The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)

Analysis of Joint Multilingual Sentence Representations and Semantic K-Nearest Neighbor Graphs

Holger Schwenk, Douwe Kiela, Matthijs Douze
Facebook AI Research
{schwenk,dkiela,matthijs}@fb.com

Abstract

Multilingual sentence and document representations are becoming increasingly important. We build on recent advances in multilingual sentence encoders, with a focus on efficiency and large-scale applicability. Specifically, we construct and investigate the k-nn graph over the joint space of 566 million news sentences in seven different languages. We show excellent multilingual retrieval quality on the UN corpus of 11.3M sentences, which extends to the zero-shot case where we have never seen a language. We provide a detailed analysis of both the multilingual sentence encoder for twenty-one European languages and the learned graph. Our sentence encoder is language agnostic and supports code switching.

1 Introduction

Multilingual representations, at the word, sentence or document level, are becoming increasingly important. The main motivation for such joint multilingual representations is the desire to transfer NLP applications across languages, using no or only a limited amount of resources for the target language. This is particularly useful in cases where the target language is low-resourced, as is the case for most languages in the world. Multilingual representations tend to be evaluated on cross-lingual document classification, where the standard task is based on the Reuters corpus (Klementiev, Titov, and Bhattarai 2012; Schwenk and Li 2018) or sentiment analysis (Chen et al. 2016). More recently, SemEval-17 (Bethard et al. 2017) organized a task on multi- and cross-lingual semantic word similarity, as well as a task on semantic textual similarity for multiple monolingual and a few cross-lingual pairs. These examples illustrate that multilingual representation learning has mostly been focused on constructing optimal representations only for specific tasks, and has often been limited to two languages only. Recently, a natural language inference task (NLI) was extended to fifteen languages (Conneau et al. 2018).

At the same time, a large body of research has developed that concerns itself with learning *"universal representations"* for a single language, usually English. The goal of such representations is to perform well on many tasks. On the one hand, unsupervised approaches have shown to be promising, e.g. (Kiros et al. 2015; Peters et al. 2018), due to their ability to take advantage of huge corpora.On the other hand, it was shown that state-of-the-art results can be obtained with supervised training on an NLI corpus (Conneau et al. 2017). Extending this idea, multi-task learning was successfully applied to learning universal sentence representations, e.g. (Subramanian et al. 2018; Cer et al. 2018).

In this paper we learn universal multilingual sentence representations for twenty-one languages, and show how they can be applied to a variety of problems efficiently and at a large scale, going beyond the traditional limited set of evaluations. The major contributions of this work are:

- we train a joint multilingual sentence embedding with a shared byte-pair-encoding for 21 European languages;
- we construct a large k-nn graph over the joint space of 566M sentences in seven languages.
- we provide a detailed quantitative analysis comprising important open questions such as:
  - how many BPE tokens are shared between languages?
  - to what extent does this overlap reflect the linguistic properties of the languages and their categorization into language families?
  - how dense are the k nearest neighbors?
  - is it possible to extract trajectories between sentences in the graph that are linguistically plausible?
  - does the model support code-switching, i.e. words from multiple languages in one sentence?
- we show that our multilingual encoder outperforms previous work on large scale similarity search: we achieve a precision@1 of 83.3 on the reconstruction of the UN corpus of 11.3M English/French sentences, in comparison to P@1 of 48.9 obtained by (Guo et al. 2018);
- we define new quantitative evaluation tasks to analyze the generalization behavior of multilingual sentence embeddings with respect to unseen domains and languages;
- we show that our system is able to handle zero-shot transfer to several linguistically related languages without using any resources of those languages;
- the code used in this paper is freely available in the LASER toolkit (**L**anguage **A**gnostic **SE**ntence

Representations).[1] We also make available the entire k-nn graph over all 566M sentences.

This paper is organized as follows. We first summarize related work, and then describe our approach for learning a multilingual joint sentence representation space and describe efficient ways to calculate an k-nn graph over 566M sentences. Section 4 provides a detailed qualitative analysis, in order to more closely examine the learned representations. Full quantitative results, including comparison with other works, are then presented in section 5. Finally, we conclude this paper with directions for future research.

2 Related Work

There is an increasing body of research on the topic of learning multilingual sentence representations, for instance (Hermann and Blunsom 2014; Pham, Luong, and Manning 2015; Zhou, Wan, and Xiao 2016; Chandar et al. 2013; Mogadala and Rettinger 2016). In this paper, we build upon the approach of (Schwenk and Douze 2017), who experiment with multiple sequence encoders and decoders, trained with N-way aligned corpora from the machine translation community. The paper shows that the cosine distance in the joint embedding space appears to be proportional to the linguistic similarity of the sentences, independently of the original languages that the sentences were written in. Their analysis was limited to multilingual similarity search and paraphrasing. In follow up work (Schwenk 2018), one single shared BiLSTM encoder for all input languages was used, similar to ideas in multilingual neural MT (Johnson et al. 2016). However, this system was only evaluated on the task of filtering and mining parallel data for neural machine translation. A very similar approach for bitext mining was proposed in (España-Bonet et al. 2017). A joint NMT system with attention is trained on several languages pairs, similar to (Johnson et al. 2016), including a special token to indicate the target language. After training, the sum of the encoder output states is used to obtain a fixed size sentence representation. More recently, (Guo et al. 2018) train bilingual sentence embeddings and report bitext filtering performance and reconstruction of the UN corpus. Other approaches which use the distance between sentence representations in two different languages include (Grégoire and Langlais 2017; Bouamor and Sajjad 2018; Hassan and et al. 2018)

We are not aware of other works concerned with the in-depth analysis of the generalization behavior of multilingual sentence embeddings, nor of the construction and analysis of large scale k-nn graphs in a multilingual space.

3 Approach

In this section, we describe the process for learning universal multilingual sentence representations, as well as how we construct the k-nn graph over the joint space.

3.1 Multilingual sentence representations

In this paper, we build on the encoder/decoder approach of (Schwenk 2018), and its open-source LASER implementation. The basic architecture is the same as used in (Artetxe and Schwenk 2018), see Figure 1. We use a single shared BiLSTM encoder for all twenty-one languages. The word embeddings are 384 dimensional, and the BiLSTM uses five layers of size 512, respectively. Dropout is set to 0.1. The 1024-dimensional sentence embedding is obtained by max-pooling over the BiLSTM outputs. The decoder is a 5-layer LSTMs with 2048-dimensional hidden layers, and shared for two target languages: English and Spanish. An additional 32-dimensional embedding layer is used to give the language information to the decoder.

We use byte-pair encoding (BPE) (Sennrich, Haddow, and Birch 2016) with 40k merge operations to learn a joint vocabulary for all the twenty-one languages.[2] Larger vocabularies achieve only slight improvements. As such, we hope to learn better multilingual models, and to be able to generalize to languages not seen during training. In contrast to (Johnson et al. 2016), we do not use a special input token to indicate the target language. Our joint encoder has no information at all on the encoded language, or what will be done with the sentence representation.

The purpose of this work is not to train multilingual joint sentence embeddings for as many languages as possible, but to study several properties of the joint embedding space. We trained our model on the twenty-one languages of the Europarl corpus (Koehn 2005). These cover several and diversified language families:

- **Germanic:** English (en), Danish (da), Dutch (nl), German (de) and Swedish (sv);
- **Romance:** French (fr), Italian (it), Portuguese (pt) and Romanian (ro) and Spanish (es);
- **Slavic:** Bulgarian* (bg), Czech (cs), Polish (pl), Slovak (sk) and Slovenian (sl);
- **Baltic:** Latvian (lv) and Lithuanian (lt);
- **Uralic:** Estonian (et), Hungarian (hu) and Finish (fi);
- **Hellenic:** Greek* (el)

An important aspect of this work is the analysis of the generalization behavior to new languages (see Sections 5.2 and 5.4). For this purpose, we evaluate additional languages which are in the same linguistic family as some of the trained ones:

- **Germanic:** Afrikaans (af), Norwegian Bokmål (no) and Norwegian Nynorsk (nn);
- **Romance:** Catalan (ca), Chavacano (cbk) and Galician (gl);
- **Slavic:** Bosnian* (bs), Croatian (hr), Macedonian* (mk), Russian* (ru) and Serbian* (sr)

We would like to stress that no resources of these generalization languages are used during training, not even monolingual resources. There are two official written standards of Norwegian: *"Bokmål"* is the preferred one, while *"Nynorsk"* is used by an estimated 15% of the population. Galician is spoken by more than two million people in North West Spain, and Chavacano by 600 thousand people in the

[1] https://github.com/facebookresearch/LASER

[2] We use https://github.com/glample/fastBPE

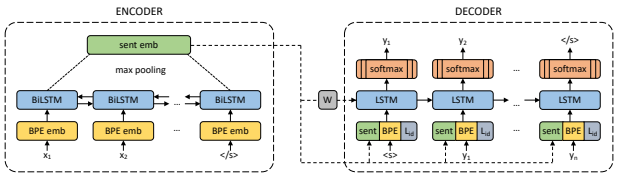

Figure 1: Architecture of our approach to learn joint multilingual sentence embeddings, based on (Artetxe and Schwenk 2018).

Philippines. All languages with an asterisk use the Cyrillic script. We transliterate these languages into the Latin script using an open-source tool.[3] The shared encoder can handle multiple scripts, but a common script is needed for our analysis of the joint BPE vocabulary (see Section 4.1). Training is done using bitexts, separately aligned with two target languages: English and Spanish. We use the aligned texts available on the OPUS web site (Tiedemann 2012).[4] This implements multi-task training with two target languages. For each mini-batch, one language pair is selected.

3.2 Creating the multilingual k-nn graph

In the framework of the WMT evaluations,[5] large collections of monolingual news texts are provided. Those can be used for language modeling, for the purpose of mining parallel data, see for instance (Schwenk 2018), or for back-translation, e.g. (Edunov et al. 2018). We use these freely available texts to construct our multilingual k-nn graph, covering the following languages: Czech, English, Estonian, French, Finish, German and Spanish.

We performed some basic cleaning of the data: removal of sentences with email addresses or references to WEB pages, and sentences with less than 4 or more than 60 words. This is motivated by the fact that it is very unlikely to find similar long sentences. This totals to 566M sentences and roughly 10 billion words (see Table 1). For each sentence, we create the k-nn graph by finding the 20 closest sentences in the embedding space, i.e., k=20. These k closest sentences can come from different languages. Calculating, storing and searching in this graph is a significant computational challenge. Our sentence embeddings are of dimension 1024 and are stored as floating point numbers with 4 bytes. All the embeddings thus require $566M \times 1024 \times 4 = 2.4$ TB for storage. A brute force approach to obtain the k-nn graph, i.e., calculating the L_2-distance between all vectors and keeping the k smallest values, would require at least $566M^2 \times 1024 \approx 3.3 \times 10^{20}$ floating point operations.

We tackle this computational challenge with the highly optimized FAISS library for efficient similarity search and clustering of dense vectors (Johnson, Douze, and Jégou 2017).[6] FAISS is mainly used for indexing and searching in huge image collections, but it can operate on any type of object represented by fixed-size vectors. It offers a collection of methods for reducing memory requirements and speeding up searching over huge indices. In general, one has to aim for a compromise between storage size, speed and search errors. One straight-forward possibility would be to perform PCA to reduce the dimension of the vectors. However this quickly leads to high search errors and is unlikely to offer space reductions larger than 10 times. After several experiments with the available compression, quantization and search algorithms available in FAISS, we found a good compromise with the following setting: the collection of vectors is split up with k-means into 16 384 well-balanced clusters, and compressed with OPQ (Ge et al. 2013) to 32 bytes. This corresponds to the index type "`OPQ32,IVF16384,PQ32`" in FAISS terms. By these means, we were able to reduce the storage requirement of the index over the 566M sentences to 22 GB, i.e. a hundred times smaller than all the embedding vectors. This index can be easily loaded into main memory of a standard server. The k-nn graph is built in brute-force fashion: each vector in turn is used as a query. To query a vector, a subset of 512 clusters is visited and query-to-distance codes are computed without decompressing the codes (Jegou, Douze, and Schmid 2011).

The calculation of the 566M sentence embeddings took about 100h on GPU (which can be run in parallel by splitting the data), and the creation of the compressed index needed 12h on a multi-threaded CPU. All the distances of the 20-nn graph are calculated in a distributed way on 4 GPUs. This required about 55h of compute time. This k-nn graph is freely available in the LASER toolkit, together with tools to manipulate and search the graph.

Lang.	**en**	**de**	**fr**	**es**	**cs**	**fi**	**et**
# sents	186M	251M	39.0M	12.3M	61.8M	13.5M	2.7M
# words	4.2G	4.5G	929M	166M	104M	335M	45M

Table 1: Size of WMT'18 monolingual corpora (limited to lower cased sentences with 4–60 words).

4 Qualitative Analysis

In this section, we provide a qualitative analysis of the learned multilingual sentence representations. Having obtained a better understanding of the sentence space, the next section quantitatively evaluates its properties for retrieval.

We first analyze the joint BPE vocabulary and show its potential for understanding differences between languages: is there an overlap among the tokens used for each language ?; to what extent does this overlap reflect the linguistic properties of the languages and their categorization into language families? Second, we believe that the k-nn graph over 566M sentences in seven languages will be a very useful resource to study relations between sentences and languages. We study the distribution of the distances in the k-nn graph, and subsequently examine a new algorithm for *"warping"* from one sentence to another, making only small linguistic changes at each step, leading to similarity trajectories through semantic space.

[3] https://pypi.org/project/transliterate/
[4] http://opus.nlpl.eu/Europarl.php
[5] http://www.statmt.org/wmt18/translation-task.html
[6] https://github.com/facebookresearch/faiss

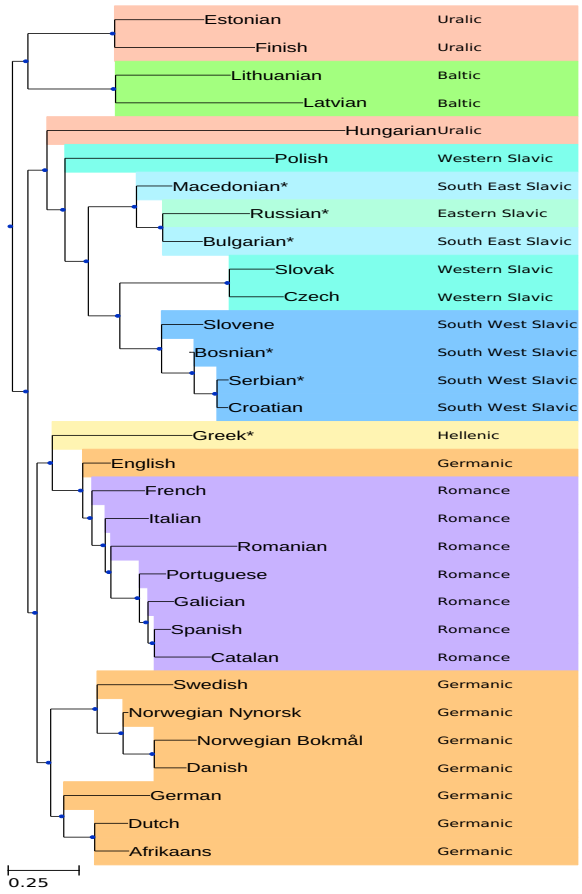

Figure 2: Graphical visualization of a phylogenetic tree over the symmetrized KL-divergence between the BPE vocabularies for each language pair. An asterisk at the language name means that the texts are romanized.

4.1 Joint BPE vocabulary

Our shared multilingual sentence encoder uses a joint multilingual BPE vocabulary. It is calculated by concatenating the training corpora of all the twenty-one languages and finding the most frequent character n-grams. We are interested in knowing whether the vocabularies for each languages are totally separate, i.e. each one has 40 k / 21 $\approx$ 1900 BPE tokens, or whether some tokens are used in several languages. All the statistics in this section are calculated on ten million sentences of Common Crawl data. Please remember that we apply romanization for the languages which use a Cyrillic script (Greek, Bosnian, Bulgarian, Macedonian, Serbian and Russian).

The number of unique BPE tokens per language varies between 11.9k (Slovak) and 13.9k (Finish), with less than 100 tokens appearing in each language only. This clearly indicates that almost all BPE tokens are used by at least two languages. Aiming at a detailed comparison, we calculated the symmetrized KL-divergence between the distributions over the BPE tokens for all language pairs, applied a bottom-up neighbor joining algorithm and build a

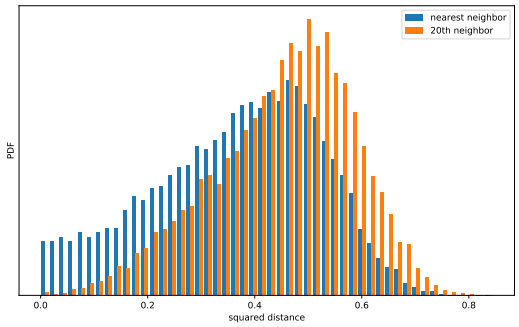

Figure 3: Distribution of squared distances between sentence embeddings, for a sample of 50k sentences, it shows the distance to the nearest sentence and the 20$^{\text{th}}$ nearest sentence.

"phylogenetic tree". The length of the branches reflects the statistical closeness. As can be seen in Figure 2, the observed statistical similarities based on the BPE vocabulary distributions correspond surprisingly well to the linguistically defined languages families. All Germanic and Romance languages are clustered accordingly, and we even recover the various sub-categories of Slavic languages. Within one language family, we also see intuitive associations, like Dutch/Afrikaans, Northern and Southern Germanic languages, Spanish/Galician or Czech/Slovak. The only exception is Hungarian which is considered as an Uralic language but clustered with Slavic languages in our tree. The linguistic classification of Hungarian does not seem to be unanimous between linguists – an affinity to the Turkic language family has also been proposed. A detailed comparison of our statistical clustering of the languages into a tree structure with linguistic theories is beyond the focus of this paper.

4.2 Distance distributions

Figure 3 shows the distribution of distances between sentences. Distances of 0 were removed, as these are just occurrences of the same sentence with minor changes in punctuation or unicode normalization artifacts. Distances below 0.1 correspond to tiny changes in word choice "*But it comes at a cost*" $\leftrightarrow$ "*But this comes at a cost*".

Around 0.2 corresponds to more important changes, like "*In Europa sind wir Spitze*" $\leftrightarrow$ "*Wir sind zurück an der Spitz*". Above 0.3 usually means the sense is likely to be different like "*How were these sales considered illegal ?*" $\leftrightarrow$ "*The sale of these products is illegal .*".

4.3 Linguistic trajectories

Once the multilingual k-nn graph is constructed, we can find paths between two arbitrary sentences, of the same or in different languages. Given a source and destination sentence, S and D respectively, the shortest path connecting these two sentences is defined by $P = \{p_1, \ldots, p_n\}$ with $p_1 = S$ and $p_n = D$ such that

$$\min_{P} \max_{i=1..n} \text{Dist}(p_i, p_{i+1}),$$

i.e., we want to find a path that follows nearby sentences, such that the largest step in the path is as small possible.

Example 1

en	It's just painful."
en	It's painful."
en	It's been painful."
en	That was painful."
en	That's been tricky."
en	It's tricky."
en	It is gorgeous."
en	It's gorgeous.'
en	But it's gorgeous.'
en	But it was foul."
en	But it was fun."

Example 2

en	5 slices of lemon
en	3 slices of lemon
en	3 teaspoons lemon zest
en	1 teaspoon lemon zest
de	1 großer Bund Schnittlauch
de	1 kleines Bund Schnittlauch
en	1 small bouquet garni
en	1 small butternut squash

Example 3

fr	Le but est de montrer qu'ensemble, on y arrive mieux
de	Ziel ist es, gemeinsam für Verbesserungen zu sorgen.
de	Ich bin überzeugt, dass wir gemeinsam etwas weiterbringen können.
de	Ich könnte mir vorstellen, dass wir mal etwas zusammen unternehmen.
de	Da habe ich mir gleich gedacht, dass man doch zusammen etwas machen könnte.
de	Sie überlegten, ob sie zusammen etwas machen könnten.
de	Sie müsse entscheiden, ob etwas unternommen würde.
fr	Il pourra par la suite déterminer si des actions sont nécessaires.
de	Dann werde man sehen, ob es Handlungsbedarf gebe.
de	Dann wird klar werden, ob überhaupt Bedarf besteht.
de	Wir werden erst einmal prüfen, ob Bedarf besteht.
fi	Palaamme asiaan myöhemmin, jos tarvetta ilmenee.
en	This is something I can use should the need arise.

Example 4

en	Sometimes I don't know whether I'm coming or going.
en	I don't know whether I'm coming or going.
en	I don't know whether to go or not.
de	Ich muss überlegen, ob ich gehe oder nicht.
en	It depends if I score or not.
en	Tell me if I'm crazy or not.
en	And tell me if I'm crazy.
de	Mal gucken, ob ich das hinbekomme.
de	Oh Gott, wenn ich das geahnt hätte.
cs	Určitě, právě kvůli tomu si ho pamatuju.
cs	Stále je mi do pláče, když si na to vzpomenu.
cs	Měl bych jí zavolat, připomenout se.
fi	Luistoa riitti, hän muistelee.
fi	Olihan se aika luksusta, Niko muistelee.
fi	Se oli kovaa aikaa, Nagelsmann muistelee.
cs	Už je to dlouho a málokdo si vzpomene.
cs	Tělo si pamatuje a pamatuje si dlouho.
fr	Et lorsque ça t'arrive, tu t'en rappelles longtemps.
en	You go step by step and finally when you look back it's a long time.
en	You get past that and after a while you don't think about it.
en	You forget that when you haven't seen him in a while.
en	Make it again, if you have forgotten it for a while.
de	Oder so lange singt, bis man es vergisst.
de	So lange, bis er irgendwann begann, sich selbst zu vergessen.
en	So before he starts, he takes a moment to remember that.
en	And when things occasionally go awry, take a moment to remember this.
de	Manchmal werden wir, wenn wir etwas vorgeblich Neuem begegnen, an etwas ziemlich Altes erinnert.
en	But once in a great while, someone says something truly memorable.

Figure 4: Examples of linguistic trajectories found in the public WMT news texts by the shortest connected path algorithm. For each example, the first sentence is the starting point and the last sentence is the target.

This algorithm was first proposed for a k-nn graph over images. It has been shown to find smooth transitions between images (see Figure 6 in (Johnson, Douze, and Jégou 2017)). What might constitute a *"smooth transition"* or *"a small change"* between sentences in NLP, when handling sequences of discrete units, is less straightforward. Intuitively, one would probably accept that replacing a word by a synonym or paraphrasing a sequence of words constitutes a small enough linguistic change. However, if we want to find a path between two sentences that have a very different meaning, coming up with several consecutive *small linguistic changes* that together form the trajectory between them is less obvious. Figure 4 shows some examples of these linguistic trajectories in our embedding space. It is important to remember that we do not use a generative model: the algorithm tries to find the nearest sentence in the pool of available sentences. This naturally favors sentences in languages with more indexed sentences (in decreasing order: German, English, Czech, French, . . .).

In the first example, we request a path between two English sentences with opposite meaning. The algorithm finds a path, using the available sentences, which "slowly" changes the sentiment from negative to positive. We are also able to connect two recipes, gracefully switching the amounts and ingredients (example 2). In the third example, the two sentences are in a different language, French and English. The path uses mainly German sentences, probably because they are simply more frequent which makes it more likely to find a linguistically close one. In the the forth example on the right, we ask for the best path between two very different English sentences. Note that the algorithm switches many times between five different languages. This suggests that our embeddings are language agnostic. In most cases, there is only a small linguistic difference between subsequent sentences. Each sub-path between two sentences is also a shortest path.

We believe that the paths between sentences in the graph can shed new light on the question of semantic sentence similarity. While there are benchmarks such as STS (semantic textual similarity) that measure agreement with human similarity ratings, the question of what makes two sentences more or less similar remains poorly understood. It may well be that we can get a better understanding of the similarity between two sentences by observing these trajectories.

Query:	The law will be debated next week. (en)
0.159	The motion will be debated next week. (en)
0.175	Le texte sera examiné la semaine prochaine. (fr)
0.176	Das Gesetz soll in der kommenden Woche beschlossen werden. (de)
0.182	Das Gesetz soll kommende Woche erneut diskutiert werden. (de)
0.186	Parliament will debate the new legislation next week. (en)
0.188	It will be voted on next week. (en)
0.189	Debate on the resolution is expected next week. (en)
0.190	It is to be voted on next week. (en)
0.195	The chamber will begin debating legislation next week. (en)
0.196	Das Gesetz soll in der kommenden Woche verabschiedet werden. (en)
0.197	Sopimus vahvistetaan ensi viikolla. (fi)
0.201	Darüber wird in der kommenden Woche entschieden. (de)
0.202	Entschieden wird in der kommenden Woche. (de)
0.205	They will vote on the matter next week. (en)
0.206	Raportti julkaistaan ensi viikolla. (fi)
0.207	The law is expected to be ratified next week. (en)
0.207	The matter will be heard next week. (en)

Figure 5: Examples of multilingual paraphrases found in the public WMT news texts (1st column: multilingual distance).

4.4 Paraphrasing

Another application of the multilingual joint embedding is paraphrasing: we simply need to consider the k-nn sentences and threshold the distance. On the one hand, we cannot guarantee that close sentences exist, i.e. to find paraphrases for all possible queries, since we do not use a generative model like most other approaches to paraphrasing. On the other hand, we can be sure to produce valid *"existing"* sentences (given of course that the indexed corpus is not too noisy). Figure 5 shows an example of paraphrasing. The k nearest neighbors include sentences in other languages, which are all valid translations. One can clearly see that the differences between the sentences go well beyond changing isolated words with synonyms.

4.5 Code switching

In many NLP applications the language may change within a short section, sentence or even within words, e.g. for citations, foreign named entities, loanwords or when multilingual people communicate on social media. Such code-switching—i.e. the tendency for multilingual speakers to alternate between multiple languages or language varieties—poses important challenges for NLP systems: traditional monolingual techniques quickly deteriorate with input from mixed languages. Even for well-known problems such as POS-tagging and language identification, which the community often considers "solved", performance deteriorates proportional to the degree of code-switching in the data (Aguilar et al. 2018).

Our shared BiLSTM encoder is jointly trained on all twenty-one languages, which should allow for it to gracefully handle code-switching. It was already observed in the framework of multilingual NMT that such encoders are able to handle code-switching within sentences (Johnson et al. 2016). We are not aware of a well-established test set to evaluate code-switching at the sentence level, and it is not straightforward to construct one. Instead, we push this idea to the limit and test sentences which combine words of up to seven languages. Figure 6 part 2 shows several sentences in which we have replaced the English words arbitrarily with words in the other known languages. For each sentence the original English sentence was still the closest in the huge k-nn graph of 566M sentences. In reality, code-switching is unlikely to occur in such extreme combinations, and will be restricted to two or three languages at most. The examples illustrates how robust the approach is to code-switching, which makes it an interesting direction for learning representations on noisy user-generated data and real-world use-cases.

Src/Tgt	cs	de	en	es	fr	avg
cs	–	0.83	0.99	0.79	0.83	0.86
de	0.75	–	0.99	0.79	1.07	0.90
en	0.63	0.79	–	0.75	0.67	0.71
es	0.83	0.79	0.99	–	0.91	0.88
fr	0.79	1.03	0.83	0.75	–	0.85
avg	0.75	0.86	0.95	0.77	0.87	**0.84**

Table 2: Europarl 2009 news test set: average similarity search error rates in percent for all language pairs.

5 Quantitative Experiments

In this section, we provide a quantitative evaluation of several important properties of multilingual sentence embeddings. We first give baseline results of the proposed approach for performing multilingual similarity search on in-domain data, and compare our approach to published results for the reconstruction of the UN corpus. We subsequently analyze the transfer to an informal domain and new languages.

5.1 Multilingual similarity search

We perform a first evaluation with multilingual similarity search on the WMT 2009 news test set of 5,000 sentences which are 5-way parallel. For each sentence, we search the closest one in a different language. If this sentence is not identical to the official translation, an error is counted.

Our results suggest that the similarity error rate is very low and remarkably homogeneous for all language pairs. We achieve an average error rate over the 20 language pairs below 0.9% (see Table 2). Note that the table is not symmetric, e.g. encoding sentences in Czech and searching for the closest English sentence yields an error rate of 0.99%, while we obtain 0.63% in the reverse setting.

5.2 Reconstructing the United Nations corpus

In order to compare the performance of our sentence embeddings with published results, we also report the reconstruction error of the whole UN corpus of 11.3M sentences

1) Query in all nine trained languages :

en	Can you buy a home for less than half a million in San Francisco ?
de	Können Sie ein Haus für weniger als eine halbe Million in San Francisco kaufen?
da	Kan du købe et hjem for mindre end en halv million i San Francisco?
nl	Kun je een huis kopen voor minder dan een half miljoen in San Francisco?
fr	Pouvez-vous acheter une maison pour moins d'un demi-million à San Francisco?
es	¿Puedes comprar una casa por menos de medio millón en San Francisco?
it	Puoi comprare una casa per meno di mezzo milione a San Francisco?
pt	Você pode comprar uma casa por menos de meio milhão em San Francisco?
fi	Voitko ostaa talon vähemmän kuin puoli miljoonaa san franciscoa?

2) Multiple languages in one sentence (code-switching)

mix ¿Puedes comprar (es) een huis (nl) for mindre end (da) medio millón (es) in San Francisco kaufen? (de)

mix Voitko (fi) comprar (es) ein (de) huis (nl) per meno (it) than (en) en halv (da) miljoen (nl) a San Francisco?

Figure 6: Example of multilingual queries. The original English sentence was translated by an online MT system. 1) Query in eight trained languages; 2) mixing multiple languages within one sentence (code-switching). In all examples the closest sentence in the multilingual space of 566M is the original English sentence.

(Ziemski, Juncys-Dowmunt, and Pouliquen 2016), as defined by (Guo et al. 2018). We follow the methodology of that work and report precision at 1 for textual matches. These results are summarized in Table 3. Our method outperforms the previously published results by a large margin. The method proposed in (Guo et al. 2018) learns joint sentence embeddings for two languages only, usually English and a foreign language. This does not guarantee that sentences in two foreign languages, i.e. French and Spanish, are also close in the embedding space. The authors do not report results for that language pair. We also achieve a very high precision of 83.65 for that language pair. Finally, we tested the generalization performance of our approach on a language which was not used during training: Russian, an Eastern Slavic language using the Cyrillic script. We achieve a precision of about 60 for all language pairs with Russian. This is quite remarkable, given the fact that we used no Russian resources at all to train the sentence embeddings, not even monolingual ones.

5.3 Domain transfer to informal language

The results in Table 2 and 3 have been obtained on a test set of the same domain as the training data. In practice, this is rarely the case, so it is important to know how well the system generalizes to a different type of domain.

In need of a freely available corpus that provides texts in English, from a different domain than formal parliament language, and with high quality translations to many languages, we decided to use the Tatoeba corpus.[7] Tatoeba is a collection of English sentences and community-provided translations into more than 300 languages. However, the amount of available data varies a lot, from a couple of sentences (e.g. Sinhala), to several hundreds of thousands (e.g. Turkish, Russian, Italian or Japanese). More than 1 000 sentences

[7]https://tatoeba.org/eng/

	(Guo et al. 2018)	Proposed method
EN-FR	48.90	**83.30**
EN-ES	54.94	**85.40**
ES-FR	n/a	**83.65**
EN-RU	n/a	*65.84*
ES-RU	n/a	*62.80*
FR-RU	n/a	*59.89*

Table 3: Results on UN corpus reconstruction for trained languages, and zero-shot generalization to Russian (P@1).

are available for 72 languages, and at least 100 sentences for 112 languages. This makes Tatoeba a very interesting resource to evaluate highly multilingual sentence embeddings. We performed the following pre-processing steps:

- exclude sentences with WEB addresses or emails;
- keep only sentences with at least three words before tokenization;
- remove duplicates, either in source or target.

In Table 4, we first provide the similarity error for all the languages on which our system was trained on. All the results reported in this paper use a test set of 1 000 sentences of the Tatoeba corpus. We will make this test set available to encourage the publication of comparative results. It is important to note that the English source texts are not the same for all the target languages and the numbers are not necessarily comparable between the languages.

The average error rate is higher than on the Europarl test set (cf. Table 2). This can probably be explained by the fact that sentences in the Tatoeba corpus are shorter, with an average of 7 words compared to Europarl's average of more

Germanic				Romance					Hellenic
de	da	nl	sv	es	fr	it	pt	ro	el
2.8	7.7	6.3	6.8	5.8	3.3	5.5	4.1	10.7	15.4

Slavic					Baltic		Uralic		
bg	cs	pl	sk	sl	lt	lv	et	fi	hu
17.5	12.8	13.1	11.6	15.2	14.6	14.0	15.4	12.4	18.7

Table 4: Tatoeba corpus: similarity error rates between English and the twenty-one trained languages.

than 20 words. Short sentences are easier to confuse, as they may differ by only a few words. It was also observed in neural machine translation that LSTMs tend to perform best on sentences with a similar length distribution as the training corpus, see for instance (Kocmi and Bojar 2017). We have trained the multilingual model on the proceedings of the European Parliament only, which consists of rather formal and polite language and so is from a different domain. As an example, it can be observed that the second person singular is rarely used in morphologically rich languages like German (*"du"*) or French (*"tu"*). This is typically not the kind of language we are faced with in many multilingual NLP application, in particular when social media are involved. We provide preliminary experimental evidence that our multilingual sentence encoder generalizes to a non-formal language as well. Training on parallel data of multiple sources is likely to improve the generalization performance of the multilingual embeddings. This is left for future research.

5.4 Zero-shot transfer to new languages

We now turn to the question of generalization with respect to languages that were never seen by the system. We consider the zero-shot setting, i.e. no resources for these languages have been used during training. We have shown in section 4.1 that many BPE tokens of our 40k joint vocabulary are shared among all languages. Our hope is that these generic BPE building blocks, together with our single shared encoder that is trained on several languages, will allow our system to handle additional languages of the same family, but that have not at all been seen during training. It is important to point out that we are using no resources whatsoever from these additional languages, not even monolingual texts. We are not aware of previous work in this direction. This is as a pilot study of the generalization capacities of a shared BiLSTM encoder trained simultaneously on several languages, a useful future direction for the fields of historical and comparative linguistics (Jaeger 2018).

Table 5 shows the similarity error rates on eleven generalization languages, with 1 000 sentences each. Thus, a random choice of the closest sentence would yield and error rate of 99.9%. The two Germanic languages, Afrikaans and Norwegian, achieve error rates below 50%. Norwegian Bokmål (no), the most used written form, performs significantly better, with an error rate of 30%. For Romanic languages, we observe reasonable performance on Catalan (ca=47.7%) and

Germanic			Romance			Slavic				
af	no	nb	ca	cbk	gl	bs*	hr	mk*	ru*	sr*
49.3	30.7	43.9	47.7	50.9	20.1	36.4	42.0	45.4	49.2	44.5

Table 5: Tatoeba corpus zero-shot transfer: similarity error between English and languages never seen during training.

Chavacano (cbk=50.9%). Galician achieves an similarity error rate of 20%. Finally, we tested five languages which belong to the family of Slavic languages: Bosnian, Croatian, Macedonian, Russian and Serbian. All achieve similarity search error rates of roughly 40%, the best being Bosnian with 36.4%, and Russian achieving the highest error rate with 49.2%. This can be compared to our results for English-Russian on the UN corpus: an error rate below 35% on 11.3M sentences (see Table 3). This clearly shows that the big difference in the length distribution of the training and testing corpus seems to have an important impact, or more generally speaking, the domain mismatch.

As an ablation experiment, we tested five languages that use a Latin script, but which are not at all related to the ones the system was trained on: Turkish (Turkic language); Vietnamese, Indonesian and Tagalog (Austronesian languages); and Esperanto (artificial language). The similarity error is higher than 95% for Turkish and the three Austronesian languages, as expected. These languages have no linguistic similarities to the languages our encoder is trained on and we can not expect that the encoding of a sentence is closed to the English translation. Esperanto is an artificial language, but has some overlap with European languages. This leads to a similarity error rate of 71%.

6 Conclusion

We have trained a single BiLSTM sentence encoder on all twenty-one languages of the Europarl corpus using a small joint 40k BPE vocabulary. We embedded more than 500 million news sentences in seven languages in the same joint space and calculated the k-nn graph over the *"linguistically closest"* sentences. Our experiments illustrated several important applications: large-scale paraphrasing, arbitrary code-switching between many languages, and the capacity to handle similar languages without the need to use any resource in that language. An analysis of the shared BPE vocabulary allowed us to recover all language families. Additionally, we introduced a new technique for analyzing the semantic similarity of sentences by studying the path that connects them in the k-nn graph. The source code, trained networks, and the whole multilingual k-nn graph are freely available in the LASER toolkit.[8] We believe that these graphs have many applications, in particular when scaling up to more languages and sentences.

Acknowledgments

We would like to thank Mikel Artetxe and Daniel Haziza for help with the implementation of the tools.

[8] https://github.com/facebookresearch/LASER

References

Aguilar, G.; AlGhamdi, F.; Soto, V.; Diab, M.; Hirschberg, J.; and Solorio, T. 2018. Overview of the CALCS 2018 Shared Task: Named Entity Recognition on Code-switched Data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*.

Artetxe, M., and Schwenk, H. 2018. Margin-based parallel corpus mining with multilingual sentence embeddings. In *arXiv:1811.01136*.

Bethard, S.; Carpuat, M.; Apidianaki, M.; Mohammad, S. M.; Cer, D.; and Jurgens, D. 2017. Semeval-2017.

Bouamor, H., and Sajjad, H. 2018. H2@BUCC18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In *BUCC*.

Cer, D.; Yang, Y.; yi Kong, S.; Hua, N.; Limtiaco, N.; John, R. S.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; Sung, Y.-H.; Strope, B.; and Kurzweil, R. 2018. Universal sentence encoder. In *arXiv:1803.11175*.

Chandar, S.; Khapra, M. M.; Ravindran, B.; Raykar, V.; and Saha, A. 2013. Multilingual deep learning. In *NIPS DL wshop*.

Chen, X.; Sun, Y.; Athiwaratkun, B.; Cardie, C.; and Weinberger, K. 2016. Adversarial deep averaging networks for cross-lingual sentiment classification. *arXiv:1606.01614*.

Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; and Bordes, A. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, 670–680.

Conneau, A.; Lample, G.; Rinott, R.; Williams, A.; Bowman, S. R.; Schwenk, H.; and Stoyanov, V. 2018. XNLI: Evaluating cross-lingual sentence representations. In *EMNLP*.

Edunov, S.; Ott, M.; Auli, M.; and Grangier, D. 2018. Understanding back-translation at scale. In *EMNLP*, 489–500.

España-Bonet, C.; Ádám Csaba Varga; Barrón-Cedeño, A.; and van Genabith, J. 2017. An empirical analysis of nmt-derived interlingual embeddings and their use in parallel sentence identification. *IEEE Journal of Selected Topics in Signal Processing* 1340–1348.

Ge, T.; He, K.; Ke, Q.; and Sun, J. 2013. Optimized product quantization for approximate nearest neighbor search. In *CVPR*, 2946–2953.

Grégoire, F., and Langlais, P. 2017. Bucc 2017 shared task: a first attempt toward a deep learning framework for identifying parallel sentences in comparable corpora. In *BUCC*, 46–50.

Guo, M.; Shen, Q.; Yang, Y.; Ge, H.; Cer, D.; Abrego, G. H.; Stevens, K.; Constant, N.; Sung, Y.-H.; Strope, B.; and Kurzweil, R. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *arXiv:1807.11906*.

Hassan, H., and et al. 2018. Achieving human parity on automatic chinese to english translation. In *arXiv:1803.05567*.

Hermann, K. M., and Blunsom, P. 2014. Multilingual models for compositional distributed semantics. In *ACL*, 58–68.

Jaeger, G. 2018. Computational historical linguistics. *arXiv journal preprint* 1805.08099.

Jegou, H.; Douze, M.; and Schmid, C. 2011. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence* 33(1):117–128.

Johnson, J.; Douze, M.; and Jégou, H. 2017. Billion-scale similarity search with GPUs. *arXiv:1702.08734*.

Johnson et al., M. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. In *arXiv:1611.04558*.

Kiros, R.; Zhu, T.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. In *NIPS*, 3294–3302.

Klementiev, A.; Titov, I.; and Bhattarai, B. 2012. Inducing crosslingual distributed representations of words. In *Coling*.

Kocmi, T., and Bojar, O. 2017. Curriculum learning and minibatch bucketing in neural machine translation. In *WMT*.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*.

Mogadala, A., and Rettinger, A. 2016. Bilingual word embeddings from parallel and non-parallel corpora for cross-language classification. In *NAACL*, 692–702.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *NAACL*, 2227–2237.

Pham, H.; Luong, M.-T.; and Manning, C. D. 2015. Learning distributed representations for multilingual text sequences. In *Workshop on Vector Space Modeling for NLP*.

Schwenk, H., and Douze, M. 2017. Learning joint multilingual sentence representations with neural machine translation. In *ACL workshop on Representation Learning for NLP*.

Schwenk, H., and Li, X. 2018. A corpus for multilingual document classification in eight languages. In *LREC*, 3548–3551.

Schwenk, H. 2018. Filtering and mining parallel data in a joint multilingual space. In *ACL*, 228–234.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural machine translation of rare words with subword units. In *ACL*, 1715–1725.

Subramanian, S.; Trischler, A.; Bengio, Y.; and Pal, C. J. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. In *ICLR*.

Tiedemann, J. 2012. Parallel data, tools and interfaces in opus. In *LREC*.

Zhou, X.; Wan, X.; and Xiao, J. 2016. Cross-lingual sentiment classification with bilingual document representation learning. In *ACL*.

Ziemski, M.; Juncys-Dowmunt, M.; and Pouliquen, B. 2016. The united nations parallel corpus v1.0. In *LREC*.