ArbEngVec : Arabic-English Cross-Lingual Word Embedding Model

Raki Lachraf, El Moatez Billah Nagoudi, Youcef Echahid, Ahmed Abdelali, Didier Schwab

ArbEngVec : Arabic-English Cross-Lingual Word Embedding Model

Raki Lachraf
Echahid Hamma Lakhdar University,
El Oued, Algeria
raki.lachraf@univ-eloued.dz

El Moatez Billah Nagoudi
Echahid Hamma Lakhdar University,
El Oued, Algeria
LIM laboratory, Laghouat
moatez-nagoudi@univ-eloued.dz

Youcef Ayachi
Echahid Hamma Lakhdar University
El Oued, Algeria
youcef.ayachi@univ-eloued.dz

Ahmed Abdelali
Hamad Bin Khalifa University
Qatar Computing Research Institute
Doha, Qatar
aabdelali@qf.org.qa

Didier Schwab
LIG-GETALP
Univ. Grenoble Alpes,
France
didier.schwab@imag.fr

Abstract

Word Embeddings (WE) are getting increasingly popular and widely applied in many Natural Language Processing (NLP) applications due to their effectiveness in capturing semantic properties of words; Machine Translation (MT), Information Retrieval (IR) and Information Extraction (IE) are among such areas. In this paper, we propose an open source ArbEngVec which provides several Arabic-English cross-lingual word embedding models. To train our bilingual models, we use a large dataset with more than 93 million pairs of Arabic-English parallel sentences. In addition, we perform both extrinsic and intrinsic evaluations for the different word embedding model variants. The extrinsic evaluation assesses the performance of models on the cross-language Semantic Textual Similarity (STS), while the intrinsic evaluation is based on the Word Translation (WT) task.

1 Introduction

Distributed word representations in vector space (Word Embeddings) are one of the most successful applications in deep learning for capturing the semantic and syntactic properties of words. Lately, many NLP tasks have been enriched using tools based on Mono and Cross-Lingual word embedding models. For instance, Mono-Lingual Word Embeddings (MLWE) have been widely used in information retrieval (Vulić and Moens, 2015a), text classification (Lai et al., 2015), semantic textual similarity (Kenter and De Rijke, 2015; Nagoudi and Schwab, 2017) and plagiarism detection (Nagoudi et al., 2018).

Cross-Lingual Word Embeddings (CLWE) is a more challenging task because the knowledge is transferred between two or more different languages (Doval et al., 2018). Recently, cross-lingual word embeddings was used to address several issues, e.g. machine translation (Zou et al., 2013), cross-language information retrieval (Vulić and Moens, 2015a; Zhou et al., 2012), cross-language semantic similarity (Ataman et al., 2016; Nagoudi et al., 2017b) and plagiarism detection across multiple languages (Ferrero et al., 2017; Barrón-Cedeño et al., 2013). Many cross-lingual word embedding models in natural language have been developed, particularly for English, but Arabic did not get that much of interest.

In this paper, we propose six Arabic-English cross-lingual word embedding models[1]. To train these models, we have used a large collection with more than 93 million pairs of parallel Arabic-English sentences.

The rest of this paper is organised as follows: in section 2 we provide a quick overview of work related to the cross-lingual word embedding models. We describe our dataset collection and the preprocessing process in Section 3. Section 4 presents our proposed cross-lingual models. Section 5 presents the evaluation results. Section 6 concludes the paper with our main findings and points to possible directions for future work.

2 Related works

While we focus on the cross-lingual word embedding models, the interested reader may refer to a number of research studies on the subject of mono-lingual word embeddings in general (Collobert and Weston, 2008), (Turian et al., 2010), (Mnih and Hinton, 2009), (Mikolov et al.,

[1] All models can be downloaded from : https://github.com/RnonymousTrain/ArbEngVec.git

2013c,b) and (Peters et al., 2018).

In the cross-lingual context, several word embedding models are proposed. Blunsom and Hermann (2014) introduced a Bilingual Compositional Model (BiCVM). Leveraging from the fact that aligned sentences have the same meaning. BiCVM is based on a sentence-aligned corpus to learn the bilingual word embedding vectors.

Vulić and Moens (2015b) introduced a Bilingual Word Embedding Skip-Gram (BWESG), this model is constructed through three main steps: i) prepare a Skip-Gram Negative Sampling (Mikolov et al., 2013b) architecture that deals with document aligned comparable data, ii) provide bilingual document pairs, iii) shuffle each pair producing pseudo-bilingual document that serves as the architecture's input which is to be trained.

Luong et al. (2015) proposed a Bilingual Skip-Gram model (BiSKip). BiSKip uses the Skip-Gram of (Mikolov et al., 2013b) to train two different languages at the same time by manipulating the Skip-Gram architecture to obtain two pivots and two contexts and provide a training session for each combination. Choosing two Germanic languages (English and German) made it easier to predict target language's appropriate pivot and context for the ones from source language by simply aligning the target words at position $[i * T/S]$ with source words at position i where S and T are source and target sentence lengths respectively.

Chen et al. (2018) presented an Adversarial Deep Averaging Network (ADAN) for cross-lingual sentiment classification. In fact, they trained many bilingual WE models, one of them was trained using the United Nations (UN) English-Arabic parallel aligned corpus (Ziemski et al., 2016) and Bilingual Bag-of-Words without Alignments (BilBOWA) (Stephan Gouws, 2015). Additionally, ADAN replaces the softmax and regularization terms by a less costly alternatives.

Recently, Devlin et al. (2018) have proposed a deep learning method called Bidirectional Encoder Representations from Transformers (BERT) based on overcoming the limitations of *next* and *previous* token prediction procedures benefiting from Masked Language Modeling (MLM) (Taylor, 1953) by masking 15% of the sentence tokens fed into the architecture alongside the transformer encoder (Vaswani et al., 2017). Devlin et al. (2018) have extended their work by applying the same architecture in a Wikipedia corpora of 104 different languages, requiring not a single alignment signal and realising, if not outperforming, state-of-the-art score in many NLP tasks such as Part Of Speech Tagging and Named Entity Recognition. However, BERT demands significantly more machine effort (Wu and Dredze, 2019). Table 1 summarises the cross-language embedding models mentioned above according to the architecture and used corpus, the target languages and the evaluation methods.

3 Dataset Collection

3.1 Corpus Used

The main objective of this work is to provide an efficient Arabic-English cross-lingual word embedding models across different text domains. Indeed, we used a large dataset of parallel Arabic-English sentences mainly extracted from the Open Parallel Corpus Project[2] (OPUS) (Tiedemann, 2012). OPUS contains 90 languages, and more than 2.7 billion parallel sentences. This corpus consists of data from multiple domains and sources including: MultiUN Corpus (Daniel Tapias, 2010), OpenSubtitles (Creutz, 2018), Tanzil (ZarrabiZadeh, 2007), News-Commentary, United Nations (UN) (Ziemski et al., 2016), Wikipedia, TED 2013[3], GNOME[4], Tatoeba[5], Global Voices[6], KDE4[7] and Ubuntu[8] corpus. To train our models, we extract more than 93.9 million parallel sentences of Arabic-English from whole collection, this alignment contains more than 800 million Arabic tokens and 1 billion for English. More details about our dataset are given in Table 2.

3.2 Preprocessing and Normalization

Preprocessing is an important step in building any word embedding model as it can potentially significantly affect the end results. We first remove the punctuation marks, non letters, URLs, emojis and emoticons from the Arabic and English sentences. Additionally, we normalize Arabic sentences using the preprocessing suggested by Nagoudi et al. (2017a):

[2] http://opus.nlpl.eu/
[3] http://www.casmacat.eu/corpus/ted2013.html
[4] https://l10n.gnome.org
[5] www.tatoeba.org
[6] https://globalvoices.org/
[7] http://i18n.kde.org
[8] https://translations.launchpad.net

CLWE Models	Corpus Used	Arch.	Languages	Evaluation
BiCVM (Hermann and Blunsom, 2014)	Europarl (Koehn, 2005), TED (Cettolo et al., 2012), RCV (Lewis et al., 2004)	CVM	English, German, French, Arabic, Spanish, Italian, Dutch, Brazilian	Cross-lingual classification
BiSKip (Luong et al., 2015)	UN corpus Koehn (2005)	Skip-Gram	English, German	Mono and bilingual word similarity, cross-lingual classification
BWESG (Vulić and Moens, 2015b)	UN corpus Koehn (2005)	Skip-Gram	English, Dutch	Mono and cross-lingual ad-hoc retrieval
BilBOWA (Stephan Gouws, 2015)	RCV (Lewis et al., 2004), WMT11 (2011)	CBOW	English, German, Spanish	Word translation, cross-lingual classification
ADAN (Chen et al., 2018)	UN corpus (Ziemski et al., 2016)	Skip-Gram	English, Arabic, Chinese	Domain Adaptation and Machine Translation
mBERT (Devlin et al., 2018)	Large Wikipedia Corpora	BERT	104 Languages (including Arabic)	POS Tagging and NER...etc

Table 1: Different cross-language word embedding models

1. The letters أ ، إ ، آ are replaced with ا while the letter ة is replaced with ه. Also, The letter ى followed by ء replaced with ئ.
2. We converted elongated words back to their original form, example : معاااااهدة, which means *treaty* in English, and الجزائـــــــر , which means *Algeria* will be converted to معاهدة, الجزائر.
3. In addition, we remove the stop-words from Arabic and English sentences.

4 Building ArbEngVec Models

4.1 Used Architectures

In Mikolov et al. (2013a) all the word embedding models (Collobert and Weston, 2008), (Turian et al., 2010), (Mnih and Hinton, 2009), (Mikolov et al., 2010), (Mikolov et al., 2013c) and (Mikolov et al., 2013b) have been compared and evaluated, and they show that CBOW (Mikolov et al., 2013c) and Skip-Gram (Mikolov et al., 2013b) models are significantly faster to train with better accuracy. Accordingly, we used the CBOW and Skip-Gram to build our Arabic-English cross-lingual word embedding models.

The CBOW (Mikolov et al., 2013c) and Skip-Gram (Mikolov et al., 2013b) are two shallow neural network architectures with a single hidden layer that learns similar vector representations for words with similar distributional properties. The CBOW model, predicts a targeted word w_t according to the context in which w_t appears by using a window of contextual words. While the Skip-Gram model, predicts the words around the word w_t (Mikolov et al., 2013a), as illustrated in figure 1.

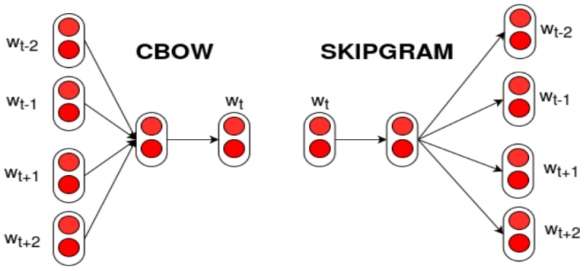

Figure 1: Architecture of CBOW and Skip-gram as described in (Mikolov et al., 2013b)

4.2 Proposed Models

In this section, we present our proposed ArbEngVec models. In order to learn our models, we have relied basically on shuffling the corpus as in Vulić and Moens (2015b), with one

Corpus	Content	Documents	Sentences	Ar Words	En Words
MultiUN Corpus.	The official documents of the United Nations (UN)	67617	10.6M	263.1M	289.6M
OpenSubtitles.	A new collection of translated movie subtitles	104325	81.4M	501.5M	695.9M
Tanzil.	A collection of Quran translations	30 Quran Party	0.2M	7.9M	5.6M
News-Comment	A parallel corpus of News Commentaries provided by WMT for training Statistical Machine Translation (SMT)	7185	0.6M	15.4M	15.5M
UN.	A collection of translated documents from the United Nations originally	1	74.1k	3.3M	3.7M
Wikipedia.	A corpus of parallel sentences extracted from Wikipedia	1	0.2M	3.2M	3.5M
TED2013.	A parallel corpus of TED talk subtitles provided by CASMACAT	1	0.2M	2.4M	3.0M
GNOME.	A parallel corpus of GNOME localization files	1313	0.5M	2.4M	2.6M
Tatoeba.	A collection of translated sentences from Tatoeba	1	13.0k	90.1k	3.6M
GlobalVoices.	A parallel corpus of news stories from the web site Global Voices	7017	93.9k	2.1M	3.0M
KDE4.	A parallel corpus of KDE4 system messages	784	0.1M	0.7M	0.8M
Ubuntu.	A parallel corpus of the Ubuntu Dialogue Corpus	299	56.3M	0.2M	0.5M
EUbookshop.	Corpus of documents from the EU bookshop an online service and archive of publications from various European institutions	30	1.7k	80.0k	0.4M
Total	All the corpus used and extracted from OPUS	188606	93.9M	802.3M	1.0G

Table 2: Some statistics about the used dataset (Tiedemann, 2012)

	CBOW					Skip-Gram				
#Modes	**Top1**	**Top2**	**Top3**	**Top5**	**Top10**	**Top1**	**Top2**	**Top3**	**Top5**	**Top10**
Parallel	0.1%	0.5%	0.7%	1.2%	2.1%	2.8%	4.5%	6.1%	6.1%	9.3%
W. by W.	4.1%	11.3%	17.4%	25.3%	37.2%	60.6%	73.5%	78.3%	86.8%	92.4%
Random	57.7%	71.4%	79.2%	85.3%	90.5%	62.4%	74.2%	78.4%	87.5%	**93.8%**

Table 3: Intrinsic evaluation results of ArbEngVec models

major difference choosing sentence-aligned parallel data rather than their comparable document-aligned choice. Indeed, we propose to use three methods for learning our models: *Parallel Mode*, *Word by Word Alignment Mode* and *Random Shuffling Mode*.

4.2.1 Parallel Mode

To make clear that shuffling methods adds cross-lingual improvements, we decided to train a model without any alignment. For example, let S_{ar} and S_{en} be Arabic and English sentences:

S_{ar} = "الولدان الصغيران شقيقان".

S_{en} = " *The young boys are brothers*".

The pair (S_{ar}, S_{en}) were fed directly to the training as follows: "*young, boys, brothers,* الولدان, الصغيران, شقيقان".

4.2.2 Word by Word Alignment Mode

The second method used on the same corpus type with aligning pairs *word by word* and paying attention to sentences length and start aligning with the longest (the short sentence words will be surrounded with those of the long sentence). This method supports using pairs with almost equal lengths. In this situation, stop-words removal preprocessing step is highly blessed. We shall continue with the sentences of the previous example, the input of the training is : " *young,* الولدان, *boys,* الصغيران, *brothers,* شقيقان".

4.2.3 Random Shuffling Mode

In this method, we put each pair of bilingual sentences as a list that contains their words and shuffle it **randomly** and separately from the rest of the corpus to have a list of combined English-Arabic tokens. As shown in our example : " *young,* الصغيران, الولدان, *boys, brothers,* شقيقان".

4.3 Parameters and Training Environment

Training word embedding models require the choice of some parameters affecting the resulting vectors. For our CBOW models we have used recommended parameters values proposed by (Mikolov et al., 2013c). Thus, we set the $vector\ size$ to 300, the $window$ = 5, and $Frequency\ threshold$ = 100. Regarding the Skip-gram models we have chosen Negative Sampling with $negative$ = 5 instead of Hierarchical Softmax. Worth mentioning that all models were trained on 10 epochs with Řehřek and Sojka (2011) GenSim tool.

Concerning the training environment, we have used *Google Colaboratory*[9] research project (also known as *Colab*) for training our model variants. It is a perfectly prepared developing environment with no requirements but a browser. This environment provides a free 12 GB of GPU, also access to *Google Drive* personal account for saving and loading files and there are many other services that can be plugged into it[10].

5 Evaluation

Usually multilingual models go against two aspects of evaluation methodology: maintain monolingual aspect and provide the other cross-lingual. Clearly for us, after creding on the shuffle we lost the former willingly to stick around the latter. Preserving the model's monolingual behaviour requires keeping words in a semantic meaningful order, which is exactly what happens with our first parallel (non-shuffling) model with completely skewed cross-lingual aspect. To clarify that, we have evaluated our models through Semantic Textual Similarity as extrinsic, and Word Translation as intrinsic.

5.1 Intrinsic Evaluation

In this step, we basically focused on word translation following (Stephan Gouws, 2015) evaluation procedure, so we generated a 1000 tuples starting with choosing random 1000 words from the model vocabulary. Then, we find their k-$closest$ (k most similar) cross-lingual words based on the cosine similarity in our six ArbEngVec models. In fact, we have used five different values of k to generate the 1-$closest$, 2-$closest$, 3-$closest$, 5-$closest$ and 10-$closest$ words. For example, Table 4 shows the 5-$closest$ words of مالزيا and $weapons$ in our *random Skip-Gram* model. Afterwards, we calculate the accuracy of each range, which has been calculated by giving a value 1 to each word couple that represents a translation, we make sure that the word provided by our model is a translation with comparing it to Google Translate API's bag of words, if this comparison comes negative we compare manually, if also manual comparison comes negative we give negative score 0. Eventually we count the average of the 1000 scores. Results of the six studied models are provided in Table 3.

Discussion. Parallel results were so dim bilingually as Table 3 shows, but monolingual aspect was preserved especially in CBOW variant. This fact is illustrated in Table 5, the same 5-$closest$ words of مالزيا and $weapon$ using Parallel CBOW model. Switching to *word by word* alignment method, both variants gave promising results and notably Skip-gram's by an average of 59.26% from CBOW, and these are a consequence of getting word translation pairs at the context window range but still since Arabic and English are structurally different this alignment method had its inconvenience. Arriving to *random shuffle* variants which have given the best results and again Skip-Gram with average of 2.44% better than CBOW.

5-closest (مالزيا)	5-closest (weapons)
malaysia, قرغيزيستان, منغوريا, تونغا, كودت	الأسلحة, أسلحة, الدمار, mass, indiscriminite

Table 4: A sample of 5-$closest$ words of مالزيا and $weapons$ in our Random Skip-Gram model

5-closest (مالزيا)	5-closest (weapons)
مدغشقر, المكسيك, نيبال, نيجريا, ليسوتو	arms, weaponry, warheads, missiles, arsenals

Table 5: A sample of 5-$closest$ words of مالزيا and $weapons$ in our Parallel CBOW model

[9] https://colab.research.google.com/

[10] All scripts used for training our models in *Colab* are available on : https://colab.research.google.com/drive/ 1QejCcJ9Ofp07Xw8 IIpr8IeUFbCUy9XQ3

5.2 Extrinsic Evaluation

Extrinsic evaluating means surveilling the model performance under real-world Natural Language Processing tasks use. Our choice fell on Semantic Sentences Similarity (STS) task. To estimate the semantic similarity between the Arabic-English sentences, we have used the WE-based approach proposed by Nagoudi et al. (2017b) jointly with our ArbEngVec models. In fact, we have had STS2017-Eval[11] datasets drawn from the shared taskSemEval-2017 Task1: STS Cross-lingual Arabic-English (Cer et al., 2017). The sentence pairs of STS2017-Eval have been manually labelled by five annotators, and the similarity score is the average of the annotators judgments. Afterwards, in order to evaluate the performance of each model, we calculate Pearson correlation between our assigned semantic similarity scores and human judgement. Table 6 reports the results of the six studied models.

# Modes	**CBOW**	**Skip-Gram**
Parallel.	6.3%	18.1%
W. by W.	49.4%	73.6%
Random.	52.8%	**75.7%**

Table 6: Extrinsic evaluation results of ArbEngVec models

Discussion. These results indicate that when the *parallel* alignment is used the correlation rate gets very low in both architectures. This is due to the distance of every word and its translation in the parallel sentences pair shape. However, when applying the *word by word* alignment the correlation rate is clearly outperformed to 49.4% and 73.6% with the CBOW and Skip-Gram model respectively. Additionally, the observed results indicate that the *random shuffling* method with Skip-Gram model is the best performing method with a correlation rate of 75.7%.

5.3 Models Visualization

As part of the discussion, we have chosen to illustrate our models using *pyplot* scatters with Maaten and Hinton (2008) *t-SNE* algorithm. We provide these visualizations by choosing 20 arbitrary words from our vocabulary, run 4-*closest* similarity to each word and finally project all of them on the 2-dimensional plot. Starting with *parallel* mode models, charts show that distance between Arabic markers are distant from others of English comparing to those of the same language. Same thing can be said on the situation that concerns *word by word* method CBOW variant with less distant languages but still marker bags most often do not include translation pairs. Eventually, *random* variant charts make it clear that close markers include translation pairs alongside mono and cross-lingual similarities, six model charts are in figure 2. Especially for Skip-Gram variant, supposedly that t-SNE feature reduction procedure got rid of both language characteristics, as figure 3 shows, words and their translations most often appear next to each other.

6 Conclusion

In this paper, we have presented the open source project named ArbEngVec. This project provides several Arabic-English cross-lingual word embedding models. The embedding models are learned through a large dataset of parallel Arabic-English sentences. Additionally, we evaluated the ArbEngVec models via extrinsic and intrinsic evaluations. In the extrinsic evaluation, we used the cross-language semantic similarity task to test the capability of our models to capture the semantic and syntactic properties of words in two different languages. While in the intrinsic evaluations, we employed the embedding vectors to evaluate the word translation task.

As future work, we are going to use these models with those of other classical NLP techniques, including word sense disambiguation, named entity recognition to make more improvement in the Arabic-English cross-language semantic similarity and plagiarism detection. We also are going to aim on finding better word alignment methods to improve features capturing regarding the transfer between Semitic and Germanic languages.

References

Duygu Ataman, Jose GC De Souza, Marco Turchi, and Matteo Negri. 2016. Fbk hlt-mt at semeval-2016 task 1: Cross-lingual semantic similarity measurement using quality estimation features and compositional bilingual word embeddings. In *Proceed-*

[11] http://alt.qcri.org/semeval2017/task1/index.php?id=data-and-toolsb

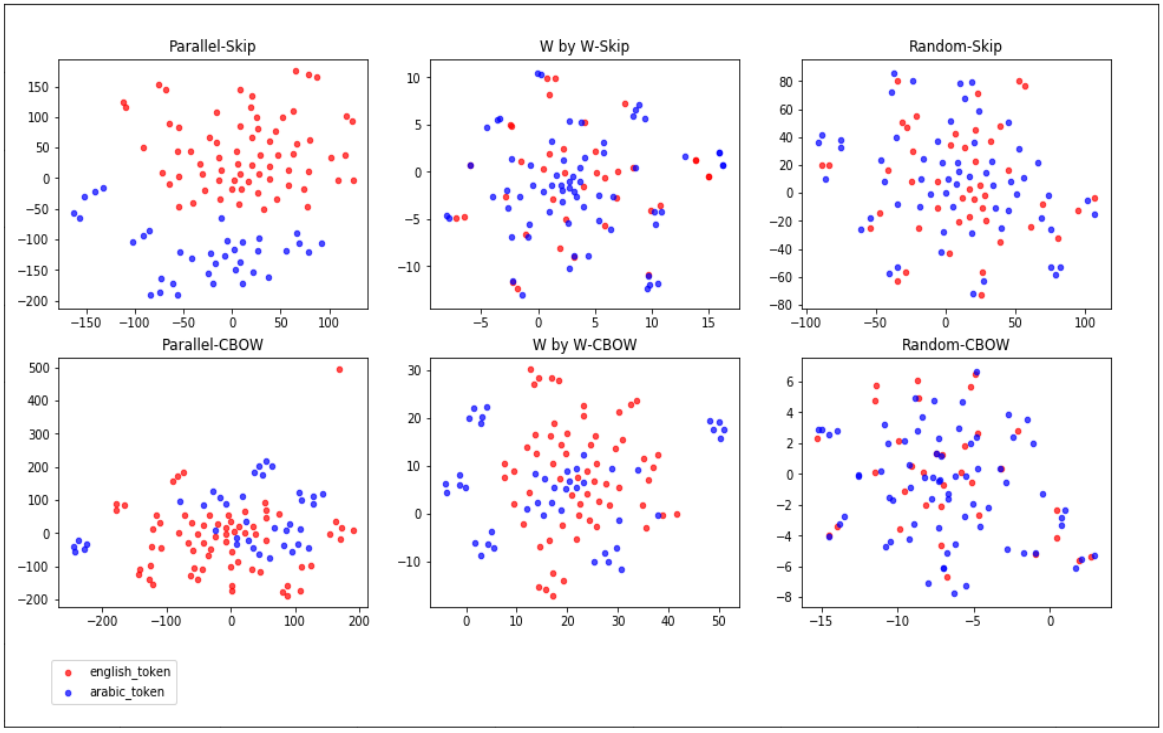

Figure 2: Charts of the model's six variants

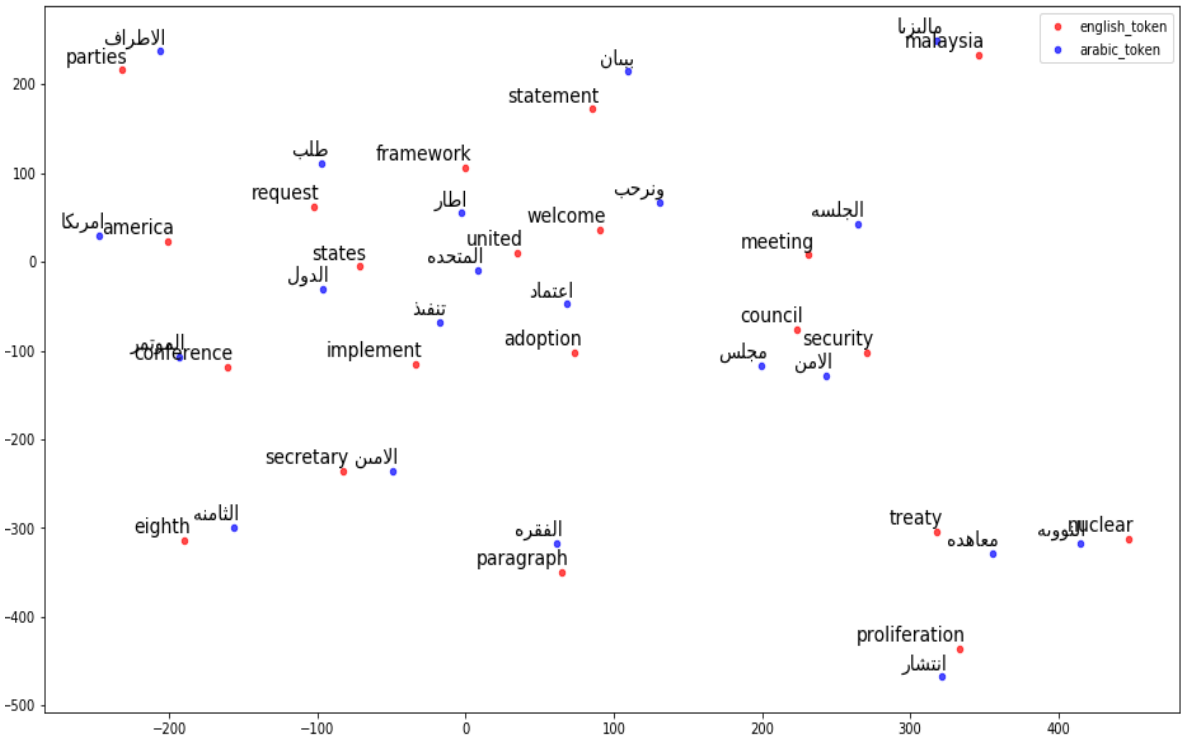

Figure 3: Chart of Random Skip-Gram model

ings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 570–576.

Alberto Barrón-Cedeño, Parth Gupta, and Paolo Rosso. 2013. Methods for cross-language plagiarism detection. *Knowledge-Based Systems*, 50:211–217.

Phil Blunsom and Karl Moritz Hermann. 2014. Multilingual models for compositional distributional semantics.

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. ACL.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Conference of European Association for Machine Translation*, pages 261–268.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. *arXiv preprint arXiv:1809.06142*.

Stelios Piperidis Jan Odjik Joseph Mariani Bente Maegaard Khalid Choukri Nicoletta Calzolari Daniel Tapias, Mike Rosner. 2010. Multiun: A multilingual corpus from united nation documents.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yerai Doval, Jose Camacho-Collados, Luis Espinosa-Anke, and Steven Schockaert. 2018. Improving cross-lingual word embeddings by meeting in the middle. *arXiv preprint arXiv:1808.08780*.

Jérémy Ferrero, Frédéric Agnes, Laurent Besacier, and Didier Schwab. 2017. Usingword embedding for cross-language plagiarism detection. *arXiv preprint arXiv:1702.03082*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. *arXiv preprint arXiv:1404.4641*.

Tom Kenter and Maarten De Rijke. 2015. Short text similarity with word embeddings. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1411–1420. ACM.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. pages 151–159.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *In: ICLR: Proceeding of the International Conference on Learning Representations Workshop Track*, pages 1301–3781.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3.

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

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Hlt-naacl*, volume 13, pages 746–751.

Andriy Mnih and Geoffrey E Hinton. 2009. A scalable hierarchical distributed language model. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1081–1088. Curran Associates, Inc.

El Moatez Billah Nagoudi, Jérémy Ferrero, and Didier Schwab. 2017a. Lim-lig at semeval-2017 task1: Enhancing the semantic similarity for arabic sentences with vectors weighting. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 134–138.

El Moatez Billah Nagoudi, Jérémy Ferrero, Didier Schwab, Hadda Cherroun, et al. 2017b. Word embedding-based approaches for measuring semantic similarity of arabic-english sentences. In *International Conference on Arabic Language Processing*, pages 19–33. Springer.

El Moatez Billah Nagoudi, Ahmed Khorsi, Hadda Cherroun, and Didier Schwab. 2018. A two-level plagiarism detection system for arabic documents. *Cybernetics and Information Technologies*, 20.

El Moatez Billah Nagoudi and Didier Schwab. 2017. Semantic similarity of arabic sentences with word embeddings. In *Third Arabic Natural Language Processing Workshop*, pages 18–24.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Radim Řehřek and Petr Sojka. 2011. Gensimstatistical semantics in python. *statistical semantics; gensim; Python; LDA; SVD*.

Greg Corrado Stephan Gouws, Yoshua Bengio. 2015. Bilbowa: Fast bilingual distributed representations without word alignments.

Wilson L Taylor. 1953. cloze procedure: A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.

Jorg Tiedemann. 2012. Parallel data, tools and interfaces in opus.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz 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. 2015a. 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*, pages 363–372. ACM.

Ivan Vulić and Marie-Francine Moens. 2015b. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. pages 363–372.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

Francisco Zamora-Martinez and Maria Jose Castro-Bleda. 2011. Ceu-upv english-spanish system for wmt11. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 490–495. Association for Computational Linguistics.

Hamid ZarrabiZadeh. 2007. Tanzil project. *URL: http://tanzil. net/wiki/Tanzil_Project*.

Dong Zhou, Mark Truran, Tim Brailsford, Vincent Wade, and Helen Ashman. 2012. Translation techniques in cross-language information retrieval. *ACM Computing Surveys (CSUR)*, 45(1):1.

Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Lrec*.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.