Interpretation Support of Terms while Browsing in Slovak Language

Róbert HORVÁTH*

Slovak University of Technology
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
roberthorvath89@gmail.com

Today, Web is an inseparable part of our everyday life. When accessing web pages, it is necessary to fully understand their meaning in order to take full advantage of information contained within. In many technical articles there are words and phrases, whose meaning is unknown to many users. Most users look for the explanations of phrases in online dictionaries – they have to open a new window with the dictionary and manually enter the corresponding phrases. Furthermore, when considering polysemy, they need to identify the correct meaning. This is not very fast neither comfortable. Another possibility for users is to use web browser extensions that can automate the whole process. Typically, extensions are made to show the main explanation of words in tooltips (e.g., Google Dictionary, Dictionary Lookup). Their drawback being that they are not created to support Slovak language.

The goal of our work is to create a tool that will provide explanations of Slovak words during web browsing by automatically looking up their definitions in available online dictionaries. If a word has multiple meanings we need to perform word sense disambiguation [1]. In order to keep the process automatic, we aim only for unsupervised methods without the need for human help. We combine existing methods of text processing and text comparison to acquire correct term definitions. In order to evaluate our approach to word definition acquisition, we have decided to create a web browser extension. The extension utilizes a web service designed for lemmatization and text processing. When a user finds an unknown word while browsing the Web, he simply selects the word and the extension displays the most probable meaning of the word by showing a word definition.

Our approach to acquire term definition consists of the three main steps (see Figure 1):

1. web page part (a selected term and its neighborhood) text pre-processing,
2. potential definition lookup,
3. correct definition selection.

* Supervisor: Marián Šimko, Institute of Informatics and Software Engineering

Spring 2011 PeWe Workshop, April 1, 2011, pp. 7–8.

Input data gathered from a web page contains some information not relevant for finding of meaning or similarity measurement like HTML tags, numbers, symbols, etc. Such data (e.g., HTML tags, stop words) are removed and each word is transformed into its basic form in order to enable further comparison.

One of the most important parts of our approach is word definition lookup. There are many types of online services able to provide such functionality. We consider online dictionaries, but search engines can be utilized as well. Online dictionaries typically need words in basic form – *lemma* as an input. Due to the polysemy issue present in every natural language, input often matches more than one definition. Thus, the output of this step is typically a list of potential definitions. We utilize existing online dictionaries that represent extendable and configurable options for users.

Correct word definition selection considering its context (as means to resolve polysemy) is done via similarity calculation. In this step we calculate cosine similarity [2] between the selected word context (i.e., the textual neighbourhood of the word) and potential definitions. We expect words describing the correct meaning to be located close to the unknown word. Experiments we had conducted have shown that *paragraph* was the best neighbourhood option (80% success rate in comparison with sentence – 60% and the whole page text – 70%, tested for 20 homonyms).

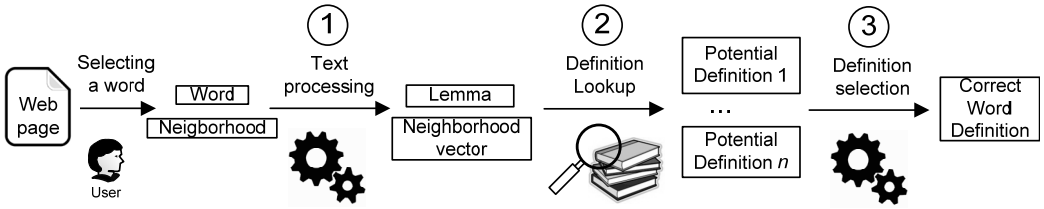

Figure 1. Process of acquiring term definition.

In the current stage of our work, we have already have created a browser extension, which provides users with context based term definitions. We plan to conduct small experiments with a select group of users to gather feedback and evaluate the extension.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Paralič, J., Furdík, K., Tutoky, G., Bednár, P., Sarnovský, M., Butka, P., Babič, F. Preprocessing text data. In *Mining knowledge from texts*, Technical University Košice, pp. 15-65, 2010 (in Slovak).

[2] Krajči, S., Novotný, R. Finding common base of Slovak words employing common suffix. In *Proc. of 1st Workshop on Intelligent and Knowledge oriented Technologies*, pp. 99-101, 2007 (in Slovak).