Are the Best Multilingual Document Embeddings simply Based on Sentence Embeddings?

Sonal Sannigrahi,[1] **Josef van Genabith,**[1,2] **Cristina España-Bonet**[2]
[1]Saarland University, Saarland Informatics Campus, Germany
[2]German Research Center for Artificial Intelligence (DFKI)
sosa00001@stud.uni-saarland.de
{cristinae, Josef.Van_Genabith}@dfki.de

Abstract

Dense vector representations for textual data are crucial in modern NLP. Word embeddings and sentence embeddings estimated from raw texts are key in achieving state-of-the-art results in various tasks requiring semantic understanding. However, obtaining embeddings at the document level is challenging due to computational requirements and lack of appropriate data. Instead, most approaches fall back on computing document embeddings based on sentence representations. Although there exist architectures and models to encode documents fully, they are in general limited to English and few other high-resourced languages. In this work, we provide a systematic comparison of methods to produce document-level representations from sentences based on LASER, LaBSE, and Sentence BERT pre-trained multilingual models. We compare input token number truncation, sentence averaging as well as some simple windowing and in some cases new augmented and learnable approaches, on 3 multi- and cross-lingual tasks in 8 languages belonging to 3 different language families. Our task-based extrinsic evaluations show that, independently of the language, a clever combination of sentence embeddings is usually better than encoding the full document as a single unit, even when this is possible. We demonstrate that while a simple sentence average results in a strong baseline for classification tasks, more complex combinations are necessary for semantic tasks. Our code is publicly available.[1]

1 Introduction

Semantic representations, especially embeddings, are crucial for natural language processing (NLP). In fact, the field has exploded since the success of dense word embeddings (Mikolov et al., 2013). For some tasks like finding semantic or syntactic relations among words, high quality word embeddings are enough. Other tasks, like question classification or paraphrase detection, benefit from sentence embeddings. Finally, lots of tasks deal with documents: summarisation, document classification, question answering, etc. Document representations are difficult to be learned, especially multilingually, given the amount of available training data and the length of each training instance.

For these reasons, document embeddings usually resort to sentence embeddings. Since some of the state-of-the-art techniques for language modelling and sentence embeddings are based on self-attention architectures such as BERT (Devlin et al., 2019), and self-attention scales quadratically with the input length, one cannot afford arbitrarily long inputs. Training is usually constrained to input fragments up to 512 tokens (subunits). This limit goes well beyond an average sentence length and can cover several paragraphs. However, full documents can be significantly longer. The average length of a Wikipedia article in English is 647 words (not subunits) for example,[2] and the average for two of the tasks that we consider in this work, document alignment and ICD code classification, is around 800 words, with documents up to 40k words.

In order to be able to process long inputs, more efficient architectures such as Linformer (Wang et al., 2020), Big Bird (Zaheer et al., 2020) or Longformer (Beltagy et al., 2020) implement sparse attention mechanisms that scale linearly instead of quadratically. These architectures accept at least 4096 input tokens. With this length, one can embed most Wikipedia articles, news articles, medical records, etc. These architectures are available as pre-trained models in English[3] and can be fine-tuned for NLP tasks such as document classification, question answering or summarisation. How-

[1]https://github.com/sonalsannigrahi/Document_Embeddings

[2]https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia
Consulted on Feb. 2023.

[3]https://huggingface.co

Findings of the Association for Computational Linguistics: EACL 2023, pages 2306–2316
May 2-6, 2023

ever, multilingual or non-English versions are rare. For most languages, it is not just a matter of training a model from scratch, but the amount of documents is just not enough to train high quality models.

LASER (Artetxe and Schwenk, 2019; Heffernan et al., 2022), Sentence BERT (Reimers and Gurevych, 2019, 2020) and LaBSE (Feng et al., 2022) are representative and state-of-the-art models which largely adapt language models to be used as task-independent sentence representations. These models are available as pre-trained models and, contrary to the long sequence models introduced before, they are multilingual. LASER, which is not transformer-based, allows longer inputs.

These observations explain why the two main approaches to obtain multilingual (or non-English) document embeddings are simply (*i*) truncating the input to 512 tokens and feeding it into a sentence-level encoder or (*ii*) splitting the document in shorter fragments and then combine their embeddings. There are few works that do a systematic comparison among methods. Park et al. (2022) perform a systematic study for document classification in English and found that the most sophisticated models such as Longformer do not always improve on a baseline that truncates the input to fit it into a fine-tuned BERT. The results mostly depend on how the information is distributed along a document and therefore varies from dataset to dataset.

In this work we explore multilingual document-level embeddings in three tasks in detail: *document alignment*, a bilingual semantic task; *ICD code (multi-label) classification* in 2 languages; and *cross-lingual document classification* in 8 languages. We compare input token number truncation, sentence averaging as well as some simple windowing and in some cases new augmented and learnable approaches. Our results show that a simple sentence average is a very strong baseline, even better than considering the whole document as a single unit, but that positional information is needed when the distribution of information across a document is not uniform.

2 Related Work

Word embeddings have been exceptionally successful in many NLP applications (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017). Subsequent works developed methods to learn continuous vector representations for longer sequences such as sentences or even documents. Skip-thought embeddings (Kiros et al., 2015) train an encoder–decoder architecture to predict surrounding sentences. Conneau et al. (2017) showed that the task on which sentence representations are learnt significantly impacts their quality. InferSent (Conneau et al., 2017), a Siamese BiLSTM network with max pooling, and Universal Sentence Encoder (Cer et al., 2018), a transformer-based network, are trained over the SNLI dataset which is suitable for learning semantic representations (Bowman et al., 2015).

These methods primarily work on a single language but as multilingual representations have attracted more interest, sentence-level embeddings have been extended to obtain a wider language coverage. Artetxe and Schwenk (2019) (LASER) learn joint multilingual sentence representations for 93 languages based on a single BiLSTM encoder with a shared BPE vocabulary trained on publicly available parallel corpora. However, this architecture was shown to underperform in high-resource scenarios (Feng et al., 2022). LASER is especially interesting for our work as, being LSTM-based, it does not have the 512-length constraint. Li and Mak (2020) introduce T-LASER, which is a version of LASER that uses a transformer encoder in place of the original bidirectional LSTM. However, this model was tested only on the Multilingual Document Classification (MLDoc) corpus (Schwenk and Li, 2018), which does not have significantly long documents. Similarly, Reimers and Gurevych (2019) (sBERT in the following) extended a transformer-encoder architecture, BERT, by using a Siamese network with cosine similarity for contrastive learning in order to derive semantically meaningful sentence representations. More recently, Feng et al. (2022) (LaBSE) explored cross-lingual sentence embeddings with BERT by introducing a pre-trained multilingual language model component and show that on several benchmarks, their method outperforms many state-of-the-art embeddings such as LASER.

While sentence-level representations have been widely explored in literature, document-level representations are less well-explored. The earliest approaches in learning document-level vector representations included an extension of the Word2Vec algorithm named Doc2Vec (Le and Mikolov, 2014) with two variants proposed, a bag-of-words and a skip-gram based model. However, while these methods worked well at the word-level,

the document-level counterpart led to issues in scaling due to large vocabulary sizes (Lau and Baldwin, 2016). Due to these limitations, further works have attempted to improve the computational bottlenecks involved with training on long sequences such as documents. Linformer (Wang et al., 2020) is a transformer-based architecture with linear complexity due to a sparse self-attention mechanism making it significantly more memory- and time-efficient in comparison with the original transformer (Vaswani et al., 2017). Works such as Big Bird (Zaheer et al., 2020) and Longformer (Beltagy et al., 2020) introduced a sparse attention mechanism and localised global attention respectively. BigBird is able to handle sequences of up to 4,096 tokens and Longformer scales linearly with the sequence length, with experiments on sequences of length upto 32,256. To the best of our knowledge, to date not much has been done to extend them beyond English. Shen (2021) and Romero (2022) made available Chinese and Spanish Longformer models, respectively, while Sagen (2021) trained a multilingual version starting from a RoBERTa checkpoint and not from scratch. We use Longformer as a comparison system in our experiments but we do not consider the multilingual model given that multilinguality was achieved by fine-tuning on question answering data and we do not explore this task.

3 Sentence Embeddings

We use three multilingual sentence-level embedding models that cover different languages, architectures and learning objectives:

LASER (Schwenk and Douze, 2017; Artetxe and Schwenk, 2019) uses max-pooling over the output of a stacked BiLSTM-encoder. The encoder is extracted from an encoder–decoder machine translation setup trained on parallel corpora over 93 languages. Since it is not based on transformers but on LSTMs, the maximum number of input tokens can in principle be arbitrary and is set to 12,000.

LaBSE Feng et al. (2022) train a multilingual BERT-like model with a masked LM and translation LM objective functions. A dual-encoder transformer is initialised with the model and fine-tuned on a translation ranking task. The final model covers 109 languages. The maximum number of input tokens is 512.

sBERT Reimers and Gurevych (2019) use the output of BERT-base with mean pooling to create a fixed-size sentence representation. A Siamese-BERT architecture trained on NLI is used to obtain the final sentence-embedding model. The maximum number of input tokens is 512, with a default value of 128. We use the multilingual version (Reimers and Gurevych, 2020).

4 Document Embeddings

We divide our approaches to build document embeddings into three families: in *(i) Document Excerpts,* we feed token sequences as they are directly into LASER, LaBSE and sBERT to obtain a document-level representation, in *(ii) Sentence Weighting Schemes,* we divide documents into sentences represented using base sentence embeddings and then explore different combination and weight strategies to obtain document embeddings, in *(iii) Windowing Approaches,* we study different distributions to learn document-level positional and semantic information.

(i) Document Excerpts

All Tokens: The full document is fed into the system (no truncation). We explore this option only with LASER which does not have the 510-token-length restriction[4] and when possible (English, Spanish and Chinese) with Longformer.

Top-N Tokens: The document is truncated to the first $n = 510$ tokens.

Bottom-N Tokens: The last $n = 510$ tokens are fed into the system.

Top-N + Bottom-M Tokens: We select $N = 128$ and $M = 382$ to use the first N and last M tokens of the documents. These values are based on empirical explorations by Sun et al. (2019).

(ii) Sentence Weighting Schemes

Sentence Average: Each base sentence embedding (obtained with LASER, LaBSE or SBERT) is given a uniform weight. This computes the vanilla average embedding vector of all sentences in the document.

[4]That is the maximum length of tokens accepted by transformer-style embedding models, 512 without the `[CLS]` and `[SEP]` tokens.

Top/Bottom-Half Average: Only the top (bottom) half of the sentences in the document are considered for averaging.

TF-IDF Weights: We compute TF-IDF scores for all terms in a document, and average their values at sentence level. The base sentence embeddings (LASER, LaBSE, SBERT) are then weighted by the normalised value of the TF-IDF averages. Following Buck and Koehn (2016b), we use different TF-IDF computations based on variations of term frequency tf and inverse document frequency idf definitions. For words w in a document d belonging to a collection D we report results using:

$$tf_2(w,d) = \text{freq}(w,d) \quad (1)$$

$$tf_4(w,d) = 0.4 + 0.6\frac{\text{freq}(w,d)}{\max_{\tilde{w}}\text{freq}(\tilde{w},d)} \quad (2)$$

$$idf_4(w,d) = \log(1 + \frac{|D|}{df(w,|D|)}), \quad (3)$$

with $df(w,D) = |\{d \in D | w \in d\}|$, and

$$tf_i idf_j(S_k) = \frac{\sum_{w \in S_k} tf_i(w,d) idf_j(w,d)}{\#w_k}, \quad (4)$$

where S_k is a sentence in a given document d, and $\#w_k$ is the number of words in sentence S_k.

The weights of these models are fixed for the static tasks and used as initialisation when training a classifier.

(iii) Windowing Approaches

TK-PERT: Thompson and Koehn (2020) introduced a windowing approach that weights the contribution of each sentence according to the modified PERT function (Vose, 2008) and a down-weighting function for boilerplate text. The latter was introduced to deal with webpages but it can be ignored for other types of documents. The smoothed overlapping windowing functions based on a cache of the PERT distribution (PERT-cache) encode fine-grained positional information into the resultant document vector.

A document with N sentences $S_{i|i\in\{0,\ldots,N-1\}}$ is split uniformly into J parts and the final representation D for a document is given by a concatenation of normalised position-weighted (via PERT) sub-vectors where each sub-vector D_j is

$$D_j = \sum_{n=0}^{N-1} \text{emb}(S_n) P_j(n) B(S_n), \quad (5)$$

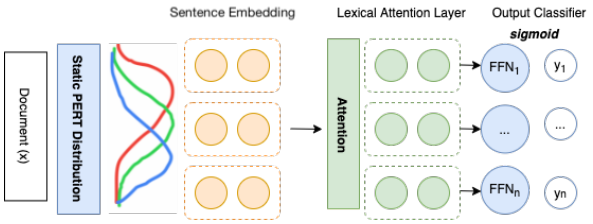

Figure 1: ATT-PERT model for classification. A static modified PERT distribution is used to extend the sentence embeddings to documents. Afterwards, an attention-weighted classifier is learnt.

emb is the (LASER, LaBSE, SBERT) embedding of sentence n, P is the modified PERT function for part j and B is a boilerplate function if there is one. In cases when no boilerplate text is present, we set it to 1.

Following Thompson and Koehn (2020) setting for the modified PERT distribution, we use $J = 16$ and set its shape parameter to $\gamma = 20$.

TF-PERT: is a new extension of TK-PERT to further incorporate semantics. PERT focuses on positional information encoded in the document while TF-IDF focuses on the semantic information, therefore a combined metric would likely be able to consider both features. We combine the two contributions with a multiplication at sentence level:

$$D_j = \sum_{n=0}^{N-1} \text{emb}(S_n) P_j(n) B(S_n) tfidf(S_n), \quad (6)$$

where we use the same notation as in Eqs. 4 and 5.

ATT-PERT: is a new extension of TK-PERT to further incorporate a global learnable attention. Figure 1 illustrates the basic architecture. The PERT distribution encodes global positional information of the document. By adding an attention layer over it, we introduce a *global attention* that weights the different parts of the document and that is combined with the standard *local attention* at word level performed by the sentence encoder. Mathematically,

$$D_j = \sum_{n=0}^{N-1} \text{emb}(S_n) P_j(n) a_j(n), \quad (7)$$

where S_n refers to the sentence embedding that has been trained for a classification task and $a_j(n)$ is the respective global attention weight.

In TK-PERT, the static PERT distribution is multiplied by the fine-tuned sentence embeddings. In

	# Documents		Length	
	Train	Test	Avg.	Max.
Document Alignment, WMT2016				
English	349k	682k	737	43.3k
French	225k	522k	842	45.2k
Web Domains	49	203	-	-
Gold Pairs	1624	2402	-	-
Multi-label Classification, ICD Code Classification				
Spanish	1001	1600	792	4352
German	8385	407	876	2249
Document Classification, MLDoc				
English	10k	4000	275	576
German	10k	4000	342	675
French	10k	4000	445	782
Italian	10k	4000	376	765
Spanish	9458	4000	354	778
Japanese	10k	4000	327	897
Russian	5216	4000	235	967
Chinese	10k	4000	562	983

Table 1: Number of documents and average tokenised document length in sentencepiece units (prior to boilerplate downweighting for Document Alignment) for the three tasks used in the experiments.

contrast, in ATT-PERT, the distribution is multiplied with the embeddings prior to training a classifier without freezing the embedding layer, as this allows the positional weights in the PERT distribution to be trained for the specific task.

ATT-TF-PERT: is a new extension of TF-PERT to further incorporate a global learnable attention as in ATT-PERT. In this configuration, we learn combined TF-IDF-PERT weighted embeddings whose attention weights are further updated while training the classifier. We use the same *global attention* $a_j(n)$ as in ATT-PERT, however here it is multiplied with both the TF-IDF weight of the sentence $tfidf_j(w, S_n)$ as computed in the TF-IDF set up and the PERT distribution $P_j(n)$ as in TK-PERT:

$$D_j = \sum_{n=0}^{N-1} \text{emb}(S_n) P_j(n) a_j(n) tfidf(S_n). \quad (8)$$

5 Evaluation Tasks

We apply the different configurations discussed above across the following tasks:

Bilingual Document Alignment aims at aligning documents from two collections in language L1 and language L2 according to whether they are parallel or comparable. In our experiments, we use the data given for the WMT 2016 Shared Task on Bilingual Document Alignment to align French web pages to English web pages for a given crawled webdomain (Buck and Koehn, 2016a). In these experiments we do not perform any learning using the training data, but just estimate document-level semantic similarity between the pairs of documents in the test set. To compute this, we find the top K=32 candidate translations using approximate nearest neighbor search via FAISS[5] as in (Buck and Koehn, 2016a). We use cosine similarity to quantify semantic similarity on the document embeddings.

Multi-label ICD Code Classification aims at assigning one or more ICD-10 codes to medical-domain texts (electronic health records). Here there can be an arbitrary number of ICD-10 codes assigned to the input text. In particular, out of all the possible ICD-10 Codes, 4 account for more than 90% of the documents, making this an imbalanced classification task and leading to the 'tail end problem' (Chapman and Neumann, 2020). We use the CLEF eHealth 2019 task for German non-technical summaries (Neves et al., 2019) and CANTEMIST-CODING (Miranda-Escalada et al., 2020) for Spanish electronic health records. Here, we learn a weighted-attention classifier layer (Lee et al., 2022) on top of the base document embeddings consisting of a feed-forward neural network with a single hidden layer of 10 units.

Cross-lingual Document Classification aims at classifying documents in a set of predefined categories in a language (usually English) and then transfer the model to unseen languages. We use the MLDoc dataset for this purpose (Schwenk and Li, 2018). The corpus contains 1,000 development documents and 4,000 test documents in eight languages (English, German, French, Italian, Spanish, Japanese, Russian and Chinese), divided in four different genres with uniform class priors. For zero-shot transfer, we train a classifier on top of the multilingual document representations estimated as described in Section 4 by using only the English training data and the hyperparameters optimised in Artetxe and Schwenk (2019). Similar to the previous classification task, we use a feed-forward neural network with one hidden layer with 10 units.

[5] https://github.com/facebookresearch/faiss

	LASER	LaBSE	sBERT
All tokens	**81.2**$^{+0.3}_{-0.4}$	—	—
Top-510 tokens	70.8$^{+0.2}_{-0.3}$	71.2$^{+0.5}_{-0.4}$	72.3$^{+0.2}_{-0.4}$
Bottom-510 tokens	65.8$^{+0.5}_{-0.3}$	66.3$^{+0.7}_{-0.8}$	67.1$^{+0.6}_{-0.7}$
Top-128 + Bot-312	75.3$^{+0.5}_{-0.5}$	**76.1**$^{+0.3}_{-0.5}$	**74.2**$^{+0.3}_{-0.3}$
Sentence Average	81.8$^{+0.7}_{-0.5}$	83.4$^{+0.6}_{-0.5}$	82.3$^{+0.4}_{-0.6}$
Top-Half Avg.	82.2$^{+0.3}_{-0.5}$	81.3$^{+0.6}_{-0.8}$	81.7$^{+0.7}_{-0.6}$
Bottom-Half Avg.	67.8$^{+0.8}_{-0.7}$	66.5$^{+0.4}_{-0.3}$	65.3$^{+0.5}_{-0.4}$
TF-IDF Weighted			
$tf_2 - idf_4$	80.2$^{+0.7}_{-0.4}$	80.5$^{+0.7}_{-0.6}$	79.3$^{+0.2}_{-0.4}$
$tf_4 - idf_4$	**86.3**$^{+0.5}_{-0.4}$	**87.2**$^{+0.3}_{-0.4}$	**85.4**$^{+0.6}_{-0.5}$
TK-PERT (Euclidean)	93.2$^{+0.7}_{-0.8}$	93.5$^{+0.6}_{-0.5}$	92.8$^{+0.5}_{-0.4}$
TK-PERT (cosine)	**96.4**$^{+0.6}_{-0.5}$	**94.2**$^{+0.5}_{-0.4}$	**95.3**$^{+0.8}_{-0.9}$
TF-PERT (cosine)	93.4$^{+0.5}_{-0.3}$	92.5$^{+0.3}_{-0.4}$	93.1$^{+0.4}_{-0.4}$

Table 2: Document recall on WMT-16 Shared Task on English–French document alignment. Best score for each family is in bold.

We use this classifier on top on the multilingual embeddings to evaluate the system on the remaining languages.

Table 1 shows the statistics for the datasets used in the three tasks as well as an average length of training instances in terms of sentencepiece tokens.[6] The average document length in the document alignment and ICD code classification tasks is larger than 512 tokens, making the usage of sentence embeddings alone insufficient. This is not the case for document classification, but we still consider it in order to compare the different approaches and add a highly multilingual setting.

6 Results and Discussion

Thompson and Koehn (2020) empirically obtained the best trade-off between accuracy and inference time when using PCA-reduced sentence embeddings of 128 dimensions in the bilingual document alignment task. We performed equivalent experiments with 128 and 256 dimensions for selected configurations in the three tasks and confirmed the trend. As we obtained no major gains in using more dimensions, we report all the results for the three tasks with 128-dimensional sentence embeddings.

We report confidence intervals at 95% confidence level using bootstrap resampling with 1000 samples for document alignment, 500 samples for ICD code classification and 1000 samples for document classification.

[6] https://github.com/google/sentencepiece

Bilingual Document Alignment quality ranges from 65% to 96% recall depending on the document embedding method. Table 2 shows the results obtained for all the configurations considered. A simple sentence average achieves a recall around 82% (depending on the sentence embedding used). When using LASER, the only method that allows the comparison, the recall with sentence average is larger but not statistically significantly over embedding the full document as a single unit (81.8% vs 81.2%). Taking a token-based excerpt of the document is 10 percentage points below sentence-averaging the same excerpt. The information in webpages seems to be more densely distributed towards the top of the page. Looking at the top-half versus the bottom-half of the sentences of the webpages, there is a 17% reduction in the scores obtained. In these unweighted and average configurations in both the token and sentence-based methods, we do not encode any positional information: sentence order and semantic relevance is not considered in the final document embeddings. However, intuitively, these factors are indicative of each sentence's contribution to the larger document embedding. In order to incorporate semantic relevance into our final embeddings, we consider the weighted average using TF-IDF. We explore several TF-IDF forms and obtain a difference of 7% on average among them. Table 2 shows the 2 most promising ones. With the best option ($tf_4 - idf_4$), TF-IDF weighting improves between 3 and 5 percentage points with respect to the sentence averaging which uses uniform weights. We use $tf_4 - idf_4$ for the next experiments when required as these formulae empirically performed the best. To include sentence order, we use the PERT-window based approach. TK-PERT outperforms all other methods by a margin of 11.7%. This result attests the relevance of contextual information, sentence order, and positional importance. Although we find improvements over the baseline models by introducing TF-IDF weights and the PERT distribution, a combination of the two in TF-PERT does not lead to further improvements.

The other dimension of the study, the particulars of sentence embeddings, is less important to the recall. LASER, LaBSE and sBERT achieve similar results. As we are working with French and English documents, both languages being high-resource, all base sentence embeddings are high-quality and therefore they do not impact the final

	LASER		LaBSE		sBERT	
	de	**es**	**de**	**es**	**de**	**es**
All tokens	$\mathbf{73.1}^{+0.5}_{-0.6}$	$\mathit{18.4}^{+0.4}_{-0.3}$	–	–	–	–
Top-510 tokens	$65.6^{+0.8}_{-0.7}$	$16.5^{+0.5}_{-0.8}$	$68.2^{+0.5}_{-0.5}$	$\mathit{19.2}^{+0.6}_{-0.4}$	$63.2^{+0.7}_{-0.8}$	$\mathit{18.3}^{+0.5}_{-0.6}$
Bottom-510 tokens	$67.8^{+0.4}_{-0.9}$	$17.5^{+0.4}_{-0.9}$	$66.7^{+0.8}_{-0.6}$	$17.4^{+0.7}_{-0.6}$	$61.5^{+0.8}_{-0.6}$	$16.8^{+0.5}_{-0.7}$
Top-128 + Bot-312	$66.4^{+0.8}_{-0.6}$	$17.2^{+0.8}_{-0.7}$	$\mathit{69.1}^{+0.7}_{-0.9}$	$18.7^{+0.7}_{-0.8}$	$\mathit{64.8}^{+0.7}_{-0.5}$	$17.5^{+0.6}_{-0.8}$
Sentence Average	$\mathit{72.1}^{+0.9}_{-0.8}$	$17.0^{+0.7}_{-0.6}$	$\mathbf{74.5}^{+0.8}_{-0.9}$	$\mathit{24.2}^{+0.8}_{-0.6}$	$\mathbf{68.9}^{+0.7}_{-0.6}$	$\mathit{20.3}^{+0.8}_{-0.4}$
Top-Half Avg.	$68.4^{+0.7}_{-0.9}$	$16.5^{+0.5}_{-0.6}$	$68.3^{+0.4}_{-0.8}$	$18.9^{+0.5}_{-0.5}$	$61.5^{+0.7}_{-0.6}$	$16.4^{+0.8}_{-0.6}$
Bottom-Half Avg.	$63.1^{+0.7}_{-0.6}$	$15.8^{+0.8}_{-0.7}$	$67.4^{+0.8}_{-0.9}$	$15.2^{+0.6}_{-0.7}$	$58.6^{+0.9}_{-0.8}$	$17.9^{+0.7}_{-0.6}$
TF-IDF Weighted	$65.3^{+0.5}_{-0.4}$	$\mathit{17.2}^{+0.7}_{-0.8}$	$68.2^{+0.9}_{-1.0}$	$19.2^{+0.9}_{-0.7}$	$63.2^{+0.6}_{-0.8}$	$18.3^{+0.7}_{-0.6}$
TK-PERT	$68.2^{+0.8}_{-0.6}$	$22.1^{+0.7}_{-0.4}$	$70.1^{+0.8}_{-0.7}$	$20.1^{+0.7}_{-0.4}$	$65.2^{+0.8}_{-0.6}$	$19.5^{+0.7}_{-0.8}$
TF-PERT	$68.5^{+0.4}_{-0.3}$	$23.4^{+0.6}_{-0.6}$	$68.6^{+0.3}_{-0.7}$	$21.3^{+0.5}_{-0.4}$	$65.4^{+0.6}_{-0.7}$	$18.7^{+0.4}_{-0.3}$
ATT-PERT	$\mathit{70.7}^{+0.7}_{-0.9}$	$\mathbf{32.2}^{+0.7}_{-0.4}$	$72.1^{+0.8}_{-0.6}$	$\mathbf{30.1}^{+0.7}_{-0.8}$	$\mathit{66.3}^{+1.1}_{-1.3}$	$\mathbf{27.4}^{+0.8}_{-0.7}$
ATT-TF-PERT	$70.3^{+0.5}_{-0.4}$	$31.4^{+0.8}_{-0.7}$	$\mathit{73.2}^{+0.4}_{-0.8}$	$29.7^{+0.6}_{-0.5}$	$66.1^{+0.9}_{-0.8}$	$27.1^{+0.5}_{-0.6}$

Table 3: F1 scores for the Multi-label ICD code classification task for German (de) and Spanish (es) documents. Best scores are in bold, and best scores per family are in italics.

model strongly in a consistent way.

Multi-label ICD Code Classification shows the same trend with respect to different sentence embeddings as above for German and Spanish, with a slight preference towards LaBSE embeddings. Table 3 shows the results for this task. There is a large discrepancy between the scores for the German and the Spanish datasets, as already noticed by the evaluations in the original corresponding shared tasks. The classification in Spanish achieves much lower results probably because of a very small training corpus. Our results indicate that the information is spread throughout documents in this case. The difference between only using the top of the document and only using the bottom part is small, and using the whole document either by sentence averaging or considering it a single unit is always better than any of its parts at a 95% significance level. Semantic (TF-IDF) and positional (TK-PERT) information is less relevant. For the German task, either considering the full document as a whole (*All tokens*) or averaging all the sentences gives the highest performance. For the Spanish task, even with a very low overall quality, learning specific weights for different parts of the document (ATT-PERT) boosts the quality. Comparing ATT-PERT with TK-PERT, we find that the trainable alternative performs better for all languages and base embeddings considered, however, the improvements are not statistically significant for all base embeddings in the case of German. In general, the windowing approaches that combine semantics with position (TF-PERT and ATT-TF-PERT) do not perform significantly better than the pure positional methods (TK-PERT and ATT-PERT). This can be explained by looking a concrete example. Figure 2 shows the distribution of weights across a document from the CANTEMIST health record corpus for 8 configurations based on LASER embeddings. The example shows that the effect of the $tfidf$ component in ATT-TF-PERT (configuration 7) is equivalent to move weight mass from ATT-PERT (configuration 6) into TF-IDF (configuration 3). When this happens, the result is a score in the middle of the way between ATT-PERT and TF-IDF. In this document, a medical diagnostic evaluation is detailed and includes patient information, past diagnoses, family medical history, as well potential evolution of the disease. We observe that while the 'Sentence average' configuration places largely equivalent weights on all the sentences, the TF-IDF weights place more emphasis on the beginning and end of the document which stores information about the patient and the evolution of the disease respectively. This behaviour is similar to the one exhibited by the PERT family of methods: the weight pattern observed for configurations 3-7 remain quite consistent but vary in their intensity.

Cross-lingual Document Classification data allows us to test the embedding methods on 8 languages (Table 4). The languages belong to three families, Indo-European (Germanic, Romance and Slavic), Japonic and Sino-Tibetan. All languages are high-resourced and included in our pre-trained sentence representation models. MLDoc documents are shorter than 1,000 tokens with an average length of 275 tokens for English and 562 for Chinese; the other languages stay in the middle.

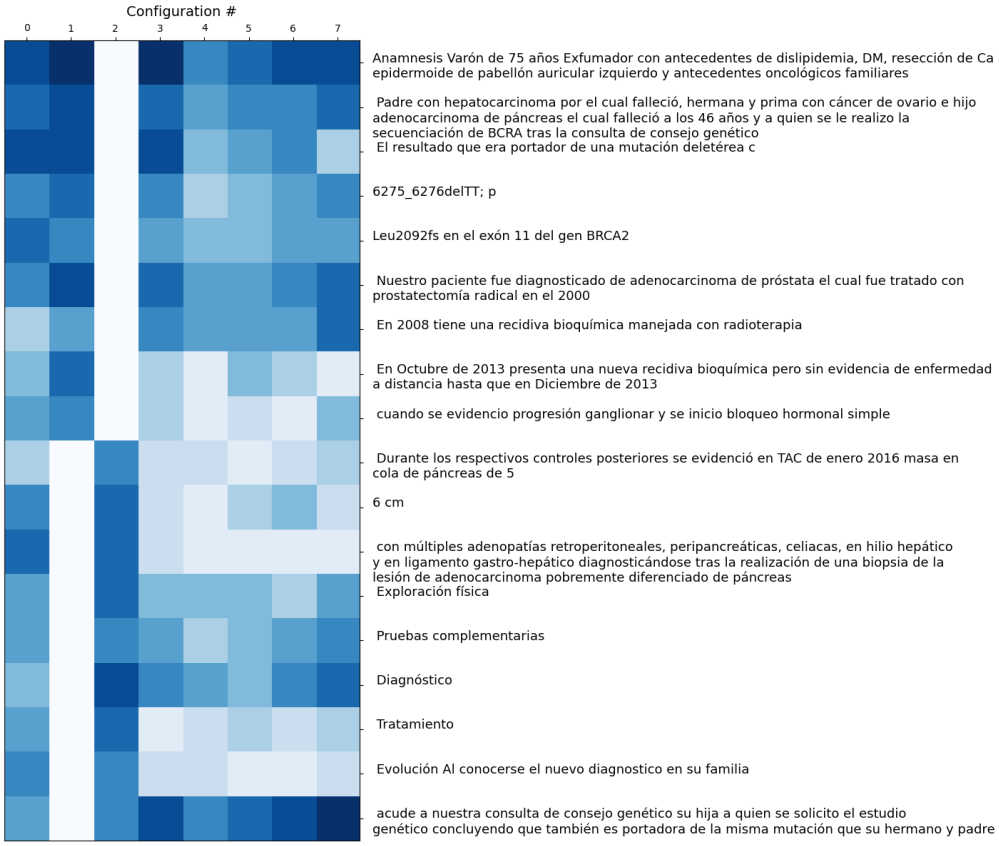

Figure 2: Sentence weights for an example document with LASER embeddings and the configurations: 0-Sentence average, 1-Top-Half, 2-Bottom-Half, 3-TF-IDF weighted, 4-TK-PERT, 5-TF-PERT, 6-ATT-PERT, 7-ATT-TF-PERT.

Given that length, the methods that use different 510-sized excerpts of the documents do not differ much as all the excerpts are —for most of the documents— the same.

Accuracies in Table 4 show that the documents convey slightly more meaning at the top part than at the bottom (*Top-Half Avg.* vs *Bottom-Half Avg.*). The sentence average is a very strong baseline and, for half of the languages (English, German, Russian and Chinese), this is statistically significantly better at 95% confidence level than treating the document as a single unit with LASER. The TF-IDF version is worse than the simple sentence average except for Japanese. Japanese has the lowest accuracy for all the languages and a high difference between the information at the top and the bottom of its documents. In general, position (TK-PERT) is more important than semantics (TF-IDF) and learning task-specific weights (ATT-PERT) further increases accuracy. Additional experiments with TF-PERT and ATT-TF-PERT do not show statistically significant improvements over their counterparts TK-PERT and ATT-PERT, similarly to the trend observed in the previous tasks. For English, Chinese and Spanish, we are further able to compare the performance of pre-trained large-input transformers. Longformer achieves 92.3% of accuracy for English, which is 4.1% better than the 88.7% that LASER achieves in the *All tokens* configuration and about 2% better than the best performing architecture, the sentence average of LaBSE embeddings (90.9%). However, the latter is not statistically significant at 95% confidence level. The result is different for Chinese and Spanish. In both cases, considering all tokens with LASER and sentence average are better than Longformer, although the difference is not statistically significant for Spanish. This indicates that smaller amounts of training data can prevent native full document-level embeddings to be extended to languages other than English.

7 Summary and Conclusions

In this work, we studied effective methods for developing multilingual document-level representations. We used state-of-the-art sentence-level embeddings as basic units and systematically compare different pooling methods to evaluate these representations at the document level. We performed

		en → xx							
		en	**de**	**es**	**fr**	**it**	**ja**	**ru**	**zh**
Longformer		***92.3***$^{+0.7}_{-0.8}$	–	76.9$^{+0.6}_{-0.7}$	–	–	–	–	68.5$^{+0.4}_{-0.5}$
LASER	All tokens	88.7$^{+1.1}_{-0.8}$	83.6$^{+0.5}_{-0.4}$	77.4$^{+0.9}_{-0.8}$	78.1$^{+0.7}_{-0.8}$	65.1$^{+0.6}_{-0.7}$	61.8$^{+0.6}_{-0.4}$	66.6$^{+0.5}_{-0.6}$	70.1$^{+0.9}_{-0.8}$
	Sentence Average	89.9$^{+0.9}_{-0.8}$	84.8$^{+0.7}_{-0.6}$	77.3$^{+0.9}_{-0.7}$	77.9$^{+0.5}_{-0.9}$	*64.9*$^{+0.4}_{-0.8}$	60.3$^{+0.8}_{-0.7}$	67.8$^{+0.8}_{-0.9}$	**71.9**$^{+0.8}_{-0.7}$
	Top-Half Avg.	86.4$^{+0.3}_{-0.9}$	83.5$^{+0.4}_{-0.5}$	75.8$^{+0.9}_{-0.6}$	76.2$^{+0.8}_{-0.5}$	63.2$^{+0.7}_{-0.9}$	56.5$^{+0.7}_{-0.6}$	64.1$^{+0.7}_{-0.8}$	67.5$^{+0.8}_{-0.7}$
	Bottom-Half Avg.	83.2$^{+0.4}_{-0.6}$	81.4$^{+0.7}_{-0.8}$	71.2$^{+0.7}_{-0.6}$	70.5$^{+0.8}_{-0.9}$	59.2$^{+0.5}_{-0.4}$	50.4$^{+0.6}_{-0.7}$	56.2$^{+0.6}_{-0.4}$	60.3$^{+0.6}_{-0.7}$
	TF-IDF Weighted	86.3$^{+0.8}_{-0.8}$	85.1$^{+0.5}_{-0.8}$	75.3$^{+0.7}_{-0.4}$	74.1$^{+0.7}_{-0.8}$	56.4$^{+0.5}_{-0.7}$	61.4$^{+0.6}_{-0.8}$	60.2$^{+0.7}_{-0.6}$	71.5$^{+0.4}_{-0.5}$
	TK-PERT	89.1$^{+0.4}_{-0.7}$	85.2$^{+0.6}_{-0.6}$	75.6$^{+0.8}_{-0.7}$	78.2$^{+0.8}_{-1.1}$	63.6$^{+0.9}_{-0.7}$	62.3$^{+0.8}_{-0.4}$	**67.8**$^{+0.6}_{-0.7}$	71.1$^{+0.4}_{-0.6}$
	TF-PERT	88.7$^{+0.6}_{-0.5}$	84.8$^{+0.8}_{-0.6}$	75.4$^{+0.5}_{-0.4}$	77.9$^{+0.6}_{-0.4}$	61.2$^{+0.9}_{-0.8}$	61.8$^{+0.3}_{-0.4}$	67.2$^{+0.5}_{-0.5}$	70.8$^{+0.6}_{-0.5}$
	ATT-PERT	*89.2*$^{+0.7}_{-0.8}$	**86.2**$^{+0.6}_{-0.5}$	**77.5**$^{+0.8}_{-0.7}$	*79.1*$^{+1.0}_{-0.8}$	64.0$^{+0.3}_{-0.9}$	62.5$^{+0.6}_{-0.4}$	66.2$^{+0.8}_{-0.9}$	71.3$^{+0.7}_{-0.6}$
	ATT-TF-PERT	88.5$^{+0.6}_{-0.5}$	86.0$^{+0.4}_{-0.3}$	76.7$^{+0.4}_{-0.5}$	78.9$^{+0.5}_{-0.5}$	63.8$^{+0.5}_{-0.6}$	**62.8**$^{+0.5}_{-0.4}$	66.5$^{+0.4}_{-0.7}$	70.5$^{+0.3}_{-0.5}$
LaBSE	Sentence Average	**90.9**$^{+0.6}_{-0.3}$	85.2$^{+0.8}_{-0.7}$	75.6$^{+0.5}_{-0.8}$	**79.9**$^{+0.5}_{-0.3}$	**66.9**$^{+0.9}_{-0.6}$	58.3$^{+0.7}_{-0.6}$	65.4$^{+0.5}_{-0.5}$	70.1$^{+0.5}_{-0.6}$
	Top-Half Avg.	86.1$^{+0.7}_{-0.8}$	80.5$^{+0.5}_{-0.9}$	73.2$^{+0.7}_{-0.8}$	76.5$^{+0.9}_{-0.7}$	62.5$^{+0.6}_{-0.8}$	56.1$^{+0.7}_{-0.6}$	61.8$^{+1.0}_{-0.9}$	67.3$^{+0.6}_{-0.7}$
	Bottom-Half Avg.	85.4$^{+1.2}_{-1.1}$	78.7$^{+0.5}_{-0.6}$	71.4$^{+0.6}_{-0.7}$	73.3$^{+0.8}_{-0.6}$	59.6$^{+0.4}_{-0.7}$	50.7$^{+0.3}_{-0.8}$	58.9$^{+0.7}_{-0.6}$	61.4$^{+0.9}_{-1.1}$
	TF-IDF Weighted	86.2$^{+0.2}_{-0.6}$	84.1$^{+0.5}_{-0.4}$	73.9$^{+0.6}_{-0.3}$	77.1$^{+0.3}_{-0.4}$	62.6$^{+0.2}_{-0.5}$	59.3$^{+0.3}_{-0.6}$	65.4$^{+0.5}_{-0.4}$	68.1$^{+0.8}_{-0.7}$
	TK-PERT	87.1$^{+0.5}_{-0.9}$	83.6$^{+0.8}_{-0.6}$	75.8$^{+0.5}_{-0.4}$	79.1$^{+0.7}_{-0.8}$	62.5$^{+0.3}_{-0.8}$	60.0$^{+0.6}_{-0.7}$	64.9$^{+0.6}_{-0.4}$	70.6$^{+0.7}_{-0.6}$
	TF-PERT	86.2$^{+0.5}_{-0.4}$	84.7$^{+0.4}_{-0.7}$	77.3$^{+0.7}_{-0.6}$	76.3$^{+0.6}_{-0.5}$	62.8$^{+0.5}_{-0.5}$	61.2$^{+0.7}_{-0.6}$	64.5$^{+0.6}_{-0.5}$	69.2$^{+0.5}_{-0.6}$
	ATT-PERT	88.9$^{+0.8}_{-0.6}$	84.3$^{+0.7}_{-0.8}$	*77.3*$^{+0.5}_{-0.5}$	79.4$^{+0.7}_{-0.9}$	63.8$^{+0.6}_{-0.7}$	*62.2*$^{+0.8}_{-0.5}$	*65.9*$^{+0.7}_{-0.9}$	*71.2*$^{+0.8}_{-0.7}$
	ATT-TF-PERT	88.4$^{+0.4}_{-0.3}$	*85.4*$^{+0.9}_{-0.6}$	77.2$^{+0.5}_{-0.4}$	78.2$^{+0.4}_{-0.5}$	65.7$^{+0.5}_{-0.3}$	61.3$^{+0.7}_{-0.5}$	65.3$^{+0.7}_{-0.8}$	67.4$^{+0.6}_{-0.8}$
sBERT	Sentence Average	85.1$^{+0.6}_{-0.7}$	85.2$^{+0.6}_{-0.7}$	75.7$^{+0.6}_{-0.8}$	*78.2*$^{+0.6}_{-0.7}$	*64.5*$^{+0.7}_{-0.5}$	60.4$^{+0.8}_{-0.6}$	*66.4*$^{+0.8}_{-0.7}$	*69.5*$^{+0.8}_{-0.7}$
	Top-Half Avg.	83.2$^{+0.8}_{-0.6}$	84.1$^{+0.7}_{-0.6}$	71.3$^{+0.5}_{-0.6}$	76.5$^{+0.8}_{-0.6}$	60.8$^{+0.7}_{-0.9}$	60.4$^{+0.9}_{-1.2}$	62.8$^{+0.8}_{-0.7}$	63.5$^{+0.9}_{-0.8}$
	Bottom-Half Avg.	80.6$^{+0.7}_{-0.6}$	81.3$^{+0.5}_{-0.8}$	66.5$^{+0.4}_{-0.4}$	70.1$^{+0.6}_{-0.4}$	56.5$^{+0.4}_{-0.8}$	58.7$^{+0.5}_{-0.6}$	56.1$^{+0.7}_{-0.6}$	60.5$^{+0.5}_{-0.4}$
	TF-IDF Weighted	84.2$^{+0.4}_{-0.5}$	82.8$^{+0.5}_{-0.4}$	75.1$^{+0.6}_{-0.7}$	74.3$^{+0.4}_{-0.6}$	63.2$^{+0.3}_{-0.2}$	61.2$^{+0.4}_{-0.5}$	63.4$^{+0.5}_{-0.3}$	65.8$^{+0.7}_{-0.6}$
	TK-PERT	86.2$^{+0.6}_{-0.7}$	84.1$^{+0.8}_{-0.7}$	73.9$^{+0.6}_{-0.6}$	77.1$^{+0.8}_{-0.6}$	62.6$^{+0.6}_{-0.8}$	59.3$^{+0.7}_{-0.5}$	65.4$^{+0.8}_{-0.6}$	68.1$^{+0.6}_{-0.7}$
	TF-PERT	85.8$^{+0.5}_{-0.4}$	83.7$^{+0.2}_{-0.4}$	72.7$^{+0.6}_{-0.5}$	76.5$^{+0.4}_{-0.3}$	62.0$^{+0.6}_{-0.5}$	60.4$^{+0.3}_{-0.6}$	64.3$^{+0.4}_{-0.8}$	68.2$^{+0.7}_{-0.8}$
	ATT-PERT	*88.5*$^{+0.7}_{-0.6}$	*85.8*$^{+0.5}_{-0.5}$	*76.2*$^{+0.8}_{-0.4}$	77.4$^{+0.5}_{-0.6}$	62.1$^{+0.6}_{-0.7}$	60.8$^{+0.3}_{-0.6}$	66.1$^{+0.7}_{-0.4}$	*69.5*$^{+0.8}_{-0.6}$
	ATT-TF-PERT	85.6$^{+0.5}_{-0.6}$	84.3$^{+0.3}_{-0.4}$	75.1$^{+0.6}_{-0.6}$	76.8$^{+0.5}_{-0.6}$	61.3$^{+0.8}_{-0.5}$	*62.7*$^{+0.4}_{-0.5}$	65.8$^{+0.5}_{-0.3}$	66.4$^{+0.6}_{-0.6}$

Table 4: Accuracy for MLDoc classification on the zero-shot transfer task. Best results per language are shown in bold and per family in italics.

exhaustive evaluations across three sentence embeddings models, three tasks and eight languages.

Our experiments show that specific **base sentence embedding models** (LASER, LaBSE, sBERT) do not impact the performance of the document-level embeddings much. We observe similar performance amongst them across all experiments. However, it is to be noted that we experiment with languages that while being morphologically distinct, are well resourced and covered by the three base sentence-embedding models. It would be interesting to explore how models behave when embeddings have a lower quality. For this, one would need to create evaluation datasets at the document level for low-resourced languages but this is out of the scope of this work.

We observed that a simple sentence average is a very strong **pooling strategy**, specially for classification tasks. Positional and contextual information is more important than semantic information for the final performance as exemplified by the fact that PERT-based weightings perform better than TF-IDF's in all the tasks. When combining both, positional and semantic information, we do not observe statistically significant improvements with respect to only including positional information. For the classification tasks which include a learnable layer, we extend TK-PERT to ATT-PERT (and the semantic counterparts) and include global trainable attention on the positional information. This global attention is beneficial in all the cases.

The **type of document** is also relevant to chose the best method. Long documents might have the most crucial information stored in different parts. For instance, webpages have a majority of their information in the first half of the document as we observed in the document alignment task. In this case, the positional information significantly outperforms any model that does not take it into account.

Limitations

One of the main focal points of this work is multilinguality. In the presented approaches, the multilinguality of the resultant document embeddings depends solely on the language coverage and cross-

lingual transfer ability of the pre-trained sentence embeddings used as basic units. Document-level representations are as robust to new languages and scripts as the base sentence embeddings are. Cross-lingual transfer is a perpendicular dimension not studied in this work.

We introduce ATT-PERT, a new learnable approach for the combination of sentence embeddings. This model is therefore of use for tasks with a learning/fine-tuning phase but it is not intended for ready-to-use multilingual document-level embeddings in contrast to the existing pre-trained sentence-level counterparts.

Acknowledgements

This work has been supported by the German Federal Ministry of Education and Research (BMBF) under funding code 01IW20010 (CORA4NLP).

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

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

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Christian Buck and Philipp Koehn. 2016a. Findings of the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 554–563, Berlin, Germany. Association for Computational Linguistics.

Christian Buck and Philipp Koehn. 2016b. Quick and reliable document alignment via TF/IDF-weighted cosine distance. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 672–678, Berlin, Germany. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Kathryn Annette Chapman and Günter Neumann. 2020. Automatic ICD Code Classification with Label Description Attention Mechanism. In *IberLEF@ SEPLN*, volume 2664 of *CEUR Workshop Proceedings*, pages 477–488. CEUR-WS.org.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems*, 28.

Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 78–86, Berlin, Germany. Association for Computational Linguistics.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Ma-*

chine Learning Research, pages 1188–1196, Bejing, China. PMLR.

Min Seok Lee, Seok Woo Yang, and Hong Joo Lee. 2022. Weight attention layer-based document classification incorporating information gain. *Expert Systems*, 39(1):e12833.

Wei Li and Brian Kan-Wing Mak. 2020. Transformer based multilingual document embedding model. *ArXiv*, abs/2008.08567.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26:3111–3119.

Antonio Miranda-Escalada, Eulàlia Farré, and Martin Krallinger. 2020. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. In *IberLEF@SEPLN*, volume 2664 of *CEUR Workshop Proceedings*, pages 303–323. CEUR-WS.org.

Mariana Neves, Daniel Butzke, Antje Dörendahl, Nora Leich, Barbara Grune, and Gilbert Schönfelder. 2019. Non-technical Summaries (NTS) of Animal Experiments Indexed with ICD-10 Codes (Version 1.0).

Hyunji Park, Yogarshi Vyas, and Kashif Shah. 2022. Efficient classification of long documents using transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 702–709, Dublin, Ireland. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Manuel Romero. 2022. Spanish LongFormer. https://huggingface.co/mrm8488/longformer-base-4096-spanish.

Markus Sagen. 2021. Large-context question answering with cross-lingual transfer. Master's thesis, Uppsala University, Department of Information Technology.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Liu Shen. 2021. Chinese LongFormer. https://huggingface.co/schen/longformer-chinese-base-4096.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, pages 194–206. Springer.

Brian Thompson and Philipp Koehn. 2020. Exploiting sentence order in document alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5997–6007, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

David Vose. 2008. *Risk analysis: a quantitative guide*. John Wiley & Sons.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *CoRR*, abs/2006.04768.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.