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 subsumer[1] (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 PlWordNet (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 databases[2]. 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 Shima (2015) called WS4J beside Princeton WordNet can also load the Japanese WordNet (Isahara 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 PlWordNet 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 PlWordNet 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 PlWordNet 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 PlWordNet 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-

[1] A joint transitive hypernym of two synsets such that no other joint transitive hypernym of these synsets is placed below it within the hypernymy hierarchy.

[2] Databases that are organized in a similar manner to WordNet (Fellbaum, 1998); called wordnets in the rest of the paper.

mat[3], 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 PlWordNet 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

```
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 `wsim` command, separated by tab characters.

```
wsim polnet.wq < pairs
```

By default `wsim` determines similarity of a pair of words by inverting the value of the shortest path length in the hypernymy hierarchy that links the synsets containing the given words, thus for the pair *samochód* (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),
6. Lin (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 `wsim` command. For instance, to compute the Wu and Palmer measure one has to execute the command

```
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.

```
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.

```
wsim wordnet.wq -m res < pairs
```

[3]Through the JWI library (Finlayson, 2014).

[4]We assume in the following examples that all commands are invoked in the Linux shell environment.

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 `wsim` command. For every pair of senses read from the input the `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 `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)

$$\frac{2 * dep(lcs(l,r))}{dist(l,lcs(l,r)) + dist(r,lcs(l,r)) + 2 * dep(lcs(l,r))}$$

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

```
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)
  emit 2*%dlcs/(%dl + %dr + 2*%dlcs)
end
```

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 `wup_measure` the sub-function `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.hypernym*.%lcs
```

that traverses zero or more times through the `hypernym` relation from the synset `%s` to its subsumer `%lcs`. The `root_dist` function that computes the distance from a synset to the hypernymy root uses the expression

```
%A.hypernym*[empty(hypernym)]
```

[5]An interested reader may consult (Kubis, 2012).

to denote the paths from a synset `%A` through zero or more hypernymy 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 1/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 PlWordNet. 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 PlWordNet (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]

Measure	Pearson		Spearman	
	PlWN	PolNet	PlWN	PolNet
path	0.6051	0.6421	0.4658	0.6530
wup	0.6322	0.6835	0.6079	0.6902
lch	0.5981	0.6865	0.4658	0.6530
res	0.6028	0.6394	0.6452	0.6498
jcn	0.5358	0.4938	0.6114	0.6700
lin	**0.6591**	**0.7104**	**0.6520**	**0.7084**

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

PlWordNet 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.

Given the correlation coefficients for a fixed measure and the same corpus[8], it is tempting to compare the differences between the two wordnets with respect to the results on the same dataset. However, it has to be noted that although the correlation coefficients between human ratings for the 26 nouns from PL39 and measure values induced from PolNet are generally higher[9] than the corresponding coefficients derived for PlWordNet, the results are hard to interpret due to the dataset size and are not significant at the $\alpha = 0.05$ level according to the Meng, Rosenthal, and Rubin's z-test as implemented by Diedenhofen (2013).

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 measaure.

Left	Right	PlWN	PolNet
południe	sznurek	0.0000	0.0000
owoc	piec	0.2270	0.3499
kopiec	kuchenka	0.1776	0.3722
azyl	owoc	0.0000	0.0000
azyl	zakonnik	0.0000	0.0000
chłopiec	kogut	0.4873	0.6407
zakonnik	niewolnik	0.7075	0.3355
azyl	cmentarz	0.4951	0.2856
wybrzeże	las	0.6860	0.7451
kopiec	wybrzeże	0.0000	0.0000
las	cmentarz	0.0000	0.2875
jedzenie	kogut	0.7337	0.3460
wybrzeże	pagórek	0.6085	0.6446
piec	narzędzie	0.4345	0.5900
cmentarz	kopiec	0.0000	0.0000
szkło	klejnot	0.3015	0.6587
brat	chłopak	0.8422	0.6845
ptak	kogut	0.7597	0.7705
jedzenie	owoc	0.2762	0.8822
brat	zakonnik	1.0000	1.0000
piec	kuchenka	0.5384	0.3892
pagórek	kopiec	1.0000	1.0000
przewód	sznurek	0.0000	0.5180
narzędzie	przyrząd	0.9788	1.0000
chłopiec	chłopak	1.0000	1.0000
auto	samochód	1.0000	0.8661

Table 2: The values of the Lin measure for 26 pairs of nouns shared by both Polish wordnets.

Measure	Pearson	Spearman
path	0.5915	0.5537
wup	0.6896	0.6738
lch	0.6423	0.5537
res	0.6782	0.6912
jcn	0.4419	0.6517
lin	**0.7073**	**0.6932**

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

Measure	Pearson	Spearman
path	**0.8503**	0.7344
wup	0.8450	**0.8544**
lch	0.8369	0.7344
res	0.7690	0.7331
jcn	0.7045	0.7916
lin	0.8038	0.7923

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

Measure	Pearson	Spearman
path	0.7274	0.6351
wup	0.6795	0.5785
lch	**0.7373**	0.6243
res	0.6598	0.5903
jcn	0.4310	**0.6610**
lin	0.6773	0.5837

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.

nouns using measures derived from two Polish wordnets – PlWordNet 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

Bird, Steven, Ewan Klein, and Edward Loper, 2009. *Natural Language Processing with Python*. O'Reilly Media.

Budanitsky, Alexander and Graeme Hirst, 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47.

Diedenhofen, Birk, 2013. *cocor: Comparing correlations*. http://r.birkdiedenhofen.de/pckg/cocor/ (Version 1.0-0).

Fellbaum, Christiane (ed.), 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Finlayson, Mark Alan, 2014. Java Libraries for Accessing the Princeton Wordnet: Comparison and Evaluation. In *Proceedings of the 7th Global Wordnet Conference*. Tartu, Estonia.

Global WordNet Association, 2012. Global WordNet Grid. http://globalwordnet.org/global-wordnet-grid/. Access date: September 20, 2015.

Hirst, Graeme and David St-Onge, 1998. *Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms*, chapter 13. In (Fellbaum, 1998), pages 305–332.

Horak, Ales, Karel Pala, Adam Rambousek, and Martin Povolny, 2005. DEBVisDic - First Version of New Client-Server Wordnet Browsing and Editing Tool. In Petr Sojka et al. (eds.), *Proceedings of the Third International WordNet Conference – GWC 2006*. Brno, Czech Republic: Masaryk University.

Isahara, Hitoshi, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki, 2008. Development of the Japanese WordNet. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association.

Jiang, Jay J. and David W. Conrath, 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of 10th International Conference on Research in Computational Linguistics, ROCLING'97*.

Kubis, Marek, 2012. A Query Language for WordNet-like Lexical Databases. In Jeng-Shyang Pan, Shyi-Ming Chen, and Ngoc-Thanh Nguyen (eds.), *Intelligent Information and Database Systems*, volume 7198 of *Lecture Notes in Artificial Intelligence*. Springer Heidelberg, pages 436–445.

Kubis, Marek, 2014. A Tool for Transforming WordNet-Like Databases. In Zygmunt Vetulani and Joseph Mariani (eds.), *Human Language Technology Challenges for Computer Science and Linguistics*, volume 8387 of *Lecture Notes in Computer Science*. Springer International Publishing, pages 343–355.

Leacock, Claudia and Martin Chodorow, 1998. *Combining Local Context and WordNet Similarity for Word Sense Identification*, chapter 11. In (Fellbaum, 1998), pages 265–283.

Lin, Dekang, 1998. An Information-Theoretic Definition of Similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Maziarz, Marek, Maciej Piasecki, and Stanisław Szpakowicz, 2012. Approaching plWordNet 2.0. In *Proceedings of the 6th Global Wordnet Conference*. Matsue, Japan.

Miller, George A. and Walter G. Charles, 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Paliwoda-Pękosz, Grażyna and Paweł Lula, 2009. Measures of Semantic Relatedness Based on Wordnet. In *International Workshop For PhD Students*. Brno, Czech Republic. ISBN: 978-80-214-3980-1.

Patwardhan, Siddharth, Satanjeev Banerjee, and Ted Pedersen, 2003. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, volume 2588 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, pages 241–257.

Pedersen, Ted, 2010. Information Content Measures of Semantic Similarity Perform Better Without Sense-tagged Text. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10. Stroudsburg, PA, USA: Association for Computational Linguistics.

Pedersen, Ted, Siddharth Patwardhan, and Jason Michelizzi, 2004. WordNet::Similarity: Measuring the Relatedness of Concepts. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04. Stroudsburg, PA, USA: Association for Computational Linguistics.

Postma, Marten and Piek Vossen, 2014. What implementation and translation teach us: the case of semantic similarity measures in wordnets. In Heili Orav, Christiane Fellbaum, and Piek Vossen (eds.), *Proceedings of the Seventh Global Wordnet Conference*. Tartu, Estonia.

Rada, Roy, Hafedh Mili, Ellen Bicknell, and Maria Blettner, 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30.

Resnik, Philip, 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Rubenstein, Herbert and John B. Goodenough, 1965. Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633.

Shima, Hideki, 2015. ws4j – WordNet Similarity for Java. https://code.google.com/p/ws4j/. Access date: 2015-08-28.

Soria, Claudia, Monica Monachini, and Piek Vossen, 2009. Wordnet-LMF: Fleshing out a Standardized Format for Wordnet Interoperability. In *Proceeding of the 2009 international workshop on Intercultural collaboration*. New York, USA: ACM.

Stevenson, Mark and Mark A. Greenwood, 2005. A Semantic Approach to IE Pattern Induction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05. Stroudsburg, PA, USA: Association for Computational Linguistics.

Tengi, Randee I., 1998. *Design and Implementation of the WordNet Lexical Database and Searching Software*, chapter 4. In (Fellbaum, 1998), pages 105–127.

Vetulani, Zygmunt, Marek Kubis, and Tomasz Obrębski, 2010. PolNet - Polish WordNet: Data and Tools. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. Valletta, Malta: ELRA.

Wu, Zhibiao and Martha Palmer, 1994. Verbs Semantics and Lexical Selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94. Stroudsburg, PA, USA: Association for Computational Linguistics.