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

Automatic detection of Multi-Word Expressions (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.

General Structure of the Algorithm

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

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

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

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

(5) **Applying a variety of supervised ML methods** (e.g., SVM, LR, NB, MLP, C4.5, REPTREE, GA, ADABoost, …) 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.

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

Previous experience regarding stages (3) and (5)

We have written around 20 papers about studies dealing with various NLP tasks applied on a number of supervised ML methods. In our 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.

Potential Relevant Feature Sets

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

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

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

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

(4) "**Sequences of PoS Tags of words <u>before</u> the tested sequence of words**" 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.

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

(6) "**Sequences of PoS Tags of words after the tested sequence of words**" feature 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.

(7) **Function (stop) words**: e.g., normalized frequencies of function words in the 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).

(8) **Quantitative features**: These features present various quantitative and statistical 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.

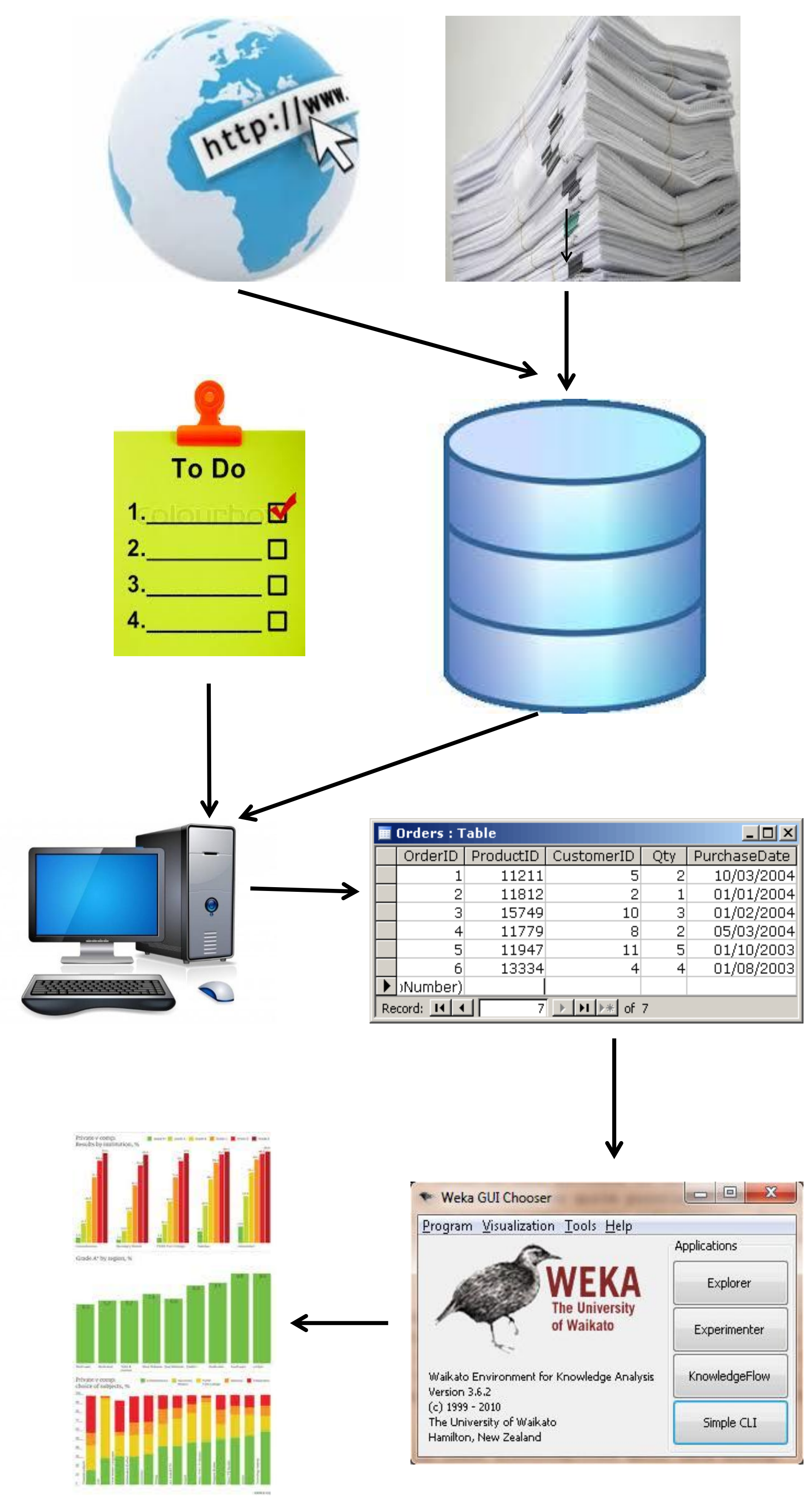
(9) **Prefix/Suffix features**: e.g., normalized frequencies of prefixes or suffixes of words tokens in the tested sequence of words.

Planned Experiments

Supervised learning of MWEs' identification using various ML methods such as: SVM, LR, NB, MLP, C4.5, REPTREE, GA, and ADABoost

- MWEs containing 2 words
  - Binary classification (MWE or not)
  - Classification according to the syntactic properties of the MWEs
- MWEs containing 3 words
  - Binary classification (MWE or not)
  - Classification according to the syntactic properties of the MWEs
- …

Graphic description of the overall structure of the proposed research



Orders : Table

	OrderID	ProductID	CustomerID	Qty	PurchaseDate
	1	11211	5	2	10/03/2004
	2	11812	2	1	01/01/2004
	3	15749	10	3	01/02/2004
	4	11779	8	2	05/03/2004
	5	11947	11	5	01/10/2003
	6	13334	4	4	01/08/2003
▶	)Number)				

Record: 7 of 7

Weka GUI Chooser

Waikato Environment for Knowledge Analysis
Version 3.6.2
(c) 1999 - 2010
The University of Waikato
Hamilton, New Zealand

Contact:

Yaakov HaCohen-Kerner

Department of Computer Science,
Jerusalem College of Technology,
21 Havaad Haleumi St., P.O.B. 16031,
9116001 Jerusalem, Israel

kerner@jct.ac.il