A Comparison of Architectures and Pretraining
Methods for Contextualized Multilingual Word Embeddings

Niels van der Heijden
ILLC, University of Amsterdam
n.vanderheijden@student.uva.nl

Samira Abnar
ILLC, University of Amsterdam
s.abnar@uva.nl

Ekaterina Shutova
ILLC, University of Amsterdam
e.shutova@uva.nl

Abstract

The lack of annotated data in many languages is a well-known challenge within the field of multilingual natural language processing (NLP). Therefore, many recent studies focus on zero-shot transfer learning and joint training across languages to overcome data scarcity for low-resource languages. In this work we (i) perform a comprehensive comparison of state-of-the-art multilingual word and sentence encoders on the tasks of named entity recognition (NER) and part of speech (POS) tagging; and (ii) propose a new method for creating multilingual contextualized word embeddings, compare it to multiple baselines and show that it performs at or above state-of-the-art level in zero-shot transfer settings. Finally, we show that our method allows for better knowledge sharing across languages in a joint training setting.

Introduction

With the aim of extending the global reach of NLP technology, much recent research has focused on the development of multilingual models. Due to the lack of annotated data and resources in many languages, these models typically rely on multilingual joint learning or zero-shot transfer learning across languages. Much of this research has utilized multilingual word embeddings (Lample et al. 2017a; Artetxe, Labaka, and Agirre 2018), that project words in multiple languages into a shared multilingual semantic space, such that translations of words across these languages appear close in the new space. Such general-purpose multilingual word embeddings then serve as a basis for joint or transfer learning focused on a particular task.

On the other hand, current best performing monolingual approaches have moved away from static to contextualized word embeddings. Such contextualisation allows the models to address the long-standing problem of polysemy and dynamically model contextual meaning variation. The first such model, CoVe (McCann et al. 2017), used a deep LSTM (Hochreiter and Schmidhuber 1997) encoder pretrained in a machine translation task to contextualize word vectors. This work paved the way for contextualized word embeddings by showing their superiority over the static ones on downstream tasks such as named entity recognition (NER) and question answering. Shortly after, ELMo (Peters et al.

2018) improved upon CoVe by making the contextualized embeddings *deep* by combining information from multiple layers of their LSTM encoder trained with language modelling objective. Finally, the Transformer-based (Vaswani et al. 2017) BERT model (Devlin et al. 2018) broke the performance records across many downstream NLP tasks, achieving up to 7.6% absolute improvement over the previous state of the art.

However, in the multilingual setting, the effects of contextualization are still relatively unexplored. One exception includes the work of Schuster et al, (2019) that generalizes the work of Lample et al, (2017a) to the ELMo model by viewing a contextual word embedding as a context-dependent shift from the (static) mean embedding of a word. The mean embeddings are aligned across languages using adversarial training. This technique, however, does not outperform the non-contextualized embedding baseline in half of the performed experiments when no supervised anchored alignment is given and does not allow for large-scale joint training across languages.

Meanwhile, recent multilingual NLP research has moved away from solely word-level representations to training "massively" multilingual sentence encoders. Artetxe and Schwenk, (2018) proposed a method for learning language-agnostic sentence embeddings (LASER) for 93 languages with a single shared encoder and limited aligned data. More specifically, they trained a deep LSTM-encoder to embed sentences in all (93) languages into a shared space such that semantically similar sentences in different languages appear close to each other in this space. Using this model the researchers advanced the state of the art on zero-shot cross-lingual natural language inference for 13 out of 14 languages in the XNLI (Conneau et al. 2018) dataset. Following these advances, Lample and Conneau, (2019) incorporated multilingual language pretraining into the BERT model (Devlin et al. 2018), coining their model XLM (cross-lingual language model), and again further improved the performance on all XNLI languages. Such multilingual sentence encoders have so far been applied and evaluated in sentence-level tasks. Yet, they provide a promising starting point for learning multilingual contextualized word representations needed for word-level classification tasks and investigating the effects of contextualisation in multilingual word-representation models.

The contributions of this paper are as follows:

- We conduct a comprehensive comparison of state-of-the-art multilingual word and sentence encoding models and pretraining methods on the tasks of NER and part of speech (POS) tagging, experimenting in both zero-shot transfer learning and joint training setting.

- We introduce a new method for learning contextualized multilingual word embeddings based on the LASER encoder and perform an in-depth analysis on its performance against multiple benchmarks in zero-shot transfer and joint training settings. We improve the previous state-of-the-art for English to German NER with 2.8 F1-points and perform at state-of-the-art level for other languages.

- We empirically show and analyze the benefit of contextual word embeddings versus static word embeddings for zero-shot transfer learning.

Related work

Monolingual word representations

After the success of static word embeddings such as word2vec (Mikolov et al. 2013) which produced general-purpose semantic encodings of words, a lot of effort has been put into contextualizing these embeddings. While static embeddings at the time improved results when applied to various NLP tasks, polysemy has remained a challenge for these models, as they represent all meanings of a word within one vector. All occurrences of a word are thus treated the same and the resulting vector is a combination of the semantics of each possible meaning of the word.

In order to incorporate the context into embeddings, Peters et al. (2017) introduced *Language Model (LM) embeddings* and showed that they improved sequence tagging performance, specifically for NER. Subsequently, Peters et al. (2018) took the LM embeddings a step further by making them deep, which resulted in performance improvements in several downstream benchmarks. Their model, ELMo, incorporates information from all layers of the network by taking a layer-wise weighted average of the embeddings. Interestingly, the authors showed that each layer of their LSTM-encoder encoded different properties of the word. The first layer captures more syntactic aspects of a word whereas the second layer captures more high-level semantic information.

Multilingual word representations

Much previous research has focused on aligning word embeddings from different languages into a language independent space (AP et al. 2014), either bilingual or multilingual. These methods either jointly train word embeddings on aligned corpora or align monolingual ones by means of post-processing. An obvious drawback of these methods is their need for aligned corpora. Hence in more recent work the focus shifted towards unsupervised alignment of word embeddings. Works such as Multilingual Unsupervised or Supervised word Embeddings (MUSE) (Lample et al. 2017b) is an example of unsupervised methods capable of aligning embeddings into a shared space, enabling easier knowledge

transfer across languages without the need for additional resources. By aligning the embedding spaces of more than 30 languages, it generates high-quality embeddings for use in multilingual semantics tasks. Because of its proven performance we use these embeddings as a baseline to compare our models to.

To the best of our knowledge, the work of Schuster et al. (2019) comes closest to ours. They present a method to align monolingual ELMo embeddings across languages by modelling such an embedding as a context-dependent shift from its mean (see equation 1, where e_i is the embedding for the work i and c is the context) and applying the linear alignment technique proposed by Mikolov, Le and Sutskever (2013) to the mean embedding of each word.

$$e_{i,c} = \overline{e}_i + \hat{e}_{i,c}, \qquad \overline{e}_i = \mathbb{E}_c[e_{i,c}] \qquad (1)$$

Whereas the authors show this approach to be a simple yet effective one, it does not allow for joint training across languages or handling code-switching. Our proposed method tries to overcome these deficiencies by sharing one encoder for all languages instead of transferring knowledge from one to another language by means of post-processing.

Zero-shot cross-lingual transfer learning

Early work on cross-lingual transfer learning used parallel corpora to create cross-lingual word clusters or exploited external knowledge bases as means of feature engineering (Täckström, McDonald, and Uszkoreit 2012). More recent approaches either exploit bilingual word embeddings to translate a dataset (Xie et al. 2018) or attempt to learn language-invariant features (Chen et al. 2019).

LASER A related promising development is that of Language-Agnostic SEntence Representations (LASER) (Artetxe and Schwenk 2018), where one of the main contributions is an encoder capable of embedding sentences in 93 languages in a shared space such that semantically related sentences are close in this space, regardless of their respective language, language family and script. At the time of release, LASER has set a new state of the art on multiple zero-shot transfer learning tasks such as XNLI (Conneau et al. 2018), indicating its success in creating language agnostic embeddings.

The LASER encoder consists of a byte-pair encoded vocabulary (BPE) (Sennrich, Haddow, and Birch 2015) followed by a 5-layer biLSTM with 512-dimensional hidden states. The final sentence embedding is obtained by applying max pooling over the hidden states of the final layer. BPE is a form of learning Subword Units (SUs) by encoding frequently occurring character n-grams as a symbol. In the case of LASER 50k of those symbols were learned together with their respective embedding.

The encoder is trained in an encoder-decoder setup in the task of machine translation, as shown in Figure 1. More specifically, a dataset was gathered by combining the Europarl, United Nations, Opensubtitles2018, Global Voices, Tanzil and Tatoeba corpora (Tiedemann 2012) comprising sentences in 93 languages translated into English and/or Spanish. The task of the encoder is to encode a sentence in
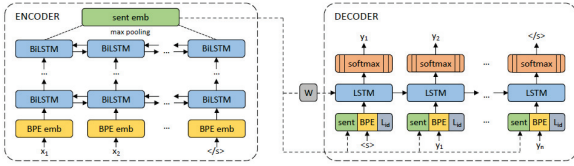
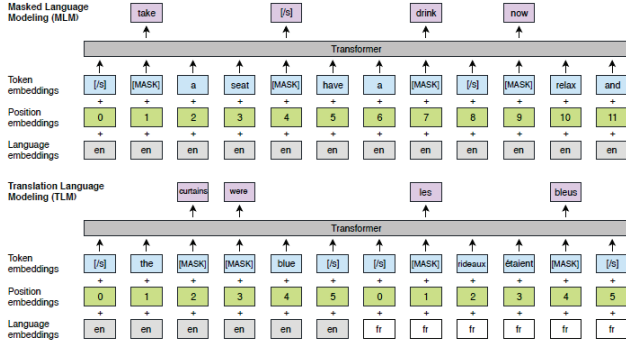Figure 1: Encoder-decoder setup for training LASER

Figure 2: Masked Language Model and Translation Language Model

a 1024-dimensional vector such that the decoder can generate the translation of the original sentence in a chosen target language. The decoder receives no information about which language is encoded by the encoder and hence it cannot distinguish between languages. This forces the encoder to create language-agnostic sentence embeddings.

The main contribution of LASER is that the the authors show that it is possible to encode numerous languages into one encoder when a shared vocabulary is learned and the training data is aligned with just two target languages. Since the LASER sentence encoder achieves very promising results on zero-shot transfer learning for sentence-level NLP tasks, we will use this model as a basis for ours. Our goal is to investigate the possibility of extracting contextualized word embeddings from an encoder trained at the sentence level. We evaluate two versions of our model and compare them to multiple baselines.

Multilingual joint learning

Multilingual joint learning has shown to be beneficial when either the target or all languages are resource-lean (Khapra et al. 2011), when code-switching is present (Adel, Vu, and Schultz 2013) or even in high-resource scenarios (Mulcaire, Swayamdipta, and Smith 2018). Often, multilingual joint training is approached by some form of parameter sharing (Johnson et al. 2017).

Methods

LASER-based contextualized embeddings

We use a pretrained LASER model to obtain the contextualized word embeddings. For this, we compare different methods, which we explain below.

BPE BOW   As the first baseline, we simply create word embeddings by averageing the BPE embeddings per word. This approach can be compared to a continuous Bag-Of-Words (BOW) approach with the BPE embeddings serving as words.

BPE GRU   As the second baseline, we introduce a GRU (Bahdanau, Cho, and Bengio 2014) encoder followed by max-pooling over time to encode the BPE embeddings into a word embedding. First, each BPE symbol is embedded by the pretrained embeddings from the LASER encoder. Then, the data is split into a [BPE_pad, N, Emb_dim] tensor with BPE_pad the length of the longest word in the batch expressed in number of Subword Units. N is the number of words in the batch and Emb_dim is the embedding dimension of the pretrained BPE symbols which equals to 320. This tensor is then fed into the GRU encoder to compute the semantics of the SUs together. The final embedding is created by applying max-pooling over time on the output of the GRU.

MUSE   As the third baseline, we consider static crosslingual word embeddings from the MUSE model (Lample et al. 2017b) as embeddings for our sequence tagger.

LASER-top   Our first proposed method incorporating the LASER LSTM encoder, which we call *LASER-top*, uses the hidden state of the final layer as a base representation of a BPE symbol. First, we apply max-pooling over the forward and the backward hidden states, $\overrightarrow{\mathbf{h}}_L^{(t)}$ and $\overleftarrow{\mathbf{h}}_L^{(t)}$, to obtain a 512-dimensional vector per BPE symbol[1]. Inspired by ELMo, we then rescale this output with a learnable scale parameter γ.

$$\mathbf{m}_L^{(t)} = \gamma \cdot maxpool(\overleftarrow{\mathbf{h}}_L^{(t)}, \overrightarrow{\mathbf{h}}_L^{(t)}), L = 5 \qquad (2)$$

As the LASER encoder is fed a sentence split into SUs its output can now be seen as a contextualized representation of the original embeddings, which is the reason why we expect this method to improve over the baselines. Then, similar to the original approach (Artetxe and Schwenk 2018), the final word embedding is created by applying max-pooling over time to all $\mathbf{m}_L^{(t)}$s belonging to a word, for each word separately.

LASER-elmo   Inspired by ELMo and hence called *LASER-elmo*, we make our multilingual contextualized embeddings *deep* by incorporating multiple layers of the LSTM encoder. In order to do this, a weighted average of the hidden states of all layers is computed by softmax-normalizing task-specific layer weights $\mathbf{s_l}$ for layer l, which are learned during training.

$$\mathbf{e}^{(t)} = \Sigma_{l=1}^5 \mathbf{s_l} \cdot \mathbf{m}_l^{(t)} \qquad (3)$$

Where $\mathbf{e}^{(t)}$ is the deep contextualized embedding for the SU at index t in the sequence, $\mathbf{m}_l^{(t)}$ is computed as in Equation 2 and s_l is the softmax-normalized layer weight. These embeddings are then used as in LASER-top to create word

---

[1]Preliminary experiments showed that max-pooling both reduced overfitting and improved computational efficiency.

embeddings.

LSTM No Pretraining   In order to verify what part of the performance of LASER-top and LASER-elmo can be attributed to the fact that the 5-layer LSTM encoder allows for modelling more complex dependencies, we replace the pretrained encoder by a randomly initialized one and retrain. As overfitting plays a major role in training our models and Peters et al, (2018) have shown a two-layer LSTM to be sufficiently powerful for sequence tagging tasks, we pick a two-layer LSTM to replace the LASER encoder. Otherwise, this baseline functions in the same way as LASER-elmo.

Transformer-based contextualized embeddings

BERT   Devlin et al. (2018) propose a novel language representation model which uses Transformers to create deep contextual representations for words. These representations are obtained by training the model on unlabeled text to predict the words at randomly chosen masked positions conditioned on both the left and the right context - the authors call this technique the Masked Language Model (MLM), see Figure 2. BERT uses WordPiece embeddings (Wu et al. 2016) with a vocabulary of 30k tokens and is trained on the BookCorpus (Zhu et al. 2015) and a Wikipedia dump together: a combined corpus of approximately 3300M words. In addition to the MLM, BERT is also trained on a next sentence prediction task in order to capture relationships between sentences. This task is phrased as a binary classification task where either two consecutive or two random sentences are sampled from the corpus. The complete pretraining procedure combines these two tasks by sampling sentences as described for the next sentence prediction task and applying both this task and the MLM.

Although BERT is not explicitly pretrained to align semantics across languages, its multilingual version[2], from which we use the cased base version in our experiments, is trained on 100+ languages and its monolingual capabilities are (near) state-of-the-art for many NLP tasks without heavily-engineered task-specific architectures.

XLM   After the success of LASER in zero-shot multilingual transfer learning Lample and Conneau (2019) proposed a similar method based on the architecture of BERT. Their contribution lies in the introduction of several new unsupervised and supervised methods for cross-lingual language model pretraining. In this work we will focus on their supervised method as it is most closely related to LASER and outperforms the former on the XNLI benchmark. This method is a multi-task setup of a slightly adjusted MLM (Devlin et al. 2018) combined with their so called Translation Language Model (TLM), see Figure 2 for a comparison of the MLM and TLM (taken from Lample and Conneau).

The TLM exploits a N-way parallel corpus of sentences to allow the model to explicitly use words from language A to predict the masked words in language B, hence encouraging the model to learn similar representations for semantically

---

[2]The multilingual version of BERT is not described in the original paper by (Devlin et al. 2018). Instead, it is described on the official GitHub page of the authors: https://github.com/google-research/bert/blob/master/multilingual.md.

		LASER	BERT	XLM
Supervised	dataset	OPUS, 223M sentences	-	XNLI, 150k sents
	task	Translation	-	TLM
Unsupervised	dataset	-	Wiki dump, 104 langs	Wiki dump, 15 langs
	task	-	MLM + next sentence	MLM

Table 1: Pretraining data and tasks per architecture

similar phrases across languages. The authors use a dataset accompanying the XNLI evaluation set which contains 10k parallel sentences in all 15 languages[3]. The TLM objective is altered with the MLM objective using Wikipedia dumps of each language.

This approach differs from the one used by Artetxe and Schwenk, (2018): firstly, no encoder-decoder structure is used to explicitly align languages in a shared space. Instead, the model is only implicitly encouraged to align languages by allowing to share knowledge across the language boundary and solving the same task independently and simultaneously for both languages. Although the performance on the XNLI dataset improved using this method, there is an obvious drawback in terms of scaling to more languages due to the N-way parallel corpus requirement.

Task-specific models

All the above methods provide a means to extract word embeddings, which then serve as input to models for downstream tasks. We experiment with two downstream tasks: NER and POS tagging. These tasks are chosen in order to evaluate the performance on both semantic (NER) and syntactic level (POS tagging).

We use the encoder of the Transformer model as described in Vaswani et al. (2017) as a sequence tagging model instead of a more commonly used RNN model in the hope to be able to transcend differences in sentence structure across languages. Specifically, we use a double layer Transformer with 2 attention-heads and 300 hidden dimensions for the query, key and value matrices as well as the feed forward network (FFN). The model is topped off by a Conditional Random Field (CRF) (Lafferty, McCallum, and Pereira 2001). For BERT and XLM, literature shows that adding a linear classification layer suffices for token-level classification tasks (Devlin et al. 2018).

Experimental setup

Data and preprocessing

We used the datasets from the CoNLL2002 (Sang 2002) and CoNLL2003 (Sang and De Meulder 2003) shared tasks, which provide data for the NER and POS in English, Spanish, Dutch and German. The data is gathered from local newspapers and is annotated with both named entities and POS tags. All datasets are approximately the same size with $\pm 15,000$ sentences to train on and $\pm 3,500$ to test on.

As the POS tags are given in language specific tags, we convert them to Universal POS tags (Petrov, Das, and McDonald 2011), leaving us with 12 POS tags. In order to eval-

---

[3]The 15 languages in the corpus are English, French, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili and Urdu

uate the ability of our methods to capture both semantic and syntactic information about the word no extra features are used to learn the model. All data is tokenized, and only punctuation normalization and lower casing has been applied in addition to that.

Training

Baseline models BPE BOW, BPE GRU and MUSE are relatively simple models and hence no intensive hyperparameter tuning is performed. Between the embedder and the sequence tagger a dropout of 0.25 is applied and within the sequence tagger a dropout of 0.15. We used Adam (Kingma and Ba 2014) as an optimizer with the default learning rate of 0.001 and applied L_2 regularization with $\lambda = 0.001$. The models were trained for a total of 15 epochs while monitoring performance on a development set and applying early stopping (Prechelt 1997) after two rounds of consecutive decreased performance.

LASER-based models During preliminary experiments we found overfitting to be a major challenge for the LASER-based models and LSTM No Pretraining, hence more sophisticated techniques compared to the baselines have been applied for training.

Firstly, instead of using Adam as optimizer, the *1cycle LR*(Smith 2018) policy is used. This policy uses the much simpler SGD optimizer with momentum and has been shown to improve generalization capabilities of neural networks while decreasing the number of epochs needed to train, a phenomenon the author calls "super convergence".

Finally, all remaining hyperparameters concerning regularization, such as dropout and L_2 regularization, are determined using Bayesian Optimization (Snoek, Larochelle, and Adams 2012).

Transformer-based models As BERT and XLM are practically the same model except for their exact hidden size and their pretraining methods, they are trained using the same method. The original work (Devlin et al. 2018) comes with a guide on how to finetune BERT for downstream NLP tasks based on a small grid search over values for the batch size, learning rate and number of epochs. Moreover, it contains optimal settings for the task of NER on the CoNLL2003 dataset, which is also part of our datasets. Apart from these specified optimal hyperparameter settings, the authors also note that BERT tends to be robust to the exact hyperparameter settings and hence we to use the specified hyperparameters for all experiments. This amounts to training for 4 epochs with a batch size of 16 using a learning rate of 5e-5.

Experiments

Zero-shot transfer learning The first set of experiments involve zero-shot transfer learning across languages. Each model is trained on the English dataset and consecutively evaluated on all other datasets, including two datasets in the low-resource languages — Hungarian (Szarvas et al. 2006) and Basque (Alegria et al. 2004) — for NER.

Joint training In order to evaluate the benefit of joint training, we considered two scenarios. In the first scenario (**A**) a quarter of the training set of each language in the CoNLL2002 and CoNLL2003 shared tasks is taken and combined into a new training set of approximately the same size as the former ones. A validation set is created in a similar fashion and used for monitoring. Each model is trained on the multilingual dataset and then evaluated on the original test sets of each language. In the second scenario (**B**) one full training set (English) was complemented with a quarter of the training sets of the remaining languages. The difference in performance between scenarios A and B can be used as a way to quantify how well each model shares knowledge across languages.

Results

Zero-shot transfer learning

Table 2 shows the F1-scores per model per language for the tasks of NER and POS tagging with the highest scores per language shown in bold and, if applicable, the state of the art underlined. Where possible, scores from monolingual training and evaluation are appended, separated by "/" for reference. All results have been tested against LASER-top for significance using the sign test with $\alpha = 0.05$ and have been found to be significant.

The highest performance scores were achieved by either BERT or LASER-top, where BERT performs the best on 6/9 tasks. Overall BERT appears to be a stronger model for learning the tasks at hand than the LSTM-based LASER-top model, as BERT achieves the highest scores in all monolingual settings except for German NER. LASER-top on the other hand is less capable of learning the task at hand in the source language, but the drop in performance when evaluating on other languages is smaller: it achieves the highest score on 2 out of 4 languages for POS tagging and advances the state of the art for German NER, indicating the added benefit of the LASER pretraining method for crosslingual knowledge sharing.

Surprisingly, XLM does not outperform BERT in any of the settings. It is worth noting that from the evaluated languages only English and German have been seen by XLM during pretraining, which explains the poor transfer learning capabilities to the remaining languages. Yet, XLM does not outperform BERT in the transfer from English to German, indicating no added benefit in the TLM method for zero-shot transfer learning across languages.

For the low-resource languages the added benefit of contextualization is evident. As expected, performance on languages from more distant language-families is lower: all but one models score higher in Hungarian than in Basque. Furthermore, pretraining on a higher number of languages appears to positively influence performance on low-resource languages, as BERT outperforms XLM by a large margin and the same holds for LASER-top and LSTM No Pretraining.

Contrary to our expectations, LASER-top consistently outperforms LASER-elmo across all tasks. This is likely to be due to overfitting: LASER-elmo achieves higher scores

Table 2: F1-scores zero-shot transfer learning appended with monolingual scores if applicable

NER	BPE BOW	BPE GRU	MUSE	LASER-top	LASER-elmo	LSTM No Pre	BERT	XLM[1]	Literature
English	0.509	0.742	0.758	0.803	0.786	0.752	**0.914**	0.892	–/0.935
Dutch	0.247/*0.441*	0.245/*0.684*	0.477/*0.64*	0.578/*0.785*	0.424/*0.75*	0.194/*0.621*	**0.587**/***0.834***	0.212/*0.792*	0.654
German	0.187/*0.55*	0.242/*0.57*	0.481/*0.603*	**0.617**/*0.679*	0.414/*0.635*	0.235/*0.591*	0.589/*0.773*	0.579/***0.813***	0.585
Spanish	0.033/*0.312*	0.167/*0.583*	0.409/*0.524*	0.511/*0.778*	0.319/*0.744*	0.1567/*0.635*	**0.618**/***0.819***	0.458/*0.803*	0.735
Basque	0.037	0.029	—[2]	0.241	0.059	0.03	**0.403**	0.0811	—[3]
Hungarian	0.023	0.043	0.387	0.414	0.208	0.073	**0.455**	0.13	—[3]
POS									
English	0.674	0.873	0.864	0.916	0.884	0.905	**0.941**	0.931	—[3]
Dutch	0.454/*0.736*	0.430/*0.931*	0.563/*0.923*	**0.726**/*0.954*	0.607/*0.952*	0.444/*0.948*	0.658/***0.961***	0.237/*0.954*	—[3]
German	0.456/*0.79*	0.466/*0.912*	0.623/*0.895*	**0.781**/*0.954*	0.665/*0.943*	0.423/*0.943*	0.722/***0.958***	0.663/*0.957*	—[3]
Spanish	0.427/*0.495*	0.450/*0.913*	0.522/*0.85*	0.715/*0.919*	0.636/*0.914*	0.367/*0.918*	**0.746**/***0.944***	0.601/*0.941*	—[3]

[1] XLM has been pretrained on the 15 XNLI languages: only English and German are amongst those languages.
[2] No MUSE embeddings available for Basque.
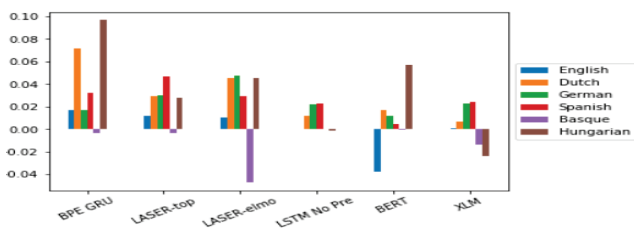[3] Not available in literature.

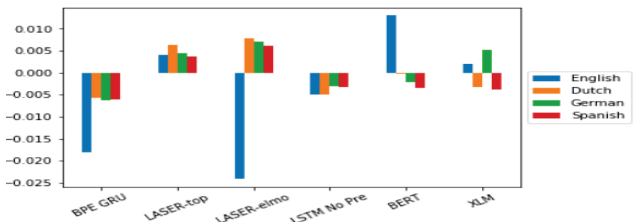Figure 3: Joint training NER

Figure 4: Joint training POS

on the training set than LASER-top in all experiments. Since the drop in performance across languages is far greater for all baseline models than for LASER-top and LASER-elmo, we attribute this improved performance to the multilingual pretraining. As the scores for LSTM No Pretraining are lower than LASER-top and LASER-elmo in the transferred languages this improved performance cannot be attributed to the added complexity from the extra layers.

Joint training

Table 3 shows the results of joint training of the four CoNLL2002 and CoNLL2003 languages. For both NER and POS tagging, BERT and XLM clearly outperform all other models, yet it is questionable whether this is the case because of the ability to share knowledge across languages or because Transformer-based models are better suited for the task. Figures 3 and 4 visualize the added benefit of joint training expressed as the difference in F1-scores compared

to the baseline. For English this baseline is the monolingual baseline whereas for the other languages the baseline is the mixed setting with a quarter of each language (scenario A) compared to the full English dataset extended with a quarter of the remaining datasets (scenario B). BPE BOW has been omitted from the graphs as its values distort the graph and is of less importance than the remaining models.

Whereas LASER-top benefits from joint training in all but one languages (Basque), this greatly differs for BERT and XLM, indicating that the pretraining method used for LASER might allow for better crosslingual knowledge sharing than the MLM and TLM methods used for BERT and XLM respectively.

When comparing BERT and XLM, it appears that XLM shares knowledge better across languages it has been pretrained on while languages from a distant language family do not benefit at all from joint training.

Discussion and error analysis

In order to further analyze the added benefit of contextualized word embeddings versus static word embeddings in the zero-shot transfer setting, we compare MUSE with LASER-top in more detail. Firstly, we look at how good the models are at identifying whether an entity is present at all by trimming down the labels to *B, I* and *O* and creating a confusion matrix in Figure 5. It can be clearly seen that LASER-top is
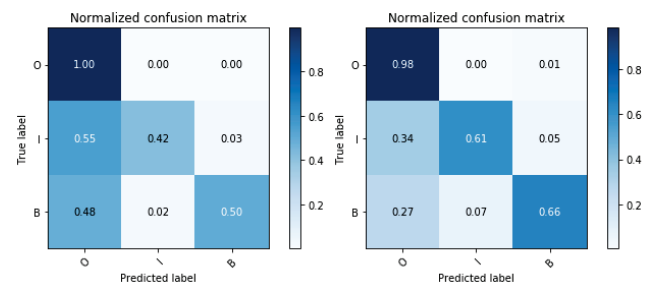

Figure 5: Dutch *BIO* confusion for MUSE, LASER-top

better at detecting an entity than MUSE.

Table 3: F1-scores joint training

NER	BPE BOW	BPE GRU	LASER-top	LASER-elmo	LSTM No Pre	BERT	XLM[1]
English	0.330+*0.096*	0.632+*0.017*	0.723+*0.012*	0.678+*0.011*	0.617-*0*	0.826-*0.038*	**0.854+*0.001***
Dutch	0.172+*0.129*	0.527+*0.071*	0.709+*0.03*	0.618+*0.045*	0.601+*0.012*	**0.804+*0.017***	0.748+*0.006*
German	0.184+*0.08*	0.497+*0.017*	0.698+*0.03*	0.545+*0.044*	0.568+*0.022*	0.799+*0.012*	**0.809+*0.022***
Spanish	0.169+*0.039*	0.491+*0.032*	0.619+*0.047*	0.584+*0.029*	0.505+*0.023*	**0.729**+*0.005*	0.724**+*0.024***
Basque	0.039+*0.006*	0.036-*0.003*	0.318-*0.004*	0.131-*0.047*	0.04+*0*	**0.502-*0.***	0.144-*0.013*
Hungarian	0.051+*0.019*	0.085+*0.097*	0.378+*0.028*	0.24+*0.045*	0.103-*0*	**0.528+*0.057***	0.254-*0.024*
POS							
English	0.771+*0.168*	0.859-*0.018*	0.887+*0.004*	0.878-*0.024*	0.865-*0.005*	0.912+*0.013*	**0.916+*0.002***
Dutch	0.807+*0.003*	0.91-*0.006*	0.927+*0.006*	0.919+*0.007*	0.912-*0.005*	**0.949-*0.***	0.928-*0.003*
German	0.884-.*003*	0.894-*0.006*	0.920+*0.005*	0.909+*0.007*	0.894-*0.003*	0.937-*0.002*	**0.94+*0.005***
Spanish	0.1-.*035*	0.876-*0.006*	0.9+*0.004*	0.886+*0.006*	0.879-*0.003*	**0.915-*0.004***	0.909+*0.002*

Note: all results are depicted as base score followed by a deviation: the base scores are the F1 scores after training using a quarter of each dataset (scenario A) and the deviation is the change in scores after training with the full English train set and a quarter of the remaining languages (scenario B).

[1] XLM has been pretrained on the 15 XNLI languages: only English and German are amongst those languages.

Furthermore, we find evidence that the difference in performance is partially attributable to the fact that LASER-top has contextualized representations for words. Two examples are given where MUSE made an error that LASER-top did not by capitalizing the errors and appending the predicted entity type.

despite winning the asian GAMES/O title two years ago , uzbekistan are in the finals as outsiders .

houston 1996-12-05 ohio state left tackle orlando PACE/O became the first repeat winner of the lombardi award thursday night when the ROTARY/O club of houston again honoured him as college football 's lineman of the year .

The incorrectly predicted words are all words that on their own would not be considered an entity of interest, but in this respective context they are part of a bigger entity span. These two sentences are examples of an often recurring pattern in the data in all evaluated languages.

Conclusion

In this work we have presented a comprehensive comparison of architectures and pretraining methods for contextualized multilingual word embeddings. We have also shown that it is possible to train a language model solely in an encoder-decoder style on the task of machine translation and consecutively use the encoder to create multilingual contextualized word embeddings. Moreover, we have shown that LASER-top outperforms (non-contextualized) baselines in multiple settings on multiple tasks and sometimes performs on par or better than BERT in the zero-shot transfer setting.

Although our results indicate that our LSTM-based model is not as well suited for downstream NLP tasks as Transformer-based models, we have empirically shown our method to be superior at sharing knowledge across languages in a joint training setting and to perform at or above state of the art in zero-shot transfer setting.

As the results of our models are not yet on par with the current state of the art in monolingual settings, a logical next step is to investigate ways to combine pretraining methods, with the aim of learning higher quality monolingual word representations while encouraging knowledge sharing across languages. For instance a multi-task setup with BERT's MLM combined with the LASER pretraining method can be explored for this purpose.

References

Adel, H.; Vu, N. T.; and Schultz, T. 2013. Combination of recurrent neural networks and factored language models for code-switching language modeling. In *Proceedings of the 51st ACL*, 206–211.

Alegria, I.; Arregi, O.; Balza, I.; Ezeiza, N.; Fernandez, I.; and Urizar, R. 2004. Design and development of a named entity recognizer for an agglutinative language. *IJCNLP-04*.

AP, S. C.; Lauly, S.; Larochelle, H.; Khapra, M.; Ravindran, B.; Raykar, V. C.; and Saha, A. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in NIPS*, 1853–1861.

Artetxe, M., and Schwenk, H. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*.

Artetxe, M.; Labaka, G.; and Agirre, E. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the ACL*, 789–798.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Chen, X.; Hassan, A.; Hassan, H.; Wang, W.; and Cardie, C. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 3098–3112.

Conneau, A.; Lample, G.; Rinott, R.; Williams, A.; Bowman, S. R.; Schwenk, H.; and Stoyanov, V. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hochreiter, S., and Schmidhuber, J. 1997. Lstm can solve hard long time lag problems. In *Advances in neural information processing systems*, 473–479.

Johnson, M.; Schuster, M.; Le, Q. V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Viégas, F.; Wattenberg, M.; Corrado, G.; et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics* 5:339–351.

Khapra, M. M.; Joshi, S.; Chatterjee, A.; and Bhattacharyya, P. 2011. Together we can: Bilingual bootstrapping for wsd. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 561–569. Association for Computational Linguistics.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lafferty, J.; McCallum, A.; and Pereira, F. C. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lample, G., and Conneau, A. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Lample, G.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2017a. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Lample, G.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2017b. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

McCann, B.; Bradbury, J.; Xiong, C.; and Socher, R. 2017. Learned in translation: Contextualized word vectors. In *Advances in NIPS*, 6294–6305.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in NIPS*, 3111–3119.

Mikolov, T.; Le, Q. V.; and Sutskever, I. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mulcaire, P.; Swayamdipta, S.; and Smith, N. 2018. Polyglot semantic role labeling. *arXiv preprint arXiv:1805.11598*.

Peters, M. E.; Ammar, W.; Bhagavatula, C.; and Power, R. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Petrov, S.; Das, D.; and McDonald, R. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Prechelt, L. 1997. Early stopping — but when?

Sang, E. F., and De Meulder, F. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Sang, E. F. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. *arXiv preprint cs/0209010*.

Schuster, T.; Ram, O.; Barzilay, R.; and Globerson, A. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. *arXiv preprint arXiv:1902.09492*.

Sennrich, R.; Haddow, B.; and Birch, A. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Smith, L. N. 2018. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*.

Snoek, J.; Larochelle, H.; and Adams, R. P. 2012. Practical bayesian optimization of machine learning algorithms. In *Advances in NIPS*, 2951–2959.

Szarvas, G.; Farkas, R.; Felföldi, L.; Kocsor, A.; and Csirik, J. 2006. A highly accurate named entity corpus for hungarian. In *LREC*, 1957–1960.

Täckström, O.; McDonald, R.; and Uszkoreit, J. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 conference of the ACL*, 477–487. Association for Computational Linguistics.

Tiedemann, J. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, 2214–2218.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in NIPS*, 5998–6008.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Xie, J.; Yang, Z.; Neubig, G.; Smith, N. A.; and Carbonell, J. 2018. Neural cross-lingual named entity recognition with minimal resources. *arXiv preprint arXiv:1808.09861*.

Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE*, 19–27.