segmented into an ordered subword sequence from the subword vocabulary S by a deterministic subword segmentation method.

To enable automatic language-agnostic segmentation across multiple languages, we focus on unsupervised segmentation methods: we work with Morfessor (Smit et al., 2014), character n-grams (Bojanowski et al., 2017) and BPE (Gage, 1994). We use the default parameters for Morfessor, and the same 3 to 6 character n-gram range as Bojanowski et al. (2017). For BPE, the number of merge operations is a tunable hyper-parameter. It controls the segmentation "aggressiveness": the larger the number the more conservative the BPE segmentation is. Following Heinzerling and Strube (2018), we investigate the values $\{1e3, 1e4, 1e5\}$: this allows us to test varying segmentation granularity in relation to different language types.

After segmentation into subwords, each subword is represented by a vector $\mathbf{s}$ from the subword embedding matrix $\mathbf{S} \in \mathbb{R}^{|S| \times d}$, where d is the dimensionality of subword embeddings. Optionally, the word itself can be appended to the subword sequence and embedded into the subword space in order to incorporate word-level information (Bojanowski et al., 2017). To encode subword order, $\mathbf{s}$ can be further enriched by a trainable position embedding $\mathbf{p}$. We use addition to combine subword and position embeddings, namely $\mathbf{s} := \mathbf{s} + \mathbf{p}$, which has become the de-facto standard method to encode positional information (Gehring et al., 2017; Vaswani et al., 2017; Devlin et al., 2019).

Finally, the subword embedding sequence is passed to a composition function, which computes the final word representation. Li et al. (2018) and Zhu et al. (2019) have empirically verified that composition by simple addition, among other more complex composition functions, is a robust choice. Therefore, we use addition in all our experiments.

Similar to Bojanowski et al. (2017); Zhu et al. (2019), we adopt skip-gram with negative sampling

Component	Option	Label
Segmentation	Morfessor	`morf`
	BPE	`bpeX`
	char n-gram	`charn`
Word token	exclusion	*w-*
	inclusion	*w+*
Position embedding	exclusion	*p-*
	additive	*p+*
Composition function	addition	*add*

Table 2: Components for constructing subword-informed word representations. In the `bpeX` label $X \in \{1e3, 1e4, 1e5\}$ denotes the BPE vocabulary size.

(Mikolov et al., 2013) as our word-level distributional model: the target word embedding is computed by our subword-informed model, and the context word is parameterized by the (word-level) context embedding matrix $\mathbf{W}_c \in \mathbb{R}^{|V| \times d_c}$.

We compare subword-informed architectures to two well-known word representation models, also captured by the general framework of Zhu et al. (2019): 1) the subword-agnostic skip-gram model from the `word2vec` package (Mikolov et al., 2013) (W2V), and 2) `fastText` (FT) (Bojanowski et al., 2017). The comparison to W2V aims to validate the potential benefit of subword-informed word representations for truly low-resource languages, while the comparison to FT measures the gains that can be achieved by more sophisticated and fine-tuned subword-informed architectures. We also compare with pretrained multilingual $BERT_{base}$ (Devlin et al., 2019) on the languages supported by this model.

3 Evaluation Tasks

Fine-Grained Entity Typing. FGET is cast as a sequence classification problem, where an entity mention consisting of one or more tokens (e.g. *Lincolnshire*, *Bill Clinton*), is mapped to one of the 112 fine-grained entity types from the FIGER inventory (Ling and Weld, 2012; Yaghoobzadeh and Schütze, 2015; Heinzerling and Strube, 2018). Since entity mentions are short token sequences and not full sentences, this semi-morphological/semantic task requires a model to rely on the subword information of individual tokens in the absence of sentence context. That is, subwords can provide evidence useful for entity type classification in the absence of context. For instance, *Lincolnshire* is assigned the type `/location/county` as -*shire* is a suffix that strongly indicates a location. Hence, FGET is well-suited for evaluating subword-informed rep-

resentations, and can benefit from the information.

Morphological Tagging. MTAG is the task of annotating each word in a sentence with features such as part-of-speech, gender, number, tense, and case. These features are represented as a set of key-value pairs. For example, *classified* is a finite (Fin) verb (V) in indicative (Ind) mood, third person, past tense, which is annotated with the morphological tag {POS=V, Mood=Ind, Person=3, Tense=Past, Verb-Form=Fin}, and the female singular third-person possessive personal pronoun *her* with the morphological tag {Gender=Fem, Number=Sing, Person=3, Poss=Yes, PronType=Prs}.

Named Entity Recognition. NER is the task of annotating textual mentions of real-world entities with their semantic type, such as person, location, and organization: e.g., *Barack Obama (person) was born in Hawaii (location)*.

4 Experimental Setup

Embedding Training: WE Data. For training word and subword embeddings, we rely on Wikipedia text for all 16 languages, with corresponding Wikipedia sizes listed in Table 1. For training embeddings in controlled low-resource settings with our 4 "simulation" languages, we sample nine data points to simulate low-resource scenarios with WE data. Specifically, we sample 10K, 20K, 50K, 100K, 200K, 500K, 1M, 2M, and 5M tokens of article text for each of the 4 languages. For the other 12 languages we report results obtained by training embedding on the full Wikipedia edition.

Task-Specific Data: Task Data. The maximum number of training instances for all languages is again provided in Table 1. As before, for 4 languages we simulate low-resource settings by taking only a sample of the available task data: for FGET we work with 200, 2K or 20K training instances which roughly correspond to training regimes of different data availability, while we select 300,[3] 1K, and 10K sentences for NER and MGET. Again, for the remaining 12 languages, we use all the available data to run the experiments. We adopt existing data splits into training, development, and test portions for MTAG (Cotterell and Heigold, 2017), and random splits for FGET (Heinzerling and Strube, 2018; Zhu et al., 2019) and NER (Pan et al., 2017).

---

[3] With a smaller number of instances (e.g., 100), NER and MGET model training was unstable and resulted in near-zero performance across multiple runs.

A large number of data points for scarcity simulations allow us to trace how performance on the three tasks varies in relation to the availability of WE data versus task data, and what data source is more important for the final performance.

Embedding Training Setup. When training our subword-informed representations, we argue that keeping hyper-parameters fixed across different data points will possibly result in underfitting for larger data sizes or overfitting for smaller data sizes. Therefore, we split data points into three groups: $[10K, 50K]$ (G1), $(50K, 500K]$ (G2) and $(500K, 5M]$ (G3), and use the same hyper-parameters for word embedding training within the same group. For G1, we train with batch size 32 for 60 epochs and set the minimum word frequency threshold to 2. For G2 the values are: $128/30/3$, and $512/15/5$ for G3. This way, we ensure that 1) the difference of the absolute data sizes can be compared within the same data group, and 2) for the corresponding data points in different groups ($10K$, $100K$, $1M$) the sample efficiency can be compared, as the models trained on these data points undergo roughly the same number of updates.[4]

Task-Specific Training Setup. For FGET, we use the dataset of Heinzerling and Strube (2018) obtained by mapping entity mentions from Wikidata (Vrandečić and Krötzsch, 2014) to their associated FIGER-based most notable entity type (Ling and Weld, 2012). For each language, we randomly sample up to 100k pairs of entity mentions with corresponding entity type and create random 60/20/20 train/dev/test splits. Our FGET model is designed after the hierarchical architecture by Zhu et al. (2019). For each entity token, we first use our subword-informed model to obtain word representations, and then feed the token embedding sequence into a bidirectional LSTM with 2 hidden layers of size 512, followed by a projection layer which predicts the entity type. We initialize our FGET model with the pretrained subword model, and fine-tune it during training. With BERT, we input the entire entity mention and then use the representation of the special [CLS] token for classification. We train with early stopping, using Adam (Kingma and Ba, 2015) with default parameters across all languages. As suggested by Wu and Dredze (2019), BERT hyper-parameters are more

---

[4]We train fastText and skip-gram from word2vec with the same number of epochs that is used to train our subword-informed models on the corresponding data points.

sensitive to smaller data sizes, so we tune them on the smallest data point with 200 training instances. We follow Wu and Dredze (2019) to select hyper-parameter candidates, i.e., $2e-5/3e-5/5e-5$ for learning rate, $16/32$ for batch size and triangular learning rate scheduler with first 10% of batches as warm-up. We do an exhaustive search on four high resource languages: EN, TR, HE, ZH and select the hyper-parameter combination with the best average score on the development sets.

For MTAG, we evaluate on the multilingual morphological annotations provided by the Universal Dependencies project (Nivre et al., 2016) and adopt the experimental protocol of Cotterell and Heigold (2017). Specifically, we cast MTAG as a sequence labeling task by treating the concatenation of all key-value pairs for a given word as the word's label. As sequence labeling model, we train a bidirectional LSTM (Hochreiter and Schmidhuber, 1997; Plank et al., 2016), with two layers of size 1024 and dropout 0.4, using early stopping on the development set. For experiments involving multilingual BERT, we fine-tune all of BERT's layers and feed the final layer into an LSTM before classification. The evaluation metric is per-label accuracy, i.e., a word's morphological tag is either predicted correctly or not, and there is no partial credit for the correct prediction of only a subset of features.

We evaluate NER performance on WikiAnn (Pan et al., 2017), a multilingual dataset which provides three-class entity type annotations which were automatically extracted from Wikipedia. We train sequence labeling models using exactly the same architectures and hyper-parameters as in MTAG, and report F1 scores. As WikiAnn does not come with predefined train/dev/test sets, we create random 60/20/20 splits.

5 Results and Discussion

Results for data scarcity simulation experiments are summarized in Figures 1-3, while the results on the remaining 12 languages for all three tasks are provided in Tables 3-4, with the best results among different configurations of subword-informed methods reported. As the first main finding, the results show that subword-informed architectures substantially outperform the subword-agnostic skip-gram W2V baseline, and the gaps are in some cases very large: e.g., see the results in Figures 1-3 for the settings with extremely scarce WE data. These scores verify the importance of subword-level knowledge
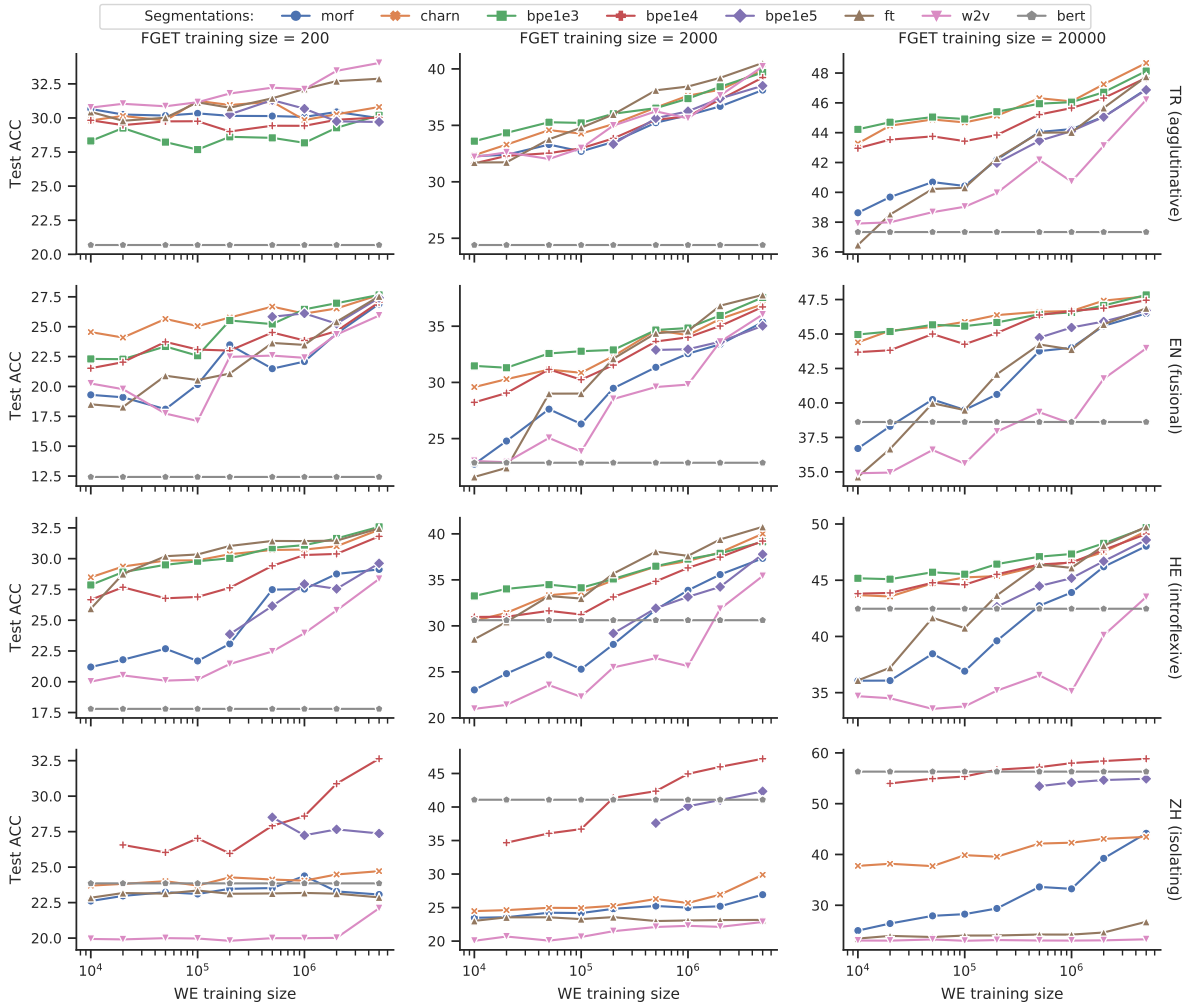
Figure 1: Test performance (accuracy) in the FGET task for different segmentation methods in data scarcity simulation experiments across 4 languages representing 4 broad morphological types, averaged over 5 runs. Some data points with Chinese (ZH) are not shown as in those cases the subword model is reduced to single characters only.

for low-resource setups.

Simulating Data Scarcity. Another general finding concerns the importance of WE data versus task data. The simulation results in Figure 1-3 suggest that both types of data are instrumental to improved task performance: This finding is universal as we observe similar behaviors across tasks and across different languages. While WE data is important, considerably larger gains are achieved by collecting more task data: e.g., see the large gains in FGET when training on 200 versus 2K entity mentions. In summary, both types of data scarcity decrease performance, but the impact of scarce task data seems more pronounced. Collecting more WE data when dealing with scarce task data leads to larger gains in the FGET task compared to MTAG or NER.

While subword models are generally better than the baselines across different data points, less aggressive segmentation models and token-based

models close the gap very quickly when increasing WE data, which is in line with the findings of Zhu et al. (2019), where morf eventually prevails in this task with abundant WE data. This again verifies the usefulness of subword-level knowledge for low-(WE) data regimes. Similar trends emerge in terms of task data, but the advantage of subword models seems more salient with more task data. The underlying task architectures start making use of subword features more effectively: this shows that subword-level knowledge is particularly useful for the three chosen morphological tasks.

Performance of BERT. An interesting analysis regarding the (relative) performance of pretrained multilingual BERT model emerges from Figures 1-3. Fine-tuned BERT displays much stronger performance in low-resources settings for the MTAG and NER tasks than for the FGET task (e.g., compare the sub-figures in the first columns of the
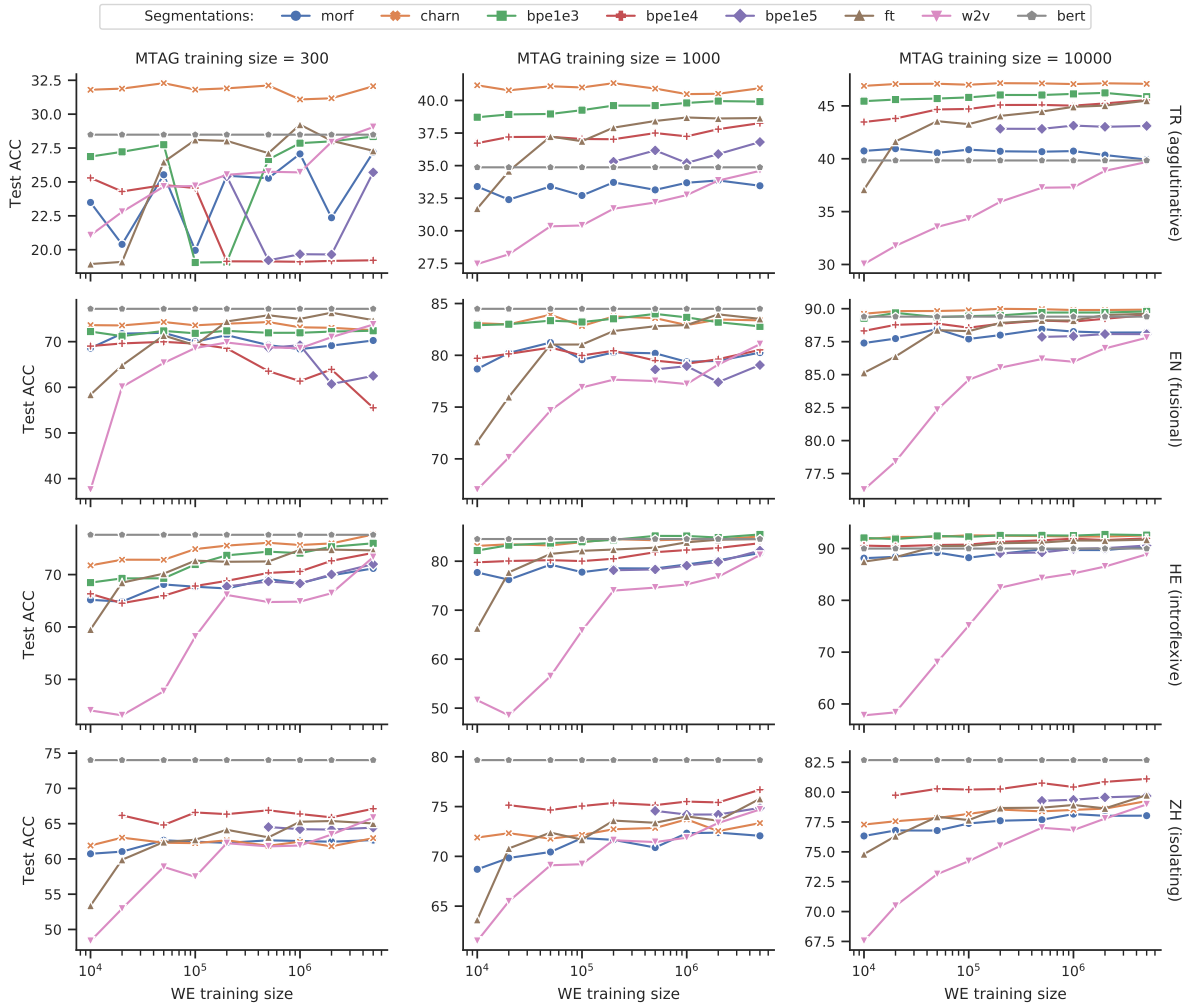
Figure 2: Test performance in the MTAG task in data scarcity simulation experiments.

corresponding figures). The explanation of MTAG and NER performance is intuitive. A pretrained BERT model encodes a massive amount of background knowledge available during its (multilingual) pretraining. However, as we supplement other subword-informed representation learning methods with more data for training the embeddings, the gap gets smaller until it almost completely vanishes: other methods now also get to see some of the distributional information which BERT already consumed in its pretraining.

BERT performance on FGET versus MTAG and NER concerns the very nature of the tasks at hand. The input data for FGET consist mostly of 2-4 word tokens (i.e., entity mentions), while MTAG and NER operate on full sentences as input. Since BERT has been pretrained on sentences, this setting is a natural fit and makes fine-tuning to these tasks easier: BERT already provides a sort of "contextual subword composition function". This stands in contrast with the other subword-informed approaches.

There, we might have good non-contextual subword embeddings and a pretrained "non-contextual" composition function, but we have to learn how to effectively leverage the context for the task at hand (i.e., by running an LSTM over the subword-informed token representations) from scratch.[5]

Truly Low-Resource Languages. The results on the 12 test languages in Table 3-4 suggest that subword-informed models are better than the baselines in most cases: this validates the initial findings from the simulation experiments. That is, leveraging subword information is important for WE induction as well as for task-specific training. The

---

[5]Another factor at play is multilingual BERT's limited vocabulary size (100K WordPiece symbols), leaving on average a bit under 1K symbols per language. Due to the different sizes of Wikipedias used for pretraining BERT, some languages might even be represented with far fewer than 1K vocabulary entries, thereby limiting the effective language-specific model capacity. Therefore, it is not that surprising that monolingual subword-informed representations gradually surpass BERT as more language-specific WE data becomes available. This finding is also supported by the results reported in Table 3.
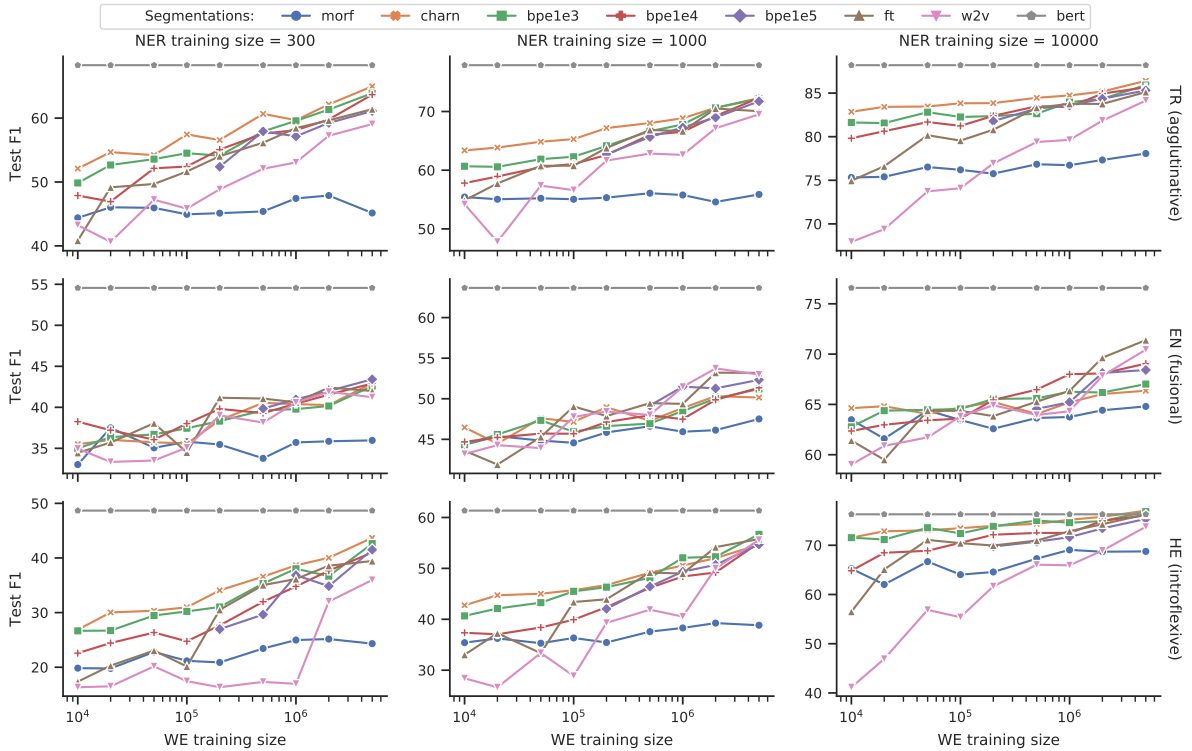
Figure 3: Test performance (F1 score) in the NER task in data scarcity simulation experiments. We do not show results for Chinese as the annotations of the Chinese NER are provided only on the character-level and thus impede experimentation with most of the subword-informed methods used in our evaluation.

gains with subword methods become larger for languages with fewer WE data (e.g., ZU, BM, GOT); this is again consistent with the previously reported simulation experiments.

Tasks, Language Types, Subword Configurations. The results further suggest that the optimal configuration indeed varies across different tasks and language types, and therefore it is required to carefully tune the configuration to reach improved performance. For instance, as agglutinative languages have different granularities of morphological complexity, it is not even possible to isolate a single optimal segmentation method within this language type. Overall, the segmentation based charn followed by BPE emerge as most robust choices across all languages and tasks. However, charn has the largest number of parameters and is slower to train compared to other segmentations, and in case of BPE its number of merge operations must be tuned to yield competitive scores.

While we do not see extremely clear patterns from the results in relation to particular language types, the scores suggest that for agglutinative and fusional languages a hybrid segmentation such as charn or a moderate one (bpe1e4, bpe1e5) is a good choice. For introflexive and isolating

languages, more aggressive segmentations seem to be also competitive in FGET and MTAG, while bpe1e4 being very effective for ZH, and charn (again) and bpe1e5 seems to be preferred in NER.

Apart from segmentation methods, we also analyzed the effect of word token embeddings (*w*+) and position embeddings (*p*+) in the subword-informed learning framework (Zhu et al., 2019) (see before Table 2 in §2), shown in Figure 4. NER can clearly benefit from both *w*+ and *p*+ and *w*+ is also useful for MTAG. However, for other tasks, the fluctuations between configurations are minimal once the segmentation has been fixed, which suggests that the most critical component is indeed the chosen segmentation method: this is why we have mostly focused on the analyses of the segmentation method and its impact on task performance in this work. Regarding the composition functions, as demonstrated in Zhu et al. (2019), more complex composition functions do not necessarily yield superior results in a range of downstream tasks. We therefore leave the exploration of more sophisticated composition functions for future work.

6   Conclusions and Future Work

We have presented an empirical study focused on the importance of subword-informed word repre-

	Agglutinative					Fusional					Intro	Isolat
	BM	BXR	MYV	TE	ZU	FO	GA	GOT	MT	RUE	AM	YO
FGET												
morf	52.43	52.47	79.11	57.79	53.00	54.43	50.77	29.90	49.48	50.38	41.82	83.43
charn	**56.09**	57.33	81.69	58.83	**56.34**	**58.44**	**52.62**	**34.02**	**54.46**	**58.59**	45.65	84.85
bpe1e3	53.61	51.30	81.13	58.73	55.41	56.04	50.74	31.55	52.79	55.57	**47.99**	**85.22**
bpe1e4	54.20	53.81	**81.93**	**59.24**	55.67	56.67	51.47	26.39	52.15	54.81	47.05	84.42
bpe1e5	-	53.80	80.00	58.13	-	56.31	51.52	-	51.52	52.52	44.74	83.39
ft	51.91	**57.96**	81.05	57.79	52.62	53.74	49.67	31.96	53.95	53.64	44.80	83.71
w2v	52.28	42.19	76.86	56.99	52.95	53.07	49.07	24.53	46.61	47.36	36.81	82.56
bert	-	-	-	49.20	-	-	47.09	-	-	-	-	81.76
NER												
morf	73.29	76.58	83.40	77.01	65.22	84.29	86.94	59.49	74.37	81.87	66.67	90.01
charn	**83.02**	**81.59**	**93.22**	**88.23**	**74.47**	**91.08**	88.95	**84.99**	**83.56**	**88.70**	**72.92**	94.68
bpe1e3	77.22	79.33	89.00	85.82	71.91	89.73	89.18	81.03	81.63	85.30	70.84	92.35
bpe1e4	76.43	79.73	89.00	85.44	65.22	89.25	88.48	70.59	80.26	86.39	64.07	92.47
bpe1e5	-	80.65	89.36	84.02	-	88.66	**89.48**	-	81.64	86.12	68.95	93.07
ft	73.29	79.81	88.57	86.88	58.16	89.48	89.18	58.16	81.64	83.54	68.29	92.58
w2v	69.57	79.66	87.50	82.97	62.37	87.81	87.99	58.56	79.43	84.21	61.37	89.57
bert	-	-	-	82.31	-	-	88.45	-	-	-	-	**95.53**

Table 3: Test accuracy for FGET and test F1 score NER for the 12 low-resource test languages. The results are obtained by training on the full WE data (except for BERT) and the full task data of the corresponding languages.
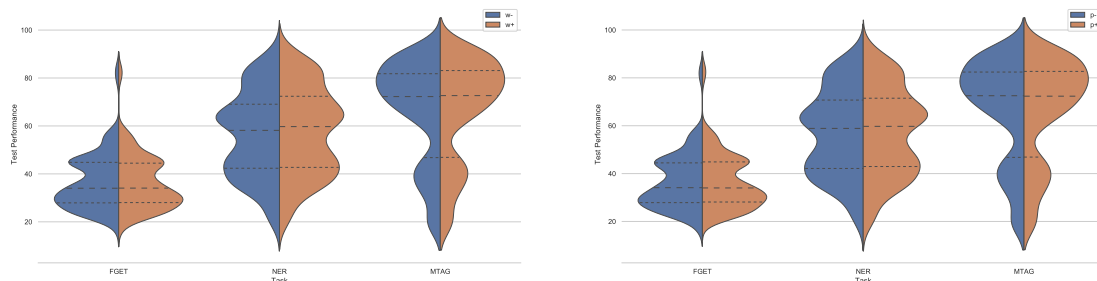

Figure 4: Comparisons of configurations with and without word token (*w-, w+*) and the position embedding (*p-, p+*). The results are obtained by collecting all data points in the data scarcity simulation for four high resource languages and the other 12 languages with both full WE data and task data.

	Agg	Fus	Int
	TE	GOT	MT
morf	87.79	76.94	92.80
charn	90.29	**85.71**	**94.39**
bpe1e3	90.01	82.07	94.16
bpe1e4	87.79	83.28	92.82
bpe1e5	87.10	-	92.51
ft	**90.85**	76.50	93.91
w2v	85.71	29.63	90.43
bert	87.45	-	-

Table 4: Test accuracy for MTAG for low-resource languages from UD where train/dev/test sets are available.

sentation architectures for truly low-resource languages. Our experiments on three diverse morphological tasks with 16 typologically diverse languages of varying degrees of data scarcity have validated that subword-level knowledge is indeed crucial for improved task performance in such low-data setups. The large amount of results reported in this work has enabled comparisons of different subword-informed methods in relation to multiple aspects such as the degree of data scarcity (both in terms of embedding training data and task-specific annotated data), the task at hand, the actual language, as well as the methods' internal design (e.g. the choice of the segmentation method). Our results have demonstrated that all these aspects must be considered in order to identify an optimal subword-informed representation architecture for a particular use case, that is, for a particular language (type), task, and data availability. However, similar paterns emerge: e.g., resorting to a segmentation method based on character n-grams seems most robust across the three tasks and across languages, although there are clear outliers. In future work, we will extend our focus to other target languages, including the ones with very limited (Adams et al., 2017) or non-existent digital footprint.

Acknowledgments

References

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of EACL*, pages 937–947.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the ACL*, 4:301–312.

Željko Agić, Barbara Plank, and Anders Søgaard. 2017. Cross-lingual tagger evaluation without test data. In *Proceedings of EACL*, pages 248–253.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of COLING*, pages 1638–1649.

Oded Avraham and Yoav Goldberg. 2017. The interplay of semantics and morphology in word embeddings. In *Proceedings of EACL*, pages 422–426.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R. Mortensen, and Jaime Carbonell. 2018. Adapting word embeddings to new languages with morphological and phonological subword representations. In *Proceedings of EMNLP*, pages 3285–3295.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, Copenhagen, Denmark. Association for Computational Linguistics.

Ryan Cotterell, Sebastian J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of NAACL-HLT*.

Ryan Cotterell and Hinrich Schütze. 2015. Morphological word-embeddings. In *Proceedings of NAACL-HLT*, pages 1287–1292.

Ryan Cotterell and Hinrich Schütze. 2018. Joint semantic synthesis and morphological analysis of the derived word. *Transactions of the ACL*, 6:33–48.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, Yonatan Belinkov, and Preslav Nakov. 2019. One size does not fit all: Comparing NMT representations of different granularities. In *Proceedings of NAACL-HLT*, pages 1504–1516.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of ACL*, pages 587–593.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of ICML*, pages 1243–1252.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of EMNLP*, pages 316–327.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of LREC*, pages 3483–3487.

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of LREC*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Yeachan Kim, Kang-Min Kim, Ji-Min Lee, and SangKeun Lee. 2018. Learning to generate word representations using subword information. In *Proceedings of COLING*, pages 2551–2561.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

András Kornai. 2013. Digital language death. *PloS One*, 8(10):e77056.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of ACL*, pages 66–75.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of EMNLP*, pages 5039–5049.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of ACL*, pages 1517–1526.

Bofang Li, Aleksandr Drozd, Tao Liu, and Xiaoyong Du. 2018. Subword-level composition functions for learning word embeddings. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 38–48.

Xiao Ling and Daniel S. Weld. 2012. Fine-grained entity recognition. In *Proceedings of AAAI*, pages 94–100.

Minh-Thang Luong and Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of ACL*, pages 1054–1063.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*, pages 1659–1666.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of ACL*, pages 1946–1958.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proccedings of NAACL-HLT*, pages 2227–2237.

Mohammad Taher Pilehvar, Dimitri Kartsaklis, Victor Prokhorov, and Nigel Collier. 2018. Card-660: Cambridge rare word dataset - a reliable benchmark for infrequent word representation models. In *Proceedings of EMNLP*, pages 1391–1401.

Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. 2017. Mimicking word embeddings using subword RNNs. In *Proceedings of EMNLP*, pages 102–112.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of ACL*, pages 412–418.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2018. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *arXiv preprint arXiv:1807.00914*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of ACL*, pages 1715–1725.

Peter Smit, Sami Virpioja, Stig-Arne Grnroos, and Mikko Kurimo. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of EACL*, pages 21–24.

Clara Vania and Adam Lopez. 2017. From characters to words to in between: Do we capture morphology? In *Proceedings of ACL*, pages 2016–2027.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS*, pages 5998–6008.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledge base. *Communications of the ACM*, 57:78–85.

Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT. *CoRR*, abs/1904.09077.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, ukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Yadollah Yaghoobzadeh and Hinrich Schütze. 2015. Corpus-level fine-grained entity typing using contextual information. In *Proceedings of EMNLP*, pages 715–725.

Zhuosheng Zhang, Yafang Huang, and Hai Zhao. 2018. Subword-augmented embedding for cloze reading comprehension. In *Proceedings of COLING*, pages 1802–1814.

Jinman Zhao, Sidharth Mudgal, and Yingyu Liang. 2018. Generalizing word embeddings using bag of subwords. In *Proceedings of EMNLP*, pages 601–606.

Yi Zhu, Ivan Vulić, and Anna Korhonen. 2019. A systematic study of leveraging subword information for learning word representations. In *Proceedings of NAACL-HLT*.