Expert Systems With Applications 92 (2018) 106–123

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Feature selection in multiword expression recognition

Senem Kumova Metin

Izmir University of Economics, Faculty of Engineering, No.156, 35330, Balçova-İzmir, Turkey

ARTICLE INFO

Article history:
Received 22 December 2016
Revised 15 September 2017
Accepted 19 September 2017
Available online 20 September 2017

Keywords:
Multiword expression
Multiword expression recognition
Learning algorithms
Feature selection

ABSTRACT

In multiword expression (MWE) recognition, there exist many studies where different learning methods are employed to decide whether given word combination is a multiword expression. The recognition methods commonly utilize a number of features that are extracted from a data source, frequently from the given text. Though the recognition methods and the features are well studied, we believe that to achieve the best possible performance with a learning method, different subsets of features should also be considered and the best performing subset must be selected.

In this paper, we propose a procedure that covers the performance comparison of well-known feature selection methods to obtain the best feature subset in MWE recognition. The evaluation tests are performed on a Turkish MWE data set and the performance is measured by precision, recall and *F1* values. The highest *F1* value =0.731 is obtained by *C4.5* classifier employing either wrapper or filtering method in feature selection. In the regarding setting(s), it is examined that the performance is increased by 1.11% compared to the setting where all features are employed in classification.

Based on the experimental results, it may be stated that feature selection improves the performance of MWE recognition by eliminating the noisy/non-effective features. Moreover, it is obvious that proposed feature selection method contributes to the overall MWE recognition system by reducing the measurement and storage requirements due to the lower number of features in classification, providing a faster and more-cost effective learning model.

1. Introduction

Multiword expressions (MWEs) are combinations of words that are conventional representations of concepts and/or facts. Those combinations are built from lexemes of sequentially ordered (uninterrupted) or interrupted units in language. Starting from Firth (1957) a number of MWE definitions (Bisht, Dhami, & Tiwari, 2006; Hoey, 1991; Manning & Schütze, 2000; Sinclair 1991) are provided that emphasize different properties of MWEs. For instance, Firth (1957) stated that MWE is the traditional co-occurrence of words. Sag, Baldwin, and Bond (2002) described MWEs as “idiosyncratic interpretations that cross word boundaries (or spaces)”. Though, the researchers tend to define the concept of MWE in different ways, there exists a common understanding/agreement that the set of MWEs encloses idioms, collocations, named-entities and domain-dependent terms. One other commonly accepted fact is that some properties indicate the presence of a MWE. Those properties are language dependency, unitization, domain-dependency and arbitrariness.

Language dependency is generally realized when an expression is translated from one language to another. In translation of MWEs, it is observed that word-to-word translation commonly fails since in different languages, different words combine to represent same concepts. For example, in English, the term “wisdom teeth” expresses the teeth that erupt between the ages of 17 and 25. Word-to-word translation of wisdom teeth results with an expression,“akıl dişleri”, that is never used in Turkish for this concept. In Turkish, the matching MWE is “yirmi yaş dişleri” which may be translated to English as “the teeth of 20 years old”.

The unitization principle in MWEs is that the words in MWEs unite building a new semantic/syntactic unit in language. This is why, when the text that includes MWEs are processed, such words must be accepted as a single unit. The most salient examples of unitization are observed in idioms since the composing words may change their meanings completely when they unite. For instance, in Turkish, “çamura yatmak” is an idiom that may be translated as “to be in mud” ignoring the unitization principle. But actually, the expression means idiomatically “not to keep up one’s word”.

Due to the domain dependency principle in MWEs, some expressions that belong to a specific domain may have a completely different meaning that cannot be extracted from the meanings of the composing words. For instance, the expression “terzi kası” in Turkish is a domain dependent MWE that means “Sartorius muscle” in medicine domain but in everyday language it may be trans-

E-mail address: senem.kumova@ieu.edu.tr

https://doi.org/10.1016/j.eswa.2017.09.047
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

lated/understood as “tailor’s muscle” rationally due to the lack of domain knowledge.

The arbitrariness property of MWEs that is firstly assigned to collocations indicates that the words arbitrarily unite to build a MWE. In other words, it cannot be explained why/how/which words unite to compose a MWE. For example, “canı tez” is a MWE in Turkish that means “impatient”, but there is no syntactic/semantic reason why “canı süratli” is not a MWE though “süratli” (in English fast) is given as the synonym of the word “tez” in dictionaries.

The ambiguity in the MWE concept and the lack of rules that recognize the MWEs directed researchers to identify MWEs based on some evidences. Those evidences are actually linguistic and/or statistical features that indicate the presence of a MWE in the given text and/or decide if the given word combination is a MWE.

In feature-based MWE recognition, firstly a feature-value that points out how close is the candidate (word combination) to be a MWE is measured from a data source (e.g. corpus, web) for regarding feature. Secondly, based on the feature-values, candidates may be assessed relative to the others or each candidate may be classified as MWE/non-MWE. In previous studies (e.g. Kumova Metin, 2016; Kumova Metin & Karaoğlan, 2010; Pecina, 2008; Tsvetkov & Wintner, 2013) many different features are reported to be effective in MWE identification. On the other hand, there exist a number of studies that several MWE features are utilized together by machine-learning methods (e.g. Pecina, 2008; Tsvetkov & Wintner, 2013). Though it is observed that the use of features together commonly increases the performance in identification of MWEs, it has also some drawbacks. One of the major drawbacks is observed when there are a large number of features. In such cases, the overall effort and the total time required in training and/or measuring the feature-values may reach to such high scores that directs the researchers to a mandatory simplification of the recognition model. One other important drawback is that when all MWE features are employed together, some features may fail in MWE identification and reduce the overall performance though they may succeed individually.

The main aim of our study is two-fold. First is to demonstrate that in feature-based MWE recognition a prior feature selection process improves the performance. Second is to present a systematic way of feature selection in order to determine the best set of features. We believe that identification of best MWE recognition feature set will lead manifold contributions to the natural language processing applications where MWE recognition is a prior task to be performed. The first contribution is that the overall performance of application will be changed in parallel to the performance increase in MWE recognition step. The second is that the total response time of application will be reduced due to the less number of features to be measured and processed. The third contribution will be on data storage. Namely, the simplified learning model and the lower number of features will require less amount of storage space. And the last is that the application will be simplified/improved in terms of coding.

In this study, a language independent feature selection procedure is proposed where well known feature selection methods; wrappers and filters; are utilized with many different learning algorithms/evaluators. In our experiments, a set of 27 statistical and 10 linguistics features are assessed with recall, precision and F-measures on a Turkish MWE data set of 8176 candidates (48.26% MWE, 51.74% non-MWE). We also presented some modifications on a group of linguistic features that are already defined for English in order to be used in MWE recognition in Turkish.

The experimental results showed that the best (reduced) feature set for both wrapper and filtering methods improves MWE identification performance when compared to the whole set of features. Furthermore, it is examined that the proposed feature selection procedure enhances the overall performance in MWE identification. To our knowledge, no systematic research exists addressing the feature selection in MWE recognition to this scope and offering a procedure to select best features in recognition. In addition, there is no study that offers a best performing feature set in Turkish MWE recognition.

This paper is organized as follows. In Section 2, we review the related work on MWE identification methods. Section 3 introduces the MWE features considered in our study. In Section 4, proposed procedure and feature selection methods are presented. Section 5 covers the experimental settings where data set, evaluation measures and set-up are explained. The experimental results are given in Section 6 and the paper is concluded in Section 7.

2. Related work

The MWE identification is defined simply as scoring the candidate word combinations from a given corpus according to their potential to be a MWE (Bouma, 2010) The identification procedure commonly includes 3 stages. Briefly, in the first stage, the candidates are selected to create a data set. Secondly, the candidates are ranked (Seretan, 2011) or classified based on the relations among the words and/or some linguistic features. The last stage includes the evaluation of identification performance. In this section, MWE recognition stages will be explained briefly and different approaches followed in each stage will be given. Table 1 presents summary information on aforementioned three stages for a number of previous works.

The preparation of candidate MWE set includes three important requirements to be satisfied. First is that a corpus that include a wide range of texts that may represent the language must be provided. In earlier studies on MWE recognition, it is explicit that due to the lack of a large corpus in different languages, the researchers tend to run their methods on English corpora. But currently, large corpora in different languages are available and this enabled researchers proposing methods specific to different languages (e.g. Kim, Yoon, & Song, 2001; Li, Lu, & Liu, 2007 respectively Korean and Chinese). The second requirement is the selection of word combinations/candidates of MWE data set. There exist several methods to select the candidates. For example, Evert and Krenn (2001) employed part of speech (POS) tags and selected the uninterrupted two-word sequences (bigrams) that are tagged as *adjective+noun* as candidates (as given in Table 1). Kumova Metin and Karaoğlan (2010) selected the candidates in data set based on statistical measures such as occurrence frequency, mutual information and chi-square similar to Pearce (2002). The third requirement in MWE data set preparation is the annotation of data set that includes both positive and negative examples. The annotation is defined simply as the procedure that the candidates are labelled as MWE (positive example) or non-MWE (negative example) by multiple judges. The purpose of employing multiple judges in annotation is simply having reliable and commonly agreed labels for the candidates (Schneider et al., 2014). The works of Pecina (2008) and Tsvetkov and Wintner (2013) may be given as examples where multiple annotators such as domain experts are employed. On the other hand, in a group of studies such as Pearce (2002) and Kumova Metin (2016) several dictionaries are used to label the candidates in order to have a more objective and reliable annotation of the data set.

In literature, there exist a number of studies where a variety of measures/features are used in ranking or classifying the MWE candidates as the second stage of MWE identification. In ranking approach, the candidates are sorted based on the predefined feature or a group of features. The expectation in ranking is that the candidates that hold the lower ranks in sorted lists have a higher potential to be a MWE compared to the candidates that