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.

© 2014 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

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
words are polysemous and can express several meanings, e.g. the word \textit{car} corresponds to 5 noun meanings in Princeton WordNet 3.1 (PWN) (Fellbaum, 1998): \textit{car 1} – a motor vehicle, \textit{car 2} – “a wheeled vehicle adapted to the rails of railroad”, \textit{car 3} – “the compartment that is suspended from an airship”, \textit{car 4} – an elevator car and \textit{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. \textit{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 \{\textit{car 1, auto 1, automobile 1, ...}\} is linked by the hypernymy relation with the synset \{\textit{motor vehicle 1, automotive vehicle 1}\} and by holonymy with the synset \{\textit{bumper 2}\}.

plWordNet 2.1 (pWN) 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. pWN 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, pWN 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 Soroa 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
of lemmas (word types) describing associations between lemmas and concepts. LKB can be represented as a graph $G = (V; E)$ where $V$ is the set of concepts and $E$ is a set of relations in the concept set, i.e $e_{i,j} \in E$ is undirected relation $(v_i, v_j)$, where $v_i, v_j \in V$. The relations correspond to the wordnet lexico-semantic relations and also weights are assigned to them. The graph based WSD builds a graph related to words included in a text fragment, next applies a ranking function to the graph, chooses concepts with the highest final ranking values from the graph and finally assigns them to the words in the text.

Agirre and Soroa 2004 [Agirre and Soora 2008, Agirre et al. 2009, Agirre et al. 2010] used the PWN relation graph as the basis for LKB. Agirre and Soora 2009 proposed the idea of using Page-Rank approach to WSD. Basic PR algorithm can be described by the state change equation: $Pr = c M Pr + (1 - c)v$ where $v$ is the vector with $N$ dimensions (i.e. graph nodes), and $Pr$ is the vector of ranking values. The initial values of the elements are set all to $1/N$. $M$ is a probability matrix of the size $N \times N$, and $M_{i,j} = 1/d_i$ if there is an edge linking $v_i$ and $v_j$ or 0 otherwise, where $d_i$ is degree of $v_i$ node in the graph, and $c$ is a damping factor (typically set to $<0.85, \ldots, 0.95>$). PR is run iteratively for predefined number of iterations.

Disambiguation of the words from a text contexts is done by: selecting the synsets that are associated with lemmas from the text context, building a subgraph of the wordnet relation on the basis of the selected synsets, running PR on the subgraph and choosing the appropriate synsets on the basis of the final node values.

This basic scheme can be extended with the use of the personalized vector $v$ prior to the disambiguation. In both cases the result is an assignment of the chosen synsets to the text words. The context can be one or many sentences. Agirre and Soroa 2004 [Agirre and Soora 2008] proposed a slightly modified PR:

$$P(v_i) = (1 - \alpha)\alpha + \sum \frac{1}{d_j} P(v_j)$$

where $\alpha$ is damping factor. They also introduced modified estimation of the initial vector $v$ values according to which the values are focused only on the nodes including lemmas identical to the text words being disambiguated. As a result, the probability mass is not distributed over the whole graph - this approach is called Personalized Page Rank (PPR). The subgraph concept links come first from the wordnet relations linking synsets selected by the text words. Links to synsets corresponding to words from the glosses of the synset (selected previously) are added. It was assumed that the glosses were sense disambiguated. In this way new concepts are added from the glosses to the subgraph which is bigger and provides more information.

In our case pLWN almost does not provide glosses and the glosses are not disambiguated. Unfortunately, the additional links derived from glosses lead to the significant improvement in the SA-WSD algorithm.

To supplement the missing information, we used Measure of Semantic Relatedness (MSR) as an additional source of links for the subgraphs. MSR assigns a numerical value to pairs of words such that words that are semantically close receive higher values than those that are not related. MSR is built on the basis of the statistical analysis of word distributions in a large corpus, i.e. words that occur in similar linguistic contexts are described by similar feature vectors and receive higher values of relatedness. For the work presented here, we used the MSR proposed by Piasecki, et al. 2009 in which word occurrences are described by a set of basic syntactic relations with other words, co-occurrence frequencies are weighed by PMI and the relatedness is calculated by the cosine measure. Using MSR, we constructed an extended LKB in which additional concepts were added on the basis of the most related words from MSR. For each word $w_i$ we acquired from MSR a list $SW_i$ of the $k=20$ most related words to $w_i$. For each word $w_j$ from the list $SW_i$ we added to the disambiguation subgraphs all synsets including $w_j$ as additional concepts and all links between those synsets and synsets already present in the subgraph, e.g. the initial list of 385370 synset relations in the text-based subgraph was expanded with 2346834 new edges in the output graph.
4. Evaluation

WSD algorithm can be evaluated directly by comparison of its decisions with the human judgments and, indirectly, by applying the algorithm in a text processing task as a tool.

4.1. Corpus based evaluation

For comparison we used a part of the KPWr corpus of Polish [Broda et al. 2012] which is sense disambiguated. This part include 1996 documents, 14 022 words and 60 unique lemmas manually sense disambiguated. WSD annotation contains information about the appropriate plWN synset e.g.:

\[
\text{\textless tok\textgreater <orth>dni</orth>}
\text{\textless lex\textgreater <base>dzień</base><ctag>subst:pl:nom:m3</ctag></lex>}
\]

where \text{"dzień\text{"} \text{"}a day\text{"} \text{"}annotates the sense of the word \text{"}dni \text{"} \text{"}days\text{"} \text{"}(lemma \text{"}dzień\text{"}).

The precision and recall of our algorithm in relation to the annotation are presented in Table 1, where \text{"}plWordnet\text{"} data source means that the subgraph included only relation added from plWN on the basis of the text word, \text{"}MSRkbes20\text{"} – the subgraph was built on the basis of the synsets corresponding to the 20 most related words and \text{"}plWwordNet+MSRkbest20\text{"} – the full extended subgraph was built.\n
\[
\begin{array}{|c|c|c|c|}
\hline
\text{Data source} & \text{Context type} & \text{Noun precision} & \text{Verb precision} \\
\hline
\text{plWordNet} & \text{Sentence} & 0.34 & 0.24 \\
& \text{Whole document} & 0.43 & 0.28 \\
\text{MSRkbest20} & \text{Sentence} & 0.38 & 0.0 \\
& \text{Whole document} & 0.37 & 0.0 \\
\text{plWordNet + MSRkbest20} & \text{Sentence} & 0.39 & 0.22 \\
& \text{Whole document} & 0.37 & 0.27 \\
\hline
\end{array}
\]

The best result is highlighted. In all configurations the recall was equal to 0.88. The precision of for verbs in the case of MSRkbes20 data source is equal to 0, because the used MSR covered only nouns. The use of the 20 most related words improved the results only in the case of sentences as contexts. A potential cause is that the first words have a high similarity, but the further do not, e.g. for word \text{"}subst:opatrzność \text{"} \text{"}providence\text{"} \text{"}the most related word is \text{"}subst:bóg \text{"} \text{"}God\text{"} (~0.25) but one of the last is \text{"}subst:ojciec \text{"} \text{"}father\text{"} (~0.096).

4.2. Evaluation by application

Evaluation by application was performed for document clustering due to the available document collection. We used two different data sources with two different representations: TF and TF/IDF with two types of features (bag-of-words of orthographic forms and plWN synsets corresponding to the text words). The first set of documents was derived from the Polish Wikipedia (http://pl.wikipedia.org). It contains 40 documents from 4 categories (10 documents per category): \text{"}Planets\text{"} (P in Table 2), \text{"}Cities\text{"} (C), \text{"}Felidae\text{"} (F) and \text{"}Canidae\text{"} animals (A). During evaluation we used CLUTO system [Tagarelli and Karypis 2008, Zhao and Karypis 2005] for
clustering. The results presented in Table 2 were achieved for RBR clustering method with $G1'$ criteria function, 4 clusters, the cosine similarity function and TF documents representation.

Table 2. Clustering results using RBR method, $G1'$ criteria function, cosine similarity and 4 clusters for Wiki dataset and synset features.

<table>
<thead>
<tr>
<th>Class</th>
<th>Size</th>
<th>ISim</th>
<th>ISdev</th>
<th>ESim</th>
<th>ESdev</th>
<th>Entropy</th>
<th>Purity</th>
<th>P</th>
<th>C</th>
<th>F</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td>0.341</td>
<td>0.060</td>
<td>0.019</td>
<td>0.004</td>
<td>0.000</td>
<td>1.000</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>0.299</td>
<td>0.045</td>
<td>0.021</td>
<td>0.006</td>
<td>0.000</td>
<td>1.000</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>0.210</td>
<td>0.023</td>
<td>0.033</td>
<td>0.010</td>
<td>0.000</td>
<td>1.000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>0.182</td>
<td>0.020</td>
<td>0.029</td>
<td>0.008</td>
<td>0.220</td>
<td>0.909</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>

$G1'=510$, Entropy: 0.062, Purity: 0.974 Accuracy = 97.5%

The purity and entropy indicator (the general and in all subclass) are almost perfect. Only one document from *Canidae* was erroneously clustered to *Felidae*. It is about hyenas that are related to *Canidae*, but they belong to the *Felidae* taxonomy. So it can be treated as correct clustering. In Table 3 the results with bag of words features for the same dataset were shown. They are worse than those for WSD synset features.

Table 3. Clustering results using RBR method, $G1'$ criteria function, cosine similarity and 4 clusters for Wiki dataset and b-o-w features.

<table>
<thead>
<tr>
<th>Class</th>
<th>Size</th>
<th>ISim</th>
<th>ISdev</th>
<th>ESim</th>
<th>ESdev</th>
<th>Entropy</th>
<th>Purity</th>
<th>P</th>
<th>C</th>
<th>F</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9</td>
<td>0.226</td>
<td>0.020</td>
<td>0.013</td>
<td>0.002</td>
<td>0.217</td>
<td>0.889</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>0.184</td>
<td>0.023</td>
<td>0.010</td>
<td>0.002</td>
<td>0.000</td>
<td>1.000</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>0.151</td>
<td>0.012</td>
<td>0.015</td>
<td>0.006</td>
<td>0.395</td>
<td>0.667</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>0.143</td>
<td>0.010</td>
<td>0.015</td>
<td>0.003</td>
<td>0.407</td>
<td>0.636</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>

$G1'=470$, Entropy: 0.256, Purity: 0.795 Accuracy = 82.5%

The second dataset comes from KPWr and includes 30 documents from 3 categories: *Government* (G in Table 4) *Science* (S) and *Technical* (T) with 10 documents in each category. The results for best configuration from the Wikipedia dataset are presented in Table 4.

Table 4. Clustering results using RBR method, $G1'$ criteria function, cosine similarity and 4 clusters, KPWr dataset and synset features.

<table>
<thead>
<tr>
<th>Class</th>
<th>Size</th>
<th>ISim</th>
<th>ISdev</th>
<th>ESim</th>
<th>ESdev</th>
<th>Entropy</th>
<th>Purity</th>
<th>G</th>
<th>S</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13</td>
<td>0.268</td>
<td>0.061</td>
<td>0.065</td>
<td>0.025</td>
<td>0.829</td>
<td>0.462</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>0.418</td>
<td>0.079</td>
<td>0.037</td>
<td>0.020</td>
<td>0.000</td>
<td>1.000</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>0.192</td>
<td>0.027</td>
<td>0.078</td>
<td>0.039</td>
<td>0.500</td>
<td>0.500</td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

$G1'=430$, Entropy: 0.616, Purity: 0.621 Accuracy = 58%

On this dataset, the obtained results were worse than those on the Wikipedia (also for bag of words feature in Table 5). Only for the government category it was possible to separate the documents. This may be due to the fact that the documents in KPWr are more complex and longer than in the Wikipedia.

Table 5. Clustering results using RBR method, $G1'$ criteria function, cosine similarity and 4 clusters, KPWr dataset and b-o-w features.

<table>
<thead>
<tr>
<th>Class</th>
<th>Size</th>
<th>ISim</th>
<th>ISdev</th>
<th>ESim</th>
<th>ESdev</th>
<th>Entropy</th>
<th>Purity</th>
<th>G</th>
<th>S</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13</td>
<td>0.497</td>
<td>0.111</td>
<td>0.105</td>
<td>0.019</td>
<td>0.648</td>
<td>0.538</td>
<td>7</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>0.320</td>
<td>0.091</td>
<td>0.048</td>
<td>0.047</td>
<td>0.234</td>
<td>0.900</td>
<td>9</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
5. Conclusions (MP)

The application of the large lexicon WSD algorithm brought very encouraging results. WSD of the limited accuracy in tests on the disambiguated corpus, showed improvement in document clustering. The WSD algorithm based on spreading activation has been easily adapted to the new language and new wordnet. So, from this perspective, even the evaluation by comparison with the human judgment can be treated as positive. In order to supplement the missing information in the plWordNet (i.e. glosses) we used a corpus-derived Measure of Semantic Relatedness which provides automatically extracted semantic associations between words. The extension of WSD algorithm with the information from MSR brought improvement for shorter contexts, but decreased the accuracy for larger contexts. There is clear need for further research on defining MSR parameters appropriate for WSD and further exploration of the corpus-derived information in SA WSD.

Acknowledgements. Polish National Centre for Research and Development project SyNaT and the European Union within European Innovative Economy Programme project NEKST POIG.01.02-14-013/09

References
