Learning Multilingual Sentence Embeddings from Monolingual Corpus

Shuai Wang[1,2,3], Lei Hou[1,2,3], Juanzi Li[1,2,3], Meihan Tong[1,2,3], and Jiabo Jiang[4]

[1] DCST, Tsinghua University, Beijing 100084, China
[2] KIRC, Institute for Artificial Intelligence, Tsinghua University
[3] Beijing National Research Center for Information Science and Technology
[4] Daqing Oilfield Information Technology Company, Beijing 100043, China
shuai-wa16@mails.tsinghua.edu.cn, houlei@tsinghua.edu.cn, lijuanzi@tsinghua.edu.cn, tongmh17@mails.tsinghua.edu.cn, jiangjb@cnpc.com.cn

Abstract. Learning multi-lingual sentence embeddings usually requires large scale of parallel sentences which are difficult to obtain. We propose a novel self-learning approach which is capable of learning multi-lingual sentence embeddings from monolingual corpora. Our assumption is that, irrelevant to languages, sentences appearing in similar contexts are similar. Thus, we first train monolingual sentence embeddings of different languages with shared parameters as initialization. Then we iteratively extract similar sentence pairs and exchange their positions regardless of languages. Through their relations to their new contexts we predict the similarities between a similar sentence pair. Our experiments show that the proposed approach outperforms existing unsupervised approaches and is competitive to supervised approaches.

Keywords: Sentence Representation · Multilingual · Unsupervised Learning.

1 Introduction

Pre-training language representation from unlabelled data is effective in many natural language processing tasks. Recently many works start to focus on sentence representation instead of word representation [10, 17, 12]. However, most of them only consider monolingual situation and fail to generalize to multi-lingual settings, but many tasks involve dealing more than one languages. Besides, many low-resource languages lack labelled data, and a unified multilingual sentence representation is helpful to deal with those languages. Hence, learning multilingual sentence embeddings is a significant research in language representation.

Currently the best performance is achieved by LASER [14, 3]. It utilizes sentence level parallel data to train machine translation model and takes the encoder as sentence features extractor. Although it achieves satisfying performance on 93 languages, the need of large scale of parallel data still limits its applications.

What's more, the direct usage of machine translation model is also unable to utilize information of adjacent sentences which is important in many tasks [7].

Since unlabelled corpora are almost infinite and easy to obtain, fully unsupervised methods have strong potential. Currently multi-lingual BERT model [7] attracts much attention, which is a simple generalization of monolingual BERT model. It takes as input unlabelled texts from different languages with shared parameters. Different languages are actually trained independently, so there is few interaction among languages. Although, with the strong capacity of BERT, it achieves good results on many cross-lingual tasks, its performance on cross-lingual tasks is strikingly worse than that on monolingual tasks.

The wide gap between multi-lingual BERT and LASER reveals the significance of interactions among languages in training multi-lingual sentence embeddings. Given only monolingual corpora, to increase connections among languages, an intuitive approach is to extract parallel sentences as seeds from unlabelled corpora and connect different languages by parallel sentences. In word level, [2] propose an iterative approach to learn cross-lingual word embeddings in an unsupervised manner. Their self-learning approach iteratively extracts a bilingual dictionary as seeds and trains the cross-lingual word embeddings according to the extracted seeds. However, the following three problems make it challenging to implement such an idea in sentence level:

- **Large number of sentences:** The number of sentences is far larger than that of words, so it is nearly impossible to traverse all sentence pairs and extract enough parallel sentences.
- **Existence of parallel sentences:** Due to the diversity of sentences, strictly parallel sentences do not necessarily exist in two corpora.
- **Generalization to multi-languages:** The approach only considers learning a mapping between two languages, but in reality an encoder which is capable of encoding multiple languages, like LASER, is far more useful.

In this paper, we propose a novel iterative approach to learn sentence embeddings from monolingual corpora. Utilizing two hypotheses: sentence-level distributional hypothesis [13] and language isomorphism [16], we assume that similar sentences appear in similar contexts even across different languages. As illustrated in Fig. 1, two pieces of texts in different languages are still coherent in semantics after two similar sentences are exchanged.

We model sentences by transformer [18] and take the mean-max pooling [22] of output hidden states as sentence representation. We build a shared word piece vocabulary of all languages and all parameters are shared among languages so that a single encoder is capable of encoding multiple languages.

The proposed approach consists of two parts, i.e., sentence-context classification and sentence coherence regression. As illustrated in the first and second row of Fig. 1, the positive instance of classification is constructed by concatenating a sentence with its context. For negative instances, we replace some sentences with random ones and also concatenate to its adjacent sentences. Then we design a classifier to distinguish them. The classification is taken as an initialization of the multi-lingual coherence regression. In regression task, we iteratively extract

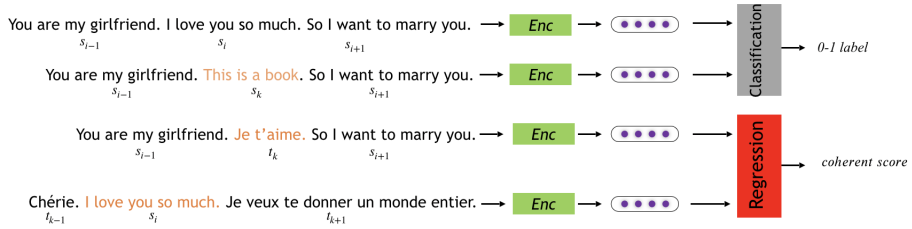

Fig. 1. The overall architecture of the proposed approach. The top part is the construction of our monolingual sentence-context classification and the bottom part is our cross-lingual coherence regression.

similar sentence pairs and exchange their positions in the original corpora and concatenate them with their new contexts. Since similar sentences are not parallel, we label concatenated sentences by a **coherence score**, which is defined as the similarity between similar sentences, as shown in the third and fourth line of Fig. 1, rather than 0-1 labels.

The experiments show that the proposed approach captures the most cross-lingual sentence information among all unsupervised approaches. Furthermore, it is capable of learning meaningful cross-lingual sentence embeddings under fully disjoint monolingual corpora.

2 Preliminaries and Framework

Our model only requires unlabelled multi-lingual sentences, and sentences are not necessarily paralleled, but should be in document-level because we need context information. Then we concatenate all documents regardless of languages altogether as our training materials.

In this paper, we are mainly dealing with sentences, so we first give a definition of a sentence.

Definition 1 (Sentence). *A sentence is defined as a sequence. The i-th sentence in a corpus is defined as* $s_i = \langle w_i^1, w_i^2, \ldots, w_i^{l_i} \rangle$*, where* l_i *is the length of the i-th sentence, and* w_i^k *is the k-th unit of the sentence i. The basic unit could be words, characters or word pieces depending on the language we are dealing.*

Now we have corpora in different languages, we concatenate them all as our training corpus D, and record the start and end index of each language. Since sentences in different languages are processed in an exactly same way, we will not mention a specific language in the following part.

We denote the concatenation of k sentences as $concat([s_1, s_2, \ldots, s_k])$, i.e., we merge them as one sentence. Our sentence encoder requires a fixed length input, so we normalize the lengths of all sentences to a fixed length $maxlen$.

Our task is to learn a multi-lingual sentence encoder which is capable of encoding similar sentences into nearby vectors. We define it as follows.

Definition 2 (Multi-lingual Sentence Encoder). *Given the training materials D, we learn a multi-lingual sentence encoder $Enc : S \to R^d$, which satisfies that given two sentences s_i and s_j, the distance between $Enc(s_i)$ and $Enc(s_j)$ reflects their similarity, where d is the dimension of sentence embeddings, S is the collection of sentences.*

Our assumption is, sentences appearing in similar contexts are similar in semantics even across different languages. To fully utilizing the isomorphism among languages, we divide the proposed approach into two stages, **Monolingual Sentence-Context Classification**, which utilizes the sentence-context relations in monolingual situation and provides an initialization for the second stages, and **Multi-lingual Coherence Regression**, as an interactive process, which generalizes the sentence-context relations to multi-lingual circumstance.

3 The Proposed Approach

In this section, we will illustrate the proposed approach, including the encoder and architecture, monolingual sentence-context classification and multi-lingual coherent regression.

Encoder and Architecture: Our sentence encoder *Enc* is a multi-layer transformer [18, 7], which is based on multi-layer self-attention. We employ exactly the same structure as in the original paper, so we omit the details. Given a sentence $s_i = \langle w_i^1, w_i^2, ..., w_i^{l_i} \rangle$, the multi-layer transformer outputs hidden states $H_i = \langle \mathbf{h}_{i1}, \mathbf{h}_{i2}, ..., \mathbf{h}_{ilen} \rangle$, and we apply mean-max pooling all hidden states as final representation of the sentence.

$$mean(H_i) = \frac{1}{len} \sum_{j=1} \mathbf{h}_{ij}; \quad max(H_i) = \max_j \mathbf{h}_{ij}$$
$$Enc(s_i) = [mean(H_i), max(H_i)]$$

The max operation selects the most salient features and the mean operation captures the general situation of the sequence, so we combine them together as our sentence representation and it is proved to be useful in [22].

3.1 Monolingual Sentence-Context Classification

As mentioned before, our approach is mainly based on sentence-level distributional hypothesis [13]. The contexts of similar sentences are similar, so in this section we train a sentence encoder utilizing the sentence-context information. We define the sentence context as follows.

Definition 3. *The context of a sentence s_i is denoted as $C_k(s_i)$, which is the collection of nearby sentences with distance less than k, i.e.,*

$$C_k(s_i) = \{s_j || i - j| \leq k\}$$

Specifically, we denote the previous sentences in the context of s_i as $C_k^-(s_i)$, and $C_k^+(s_i)$ for subsequent sentences.

We concatenate sentences with their contexts in the original corpus as positive instances. Then we replace each sentence s_i with s_{p_i}, where p_i is a random index from the whole corpus, and these random sentences are also concatenated with their current contexts as negative instances. We denote the dataset as $E = \{(x_i, y_i)\}$, where $x_i = concat([C_k^-(s_i), s_i, C_k^+(s_i)])$ or $concat([C_k^-(s_i), s_{p_i}, C_k^+(s_i)])$ and $y_i = 1$ or 0 indicate x_i is a positive or negative instance.

After concatenation these sentences are padded or clipped to a same length and then taken as input of our sentence encoder Enc. The encoded sentences are passed to a classifier denoted as M which predicts labels mentioned above. Note that the classifier here is a linear classifier because we want our sentence encoder capture more semantics. Given a sentence and its label in $(x_i, y_i) \in D$, the probability of s_i being consistent in semantics is given by

$$p(y_i = 1|x_i) = M(Enc(x_i)).$$

Our loss function is to maximize the probability of ground truth labels, i.e.,

$$L_{ml} = \sum_{(x_i,y_i)\in E} (p(y_i = 1)y_i + (1 - p(y_i = 1))(1 - y_i))$$

The intuition behind the proposed approach is that if a linear classifier can discriminate through sentence embedding whether a sentence is semantic inconsistent, then the sentence embedding should contain enough semantics.

3.2 Multi-lingual Coherent Regression

The above method is capable of capturing cross-lingual sentence information even though different languages have literally no connections between each other like multi-lingual BERT. To increase the interactions among languages, we generalize sentence level distributional hypothesis to cross-lingual setting. The linguistic isomorphism [16] assumes that, if two sentences in different languages are similar, then they should also be in similar contexts.

However, our training corpora are not assumed to contain parallel sentences, so we cannot replace a sentence with a parallel sentence in another language. In such a large corpora two arbitrary sentences are almost impossible to be similar, and they are useless for our training. Hence, we search some similar sentence pairs as follows. For the i-th place in training corpus, i.e., s_i, we randomly sample b sentences $\{s_{t_{i1}}, s_{t_{i2}}, ..., s_{t_{ib}}\}$, find the one with the largest cross-domain local scaling (CDLS for short) [6] with s_i

$$r_i = \arg\max_{l=1}^{b} CDLS(Enc(s_i), Enc(s_{t_{il}})),$$

and replace s_i with $s_{t_{ir_i}}$. Then we still concatenate new sentence and its contexts, and this time we predict a **coherent score**, which is defined as the similarity between the original sentence and the replaced sentence.

$$c_i = concat([C_k^-(s_i), s_{t_{ir_i}}, C_k^+(s_i)])$$

$$score(c_i) = dist(Enc(s_{t_{ir_i}}, Enc(s_i)),$$

where *dist* is a similarity measurement in vector space. Now we achieve our regression dataset $F = \{(c_i, score(c_i))\}$, and we also adopt a linear mapping R for the task. Mean Squared Error (MSE) is applied to optimize the regression, i.e., the loss function is

$$L_{cl} = \sum_{(c_i, score(c_i)) \in F} (score(c_i) - R(Enc(c_i)))^2$$

The retrieval process is expensive in computation, so we cannot traverse all possible sentence pairs. We randomly sample $m \times m$ sentences and choose $\min\{m \times m, 1000\}$ similar sentence pairs with the highest similarities. We repeat such selection until we obtain enough sentences pairs to ensure the interactions among languages (we use 100,000 in experiment). Intuitively, larger m corresponds higher quality training sentence pairs, but costs more computation, and we will discuss this setting in experiment.

4 Experiments

4.1 Experiment Setup

Dataset. We use Wikipedia as our training corpus and conduct experiments on English, Chinese, French, Spanish and German. We create word pieces by BPE [15] for languages other than Chinese, and we use characters for Chinese. All sentences are padded or clipped to 150.

Settings. The proposed model contains 3 layers of transformer with 8 heads in multi-head attention and the hidden dimension is 512. The learning rate is 1e-4 with linear decay and dropout rate is 0.1. The model is trained on the classification task for 1 epoch, and then is iterated on regression task for 3 times with 1 epoch for each iteration.

Baselines. We compare our proposed method with three most recent methods, multi-lingual BERT, LASER [3] and vecmap [2].

Multi-lingual BERT model is not mentioned in their original paper, but they give a brief introduction of the model in their Github repository and provide a pre-trained model. With limited computing resources, we fail to train a model as large as BERT, and the results of fine-tuning on a specific task are strikingly influenced by the capacity of model. What's more, the purpose of our experiments is to compare how much cross-lingual information the sentence encoder can capture. Hence, to decrease the influence of model capacity and make a fair comparison, we do not tune any model on specific task, and we extract features of sentences by their provided pre-trained model.

LASER is a sentence encoder supporting 93 languages trained by parallel sentences. The latest version translates every language to English and Spanish respectively and takes the encoder as their sentence encoder. It requires large scale of parallel data, so we just list its results to show the distance of our method with state-of-the-art supervised method. It is not taken into our comparison.

Vecmap is a state-of-the-art unsupervised method to learn cross-lingual word embeddings. It is robust to training corpus and even competitive to supervised methods. We take the mean and sum of cross-lingual word embeddings respectively as sentence embeddings.

We list two versions of our model. One is the model trained after the initial classification task, and the other is the final model after iteration in order to show the effect of iteration.

4.2 XNLI: Cross-lingual Natural Language Inference

Natural language inference [5, 19] is a typical task to evaluate the performance of sentence embeddings. Given two sentences, one called premise, denoted as p, and another is called hypothesis, denoted as h, this task is to predict the relation between them, including entailment, contradiction and neutral. XNLI is a multi-lingual version of natural language inference dataset, which contains 2500 development sentences and 5000 test sentences translated from English.

The training set is not translated to other languages, so in this task we train the model on English and evaluate it on other languages. We extract features by the pretrained models mentioned above and the features of each sentence pair is the concatenation of $Enc(p)$, $Enc(h)$, $Enc(p)*Enc(h)$ and $|Enc(p)-Enc(h)|$. We train a three-layer fully connected neural network, with hidden dimensions 512 and 384 respectively, on the extracted features. We do not use any regularization here and we just early stop the training on English development set. Note that BERT and LASER have statistics on this dataset, but here we use our own task-specific model to make a fair comparison.

As shown in Table 1, in monolingual evaluation, BERT obviously performs best even without fine-tune. In multi-lingual evaluation, as supervised method, LASER learns the most cross-lingual sentence information. Before the iteration, the performance of the proposed approach is much worse than multi-lingual BERT. But after iteration, the proposed approach improves strikingly especially in cross-lingual evaluations.

Table 1. Results on XNLI dataset.

Method		en	zh	fr	es	de	overall
Supervised	LASER	63.0	59.1	60.5	56.4	55.7	57.9
Unsupervised	mean	49.2	44.3	46.1	46.9	45.2	45.6
	sum	48.1	41.0	45.5	46.1	45.2	44.5
	multi-lingual bert	**67.0**	47.0	49.1	48.5	**49.8**	48.6
Our Approach	-iteration	59.1	45.4	46.7	44.8	43.2	45.0
	+iteration	59.6	**50.0**	**49.2**	**49.9**	48.7	**49.5**

Another evaluation metric here is the distance of the performance of multi-lingual task to that of monolingual task. With the huge capacity, BERT achieves striking results on monolingual evaluation, but the performance of multi-lingual

evaluation is 20% lower. LASER still performs best on this aspect. The proposed approach decreases the distance by 5% after the interaction which proves the effectiveness and significance of the interactions among languages.

4.3 RCV2: Cross-lingual Text Classification

Cross-lingual text classification is another important task for the evaluation of cross lingual sentence embeddings. In this task we train a classification model on one language and test it on other languages. RCV2 [11] is a dataset containing 487,000 articles in 13 languages. There are no parallel sentences or documents among different languages, and it has four classes. However, articles in this dataset is too long for sentence encoders, so we only take headlines of articles as the input of sentence encoders and extract features only from these titles.

Here we still follow the above settings. We first extract features by sentence encoders and then train a feed-forward neural network with one hidden layer on the features extracted. The L2 regularization of LR is tuned on the development set that is in the same language as training set. The results of this experiment are shown in Table 2. We can find that, the proposed method achieves the best performance on all datasets except English to French and German to English. Different sentence encoders perform alike in this evaluation which proves the importance of interactions among sentences.

Table 2. Results on RCV2 title dataset.

Method		en-zh	en-de	en-fr	en-es	zh-en	de-en	es-en	fr-en
Supervised	LASER	73.2	70.3	68.1	75.4	68.2	66.4	69.0	71.1
Unsupervised	mean	41.1	60.2	61.1	46.3	55.2	51.3	54.9	56.6
	sum	52.7	59.1	58.4	69.4	52.0	62.1	39.1	58.3
	multi-lingual bert	58.3	51.4	**72.3**	52.6	47.2	**57.6**	54.6	58.4
Our Approach	-iteration	58.6	50.4	56.1	51.0	48.6	54.3	55.4	57.6
	+iteration	**64.5**	**65.1**	67.6	**60.2**	**59.2**	56.7	**59.8**	**61.8**

4.4 Parameter Analysis

In this section, we investigate the impact of some important factors on the performance of sentence embeddings. The task we use here is RCV2 dataset on cross-lingual sentence classification because it is an easier task so that we can observe the influences clearly. The language pair we select is English and Chinese because they are remote and some languages, like English and German, naturally share some common word pieces under BPE.

Training Corpora. Our experiments are conducted on Wikipedia, and Wikipedia in different languages share many common contents, in which we can dig many parallel sentences. To prove that the proposed approach is robust to training corpora, we conduct an experiment on Toronto Book Corpus [23] and

Chinese Wikipedia. The two corpora are from exactly different domains and they are also in remote languages, so it is nearly impossible for a parallel signal to exist. The results of the experiment are shown in Table 2

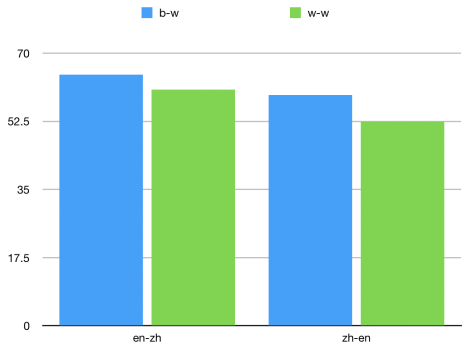

Fig. 2. Results on different Training Corpora. w-w means that two languages are both trained on Wikipedia, and b-w means that English sentences are trained on Toronto Book Corpus and Chinese sentences are trained on Wikipedia. The y-axis represents accuracy of classification.

As shown in Fig. 2, the performance of the proposed approach decreases about 6% on disjoint training corpora. Although the decrease is striking, the sentence encoder can still learn meaningful cross-lingual signals, which means the iteration process does not rely on the existence of parallel sentences, and training can work only if we can retrieve some similar sentences. Hence, our training objective is robust to the domain of training corpora.

Size of Samples. In the multi-lingual coherent regression, we retrieve some sentences to extract semi-parallel sentences. In the above experiments we set the number of sentences $m = 10,000$ for fully utilizing GPU. If the number of samples is small, we will fail to generate meaningful semi-parallel sentences and the training is completely meaningless. Otherwise, if the number is large, the model will be expensive in computation. We try some different number in the experiment, including 1, 100, 1000, 10000, 20000.

Table 3. The influence of the size of samples to the overall performance.

size	1	100	1000	10000	20000
en-zh	50.3	51.7	59.1	64.5	64.9
zh-en	51.6	50.2	52.6	59.2	59.2

As shown in Table 3, the proposed approach is sensitive to the size of samples when the size of samples is small, but when the size of samples is large enough,

increasing the size of samples does not improve the performance obviously but causes extra troubles for GPU computation. When the size of samples is too small, the training is even harmful to the performance because the cosine similarity of disjoint sentences is meaningless. Actually the training is useful only if part of sampled semi-parallel sentences are similar.

Number of Iterations. We repeat the similar sentence extraction and coherence regression process for several times. The number of iterations is an influential factor for the performance of the model. We analyze the accuracy on English to Chinese and Chinese to English cross lingual sentence classification task.

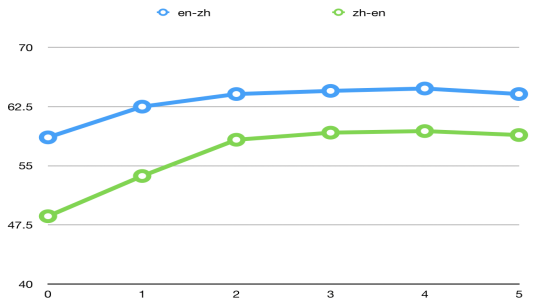

Fig. 3. the x-axis represents the number of iterations and the y-axis means the accuracy on the two datasets.

As shown in Fig. 3, the performance increases obvious in the first and the second iteration, but it almost stays fixed after 3 iterations, so we stop the training after 3 iterations. After 5 iterations the performance even starts to decrease, so too many iterations is not necessarily beneficial for the performance of sentence embeddings.

5 Related Works

This paper involves two research directions including unsupervised methods aligning languages and general purpose sentence representation. In this section, we will briefly introduce the recent progress of them.

Unsupervised Alignment of Words: GAN-based methods regard word embedding of different languages as different probability distributions, and directly align two distributions as a whole. such as [4, 20, 6, 21]. [21] employ Wasserstein-GAN to train the model and minimize earth mover's distance to refine the vectors after training. Although GAN-based methods work well in their original paper, [2] point out that they lack robustness. Iteration-based methods are more robust by this way. [1] firstly propose the iteration approach to learn across language mapping from a small seed dictionary (as small as 25 parallel words). After that, [2] create a new fully unsupervised method to generate the initial dictionary and

proceed the above iteration. Their experiments show that their approach is competitive and more robust than GAN-based methods.

General-Purpose Sentence Representation: General-purpose sentence representation methods, based on sentence level distributional hypothesis [13], usually learn a sentence encoder from a large amount of unlabelled data. The encoders can be used as sentence feature extractors to initialize other tasks. Most of such works are designed only for monolingual situations. The most simple approach is to train a sentence level log-linear model, like [9, 10, 17, 8]. In spite of the good performance, training of seq2seq is time-consuming on large datasets. [12] transform the task to a classification problem. They abandon the decoding process and convert the problem to a simple classification task (to classify if a sentence is in the context of another sentence).

6 Conclusion

We propose a novel approach to learn multi-lingual sentence embeddings from mono-lingual corpora by utilizing Language Isomorphism and sentence-level Distributional Hypothesis. Although the performance is still not competitive to supervised methods, we provide a new view of extracting and utilizing similar sentences as supervised signal.

Acknowledgement

The work is supported by NSFC projects (U1736204, 61533018, 61661146007), Ministry of Education and China Mobile Joint Fund (MCM20170301), a research fund supported by Alibaba Group, and THUNUS NExT Co-Lab.

References

1. Artetxe, M., Labaka, G., Agirre, E.: Learning bilingual word embeddings with (almost) no bilingual data. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. pp. 451–462 (2017)
2. Artetxe, M., Labaka, G., Agirre, E.: A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. pp. 789–798 (2018)
3. Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. arXiv preprint arXiv:1812.10464 (2018)
4. Barone, A.V.M.: Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In: Proceedings of the 1st Workshop on Representation Learning for NLP. pp. 121–126 (2016)
5. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 632–642 (2015)
6. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. arXiv preprint arXiv:1710.04087 (2017)

7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT 2019. pp. 4171–4186 (2018)
8. Gan, Z., Pu, Y., Henao, R., Li, C., He, X., Carin, L.: Learning generic sentence representations using convolutional neural networks. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2390–2400 (2017)
9. Hill, F., Cho, K., Korhonen, A.: Learning distributed representations of sentences from unlabelled data. In: Proceedings of NAACL-HLT 2016. pp. 1367–1377 (2016)
10. Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S.: Skip-thought vectors. In: Advances in neural information processing systems. pp. 3294–3302 (2015)
11. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: Rcv1: A new benchmark collection for text categorization research. Journal of machine learning research **5**, 361–397 (2004)
12. Logeswaran, L., Lee, H.: An efficient framework for learning sentence representations. arXiv preprint arXiv:1803.02893 (2018)
13. Sahlgren, M.: The distributional hypothesis. Italian Journal of Disability Studies **20**, 33–53 (2008)
14. Schwenk, H., Douze, M.: Learning joint multilingual sentence representations with neural machine translation. In: Proceedings of the 2nd Workshop on Representation Learning for NLP. pp. 157–167 (2017)
15. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. p. 1715–1725 (2015)
16. Storer, T.: Linguistic isomorphisms. The University of Chicago Press on behalf of the Philosophy of Science Association **19**(1), 77–85 (1952)
17. Tang, S., Jin, H., Fang, C., Wang, Z., de Sa, V.R.: Rethinking skip-thought: A neighborhood based approach. In: Proceedings of the 2nd Workshop on Representation Learning for NLP. pp. 211–218 (2017)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
19. Williams, A., Nangia, N., Bowman, S.R.: A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of NAACL-HLT 2018. pp. 1112–1122 (2017)
20. Zhang, M., Liu, Y., Luan, H., Sun, M.: Adversarial training for unsupervised bilingual lexicon induction. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. pp. 1959–1970 (2017)
21. Zhang, M., Liu, Y., Luan, H., Sun, M.: Earth mover's distance minimization for unsupervised bilingual lexicon induction. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 1934–1945 (2017)
22. Zhang, M., Wu, Y., Li, W., Li, W.: Learning universal sentence representations with mean-max attention autoencoder. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4514–4523 (2018)
23. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. arXiv preprint arXiv:1506.06724 (2015)