A semantic similarity measurement tool for WordNet-like databases

Marek Kubis
Faculty of Mathematics and Computer Science,
Adam Mickiewicz University,
ul. Umultowska 87, 61-614 Poznań, Poland,
mkubis@amu.edu.pl

Abstract
The paper describes a new framework for computing semantic similarity of words and concepts using WordNet-like databases. The main advantage of the presented approach is the possibility to implement similarity measures as concise expressions in the embedded query language. The preliminary results of using the framework to model the semantic similarity of Polish nouns are reported.

1. Introduction
Among various applications of WordNet (Fellbaum, 1998) the task of modeling semantic similarity between words attracted considerable attention within the last two decades. The WordNet-based semantic similarity measures ranging from simple path length dependent functions (Rada et al., 1989; Leacock and Chodorow, 1998), through the measures that exploit the notion of the least common subsumer1 (Wu and Palmer, 1994) to the ones that utilize information content computed over corpora (Resnik, 1995; Jiang and Conrath, 1997; Lin, 1998) have been proposed in the literature. These measures were evaluated within the task of word sense disambiguation (Patwardhan et al., 2003) and incorporated into natural language processing and information extraction systems (e.g. Budanitsky and Hirst, 2006; Stevenson and Greenwood, 2005). Despite the wide range of applications, the issue of using other wordnets in place of Princeton WordNet as resources for modeling similarity among words does not seem to gain the same level of attention. Our aim is to use PolNet (Vetulani et al., 2010) and PIWordNet (Maziarz et al., 2012) to model semantic similarity of the Polish nouns. Since we did not find a software package for measuring semantic similarity that could be easily adapted to make use of both Polish wordnets (cf. Section 2.), we decided to implement our own. Therefore, the goal of this paper is twofold. First, we present WSim — a new tool for determining degree of semantic similarity using measures computed over WordNet-like databases2. Second, we report preliminary results of using WordNet-based similarity measures to model similarity of the Polish nouns. This is, to the best of our knowledge, the first attempt to apply two wordnets developed for the same language in a shared, application-oriented task.

2. Related work
WordNet::Similarity (Pedersen et al., 2004) is a widely-cited software package that implements a range of WordNet-based semantic similarity measures. This package became a de facto standard tool for computing similarity scores using WordNet and serves as a reference point for other implementations (cf. Postma and Vossen, 2014). Unfortunately, WordNet::Similarity operates only on Princeton WordNet and is not able to load wordnets that do not conform to the internal storage format of the wn program distributed with Princeton WordNet (Tengi, 1998). The same restriction holds for Python interface to WordNet provided by the NLTK toolkit (Bird et al., 2009). The Java reimplementation of WordNet::Similarity by Shimazawa (2015) called WS4J beside Princeton WordNet can also load the Japanese WordNet (Ishahara et al., 2008). PolNet is not distributed in the Princeton WordNet conformant form and we did not find any tool that could be used to convert it to this format without a vast amount of preprocessing.

A major step in the terms of interoperability is the WordNetTools library (Postma and Vossen, 2014) which can load any wordnet that is stored in a file conforming to the Wordnet-LMF format (Soria et al., 2009). However, at the time of writing neither PolNet nor PIWordNet are released in this format. WordNetTools also accepts the files in the Global WordNet Grid format (Global WordNet Association, 2012), but we did not manage to load into it the DEBVisDic (Horak et al., 2005) conformant XML file being part of the PIWordNet distribution.

Since the replication of exact results among different software packages is not easy to achieve (cf. Postma and Vossen, 2014, sec. 5.3), we did not want to use separate tools for computing values of similarity measures for the two wordnets (e.g. NLTK for PIWordNet and WordnetTools for PolNet). Therefore, we decided to reimplement WordNet-based semantic similarity measures on top of the WQuery suite (Kubis, 2012, 2014) which is able to load both PIWordNet and PolNet. An additional advantage of this approach is the ability to modify the similarity measures by revising the concise expressions of the WQuery language (cf. Section 4.) instead of the Java code of WordNetTools which in the case of any changes would require recompilation. Furthermore, since WQuery (version 0.10) can load wordnets stored in Wordnet-LMF, DEBVisDic (Horak et al., 2005) and Princeton WordNet internal for-

---

1A joint transitive hyponym of two synsets such that no other joint transitive hyponym of these synsets is placed below it within the hyponymy hierarchy.
2Databases that are organized in a similar manner to WordNet (Fellbaum, 1998); called wordnets in the rest of the paper.
mat, we gained the ability to make direct comparisons among the values of the similarity measures computed for the lexical databases stored in all of the aforementioned formats.

3. WSim

As mentioned in the previous section, WSim is built upon the WQuery suite. Therefore, before computing the values of similarity measures, one has to convert the wordnet to the WQuery database format by the wcompile command\(^4\) from the WQuery toolkit. Since both PIWordNet and PolNet are available in the XML files compatible with the DEBVisDic editor (Horak et al., 2005) the \(-t\) deb option has to be passed to the command

\[
\text{wcompile -t deb polnet.xml > polnet.wq}
\]

Having the wordnet in the WQuery format, one can compute similarity of pairs of words (or word senses) by passing them on the standard input of the \texttt{wsim} command, separated by tab characters.

\[
\text{wsim polnet.wq < pairs}
\]

By default \texttt{wsim} determines similarity of a pair of words by inverting the value of the shortest path length in the hyponymy hierarchy that links the synsets containing the given words, thus for the pair samochéd (Eng. car) and rower (Eng. bicycle) the similarity determined with PolNet is 0.25

WSim implements six semantic similarity measures:

1. inverted length of the shortest path,
2. Wu and Palmer (1994),
3. Resnik (1995),
4. Jiang and Conrath (1997),
5. Leacock and Chodorow (1998),

Following (Pedersen et al., 2004), we denote these measures by path, wup, res, jcn, lch and lin, respectively. They can be selected by passing \(-m\) option to the \texttt{wsim} command. For instance, to compute the Wu and Palmer measure one has to execute the command

\[
\text{wsim polnet.wq -m wup < pairs}
\]

In the case of the information content dependent measures (Resnik, 1995; Jiang and Conrath, 1997; Lin, 1998) word (or sense) counts can be submitted in a file passed as an argument of the \(-c\) option, e.g.

\[
\text{wsim polnet.wq -m res -c counts < pairs}
\]

If the counts are distributed together with a wordnet (as is in the case of Princeton WordNet) the \(-c\) option can be skipped.

\[
\text{wsim wordnet.wq -m res < pairs}
\]

4. Implementation of measures

The similarity measures are implemented in WSim as functions formulated in the WQuery language (Kubis, 2012). Every function that ends with the \_measure suffix is interpreted as a similarity measure and is available through the \(-m\) option of the \texttt{wsim} command. For every pair of senses read from the input the \texttt{wsim} command determines their corresponding synsets and passes them to the function indicated by the argument of the \(-m\) option. In the case of pairs of words \texttt{wsim} returns the maximum of the similarity values computed for every pair of the senses of the submitted words.

Let us consider the Wu and Palmer measure as an example. The measure is given by the following formula (cf. Wu and Palmer, 1994; Budanitsky and Hirst, 2006)

\[
\frac{2 \cdot \text{dep}(lcs(l, r))}{\text{dist}(l, lcs(l, r)) + \text{dist}(r, lcs(l, r)) + 2 \cdot \text{dep}(lcs(l, r))}
\]

where \(l\) and \(r\) are synsets, \(lcs(l, r)\) denotes the least common subsumer of \(l\) and \(r\), \(\text{dist}\) denotes the distance between two synsets in the hyponymy hierarchy and \(\text{dep}\) returns the distance of a synset from the hyponymy root. The Wu and Palmer measure has the following implementation in WQuery

\[
\begin{aligned}
\text{function wup_measure do} \\
\%l, \%r := %A \\
\%lcs := lcs_by_depth(%l, %r) \\
\%dl := lcs_dist(%l, %lcs) \\
\%dr := lcs_dist(%r, %lcs) \\
\%dlcs := root_dist(%lcs) \\
\text{emit} 2*%dlcs/(%dl + %dr + 2*%dlcs)
\end{aligned}
\]

We will not be discussing WQuery in detail.\(^5\) In order to follow the examples its enough to understand that arithmetic expressions, variable assignments (:=) and function calls (f(...)) are interpreted in a similar manner to scripting languages such as Python. The arguments are passed to a function in the %A variable and return values are passed using the emit statement. The main advantage of using WQuery in place of a generic scripting language to implement similarity measures is the ability to use regular expressions over the semantic relation names to denote paths in the wordnet graph. In the case of \texttt{wup_measure} the sub-function \texttt{lcs_dist} that computes the distance from a synset to its least common subsumer determines the paths from a synset \%s to its subsumer \%lcs by the regular expression

\[
%s.\text{hyponym\textvisiblespace}.%lcs
\]

that traverses zero or more times through the hyponym relation from the synset \%s to its subsumer \%lcs. The \texttt{root_dist} function that computes the distance from a synset to the hyponym root uses the expression

\[
%A.\text{hyponym}\textvisiblespace.[\text{empty}(\text{hyponym})]
\]

\(^3\)Through the JW1 library (Finlayson, 2014).

\(^4\)We assume in the following examples that all commands are invoked in the Linux shell environment.

\(^5\)An interested reader may consult (Kubis, 2012).
to denote the paths from a synset $\%A$ through zero or more hypernym links to the synsets that do not have hypernyms\(^6\). We present the complete code implementing these functions below.

```
function lcs_dist do
  %s, %lcs := $A$
  emit min_size(%s.hypernym*,%lcs) - 1
end

function root_dist do
  emit min_size(
  %A.hypernym*[empty(hypernym)]) + 1
end

function min_size do
  emit distinct(min(size(%A)))
end
```

The `lcs_by_depth` function which is also called by `wup_measure` is a built-in function of WQuery that determines the least common subsumers of synsets.

The similarity functions are loaded into WSim at the beginning of execution from a designated directory. Thus, having a correspondence between arguments of `wsim` and function names and the ability to address arbitrary paths in the wordnet graph using the WQuery language, one can easily experiment with definitions of new measures. For instance, one can consider a meronymy-based variant of the `path` measure by providing to `wsim` the following function

```
function mpath_measure do
  %l, %r := $A$
  %mpaths := %l.meronym*,'meronym*,%r
  emit l/min_size(%mpaths)
end
```

### 5. Semantic similarity computation using Polish wordnets

Having a tool that accepts lexical databases stored in the DEBVisDic editor compatible format, we can compute the values of similarity measures for both Polish wordnets and compare them to the human similarity ratings. In the case of English the Rubenstein and Goodenough (1965) dataset of 65 human-rated noun pairs or its 30 pair subset from Miller and Charles (1991) are often used for the purpose of evaluating the similarity measures (e.g. Resnik, 1995; Budanitsky and Hirst, 2006; Pedersen, 2010). Paliwoda-Pękosz and Lula (2009) translated this dataset into Polish and had it rated. They also report the performance of several similarity measures on 39 pairs of the translated nouns which are covered by version 0.95 of PIWordNet. We refer to this dataset as PL39 in the rest of the paper. For the purpose of our analysis we use version 2.2 of PIWordNet (Maziarz et al., 2012) and version 3.0 of PolNet (Vetulani et al., 2010). Furthermore, in order to determine values of measures that utilize information content (i.e. Resnik, Jiang and Conrath, and Lin), we use word frequencies derived from Polish Wikipedia.\(^7\)

PIWordNet 2.2 and PolNet 3.0 cover 38 and 26 pairs of nouns from the PL39 dataset, respectively. The correlation coefficients between values of the similarity measures and the human rating of 26 noun pairs common to both wordnets are given in Table 1. One may notice that regardless of the correlation type the Lin measure performs best. We report the pairs shared by both wordnets and the corresponding values of the Lin measure in Table 2. The same measure achieves the best results in the case of all 38 word pairs covered by PLWordNet (cf. Table 3). For the purpose of comparison we also present the correlation coefficients between the values of measures computed for version 3.0 of WordNet and 26 pairs of English nouns from the Rubenstein and Goodenough dataset being counterparts of the pairs of nouns common to both Polish wordnets (Table 5). In this case the Leacock and Chodorow measure results in the highest Pearson correlation and the Jiang and Conrath measure achieves the highest Spearman’s correlation coefficient value among the analyzed measures.

<table>
<thead>
<tr>
<th>Measure</th>
<th>PIWN</th>
<th>PolNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>path</td>
<td>0.6051</td>
<td>0.6421</td>
</tr>
<tr>
<td>wup</td>
<td>0.6322</td>
<td>0.6835</td>
</tr>
<tr>
<td>lch</td>
<td>0.5981</td>
<td>0.6865</td>
</tr>
<tr>
<td>res</td>
<td>0.6028</td>
<td>0.6394</td>
</tr>
<tr>
<td>jcn</td>
<td>0.5358</td>
<td>0.4938</td>
</tr>
<tr>
<td>lin</td>
<td>0.6591</td>
<td>0.7104</td>
</tr>
</tbody>
</table>

Table 1: Correlation coefficients between the human ratings on PL39 and the measures computed on nouns common to both Polish wordnets.

\(^7\)We use the Polish Wikipedia dump from February 6, 2014.

\(^8\)In the case of information content based measures.

\(^9\)With the exception of the Pearson correlation coefficient for the Jiang and Conrath measure.

### 6. Conclusion

We presented a new framework for semantic similarity computation using wordnet-based measures. The main advantages of our tool are: the compatibility with various wordnet database formats and the possibility to implement new measures using the embedded query language. The framework was employed to model semantic similarity of

---

\(^6\)The synsets satisfying the condition empty(hypernym).

\(^7\)We use the Polish Wikipedia dump from February 6, 2014.

\(^8\)In the case of information content based measures.

\(^9\)With the exception of the Pearson correlation coefficient for the Jiang and Conrath measure.
Table 2: The values of the Lin measure for 26 pairs of nouns shared by both Polish wordnets.

<table>
<thead>
<tr>
<th>Left</th>
<th>Right</th>
<th>PIWN</th>
<th>PolNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>poludnie</td>
<td>szurek</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>owoc</td>
<td>piec</td>
<td>0.2270</td>
<td>0.3499</td>
</tr>
<tr>
<td>kopiec</td>
<td>kuchenka</td>
<td>0.1776</td>
<td>0.3722</td>
</tr>
<tr>
<td>azył</td>
<td>owoc</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>azył</td>
<td>zakonnik</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>chlopiec</td>
<td>kogut</td>
<td>0.4873</td>
<td>0.6407</td>
</tr>
<tr>
<td>zakonnik</td>
<td>niewolnik</td>
<td>0.7075</td>
<td>0.3855</td>
</tr>
<tr>
<td>azył</td>
<td>cmentarz</td>
<td>0.4951</td>
<td>0.2856</td>
</tr>
<tr>
<td>wybrzeże</td>
<td>las</td>
<td>0.6860</td>
<td>0.7451</td>
</tr>
<tr>
<td>kopiec</td>
<td>wybrzeże</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>las</td>
<td>cmentarz</td>
<td>0.0000</td>
<td>0.2875</td>
</tr>
<tr>
<td>jedzenie</td>
<td>kogut</td>
<td>0.7337</td>
<td>0.3460</td>
</tr>
<tr>
<td>wybrzeże</td>
<td>pagórek</td>
<td>0.6085</td>
<td>0.6446</td>
</tr>
<tr>
<td>piec</td>
<td>narzędzie</td>
<td>0.4345</td>
<td>0.5900</td>
</tr>
<tr>
<td>cmentarz</td>
<td>kopiec</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>szkło</td>
<td>klejnot</td>
<td>0.3015</td>
<td>0.6587</td>
</tr>
<tr>
<td>brat</td>
<td>chłopak</td>
<td>0.8422</td>
<td>0.6845</td>
</tr>
<tr>
<td>ptak</td>
<td>kogut</td>
<td>0.7597</td>
<td>0.7705</td>
</tr>
<tr>
<td>jedzenie</td>
<td>owoc</td>
<td>0.2762</td>
<td>0.8822</td>
</tr>
<tr>
<td>brat</td>
<td>zakonnik</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>piec</td>
<td>kuchenka</td>
<td>0.5384</td>
<td>0.3892</td>
</tr>
<tr>
<td>pagórek</td>
<td>kopiec</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>przewód</td>
<td>szurek</td>
<td>0.0000</td>
<td>0.5180</td>
</tr>
<tr>
<td>narzędzie</td>
<td>przyrząd</td>
<td>0.9788</td>
<td>1.0000</td>
</tr>
<tr>
<td>chlopiec</td>
<td>chłopak</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Table 3: Correlation coefficients between the human ratings on PL39 and the measures computed on all nouns covered by PIWordNet.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>path</td>
<td>0.5915</td>
<td>0.5537</td>
</tr>
<tr>
<td>wup</td>
<td>0.6896</td>
<td>0.6738</td>
</tr>
<tr>
<td>lch</td>
<td>0.6423</td>
<td>0.5537</td>
</tr>
<tr>
<td>res</td>
<td>0.6782</td>
<td>0.6912</td>
</tr>
<tr>
<td>jcn</td>
<td>0.4419</td>
<td>0.6517</td>
</tr>
<tr>
<td>lin</td>
<td>0.7073</td>
<td>0.6932</td>
</tr>
</tbody>
</table>

Table 4: Correlation coefficients between the PIWordNet and PolNet computed measures.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>path</td>
<td>0.8503</td>
<td>0.7344</td>
</tr>
<tr>
<td>wup</td>
<td>0.8450</td>
<td>0.8544</td>
</tr>
<tr>
<td>lch</td>
<td>0.8369</td>
<td>0.7344</td>
</tr>
<tr>
<td>res</td>
<td>0.7690</td>
<td>0.7331</td>
</tr>
<tr>
<td>jcn</td>
<td>0.7045</td>
<td>0.7916</td>
</tr>
<tr>
<td>lin</td>
<td>0.8038</td>
<td>0.7923</td>
</tr>
</tbody>
</table>

Table 5: Correlation coefficients between the human ratings and the measures computed for WordNet on 26 noun pairs of the Rubenstein and Goodenough dataset being counterparts of the Polish noun pairs common to both Polish wordnets.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>path</td>
<td>0.7274</td>
<td>0.6351</td>
</tr>
<tr>
<td>wup</td>
<td>0.6795</td>
<td>0.5785</td>
</tr>
<tr>
<td>lch</td>
<td>0.7373</td>
<td>0.6243</td>
</tr>
<tr>
<td>res</td>
<td>0.6598</td>
<td>0.5903</td>
</tr>
<tr>
<td>jcn</td>
<td>0.4310</td>
<td>0.6610</td>
</tr>
<tr>
<td>lin</td>
<td>0.6773</td>
<td>0.5837</td>
</tr>
</tbody>
</table>

nouns using measures derived from two Polish wordnets – PIWordNet and PolNet. The results have to be considered preliminary due to the small size of the dataset being used for the purpose of evaluation. Nevertheless, this is the first attempt to use both Polish wordnets within the context of a shared task.

As for the future, we plan to extend the framework with additional measures (e.g. Hirst and St-Onge, 1998). We also intend to create a larger evaluation set that will cover the content of PolNet more extensively.

References


