Information Processing and Management 54 (2018) 922–937

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A deep network model for paraphrase detection in short text messages

Basant Agarwal[*,a,b], Heri Ramampiaro[a], Helge Langseth[a], Massimiliano Ruocco[a,c]

[a] Department of Computer Science, Norwegian University of Science and Technology, Norway
[b] Department of Computer Science and Engineering, Swami Keshvanand Institute of Technology Management & Gramothan, India
[c] Telenor Research, Trondheim, Norway

ARTICLE INFO

Keywords:
Paraphrase detection
Sentence similarity
Deep learning
RNN
CNN

ABSTRACT

This paper is concerned with paraphrase detection, i.e., identifying sentences that are semantically identical. The ability to detect similar sentences written in natural language is crucial for several applications, such as text mining, text summarization, plagiarism detection, authorship authentication and question answering. Recognizing this importance, we study in particular how to address the challenges with detecting paraphrases in user generated short texts, such as Twitter, which often contain language irregularity and noise, and do not necessarily contain as much semantic information as longer clean texts. We propose a novel deep neural network-based approach that relies on coarse-grained sentence modelling using a convolutional neural network (CNN) and a recurrent neural network (RNN) model, combined with a specific fine-grained word-level similarity matching model. More specifically, we develop a new architecture, called DeepParaphrase, which enables to create an informative semantic representation of each sentence by (1) using CNN to extract the local region information in form of important n-grams from the sentence, and (2) applying RNN to capture the long-term dependency information. In addition, we perform a comparative study on state-of-the-art approaches within paraphrase detection. An important insight from this study is that existing paraphrase approaches perform well when applied on clean texts, but they do not necessarily deliver good performance against noisy texts, and vice versa. In contrast, our evaluation has shown that the proposed DeepParaphrase-based approach achieves good results in both types of texts, thus making it more robust and generic than the existing approaches.

1. Introduction

Twitter has for some time been a popular means for expressing opinions about a variety of subjects. Paraphrase detection in user-generated noisy texts, such as Twitter texts,[1] is an important task for various Natural Language Processing (NLP), information retrieval and text mining tasks, including query ranking, plagiarism detection, question answering, and document summarization. Recently, the paraphrase detection task has gained significant interest in applied NLP because of the need to deal with the pervasive problem of linguistic variation.

Paraphrase detection is an NLP classification problem. Given a pair of sentences, the system determines the semantic similarity

* Corresponding author at: Department of Computer Science, Norwegian University of Science and Technology, Norway.
E-mail addresses: basant.agarwal@ntnu.no (B. Agarwal), heri@ntnu.no (H. Ramampiaro), helgel@ntnu.no (H. Langseth), massimiliano.ruocco@ntnu.no (M. Ruocco).

[1] From now on referred to as Tweets.

https://doi.org/10.1016/j.ipm.2018.06.005
Received 30 November 2017; Received in revised form 11 May 2018; Accepted 14 June 2018

between the two sentences. If the two sentences convey the same meaning, then it is labelled as *paraphrase*; otherwise, it is labeled as *non-paraphrase*. Most of the existing paraphrase systems have performed quite well on clean text corpora, such as the Microsoft Paraphrase Corpus (MSRP) (Dolan, Quirk, & Brockett, 2004). However, detecting paraphrases in user-generated noisy Tweets is more challenging due to issues like misspelling, acronyms, style and structure (Xu, Ritter, Callison-Burch, Dolan, & Ji, 2014). In addition, measuring the semantic similarity between two short sentences is very difficult due to the lack of common lexical features (Kajiwara, Bollegala, Yoshida, & Kawarabayashi, 2017). Although little attention has been given to paraphrase detection in noisy short-texts thus far, some initial work has been reported on the *SemEval 2015* benchmark Twitter dataset (Dey, Shrivastava, & Kaushik, 2016; Xu, Callison-Burch, & Dolan, 2015; Xu et al., 2014). Unfortunately, the best performing approaches on one dataset doesn't seem to perform as good when evaluated against another. As we discuss later in this paper, the state-of-the-art approach for the SemEval dataset proposed by Dey et al. (2016) does not have good performance (in form of F1-score) when evaluated on the MSRP dataset. Similarly, Ji and Eisenstein (2013) is the best performing approach on the MSRP dataset, but does not perform well on the SemEval dataset. In conclusion, existing approaches are not very generic, but rather are highly dependent on the data used for training.

Focusing on the problem discussed above, the main goal of this work is to develop a robust paraphrase detection model based on deep learning techniques that is able to successfully detect paraphrasing in both noisy and clean texts. More specifically, we propose a hybrid deep neural architecture composed by a convolutional neural network (CNN) and a recurrent neural network (RNN) model, further enhanced by a novel word-pair similarity module. The proposed paraphrase detection model is composed of two main components: (1) sentence modelling and (2) pair-wise word similarity matching. First, *sentence modelling* concerns building an effective model to represent the text. To do this, we build a joint CNN and RNN architecture that takes the local features extracted by the CNN as input to the RNN. We take word embeddings as input to the CNN model. Then, after convolutions and pooling operations, the encoded feature maps are taken in sequence as input to the RNN model. The last hidden state learned by the RNN model is considered as the sentence level representation. The main rationale behind using both CNN and RNN here is that the CNN is able to learn the local features in form of important *n-grams* of the texts; whereas RNN takes words in a sequential order and is able to learn the long-term dependencies of texts rather than local features. Second, a *pair-wise similarity matching model* is used to extract fine-grained similarity information between pairs of sentences. Initially, a pair-wise similarity matrix is constructed by computing the similarity of each word in a given sentence to all the words in another sentence. We then apply a CNN onto this similarity matrix to analyse the patterns in the semantic correspondence between each pair of words in the two sentences that are intuitively useful for paraphrase identification. The idea to apply convolutions over the similarity matrix to extract the important word-word similarity pairs is motivated by how convolutions over text can extract the most important parts of a sentence.

In this paper, we show how the proposed model for paraphrase detection can be enhanced by employing an extra set of statistical features extracted from the input text. To demonstrate its robustness, we evaluate the proposed approach and compare it with the state-of-the-art models, using two different datasets, covering both noisy user-generated texts – i.e., the SemEval 2015 benchmark Twitter dataset, and clean texts – i.e., the Microsoft Paraphrase Corpus (MSRP).

The main contributions of this work can be summarized as follows:

1. We propose a novel deep neural network architecture leveraging coarse-grained sentence-level features and fine-grained word-level features for detecting paraphrases on noisy short text from Twitter. The model combines sentence-level and word-level semantic similarity information such that it can capture semantic information at each level. When the text is grammatically irregular or very short, the word-level similarity model can provide useful information; while the semantic representation of the sentence provide useful information otherwise. In this way both model-components complement each other and provide an efficient overall performance.
2. We show how the proposed pair-wise similarity model can be used to extract word-level semantic information, and demonstrate its usefulness in the paraphrase detection task.
3. We propose a method combining statistical textual features and features learned from the deep architecture.
4. We present an extensive comparative study for the paraphrase detection problem.

The rest of the paper is organized as follows: In Section 2, we formally define the problem. In Section 3, we discuss related work concerning paraphrase detection. In Section 4, we motivate our work and present our proposed solution in detail. In Section 5, we describe the experimental setup. In Section 6, we evaluate the approach and discuss the results. Finally, in Section 7, we conclude the paper and outline plans for future research.

2. Problem statement and goals

Let S_1 and S_2 be two sentences, such that $S_1 \neq S_2$. S_1 and S_2 are said to be paraphrased if they convey the same meaning and are semantically equivalent. Now, assume that we have a collection of N annotated sentence pairs (S_1^i, S_2^i), having annotations k_i, for $i = 1, 2, \ldots, N$. For a given i, k_i indicates whether the ith sentence pair is *paraphrased* or *non-paraphrased*. The problem addressed in this paper is to develop a model, which can reliably label a previously unseen sentence pair as paraphrased or non-paraphrased. In particular, we aim at answering the following main research question:

How to develop a robust and generic method for paraphrase detection?

To address the above question, this work has the following main goals: