Available online at www.sciencedirect.com

ScienceDirect

IERI Procedia 10 (2014) 38 – 44

Procedia IERI

www.elsevier.com/locate/procedia

2014 International Conference on Future Information Engineering

Distributionally Extended Network-Based Word Sense Disambiguation in Semantic Clustering of Polish Texts

Paweł Kędzia[1], Maciej Piasecki, Jan Kocoń, Agnieszka Indyka-Piasecka

Wrocław University of Technology, ul. Wybrzeże Wyspiańskiego 27, Wrocław 50-370, Poland
bSecond affiliation, Address, City and Postcode, Country

Abstract

In the paper we present an extended version of the graph-based unsupervised Word Sense Disambiguation algorithm. The algorithm is based on the spreading activation scheme applied to the graphs dynamically built on the basis of the text words and a large wordnet. The algorithm, originally proposed for English and Princeton WordNet, was adapted to Polish and plWordNet. An extension based on the knowledge acquired from the corpus-derived Measure of Semantic Relatedness was proposed. The extended algorithm was evaluated against the manually disambiguated corpus. We observed improvement in the case of the disambiguation performed for shorter text contexts. In addition the algorithm application expressed improvement in document clustering task.

Selection and peer review under responsibility of Information Engineering Research Institute

Keywords: Word Sense Disambiguation, wordnet, text classification, plWordNet

1. Introduction

Documents are commonly represented in Information Retrieval as bags of words, i.e. collections of words (words with the number of their occurrences). Linguistic structure of the text is very rarely taken into account, mostly due to the limited robustness of the natural language processing (i.e. limited precision and speed of processing). However, the bag of words model causes the loss of information even on the word level. Many

[1] * Corresponding author. Tel.: +48 71 320 42 24; fax: +0-000-000-0000.
E-mail address: pawel.kedzia@pwr.wroc.pl.

Selection and peer review under responsibility of Information Engineering Research Institute
doi:10.1016/j.ieri.2014.09.073

words are polysemous and can express several meanings, e.g. the word *car* corresponds to 5 noun meanings in Princeton WordNet 3.1 (PWN) (Fellbaum, 1998): *car* 1 – a motor vehicle, *car* 2 – "a wheeled vehicle adapted to the rails of railroad", car 3 – "the compartment that is suspended from an airship", *car* 4 – an elevator car and *car* 5 – a cable car. Improper matching of the words in the user query against their use in the documents can lead to incorrect retrieval or ranking of the results.

Word Sense Disambiguation (WSD) methods can potentially help in several Information Retrieval (IR) tasks in which the user information need specification is longer than a couple of words, e.g. in Question Answering, document classification and document clustering. However, *supervised WSD tools* (developed on the basis of supervised machine learning algorithms applied to a text corpus that was manually annotated with word senses) express relatively good accuracy but have coverage which is limited only to words annotated in the corpus. Such corpus annotation is very laborious and costly, so the typical coverage is from 100 till several thousand words at most. *Unsupervised WSD* methods are often based on the sense induction from text corpora, but their accuracy is much lower than the accuracy of the supervised methods and the coverage is still far from perfect (not all word senses are represented well enough). However, there is yet another group of the unsupervised WSD methods – algorithms that use the wordnet graph of relation (see Sec. 2) and the spreading activation scheme to find senses matching the surrounding text passages.

Our goal was to adapt a spreading activation based WSD algorithm to a wordnet different than PWN, namely Polish plWordNet, and the language different than English. Moreover, we wanted to extend the wordnet with knowledge resources acquired from the text corpus and to build a WSD tool for practical applications.

2. plWordNet – repository of lexical senses

In the spreading activation based WSD (SA-WSD) a wordnet is used in two roles: as repository of senses defining all senses per each word, and as a knowledge base describing senses by lexico-semantic relations.

A wordnet consists of *synsets*, lexical units and lexico-semantic relations. A lexical unit is a pair: a word plus sense number, e.g. *car* 2. A synset is a set of near synonyms and consists of one or more lexical units. Each synset represents a unique lexical meaning and synset identifiers can represent lexical meanings. Lexico-semantic relations represent binary meaning associations between lexical units observed in the lexical system. Lexico-semantic relations are encoded in the wordnet as relations between synsets (the basic ones) or between lexical units, e.g. the synset from PWN 3.1 {*car* 1, *auto* 1, *automobile* 1, ...} is linked by the *hypernymy* relation with the synset {*motor vehicle* 1, *automotive vehicle* 1} and by *holonymy* with the synset {*bumper* 2}.

plWordNet 2.1 (plWN) is a huge wordnet for Polish of the size close to the PWN size: ~161 000 lexical units, ~118 000 synsets and ~108 000 unique words. The lexical units are described by more than 40 different lexico-semantic relations. plWN was developed on the basis of Polish corpora and provides better coverage of the corpus vocabulary than PWN and higher relation density than PWN. However, plWN almost does not have glosses for synsets (short textual sense descriptions) that are intensively used in SA-WSD methods.

3. Distributionally extended WSD

Many approaches to the graph-based WSD were proposed, e.g. Gutiérrez et al. 2012, Tsatsaronis et al. 2010, Mihalcea and Figa 2004, Agirre and Soroa 2004, Navigli 2006, Sinha and Mihalcea 2007, Agirre and Soora 2008]. In our work we follow the Page-Rank based approach proposed by Agirre and Soroa 2004 [Agirre and Soora 2008, Agirre et al. 2009, Agirre et al. 2010]. The key concepts are: *Lexical Knowledge Base* (*LKB*) – a set of the concepts (PWN synsets) together with the relations between them and the *dictionary*