bitext we help AI understand humans

Arabic Semantic Text Similarity and Question Answering Benchmark

Introduction

Word embeddings have revolutionized the field of artificial intelligence, especially for their power to represent text as high-dimensional dense vectors in a shared vector space. These representations make it straightforward to compare different pieces of text for semantic similarity. Also different NLP tasks can benefit from these representations like question answering, semantic similarity, machine translation, conversational agents, topic modeling and others.

In this work we investigate the impact of creating better representations of text by incorporating linguistic data. As an initial step, we are analyzing the impact of lemmatization on the performance of two language understanding tasks: open domain question answering and semantic textual similarity. We focus particularly on morphologically-rich languages like Arabic (MSA, Modern Standard Arabic) in this case.

In the semantic textual similarity task, the purpose is to measure the degree to which two sentences are similar to each other by assigning an integer value denoting similarity. Whereas, in the open domain question answering task, the purpose is to extract an answer from a set of document given a specific question.

State-of-the-art question answering systems have two components: (1) the first is a retriever, which retrieve a set of relevant documents from the whole database of document, (2) the second is a reader, which is responsible for extracting a span of text from each retrieved document that answers the question at hand. In this work, we evaluated the reader component of the question answering system.

For each system, we compared using two kinds of text representations, word-based and lemma-based ones. The lemma-based representations have been created by lemmatizing text using Bitext lemmatizer. The results of our work show that using lemma-based embeddings leads to increase in performance over using word-based embeddings on all of the three datasets.

Daset

We worked with the following 3 datasets:

1. NSURL-2019 Task 8 Arabic Semantic Question Similarity dataset:

This dataset consists of 15.7k short question pairs in Modern Standard Arabic (MSA) that are annotated with either a value of "1", denoting similarity, or a value of "0", denoting non-similarity.

Developing a system for semantic question similarity calculation has several applications like detecting duplicate questions on Q/A platforms.

2. SemEval-2017 Task 1 Semantic Textual Similarity Arabic dataset:

There are several datasets published as part of the SemEval-2017 Task 1 competition. In this work, we worked with the monolingual Arabic dataset that consists of 250 pairs of sentences.

This dataset differs from the previous one in that sentences are annotated with a similarity score between 0 and 5, instead of a binary score, which is more challenging.

3. Arabic Reading Comprehension Dataset (ARCD) dataset:

This is an Arabic Question Answering dataset where the test set consists of 702 examples in the format of (context, question, and answer). The content of this dataset was extracted from Arabic Wikipedia, so questions and answers are factual in nature.

Experiments and Results

We have implemented a system using a trainable neural network for the first dataset on the semantic question similarity, whereas we used a non-learning approach for the other two datasets. Here are details on the implemented approaches and the output results.

1. NSURL-2019 Task 8 Arabic Semantic Question Similarity system:

We trained a Siamese neural network that consists of identical bidirectional LSTM layers that share parameters. The output of these layers is passed to a dense layer that predict a similarity value between 0 and 5.

We evaluated this model on the test set using the F1-score metric. The evaluation results are shown in the following table.

	F1-Score
Word embeddings	0.8115
Lemma embeddings	**0.852**

2. SemEval-2017 Task 1 Semantic Textual Similarity system:

We calculated the similarity of both word-based and lemma-based representations of sentence pairs using cosine similarity.

The evaluation is done using the Pearson correlation of our system similarity scores with human annotated scores. The following table shows the evaluation results.

	Pearson coefficients
Word embeddings	55.865
Lemma embeddings	**58.845**

3. Open domain question answering system:

We implemented an embedding reader that predicts an answer from context. We evaluated the reader using the following two metrics:

(1) Exact Match: which measures the percentage of predictions that match the ground truth answer exactly. For example, for example, if there was a question: "What is the capital of the United States?" and the system outputted: "Washington D. C.", and the ground truth was also: "Washington D. C.", this will be accounted as an exact match.

(2) F1-score.

	Exact Match	F!-Score
Word embeddings	0.1424	16.161
Lemma embeddings	**0.2849**	**17.027**

Conclusion

In this work, we evaluated the impact of using lemma-based embeddings on the performance of semantic textual similarity and open domain question answering systems.

We worked with 3 datasets that are in Modern Standard Arabic, which is a morphologically-rich language. We compared the performance when using both word-based and lemma-based embeddings in each case.

Our evaluation results show that using lemma-based embeddings leads to significant improvement in performance. The improvement in performance according to our evaluation was at least 5% on all measured metrics.

Surprisingly, the largest improvement was achieved when evaluating the QA reader using the Exact Match metric, where we were able to increase the measured score by 100%.

Select Customers