Information Processing and Management 53 (2017) 640–652

Contents lists available at ScienceDirect

Information Processing and Management

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

# Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features

Mohammad AL-Smadi*, Zain Jaradat, Mahmoud AL-Ayyoub, Yaser Jararweh

*Computer Science Department, Jordan University of Science and Technology, P.O.Box: 3030 Irbid 22110, Jordan*

ARTICLE INFO

*Article history:*
Received 1 July 2016
Revised 4 January 2017
Accepted 6 January 2017
Available online 30 January 2017

*Keywords:*
Paraphrase identification
Semantic text similarity
Semantic analysis
Arabic language
Natural language processing

ABSTRACT

The rapid growth in digital information has raised considerable challenges in particular when it comes to automated content analysis. Social media such as twitter share a lot of its users' information about their events, opinions, personalities, etc. Paraphrase Identification (PI) is concerned with recognizing whether two texts have the same/similar meaning, whereas the Semantic Text Similarity (STS) is concerned with the degree of that similarity. This research proposes a state-of-the-art approach for paraphrase identification and semantic text similarity analysis in Arabic news tweets. The approach adopts several phases of text processing, features extraction and text classification. Lexical, syntactic, and semantic features are extracted to overcome the weakness and limitations of the current technologies in solving these tasks for the Arabic language. Maximum Entropy (MaxEnt) and Support Vector Regression (SVR) classifiers are trained using these features and are evaluated using a dataset prepared for this research. The experimentation results show that the approach achieves good results in comparison to the baseline results.

## 1. Introduction

Paraphrase identification (PI) is concerned with detecting different linguistic expressions with same or similar meaning (Bhagat & Hovy, 2013). Whereas, analyzing the degree of this meaning similarity is part of the semantic text similarity (STS) task. STS measures the degree to which two linguistic items have the same or similar meaning (Agirre et al., 2014; Jurgens, Pilehvar, & Navigli, 2014). With the advances in social media, they have become one of the main sources for news stories. Users are flooded with news posts reporting similar news events. Detecting paraphrases in news posts and the degree of their semantic similarity has proven useful in many natural language processing applications in general (Madnani & Dorr, 2010) and in new events detection (first story detection) in particular (Petrović, Osborne, & Lavrenko, 2012). However, few research (if any) can be found to support PI and STS in the Arabic news tweets.

This research aims at providing a domain independent approach for PI and STS analysis of Arabic news tweets using a supervised machine learning approach. Although, tweets are considered noisy, informal and personal (Xu, Callison-Burch, & Dolan, 2015), collected tweets are posted by well-known Arabic news agencies such as Al-Jazeera[1] and Al-Arabiya[2] where

* Corresponding author.
*E-mail addresses:* maalsmadi9@just.edu.jo, smadi.mohammad@gmail.com (M. AL-Smadi).

[1] https://twitter.com/AJABreaking.

[2] https://twitter.com/alarabiya_brk.

http://dx.doi.org/10.1016/j.ipm.2017.01.002
0306-4573/© 2017 Elsevier Ltd. All rights reserved.

they post their breaking news using the Modern Standard Arabic (MSA). The proposed approach builds on lexical, syntactic and semantic computed measures and extracted features adapted from literature.

The rest of this paper is organized as follows: Section 2 sheds the light on related work for PI and STS analysis, Section 3 explains the research method and the proposed approach, Section 4 presents the approach results, Section 5 discusses the results and findings, finally, Section 6 concludes this work and presents future work plans.

## 2. Related work

The coverage of the existing literature is divided into two parts. The first part is devoted to discuss related work on the paraphrase identification task, whilst, the other part discusses related work on semantic text similarity.

Discussed related work is mainly based on - but not limited to - tasks presented at the prestigious event of the International Workshop on Semantic Evaluation (SemEval)[3] in the years 2014 and 2015. SemEval (Semantic Evaluation) is an ongoing workshop focusing on computational semantic analysis systems. In the 8th SemEval-2014, two tasks were presented focusing on STS analysis, "SemEval-2014 Task 3: Cross-Level Semantic Similarity," the task mainly evaluates the level of similarity between larger linguistic item (i.e. Document, Paragraph) and a smaller item (i.e. sentence) (Jurgens et al., 2014), and "SemEval-2014 Task 10: Multilingual Semantic Textual Similarity" (Agirre et al., 2014). In the 9th SemEval-2015 two task also were presented, "SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter (PIT)" and "SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability." SemEval-2015 Task 1 is the closest to our presented research and evaluates PI and STS degree on a sentence level (i.e. tweets). Whereas, SemEval-2015 Task 2 is a continuation of the task presented in SemEval-2014 (i.e. Task 10) were datasets for different languages (English and Spanish) were provided to support STS research.

### *2.1. Paraphrase identification (PI)*

Paraphrase identification (PI) is concerned with the ability of identifying alternative linguistic expressions of the same meaning at different textual levels (document level, paragraph level, sentence level, word level, or combination between them) (Bhagat & Hovy, 2013). Others labels can be found in literature for PI such as Paraphrase Detection (PD) and Paraphrase Recognition (PR) (Bhagat, Hovy, & Patwardhan, 2009; Qiu, Kan, & Chua, 2006; Ul-Qayyum & Altaf, 2012).

Available research on the paraphrase identification approaches can be categorized into supervised and unsupervised machine learning techniques as follows.

#### *2.1.1. Supervised techniques*

Finch, Hwang, and Sumita (2005) proposed an approach based on Bag-Of-Words (BoW) model, it used The NIST score (Doddington, 2002), Position-independent word error rate (PER) (Tillmann, Vogel, Ney, Zubiaga, & Sawaf, 1997), Word error rate (WER) (Su, Wu, & Chang, 1992), The BLEU score (Papineni, Roukos, Ward, & Zhu, 2002) and part-of-speech (POS) enhanced PER as a feature of classification using SVM. The text-preprocessing task consisted of tokenization, POS tagging and stemming of only nouns and verbs. This task enhanced the performance of 0.8%. Moreover, they used semantic similarity distance measure computed based on WordNet-based lexical relationship measures (Jiang & Conrath, 1997). The last measure got 0.6% enhancement. This approach achieved the best performance when using support vector machine (SVM) and a combination of all measures with accuracy of 74.96% on Microsoft Research Paraphrase Corpus (MRPC) test data (Dolan, Quirk, & Brockett, 2004).

Qiu et al. (2006) presented a framework of two-phase for PI. The first phase identified the common content information of the pair of sentences using similarity detection. Then these information was paired using a pairing module. This common information content was called information nuggets, it was provided in a tuple of predicated arguments form. Using a simple matching technique, the predicate arguments were compared. This approach is different from other approaches because it is focusing on dissimilarities between the pairs sentences. It achieved 72.0% accuracy on MRPC test data using SVM classifier.

Kozareva and Montoyo (2006) trained three machine-learning classifiers (SVM, K-Nearest Neighbor (K-NN) and Maximum Entropy (MaxEnt)) on features' vectors extracted from lexical and semantic attributes combination. They used cardinal number attribute, proper name attribute, Longest Common Sub-sequence (LCS) and *n*-grams as lexical features. Semantic similarity features were based on WordNet. The experiments showed that: At first when using the lexical feature set and the similarity feature set independently the best performance was achieved by the lexical feature set, but when combined the two features sets the performance enhanced by 1%. The best result between all classifier obtained with SVM.

Ul-Qayyum and Altaf (2012) provided an approach using semantic heuristic features to identify the paraphrasing. Tokenization, POS tagging and stop words removing were performed on the sentence pairs as a per-processing step. They also used monotonic and no-monotonic alignment and semantic heuristics to define feature vectors' set, then they performed machine learning techniques along with these vectors using Weka tool to predict PI. The approach achieved good accuracy according to state-of-the-art PI systems.

Eyecioglu and Keller (2015) group was one of the teams that participated in SemEval-2015 Task 1 Workshop. In their approach for PI they used a SVM along with simple lexical overlap features of (words and characters) based on *n*-grams.

[3] https://en.wikipedia.org/wiki/SemEval .