Detecting Multi-Word Expressions using Supervised ML Methods

Name: Yaakov HaCohen-Kerner

Affiliation: Jerusalem College of Technology, Jerusalem, Israel

WG3: Statistical, Hybrid and Multilingual Processing of MWEs

One possible method to automatically detect Multi-Word Expressions (MWEs) in different languages (each alone at this point) in general and in Hebrew in particular is .using supervised machine learning (ML) methods

:The general structure of the proposed algorithm is as follows

Investigation of MWEs' types and exploring their types and the specific features for (1 .each type

Indication of MWEs and their types that appear in natural language CORPORA (2 .((firstly, for one CORPUS for each language

Definition and programming of various features belonging to different sets. The (3 defined features need to be relevant in some way to the binary classification task of .whether a sequence of words is indeed a Multi-ord Expression

Computing the features for all possible sequences of words in the tested corpus. (4 For instance, we can test all possible sequences at length of 2-6 consecutive words .that are found in the same sentence

Applying a variety of supervised ML methods (e.g., SVM, LR, NB, MLP, C4.5, (5 REPTREE, GA, …) within the framework of WEKA using combinations of these feature sets or parts of them in order to obtain results as high as possible for detection of MWEs. We plan to examine various combinations of features and/or feature sets using forward hill-climbing and backward hill-climbing algorithms. The application of the ML methods in the final stage(s) can include tuning of the ML methods' .parameters

Analyzing the results and concluding about the effectiveness and appropriateness (6 of a ML paradigm for detection of Multi-Word Expressions in Hebrew and potentially .in other languages

Concerning components 3 and 5, we have rich experience (about 20 papers, including 3 JASIST papers). Most of the experience accumulated in these studies dealt with various NLP tasks based on a large number of ML methods. In these studies, we have .used hundreds of features belonging to about 10 feature sets

We have also published a few papers concerning automatic learning of key-phrases. .Key-phrases can also help to identify part of the MWEs

Examples of features sets and features that can be suitable for the desired task are as :follows

PoS Tags of individual tested words" feature sets: e.g., normalized" **(1)** .frequencies of tags (Parts of Speech) of each word in the tested sequence of words

Sequences of PoS Tags" feature sets: e.g., Normalized frequencies of different" **(2)** sequences of tags (e.g: sequences with length of two or three words) for the tested .sequences of words

PoS Tags of individual words before the tested sequence of words" feature" (3) sets: e.g., Normalized frequencies of tags of N (N>=1) words before the tested .sequence of words

Sequences of PoS Tags of words before the tested sequence of words"" (4) feature sets: e.g., Normalized frequencies of different sequences of tags (e.g: sequences with length of two or three words) for sequences before the tested .sequences of words

PoS Tags of words after the tested sequence of words" feature sets: e.g.," **(5)** Normalized frequencies of tags of N (N>=1) words after the tested sequence of .words

Sequences of PoS Tags of words after the tested sequence of words" feature" (6) sets: e.g., Normalized frequencies of different sequences of tags (e.g.: sequences with .length of two or three words) for sequences after the tested sequences of words

Function (stop) words: e.g., normalized frequencies of function words in the **(7)** tested sequence of words. Function words are words that have little lexical meaning or have ambiguous meaning, but instead serve to express grammatical relationships with other words within a sentence, or specify the attitude or mood of the speaker. Function features contain normalized frequencies of: articles (e.g.: the, a, an), pronouns (e.g., he, him, she, her), particles (e.g., if, then, well, however, thus), conjunctions (e.g., for, .(and, or, nor, but, yet, so), auxiliary verbs (be, have, shall, will, may and can

Quantitative features: These features present various quantitative and statistical **(8)** measures concerning the tested sequence of words, e.g., # of words in the tested sequence of words, # of characters in the tested sequence of words, average # of characters in a word, average # of characters in the tested sequence of words; average .# of word tokens in the tested sequence of words

Prefix/Suffix features: e.g., normalized frequencies of **prefixes or suffixes** of **(9)** .words tokens in the tested sequence of words