Fully unsupervised crosslingual semantic textual similarity metric based on BERT for identifying parallel data

Chi-kiu Lo and **Michel Simard**
NRC-CNRC
National Research Council Canada
1200 Montreal Road, Ottawa, Ontario K1A 0R6, Canada
{Chikiu.Lo|Michel.Simard}@nrc-cnrc.gc.ca

Abstract

We present a fully unsupervised crosslingual semantic textual similarity (STS) metric, based on contextual embeddings extracted from BERT – Bidirectional Encoder Representations from Transformers (Devlin et al., 2019). The goal of *crosslingual STS* is to measure to what degree two segments of text in different languages express the same meaning. Not only is it a key task in crosslingual natural language understanding (XLU), it is also particularly useful for identifying parallel resources for training and evaluating downstream multilingual natural language processing (NLP) applications, such as machine translation. Most previous crosslingual STS methods relied heavily on existing parallel resources, thus leading to a circular dependency problem. With the advent of massively multilingual context representation models such as BERT, which are trained on the concatenation of non-parallel data from each language, we show that the deadlock around parallel resources can be broken. We perform intrinsic evaluations on crosslingual STS data sets and extrinsic evaluations on parallel corpus filtering and human translation equivalence assessment tasks. Our results show that the unsupervised crosslingual STS metric using BERT without fine-tuning achieves performance on par with supervised or weakly supervised approaches.

1 Introduction

Crosslingual semantic textual similarity (STS) (Agirre et al., 2016a; Cer et al., 2017) aims at measuring the degree of meaning overlap between two texts written in different languages. It is a key task in crosslingual natural language understanding (XLU), with applications in crosslingual information retrieval (Franco-Salvador et al., 2014; Vulić and Moens, 2015), crosslingual plagiarism detection (Franco-Salvador et al., 2016a,b), etc. It is also particularly useful for identifying parallel resources (Resnik and Smith, 2003; Aziz and Specia, 2011) for training and evaluating downstream multilingual NLP applications, such as machine translation systems.

Unlike in crosslingual textual entailment (Negri et al., 2013) or crosslingual natural language inference (XNLI) (Conneau et al., 2018), which are directional classification tasks, in crosslingual STS, continuous values are produced, to reflect a range of similarity that goes from complete semantic unrelatedness to complete semantic equivalence. Machine translation quality estimation (MTQE) (Specia et al., 2018) is perhaps the field of work that is the most related to crosslingual STS: in MTQE, one tries to estimate translation quality, by comparing an original source-language text with its machine translation. In contrast, in crosslingual STS, neither the direction nor the origin (human or machine) of the translation is taken into account. Furthermore, MTQE also typically considers the fluency and grammaticality of the target text; these aspects are usually not perceived as relevant for crosslingual STS.

Many previous crosslingual STS methods rely heavily on existing parallel resources to first build a machine translation (MT) system and translate one of the test sentences into the other language for applying monolingual STS methods (Brychcín and Svoboda, 2016). Methods that do not rely explicitly on MT, such as that in Lo et al. (2018), still require parallel resources to build bilingual word representations for evaluating crosslingual lexical semantic similarity. It is clear that there is a circular dependency problem on parallel resources.

Massively multilingual context representation models, such as MUSE (Conneau et al., 2017), BERT (Devlin et al., 2019), and XLM (Lample and Conneau, 2019), that are trained in an unsupervised manner with non-parallel data from each

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 206–215
Hong Kong, China, November 3-4, 2019. ©2019 Association for Computational Linguistics

language, have shown improved performance in XNLI classification tasks using task-specific fine-tuning.

In this paper, we propose a crosslingual STS metric based on fully unsupervised contextual embeddings extracted from BERT without fine-tuning. In an intrinsic crosslingual STS evaluation and extrinsic parallel corpus filtering and human translation error detection tasks, we show that our BERT-based metric achieves performance on par with similar metrics based on supervised or weakly supervised approaches. With the availability of the multilingual context representation models, we show that the deadlock around parallel resources for crosslingual textual similarity can be broken.

2 Crosslingual STS metric

Our crosslingual STS metric is based on YiSi (Lo, 2019). YiSi is a unified adequacy-oriented MT quality evaluation and estimation metric for languages with different levels of available resources. Lo et al. (2018) showed that YiSi-2, the crosslingual MT quality estimation metric, performed almost as well as the "MT + monolingual MT evaluation metric (YiSi-1)" pipeline for identifying parallel sentence pairs from a noisy web-crawled corpus in the *Parallel Corpus Filtering* task of WMT 2018 (Koehn et al., 2018b).

To measure semantic similarity between pairs of segments, YiSi-2 proceeds by finding alignments between the words of these segments that maximize semantic similarity at the lexical level. For evaluating crosslingual lexical semantic similarity, it relies on a crosslingual embedding model, using cosine similarity of the embeddings from the crosslingual lexical representation model. Following the approach of Corley and Mihalcea (2005), these lexical semantic similarities are weighed by *lexical specificity* using inverse document frequency (IDF) collected from each side of the tested corpus.

As an MTQE metric, YiSi-2 also takes into account fluency and grammatically of each side of the sentence pairs using bag-of-ngrams and the semantic parses of the tested sentence pairs. But since crosslingual STS focuses primarily on measuring the meaning similarity between the tested sentence pairs, here we set the size of ngrams to 1 and opt not to use semantic parses in YiSi-2. In addition, rather than compute IDF weights $w(e)$ and $w(f)$ for lexical units e and f in each language directly on the texts under consideration, we rely on precomputed weights from monolingual corpora $\mathbb{E}$ and $\mathbb{F}$ of the two tested languages.

The YiSi metrics are formulated as an F-score: by viewing the source text as a "query" and the target as an "answer", precision and recall can be computed. Depending on the intended application, precision and recall can be weighed differently. For example, in MT evaluation applications, we typically assign more weight to recall ("every word in the source should find an equivalent in the target"). For this application, we give equal weights to precision and recall.

Thus, the crosslingual STS of sentences **e** and **f** using YiSi-2 in this work can be expressed as follows:

$$\begin{aligned}
v(u) &= \text{embedding of unit } u \\
s(e,f) &= cos(v(e),v(f)) \\
w\,(e) &= idf(e) = log(1 + \frac{|\mathbb{E}|+1}{|\mathbb{E}_{\exists e}|+1}) \\
w\,(f) &= idf(f) = log(1 + \frac{|\mathbb{F}|+1}{|\mathbb{F}_{\exists f}|+1}) \\
\text{precision} &= \frac{\sum\limits_{e\in\mathbf{e}} \max\limits_{f\in\mathbf{f}} w\,(e)\cdot s\,(e,f)}{\sum\limits_{e\in\mathbf{e}} w\,(e)} \\
\text{recall} &= \frac{\sum\limits_{f\in\mathbf{f}} \max\limits_{e\in\mathbf{e}} w\,(f)\cdot s\,(e,f)}{\sum\limits_{f\in\mathbf{f}} w\,(f)} \\
\text{YiSi-2} &= \frac{2\cdot \text{precision}\cdot \text{recall}}{\text{precision} + \text{recall}}
\end{aligned}$$

where $s(e,f)$ is the cosine similarity of the vector representations $v(e)$ and $v(f)$ in the bilingual embeddings model.

In the following, we present the approaches we experimented with to obtain the crosslingual embedding space in supervised, weakly supervised and unsupervised manners.

2.1 Supervised crosslingual word embeddings with *BiSkip*

Luong et al. (2015) proposed *BiSkip* (with open source implementation `bivec`[1]) to jointly learn bilingual representations from the context cooccurrence information in the monolingual data and the meaning equivalent signals in the parallel data. It trains bilingual word embeddings with the objective to preserve the clustering structures of

[1] https://github.com/lmthang/bivec

words in each language. We train our crosslingual word embeddings using bivec on the parallel resources as described in each experiment.

2.2 Weakly supervised crosslingual word embeddings with *vecmap*

Artetxe et al. (2016) generalized a framework to learn the linear transformation between two monolingual word embedding spaces by minimizing the distances between equivalences listed in a collection of bilingual lexicons (with open source implementation vecmap[2]). We train our monolingual word embeddings using word2vec[3] (Mikolov et al., 2013) on the monolingual resources and then learn the linear transformation of the two monolingual embedding space using vecmap on the dictionary entries as described in each experiment.

2.3 Unsupervised crosslingual contextual embeddings with multilingual BERT

The above two mentioned embedding models produce static word embeddings that captures the semantic space to represent the training data. The shortcoming of these static embedding models is that they provide the same embedding representation for the same word without reflecting the context variation of them being used in different sentences. In contrast, BERT (Devlin et al., 2019) uses a bidirectional transformer encoder (Vaswani et al., 2017) to capture the sentence context in the output embeddings, such that the embedding for the same word unit in different sentences would be different and better represented in the embedding space. Multilingual BERT model is trained on the Wikipedia pages of 104 languages with a shared subword vocabulary. Pires et al. (2019) showed multilingual BERT works well on different monolingual NLP tasks across different languages.

Following the recommendation in Devlin et al. (2019), we use embeddings extracted from the ninth layer of the pretrained multilingual cased BERT-Base model[4] to represent subword units in the two sentences in assessment for the crosslingual lexical semantic similarity.

3 Experiment on crosslingual STS

We first evaluate the performance of YiSi-2 on the intrinsic crosslingual STS task, before testing its ability on the downstream task of identifying parallel data.

3.1 Setup

We use data from the SemEval-2016 Semantic Textual Similarity (STS) evaluation's crosslingual track (task1) (Agirre et al., 2016b), in which the goal was to estimate the degree of equivalence between pairs of Spanish-English bilingual fragments of text.[5] The test data is partitioned into two evaluation sets: the *News* data set has 301 pairs, manually harvested from comparable Spanish and English news sources; the *Multi-source* data set consists of 294 pairs, sampled from English pairs of snippets used in the SemEval-2016 monolingual STS task, translated into Spanish.

We apply YiSi-2 directly to these pairs of text fragments, using bilingual word embeddings trained under three different conditions (details of the training sets are given in Table 1):

bivec : BWE's are produced with *bivec*, trained on WMT 2013 ES-EN parallel training data.

vecmap : BWE's are produced with *vecmap*, trained on all WMT 2013 ES and WMT 2019 EN monolingual data, using *Wikititles* as bilingual lexicon.[6]

BERT : BWE's are obtained from pre-trained multilingual BERT models.

We compare the YiSi-2 approach to direct cosine computations on sums of bilingual word embeddings (*bivec_sum*, *vecmap_sum* and *bert_sum*). We also compare our approach to an MT-based approach, in which each Spanish fragment is first machine-translated into English, then compared to the original English fragment, using English word embeddings, produced with *word2vec* trained on WMT 2019 news translation task monolingual data. Similarity is measured either as the cosine of the sums of word vectors from each fragment (*w2v_sum*), or with YiSi using monolingual embeddings as if they were bilingual (YiSi-1$_{w2v}$).

[2] https://github.com/artetxem/vecmap

[3] https://code.google.com/archive/p/word2vec/

[4] https://github.com/google-research/bert

[5] In this task, the order of languages in pairs was randomized, so that it was first necessary to detect which fragment was in which language. Here, we work from properly ordered pairs.

[6] https://linguatools.org/tools/corpora/wikipedia-parallel-titles-corpora/

Model	Training Data:				Dictionary	Embedding vocab
	lang.	domain	#sent	#words	#pairs	#words
bivec	es	WMT 2013: EU Parliament and web	3.8M	107M	—	291k
	en			102M		220k
vecmap	es	WMT 2013: News and EU Parliament	45M	1B	373k	883k
	en	WMT 2019: News	779M	13B		3M

Table 1: Statistics of data used in training the bilingual word embeddings for evaluating crosslingual lexical semantic similarity in YiSi-2.

SemEval-16 crosslingual STS

system	news	multisource
MT + monolingual STS		
UWB	**0.9062**	**0.8190**
MT+w2v_sum	0.5883	0.2021
MT+YiSi-1$_{w2v}$	0.8965	0.6212
crosslingual STS		
bivec_sum	0.5302	0.2684
vecmap_sum	0.3075	0.5398
bert_sum	0.7223	0.6071
YiSi-2$_{bivec}$	**0.8744**	0.6550
YiSi-2$_{vecmap}$	0.7854	0.7028
YiSi-2$_{bert}$	0.8723	**0.7190**

Table 2: Pearson's correlation of the system scores with the gold standard on the two test sets from the SemEval-16 crosslingual STS task.

The MT system used is a phrase-based SMT system, trained using standard resources – Europarl, Common Crawl (CC) and News & Commentary (NC) – totaling approximately 110M words in each language. We bias the SMT decoder to produce a translation that is as close as possible on the surface to the English sentence. This is done by means of log-linear model features that aim at maximizing n-gram precision between the MT output and the English sentence. More details on this method can be found in Lo et al. (2016).

3.2 Results

The results of these experiments are presented in Table 2, where performance is measured in terms of Pearson's correlation with the test sets' gold standard annotations. For reference, we also include results obtained by the UWB system (Brychcín and Svoboda, 2016), which was the best performing system in the SemEval 2016 crosslingual STS shared task. The UWB system is an MT-based system with a STS system trained on assorted lexical, syntactic and semantic features. Globally, using the YiSi metric to measure semantic similarity performs much better than sentence-level cosine ("*_sum" systems). On the *News* dataset, the best results are obtained by combining an MT-based approach with YiSi-1 using monolingual word embeddings (MT+YiSi-1$_{w2v}$), reflecting the in-domain nature of the text for the MT system. However, this is followed very closely by both the supervised BWE's (YiSi-2$_{bivec}$) and BERT (YiSi-2$_{bert}$), which yield very similar results, and clearly outperform semi-supervised BWE's (YiSi-2$_{vecmap}$). The nature of the *Multisource* translations appears to be quite different from what supervised BWE's and the MT system have been exposed to in training (YiSi-2$_{bivec}$ and MT+YiSi-1$_{w2v}$), which possibly explains their much poorer performance on this dataset. In contrast, weakly supervised BWE's and BERT behave much more reliably on this data.

Overall, while MT and supervised BWE's seem to work best with YiSi when large quantities of in-domain training data is available, the fully unsupervised alternative of using a pretrained BERT model comes very close, and behaves much better in the face of out-of-domain data.

4 Experiment on Parallel Corpus Filtering

Next, we evaluate YiSi on the task of *Parallel Corpus Filtering* (PCF). Quality – or "cleanliness" – of parallel training data for MT has been shown to affect MT quality at different degrees, and various characteristics of the data – parallelism of the sentence pairs and the grammaticality of target-language data – impact MT systems in different ways (Goutte et al., 2012; Simard, 2014; Khayrallah and Koehn, 2018).

Here, we use data from the WMT19 shared task on PCF. In this shared task, participants were challenged to find good quality translations from noisy sentence-aligned parallel corpora, for the purpose of training MT systems for translating from two low-resource languages, Nepali and Sin-

Model	Training Data:				Dictionary	Embedding vocab
	lang.	domain	#sent	#words	#pairs	#words
bivec	ne	IT and religious	563k	8M	—	34k
	en			5M		46k
vecmap	ne	wiki	92k	5M	9k	55k
	en	news	779M	13B		3M

Table 3: Statistics of data used in training the bilingual word embeddings for evaluating crosslingual lexical semantic similarity in YiSi-2.

hala, into English.[7] Both corpora were crawled from the web, using ParaCrawl (Koehn et al., 2018a). Specifically, the task is to produce a score for each sentence pair in these noisy corpora, reflecting the quality of that pair. The scoring schemes are evaluated by extracting the top-scoring sentence pairs from each corpus, then using them to train MT systems; these systems are run on test sets of Wikipedia articles (Guzmán et al., 2019), and the results are evaluated using BLEU (Papineni et al., 2002). In addition to the noisy corpora, participants are allowed to use a few small sets of parallel data, covering different domains, for each of the two low-resource languages, as well as a third, related language, Hindi (which uses the same script as Nepali). The provided data also included much larger monolingual corpora for each of English, Hindi, Nepali and Sinhala.

4.1 Setup

In these experiments, we focus on the Nepali-English corpus, and perform PCF in three steps:

1. **pre-filtering**: apply *ad hoc* filters to remove sentences that are exact duplicates (masking numbers, emails and web addresses), that contain mismatching numbers, that are in the wrong language according to the `pyCLD2` language detector[8] or that are excessively long (either side has more than 150 tokens). We also filter out all pairs where over 50% of the Nepali text is comprised of English, numbers or punctuation.

2. **scoring**: we score sentence pairs using YiSi-2.

3. **re-ranking**: to optimize vocabulary coverage in the resulting MT system, we apply a form of re-ranking: going down the ranked list of scored sentence pairs, we apply a 20% penalty to the pair's score if it does not contain at least one "new" source-language word bigram, i.e., a pair of consecutive source-language tokens not observed in previous (higher-scoring) sentence pairs. This has the effect of down-ranking sentences that are too similar to previously selected sentences.

WMT19 parallel corpus filtering		
system	1M-word	5M-word
random	1.30	3.01
Zipporah	**4.14**	**4.42**
YiSi-2_{bivec}	3.86	3.76
YiSi-2_{vecmap}	4.00	3.76
YiSi-2_{bert}	3.77	3.77

Table 4: Uncased BLEU scores on the official WMT19 PCF dev ("dev-test") sets achieved by the SMT systems trained on the 1M- and 5M-word corpora subselected by the scoring systems.

The scoring step is performed with YiSi-2, using bilingual word embeddings obtained under three different conditions (details of the various training sets used can be found in Table 3):

bivec : supervised BWE's produced using *bivec*, trained on the WMT19 PCF (clean) parallel data.

vecmap : weakly supervised BWE's are produced with *vecmap*, trained on all monolingual WMT19 PCF data, using *Wikititles* and the provided dictionary entries as bilingual lexicon.

BERT : BWE's obtained from pretrained multilingual BERT models.

As in the WMT19 PCF shared task, we evaluate the quality of our scoring by training MT systems and measuring their performance on the official test set. We used the provided software to

[7] `http://www.statmt.org/wmt19/parallel-corpus-filtering.html`

[8] `https://github.com/aboSamoor/pycld2`

extract the 1M-word and 5M-word samples from the original test corpora, using the scores of each of our systems in turn. We then trained MT systems using the extracted data: our MT systems are standard phrase-based SMT systems, with components and parameters similar to the German-English SMT system in Williams et al. (2016).

4.2 Results

BLEU scores of the resulting MT systems are shown in Table 4. For comparison, we present the results of random scoring, as well as results obtained by the Zipporah PCF method (Xu and Koehn, 2017). Zipporah combines fluency and adequacy features to score sentence pairs; adequacy features are derived from existing parallel corpora, and the feature combination (logistic regression) is optimized on in-domain parallel data. Therefore, Zipporah can be seen as a fully supervised method. The Zipporah-based MT systems were trained similarly to other systems in the results reported here.

All systems produced with YiSi-2 produce similar results. Interestingly, the MT systems produced with YiSi-2 in the 5M-word condition are not better than those of the 1M-word condition. This is possibly explained by the large quantity of noisy data in the WMT19 Nepalese-English corpus: it is not even clear that there are 5M words of proper translations in that corpus. In such harsh conditions, pre- and post-processing steps become crucially important, and deduplicating the data may even turn out to be harmful, if that means allowing more space for noise. The MT systems produced with Zipporah all achieve higher BLEU scores than YiSi-2, which may be explained by Zipporah's explicit modeling of target-language fluency. This is especially apparent in the 5M-word condition, but it may explain Zipporah's slightly better performance in the 1M-word condition as well. Overall, the benefits of supervised and weakly supervised approaches over using a pre-trained BERT model for PCF appear to be minimal, even in very low-resource conditions such as this.

5 Experiments on Translation Equivalence Error Detection

Given a text and its translation, *Translation Equivalence Error Detection* (TEED) is the task of identifying pairs of corresponding text segments whose meanings are not strictly equivalent. Note that, while in practice "translation errors" can take many forms, here, we are strictly focusing on meaning errors. In this formulation of the problem, we are also assuming that the source and target texts have been properly segmented into sentences and aligned.

The TEED problem is essentially the same as that of Parallel Corpus Filtering (PCF), discussed in the previous section. However, the usage scenario is quite different: in PCF, one is typically dealing with a very large collection of segment pairs, only a fraction of which are true translations; the PCF task is then to filter out pairs which are not proper translations, possibly with some tolerance for pairs of segments that do share partial meaning. In TEED, the data is mostly expected to be high-quality translations; the task is then to identify those pairs that deviate from this norm, even on small details.

5.1 Setup

We experiment the TEED task using a data set obtained from the Canadian government's Public Service Commission (PSC). As part of its mandate, the PSC periodically audits Canadian government job ads, to ensure that they conform with Canada's *Official Languages Act*: as such, job ads must be posted in both of Canada's official languages, English and French, and both versions must be equivalent in meaning.

Our *PSC* data set consists of 175 000 "Statement of merit criteria" paragraphs, identifying any skill, ability, academic specialization, relevant experience or any other essential or asset criteria required for a position to be filled. Of these, 3521 have been manually annotated for equivalence errors by PSC auditors. Out of the 3521 pairs, 164 (4.6%) were reported to contain equivalence errors. The majority of these errors result from missing information in one language or the other (45%). In a slightly smaller proportion (43%), we find pairs of segments that don't express exactly the same meaning – a surprisingly large proportion of this last group consists in cases where the word *and* is translated as *or* or vice-versa. The rest consist in terminology issues and untranslated segments.

We experimented applying the YiSi-2 metric to this task, using bilingual word embeddings obtained under four different conditions:

Model	Training Data: lang.	domain	#sent	#words	Dictionary #pairs	Embedding vocab #words
WMT.bivec	fr	News and EU Parliament	40.7M	1.2B	—	878k
	en			1.4B		791k
PSC.bivec	fr	Job ads	175k	3.0M	—	10k
	en			2.5M		11k
PSC.vecmap	fr	Job ads	175k	3.0M	6k	10k
	en	Job ads	175k	2.5M		11k

Table 5: Statistics of data used in training the bilingual word embeddings for evaluating translation equivalence assessment.

PSC Translation Error Detection

model	ROC AUC	mean F_1	mean F_2
PSC.bivec	**0.807**	**0.160**	**0.281**
PSC.vecmap	0.717	0.136	0.241
BERT	0.702	0.132	0.234
WMT.bivec	0.641	0.112	0.205

Table 6: Sentence-level translation error detection results on PSC test data, expressed in terms of Area under the ROC curve, mean F_1 and mean F_2.

PSC.bivec : BWE's are produced with *bivec*, trained on all unannotated PSC data.

PSC.vecmap : BWE's are produced with *vecmap*, trained on all unannotated PSC data, using *wikititles* as bilingual lexicon.

WMT.bivec : BWE's are produced with *bivec*, trained on all bilingual French-English data provided for the WMT 2015 News translation shared task.

BERT : BWE's obtained from pretrained multilingual BERT models.

Details about the training data can be found in Table 5.

5.2 Results

For these experiments, we considered an application scenario in which a text and its translation, in the form of pairs of matching segments, are scored using YiSi-2, and presented to a user, ranked in increasing order of score, so that pairs most likely to contain a translation error are presented first. The performance of the system can then be measured in terms of true and false positive rates, precision and recall, over subsets of increasing sizes of the test set. In Table 6, we report results in terms of mean F-score, with $\beta = 1$ and $\beta = 2$, and in terms of the *Area under the ROC curve* (ROC AUC), which can be interpreted as the probability that a system will score a randomly chosen faulty translation lower than a randomly chosen good translation. The ROC curves themselves can be seen in Figure 1.

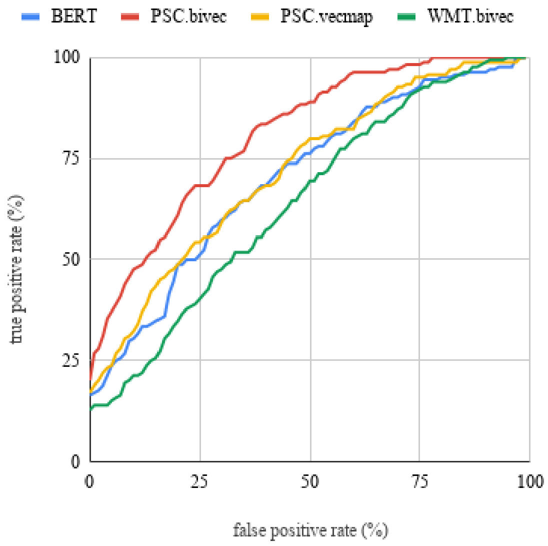

Figure 1: ROC curves of Sentence-level translation error detection results on PSC test data.

Globally, YiSi-2 clearly performs best at this task when using BWE's trained on domain-specific parallel data (PSC.bivec), even when there is very limited quantities of such data, as is the case here. However, BERT models perform comparably to vector-mapped BWE's trained with in-domain data (PSC.vecmap), and substantially better than BWE's trained on large quantities of generic, out-of-domain parallel data (WMT). We conclude that, in the absence of in-domain parallel data, for TEED applications, an unsupervised YiSi-2 method will perform at least as well as supervised methods trained on out-of-domain data.

6 Conclusion

We presented a fully unsupervised crosslingual semantic textual similarity (STS) metric, based on contextual embeddings extracted from BERT without fine-tuning. We perform intrinsic evaluations on crosslingual STS data sets and extrinsic evaluations on parallel corpus filtering and human translation equivalence assessment tasks. Our results show that the unsupervised metric we propose achieves performance on par with supervised or weakly supervised approaches. We show that the circular dependency on the existence of parallel resources for using crosslingual STS to identify parallel data can be broken.

In this paper, we have only experimented with the contextual embeddings extracted from pretrained multilingual BERT model. For domain-specific applications, such as the job advertisement domain in the PSC translation equivalence error detection task, the performance of YiSi-2 could potentially be improved by fine-tuning BERT with in-domain data, something we plan to examine in the near future. We will also want to explore the use of other multilingual context representation models, such as MUSE (Conneau et al., 2017), XLM (Lample and Conneau, 2019), etc.

Acknowledgement

We would like to thank Dominic Demers and his colleagues at the Canadian governments Public Service Commission (PSC) for providing the PSC data set and managing the human annotation in the experiments.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016a. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016b. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294.

Wilker Aziz and Lucia Specia. 2011. Fully automatic compilation of Portuguese-English and Portuguese-Spanish parallel corpora. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*.

Tomáš Brychcín and Lukáš Svoboda. 2016. UWB at SemEval-2016 task 1: Semantic textual similarity using lexical, syntactic, and semantic information. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 588–594, San Diego, California. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, EMSEE '05, pages 13–18, Stroudsburg, PA, USA. 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.

Marc Franco-Salvador, Parth Gupta, Paolo Rosso, and Rafael E. Banchs. 2016a. Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language. *Know.-Based Syst.*, 111(C):87–99.

Marc Franco-Salvador, Paolo Rosso, and Manuel Montes-y Gómez. 2016b. A systematic study of knowledge graph analysis for cross-language plagiarism detection. *Inf. Process. Manage.*, 52(4):550–570.

Marc Franco-Salvador, Paolo Rosso, and Roberto Navigli. 2014. A knowledge-based representation for cross-language document retrieval and categorization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 414–423, Gothenburg, Sweden. Association for Computational Linguistics.

Cyril Goutte, Marine Carpuat, and George Foster. 2012. The impact of sentence alignment errors on phrase-based machine translation performance. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas*.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *CoRR*, abs/1902.01382.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. *CoRR*, abs/1805.12282.

Philipp Koehn, Kenneth Heafield, Mikel L. Forcada, Miquel Esplà-Gomis, Sergio Ortiz-Rojas, Gema Ramírez Sánchez, Víctor M. Sánchez Cartagena, Barry Haddow, Marta Bañón, Marek Střelec, Anna Samiotou, and Amir Kamran. 2018a. ParaCrawl corpus version 1.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel Forcada. 2018b. Findings of the wmt 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Chi-kiu Lo. 2019. YiSi - a unified semantic mt quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation*, pages 706–712, Florence, Italy. Association for Computational Linguistics.

Chi-kiu Lo, Cyril Goutte, and Michel Simard. 2016. CNRC at SemEval-2016 task 1: Experiments in crosslingual semantic textual similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 668–673, San Diego, California. Association for Computational Linguistics.

Chi-kiu Lo, Michel Simard, Darlene Stewart, Samuel Larkin, Cyril Goutte, and Patrick Littell. 2018. Accurate semantic textual similarity for cleaning noisy parallel corpora using semantic machine translation evaluation metric: The NRC supervised submissions to the parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 908–916, Belgium, Brussels. Association for Computational Linguistics.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *NAACL Workshop on Vector Space Modeling for NLP*, Denver, United States.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. 2013. Semeval-2013 task 8: Cross-lingual textual entailment for content synchronization. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 25–33, Atlanta, Georgia, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *CoRR*, abs/1906.01502.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Comput. Linguist.*, 29(3):349–380.

Michel Simard. 2014. Clean data for training statistical MT: the case of MT contamination. In *Proceedings of the Eleventh Conference of the Association for Machine Translation in the Americas*, pages 69–82, Vancouver, BC, Canada.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón Astudillo, and André F. T. Martins. 2018. Findings of the wmt 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 702–722, Belgium, Brussels. 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. In *Advances in neural information processing systems*, pages 5998–6008.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 363–372, New York, NY, USA. ACM.

Philip Williams, Rico Sennrich, Maria Nadejde, Matthias Huck, Barry Haddow, and Ondřej Bojar. 2016. Edinburgh's statistical machine translation systems for wmt16. In *Proceedings of the First Conference on Machine Translation*, pages 399–410, Berlin, Germany. Association for Computational Linguistics.

Hainan Xu and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2945–2950, Copenhagen, Denmark. Association for Computational Linguistics.