Methodical Evaluation of Arabic Word Embeddings

Mohammed Elrazzaz
Computer Science and Engineering Department
Qatar University
Doha, Qatar
mohammed.elrazzaz@qu.edu.qa

Shady Elbassuoni
Computer Science Department
American University of Beirut
Beirut, Lebanon
se58@aub.edu.lb

Khaled Shaban
Computer Science and Engineering Department
Qatar University
Doha, Qatar
khaled.shaban@qu.edu.qa

Chadi Helwe
Computer Science Department
American University of Beirut
Beirut, Lebanon
cth05@aub.edu.lb

Abstract

Many unsupervised learning techniques have been proposed to obtain meaningful representations of words from text. In this study, we evaluate these various techniques when used to generate Arabic word embeddings. We first build a benchmark for the Arabic language that can be utilized to perform intrinsic evaluation of different word embeddings. We then perform additional extrinsic evaluations of the embeddings based on two NLP tasks.

1 Introduction

Distributed word representations, commonly referred to as word embeddings, represent words as vectors in a low-dimensional space. The goal of this deep representation of words is to capture syntactic and semantic relationships between words. These word embeddings have been proven to be very useful in various NLP applications, particularly those employing deep learning.

Word embeddings are typically learned using unsupervised learning techniques on large text corpora. Many techniques have been proposed to learn such embeddings (Pennington et al., 2014; Mikolov et al., 2013; Mnih and Kavukcuoglu, 2013). While most of the work has focused on English word embeddings, few attempts have been carried out to learn word embeddings for other languages, mostly using the above mentioned techniques.

In this paper, we focus on Arabic word embeddings. Particularly, we provide a thorough evaluation of the quality of four Arabic word embeddings that have been generated by previous work (Zahran et al., 2015; Al-Rfou et al., 2013). We use both intrinsic and extrinsic evaluation methods to evaluate the different embeddings. For the intrinsic evaluation, we build a benchmark consisting of over 115,000 word analogy questions for the Arabic language. Unlike previous attempts to evaluate Arabic embeddings, which relied on translating existing English benchmarks, our benchmark is the first specifically built for the Arabic language and is publicly available for future work in this area [1]. Translating an English benchmark is not the best strategy to evaluate Arabic embeddings for the following reasons. First, the currently available English benchmarks are specifically designed for the English language and some of the questions there are not applicable to Arabic. Second, Arabic has more relations compared to English and these should be included in the benchmark as well. Third, translating an English benchmark is subject to errors since it is usually carried out in an automatic fashion.

In addition to the new benchmark, we also extend the basic analogy reasoning task by taking into consideration more than two word pairs when evaluating a relation, and by considering the top-5 words rather than only the top-1 word when answering an analogy question. Finally, we perform an extrinsic evaluation of the different embeddings using two different NLP tasks, namely Document Classification and Named Entity Recognition.

2 Related Work

There is a wealth of research on evaluating unsupervised word embeddings, which can be can be broadly divided into intrinsic and extrinsic evalu-

[1] http://oma-project.com/res_home

Relation	(a, b)		(c, d)		#pairs	#tuples
Capital	Egypt مصر	Cairo القاهرة	Qatar قطر	Doha الدوحة	124	15252
Currency	Egypt مصر	Pound الجنيه	Qatar قطر	Riyal الريال	155	23870
Male-Female	boy ولد	girl بنت	husband زوج	wife زوجة	101	10100
Opposite	male ذكر	female أنثي	woke up أصبح	slept أمسي	110	11990
Comparative	big كبير	bigger أكبر	small صغير	smaller أصغر	100	9900
Nationality	Holland هولندا	Dutch هولندي	India الهند	Indian هندي	100	9900
Past Tense	travel سفر	traveled سافر	fight قتال	fought قاتل	110	11990
Plural	man رجل	men رجال	house بيت	houses بيوت	111	12210
Pair	man رجل	2 men رجلان	house بيت	2 houses بيتان	100	9900
ALL					1011	115112

Table 1: Summary of the Arabic Word Analogy Benchmark

ations. Intrinsic evaluations mostly rely on word analogy questions and measure the similarity of words in the low-dimensional embedding space (Mikolov et al., 2013; Gao et al., 2014; Schnabel et al., 2015). Extrinsic evaluations assess the quality of the embeddings as features in models for other tasks, such as semantic role labeling and part-of-speech tagging (Collobert et al., 2011), or noun-phrase chunking and sentiment analysis (Schnabel et al., 2015). However, all of these tasks and benchmarks are build for English and thus cannot be used to assess the quality of Arabic word embeddings, which is the main focus here.

To the best of our knowledge, only a handful of recent studies attempted evaluating Arabic word embeddings. Zahran et al. (Zahran et al., 2015) translated the English benchmark in (Mikolov et al., 2013) and used it to evaluate different embedding techniques when applied on a large Arabic corpus. However, as the authors themselves point out, translating an English benchmark is not the best strategy to evaluate Arabic embeddings. Zahran et al. also consider extrinsic evaluation on two NLP tasks, namely query expansion for IR and short answer grading.

Dahou et al. (Dahou et al., 2016) used the analogy questions from (Zahran et al., 2015) after correcting some Arabic spelling mistakes resulting from the translation and after adding new analogy questions to make up for the inadequacy of the English questions for the Arabic language. They also performed an extrinsic evaluation using sentiment analysis. Finally, Al-Rfou et al. (Al-Rfou et al., 2013) generated word embeddings for 100 different languages, including Arabic, and evaluated the embeddings using part-of-speech tagging, however the evaluation was done only for a handful of European languages.

3 Benchmark

Our benchmark is the first specifically designed for the Arabic language. It consists of nine relations, each consisting of over 100 word pairs. An Arabic linguist who was properly introduced to the word-analogy task provided the list of relations. Once the nine relations were defined, two different people collectively generated the word pairs. The two people are native Arabic speakers, and one of them is a co-author and the other is not. Table 1 displays the list of all relations in our benchmark as well as two example word pairs for each relation. The full benchmark and the evaluation tool can be obtained from the following link: `http://oma-project.com/res_home`.

Translating an English benchmark is not adequate for many reasons. First, the currently available English benchmarks contain many questions that are not applicable to Arabic. For example, comparative and superlative relations are the same in Arabic, except that the superlatives are usually prefixed with the Arabic equivalent of ”the”. Another example is the opposite relation, where some words in Arabic do not have antonyms, in which case the antonym is typically expressed by prefixing the word with ”not”. Second, Arabic has more relations compared to English. For instance, in Arabic there is the pair relation (see Table 1 for an example). Third, translating an English bench-

mark is considerably difficult due to the high ambiguity of the Arabic language.

Given our benchmark, we generate a test bank consisting of over 100,000 tuples. Each tuple consists of two word pairs (a, b) and (c, d) from the same relation. For each of our nine relations, we generate a tuple by combining two different word pairs from the same relation. Once tuples have been generated, they can be used as word analogy questions to evaluate different word embeddings as defined by Mikolov et al. (Mikolov et al., 2013). A word analogy question for a tuple consisting of two word pairs (a, b) and (c, d) can be formulated as follows: "a to b is like c to ?". Each such question will then be answered by calculating a target vector $t = b - a + c$. We then calculate the cosine similarity between the target vector t and the vector representation of each word w in a given word embeddings V. Finally, we retrieve the most similar word w to t, i.e., $argmax_{w \in V \& w \notin \{a,b,c\}} \frac{w \cdot t}{||w||||t||}$. If $w = d$ (i.e., the same word) then we assume that the word embeddings V has answered the question correctly.

We also use our benchmark to generate additional analogy questions by using more than two word pairs per question. This provides a more accurate representation of a relation as mentioned in (Mikolov et al., 2013). For each relation, we generate a question per word pair consisting of the word pair plus 10 random word pairs from the same relation. Thus, each question would consist of 11 word pairs (a_i, b_i) where $1 \leq i \leq 11$. We then use the average of the first 10 word pairs to generate the target vector t as follows: $t = \frac{1}{10} \sum_i^{10} (b_i - a_i) + a_{11}$. Finally we retrieve the closest word w to the target vector t using cosine similarity as in the previous case. The question is considered to be answered correctly if the answer word w is the same as b_{11}.

Moreover, we also extend the traditional word analogy task by taking into consideration if the correct answer is among the top-5 closest words in the embedding space to the target vector t, which allows us to more leniently evaluate the embeddings. This is particularly important in the case of Arabic since many forms of the same word exist, usually with additional prefixes or suffixes such as the equivalent of the article "the" or possessive determiners such as "her", "his", or "their". For example, consider one question which asks "ولد to بنت is like ملك to ?", i.e., "man to woman is like king to ?", with the answer being "ملكه" or "queen". Now, if we rely only on the top-1 word and it happens to be "ملكته", which means "his queen" in English, the question would be considered to be answered wrongly. To relax this and ensure that different forms of the same word will not result in a mismatch, we use the top-5 words for evaluation rather than the top-1.

4 Evaluation

We compare four different Arabic word embeddings that have been generated by previous work. The first three are based on a large corpus of Arabic documents constructed by Zahran et al. (Zahran et al., 2015), which consists of 2,340,895 words. Using this corpus, the authors generated three different word embeddings using three different techniques, namely the Continuous Bag-of-Words (CBOW) model (Mikolov et al., 2013), the Skip-gram model (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). The fourth word embeddings we evaluate in this paper is the Arabic part of the Polyglot word embeddings, which was trained on the Arabic Wikipedia by Al-Rfou et al and consists of over 100,000 words (Al-Rfou et al., 2013). To the best of our knowledge, these are the only available word embeddings that have been constructed for the Arabic language.

4.1 Intrinsic Evaluation

As we mentioned in the previous section, we use our word analogy benchmark to evaluate the embeddings using four different criteria, namely using top-1 and top-5 words when representing relations using two versus 11 word pairs. Tables 2 displays the accuracy of each embedding technique for the four evaluation criteria. Note that we consider a question to be answered wrongly if at least one of the words in the question are not present in the word embeddings. That is, we take into consideration the coverage of the embeddings as well (Gao et al., 2014).

As can be seen in Table 2, the CBOW model consistently outperforms all other compared models for all four evaluation criteria. The performance of Polyglot is particularly low since the embeddings were trained on a much smaller corpus (Arabic portion of Wikipedia), and thus both its coverage and the quality of the embeddings are much lower. As can also be seen from the table, the accuracies of all the methods are boosted when

Model	CBOW	Skip-gram	GloVe	Polyglot	CBOW	Skip-gram	Glove	Polyglot
Relation	top-1 two pairs				top-5 two pairs			
Capital	31%	26.6%	31.7%	0.4%	42.9%	40.8%	47%	1.8%
Currency	3.15%	2%	0.8%	0.4%	4.9%%	3.9%	3.7%	1.6%
Male-Female	29%	24.8%	30.8%	3.8%	45.6%	40.6	52.4%	8.3%
Opposite	7.6%	4.41%	7.3%	2.3%	15.75%	10.65%	19.8%	5.4%
Comparative	23.9%	15.7%	21.7%	1.4%	39.61%	30.95%	38.3%	4%
Nationality	29%	29.91%	25.8%	0.8%	34.65%	39.6%	32.4%	3%
Past Tense	4.3%	2.7%	4.5%	0.4%	11.4%	9.6%	16.7%	1.5%
Plural	23.3%	13.28%	19%	2.9%	45.12%	37.9%	41.9%	7.2%
Pair	8.6%	7.6%	1.8%	0.02%	23%	21.3%	5.3%	0.07%
ALL	16.3%	12.8%	14.5%	1.3%	26.6%	23.8%	26.4%	3.4%
Relation	top-1 11 pairs				top-5 11 pairs			
Capital	28.2%	28.2%	33.8	0%	48.38%	40.3%	50.8%	0.8%
Currency	3.8%	3.8%	0.64%	0.6%	7%	4.5%	2.5%	0.6%
Male-Female	29.7%	25.7%	26.7%	4.9%	48.5%	39.6%	52.4%	7.9%
Opposite	5.4%	3.6%	5.4%	2.7%	16.3%	8.1%	15.4%	3.6%
Comparative	31%	23%	25%	1%	49%	36%	39%	2%
Nationality	35%	32%	34%	1%	41%	43%	39%	4%
Past Tense	1.8%	0%	3.6%	1.8%	15.4%	9%	17.2%	3.6%
Plural	20.7%	11.7%	18%	4.5%	48.6%	39.6%	44.2%	6.3%
Pair	8%	11%	3%	0%	21%	18%	9%	0%
ALL	17.4%	14.8%	16%	1.9%	31.1%	25.4%	28.8%	2.5%

Table 2: Intrinsic evaluation of the word embeddings using different criteria

Model	Document Classification	NER
CBOW	0.948	0.800
Skip-gram	0.954	0.799
GloVe	0.946	0.816
Polyglot	0.882	0.649

Table 3: F-measure for two NLP tasks

representing a relation using 11 pairs rather just two pairs. This validates that it is indeed more appropriate to use more than two pairs to represent relations in word analogy tasks.

When considering the top-5 matches, the accuracies of the embeddings are boosted drastically, which indeed shows that relying on just the top-1 word to assess the quality of embeddings might be unduly harsh, particularly in the case of Arabic.

4.2 Extrinsic Evaluation

We perform extrinsic evaluation of the four word embeddings using two NLP tasks, namely: Arabic Document Classification and Arabic Named Entity Recognition (NER). In the Document Classification task, the goal is to classify Arabic Wikipedia articles into four different classes (person (PER), organization (ORG), location (LOC), or miscellaneous (MISC)). To do this, we relied on a neural network with a Long Short-Term Memory (LSTM) layer (Hochreiter and Schmidhuber, 1997), which is fed from the word embeddings. The LSTM layer is followed by two fully-connected layers, which in turn are followed by a softmax layer that predicts class-assignment probabilities. The model was trained for 150 epochs on 8,000 articles, validated on 1,000 articles, and tested on another 1,000 articles.

In the NER task, the goal is to label each word in a given sequence using one of the following labels: PER, LOC, ORG, and MISC, which represent different Named Entity classes. The same architecture as in the Document Classification task was used for this task as well. The model was trained for 150 epochs on 3,852 sentences and tested on 963 sentence using Columbia's University Arabic Named Entity Recognition Corpus (Columbia University, 2016). We used an LSTM neural network for both tasks since they flexibly make use of contextual data and thus are com-

monly used in NLP tasks such as Document Classification and NER.

As can be seen in Table 3, the first three methods CBOW, Skip-gram and GloVe seem to perform relatively well for both the Document Classification task as well as the NER task with very comparable performance in terms of F-measure. They also clearly outperform Polyglot when it comes to both tasks as well.

4.3 Discussion

Our experimental results indicate the superiority of CBOW and SKip-gram as word embeddings compared to Polyglot. This can be mainly attributed to the fact that the first two embeddings were trained using a much larger corpus and thus had both better coverage and higher accuracies when it comes to the word analogy task. This is also evident in the case of the extrinsic evaluation. Thus, when training word embeddings, it is crucial to use large training data to obtain meaningful embeddings.

Moreover, when performing the intrinsic evaluation of the different embeddings, we observed that relying on just the top-1 word is unduly harsh for Arabic. This is mainly attributed to the fact that for Arabic, and unlike other languages such as English, different forms of the same word exist and these must be taken into consideration when evaluating the embeddings. Thus, it is advised to use the top-k matches to perform the evaluation, where k is 5 for instance. It is also advisable to represent a relation with multiple word pairs, rather than just two as is currently done in most similar studies, to guarantee that the relation is well represented.

5 Conclusion

In this paper, we described the first word analogy benchmark specifically designed for the Arabic language. We used our benchmark to evaluate available Arabic word embeddings using the basic analogy reasoning task as well as extensions of it. In addition, we also evaluated the quality of the various embeddings using two NLP tasks, namely Document Classification and NER.

Acknowledgments

This work was made possible by NPRP 6-716-1-138 grant from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibilty of the authors.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. *arXiv preprint arXiv:1307.1662* .

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research* 12:2493–2537.

Columbia University. 2016. Arabic Named Entity Recognition Task. http://www1.ccls.columbia.edu/~ybenajiba/downloads.html.

Abdelghani Dahou, Shengwu Xiong, Junwei Zhou, Mohamed Houcine Haddoud, and Pengfei Duan. 2016. Word embeddings and convolutional neural network for arabic sentiment classification. In *International Conference on Computational Linguistics*. pages 2418–2427.

Bin Gao, Jiang Bian, and Tie-Yan Liu. 2014. Wordrep: A benchmark for research on learning word representations. *arXiv preprint arXiv:1407.1640* .

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Andriy Mnih and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems*. pages 2265–2273.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1532–1543.

Tobias Schnabel, Igor Labutov, David M Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *EMNLP*. pages 298–307.

Mohamed A Zahran, Ahmed Magooda, Ashraf Y Mahgoub, Hazem Raafat, Mohsen Rashwan, and Amir Atyia. 2015. Word representations in vector space and their applications for arabic. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pages 430–443.