УДК 004.623

A COMPARISON BETWEEN QUANTUM SEMANTIC, LSTM AND GRU MODELS IN FINDING ENTANGLEMENT OF NAME ENTITY'S WORDS

Шакер Алаа, Альдарф Алаа (Национальный исследовательский университет ИТМО)

Научный руководитель – проф. ФПИиКТ, Бессмертный И.А.

(Национальный исследовательский университет ИТМО)

Аннотация. This work aims to provide a comparison among three models: Quantum-semantic model, Long Short Term Memory (LSTM), and Gated Recurrent Units (GRU). On the side of LSTM and GRU models, we check their ability to correctly find the name entity containing pair-words. On another side, we study the ability of the quantum-semantic model to correctly detect the entanglement between name’s words.

Введение. As Known the LSTM and GRU are deep learning models used in Named Entity Recognition (NER) tasks, and the name of the entity may be consisting of more than one word, which leads us to the entanglement between words consist that name. In this work, we try to check the ability of three models to define entanglement existence between the words of name entity in the text or not. In this experiment, an Arabic Dataset consisting of more than thirty-six thousand labeled words was used to train the LSTM and GRU models. Moreover, we use a one-hundred name entity consisting of a couple of words to check the ability of models to correctly find them within the target text.

Основная часть.

1. Quantum-semantic model: The algorithm and their steps are as follows: first getting the pair-words consist the named entity (query) and text, then processing text and query, after that building the Hyperspace Analogue to Language (HAL) matrix, and extracting three victors D vector "vector represents the whole text", Dw1 and Dw2 represent the two words of query from this matrix.

The next step is to get orthogonal bases based on Dw1 and Dw2, for getting these bases, we use the rule Gramm-Schmidt, by applying this converting rule on {Dw1, Dw2} and getting the first orthogonal bases $\{\ u_1, u_2\}$, and applying Gramm-Schmidt on {Dw2,Dw1}, and getting the second orthogonal bases$\{\ v_1, v_2\}$ then project D vector on these bases $\{\ u_1, u_2\}$,$\{\ v_1, v_2\}$.

Last step applying Bell test inequity (CHSH), for detecting if there is entanglement between these two words in text or not.

If the result is between $\{2,\ 2{*}\sqrt{2}\}$ that means the entanglement between the two words of query in the text, and that means this text is relevant to the subject of user search. Under 2 that means, there is no entanglement, and this text is not relevant to the subject search.

2. Deep learning model: Our deep learning model consists of four layers. The first layer is the embedding that converts an input word to a vector; after that LSTM or GRU layer has the availability to deal with sequence data. The third layer is the dense layer that applies the Relu activation function; then the output layer uses the Logsoftmax function.

Результат. This experiment uses an Arabic NER dataset to train the NER deep learning model containing thirty-six thousand labeled words. During the experiment, deep learning was trained on different iterations numbers. And to check the ability of models to find the desired pair words in the text, we use eight Arabic texts on different domains with a list containing one hundred-nine pair names entities. The results show that the quantum-semantic model can find fifty named entities containing pair of words, which means about 45.8% of total names contained in the list. On the other hand, the best result of the LSTM model is to identify forty names with two-thousand iterations, while the best development of the GRU model is to identify forty-two names also with two-thousand iterations which equals to 38.5 %.

Выводы. This work presents a comparison between three models, the Quantum semantic model which is an unsupervised model, and LSTM, GRU models which are supervised models. The result shows that the quantum model has the ability to detect whether the pair words "the name of entity containing two words" are entanglement or not in the text, that help us to develop a better retrieval information algorithms to retrieve top relevant texts to the search topic.