Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663

www.czasopisma.uni.lodz.pl/foe/
6(351) 2020

DOI: https://doi.org/10.18778/0208-6018.351.06

Jerzy Korzeniewski

University of Łódź, Faculty of Economics and Sociology, Department of Demography, Łódź, Poland
jerzy.korzeniewski@uni.lodz.pl

A Modification of the Leacock-Chodorow Measure of the Semantic Relatedness of Concepts

Abstract: The measures of the semantic relatedness of concepts can be categorised into two types: knowledge-based methods and corpus-based methods. Knowledge-based techniques make use of man-created dictionaries, thesauruses and other artefacts as a source of knowledge. Corpus-based techniques assess the semantic similarity of two concepts making use of large corpora of text documents. Some researchers claim that knowledge-based measures outperform corpus-based ones, but it is much more important to observe that the latter ones are heavily corpus dependent. In this article, we propose to modify the best WordNet-based method of assessing semantic relatedness, i.e. the Leacock-Chodorow measure. This measure has proven to be the best in several studies and has a very simple formula. We asses our proposal on the basis of two popular benchmark sets of pairs of concepts, i.e. the Ruben-Goodenough set of 65 pairs of concepts and the Fickelstein set of 353 pairs of terms. The results prove that our proposal outperforms the traditional Leacock-Chodorow measure.

Keywords: text mining, WordNet network, semantic relatedness, Lecock-Chodorow measure

JEL: C39, C65, Z13

1. Introduction

The wish to determine semantic relatedness or its inverse, semantic distance, between two words, terms or, more broadly, two lexical concepts is a problem that dominates many tasks of natural language processing such as document summarisation, information retrieval, information extraction, word sense disambiguation, machine text translation, thesaurus creation, and the automatic correction of errors in texts. Many of these tasks require a numerical measure of the semantic relatedness between two arbitrary terms. For example, in information retrieval, we are in need of such assessments in order to expand the query words; facing the problem of word sense disambiguation, we need them in order to choose an appropriate meaning of a word. It is of substantial importance to note that semantic relatedness is a more general notion than similarity; similarterms are semantically related due to their similarity (*football – rugby*), but dissimilar terms may also be semantically related due to relationships such as antonymy (*cold – heat*), or meronymy (*car – motor*), or by any kind of frequent association (*water – fire, goalkeeper – football, rain – umbrella*). The aforementioned computational tasks usually make use of relatedness rather than similarity.

However, it is not certain how to assess many available approaches that have been designed for measuring semantic relatedness. The most widely accepted approach is to assess the quality of methods by checking how they mimic human judgement on the relatedness of a given pair of terms. Therefore, some benchmark data sets should be required to make any research feasible. We use two popular data sets in our research. One of two major groups of methods of determining semantic relatedness, i.e. the group of knowledge-based methods, has to refer to some kind of dictionary, thesaurus, or similar source. It is not certain which one is the best, especially if we take into account the special area with which a given study is concerned. In our research, we will refer to probably the most comprehensive database of the English language, namely the WordNet. We give a short description of this database in the next section.

The remaining part of the article is organised as follows. Section 2 contains the description of WordNet. In Section 3, we present an overview of existing approaches, in Section 4 we propose a modification of the Leacock-Chodorow measure, and in Section 5 we present its evaluation. In Section 6, some concluding remarks are given.

2. WordNet description

WordNet is a large lexical database of the English language which was devised at Princeton University. In WordNet, nouns, verbs, adjectives and adverbs are grouped into sets of synonyms (synsets), each expressing a distinct concept.

In WordNet 3.0, there are 147,278 concept nodes, 70% of which are nouns. The backbone of the relations between them is constituted by hypernymy and hyponymy (accounting for almost 80% of relations). Apart from these two, synonymy, antonymy and meronymy (6 types) are used. At the top of the hierarchy, there are 25 abstract concepts termed unique beginners (see Figure 1). The maximum depth of the noun hierarchy is 16 nodes (17 if the theoretical top root is included).

{act, action, activity}
{artefact}
{animal, fauna}
{attribute, property}
{body, corpus}
{cognition, knowledge}
{communication}
{event, happening}
{feeling, emotion}
{natural object}
{plant, flora}
{natural phenomenon}
{possession}
{process}
{quantity, amount}
{relation}
{shape}
{state, condition}
{food}
{substance}
{time}
{group, collection}
{location, place}
{motive}

Figure 1. List of 25 unique beginners for nouns in WordNet
Source: Fellbaum, 1998

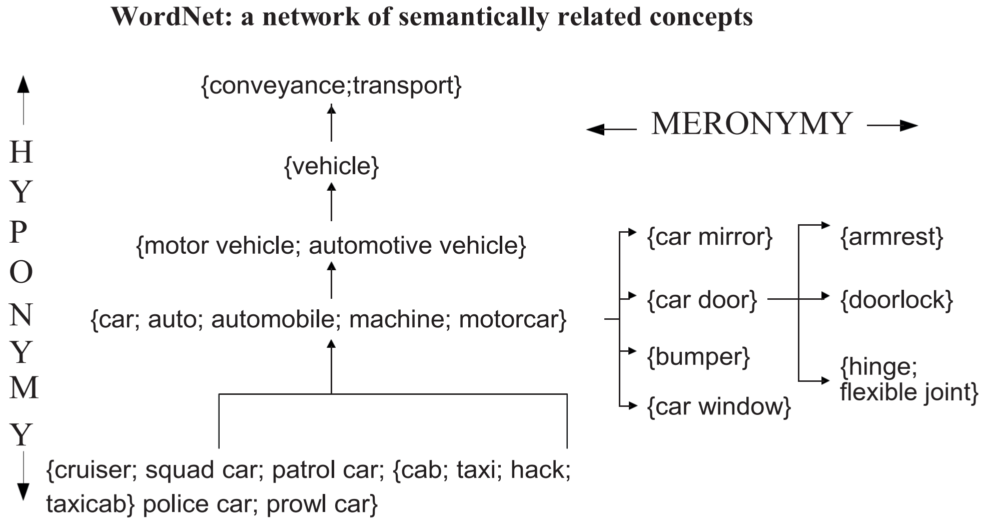
WordNet: a network of semantically related concepts
HYPONYMY
{conveyance;transport}
{vehicle}
{motor vehicle; automotive vehicle}
{car; auto; automobile; machine; motorcar}
{cruiser; squad car; patrol car; {cab; taxi; hack; taxicab} police car; prowl car}
MERONYMY
{car mirror}
{car door}
{bumper}
{car window}
{armrest}
{doorlock}
{hinge; flexible joint}

Figure 2. Exemplary structure of the WordNet network
Source: Fellbaum, 1998

In Figure 2, an exemplary structure of concepts connected with the word *car* is presented with some relations between these concepts. Mining WordNet can be made easier by applying packages or programming platforms. In our research, the *nltk* package (Bird, Loper, Klein, 2009) was used extensively.

3. Overview of existing approaches

The measures of the semantic relatedness of terms can be categorised into two types: knowledge-based methods and corpus-based methods. Knowledge-based techniques make use of man-created dictionaries, thesauruses and other artefacts as a source of knowledge. Corpus-based techniques assess semantic relatedness making use of a large corpus of text documents. Generally, there is no agreement on whether knowledge-based measures outperform corpus-based ones, but, what is crucial in our opinion, the latter ones are heavily corpus dependent, and thus unsettled. Budanitsky and Hirst (2006) provide a comparison of five different measures of either similarity (or distance) or relatedness of pairs of concepts. Let us first concentrate on knowledge-based methods. In the formulas given below, we use the following notation: $len(c_i, c_j)$ – the shortest path between concept c_i and concept c_j; $depth(c_i)$ – the taxonomy depth of concept c_i, i.e. the length of the path from the root of the taxonomy to concept c_i; $lso(c_i, c_j)$ – the lowest common subsume (i.e. hypernym) of both concepts c_i and c_j. Hirst and St-Onge (1998) propose the following relatedness measure:

$$rel_{HS}\left(c_i,c_j\right)=C-len\left(c_i,c_j\right)-k\cdot turns\left(c_i,c_j\right). \tag{1}$$

In this formula, $turns(c_i, c_j)$ is the number of the direction changes on the path from c_i to c_j. Symbols C and k are constants in the aforementioned research: $C = 8$, $k = 1$. Leacock and Chodorow (1998) propose the following similarity measure:

$$sim_{LC}\left(c_i,c_j\right)=-log_2\frac{len\left(c_i,c_j\right)}{2\cdot\max\_depth}. \tag{2}$$

A popular (available in the *nltk* computer package) measure of similarity is the Wu and Palmer (1994) formula:

$$sim_{WP}\left(c_i,c_j\right)=\frac{2\cdot H}{N_1+N_2+H}, \tag{3}$$

where N_1 and N_2 is the number of "is-a" links from, respectively, c_i and c_j to $lso(c_i, c_j)$, and H is the number of "is-a" links from $lso(c_i, c_j)$ to the root of the taxonomy.

In order to provide some kind of comparison basis, we present the results of the Budanitsky and Hirst (2006) research (see Table 1) along with three corpus based measures. The idea of this group of methods is to use a measure of the information content (IC) of concept c in the form of the following logarithm in base 2 of the likelihood $p(c)$ of the occurrence of concept c:

$$IC(c) = -\log p(c)\,. \tag{4}$$

Thus, the formulas of the three measures are as follows. The Resnick (1995) similarity measure:

$$sim_R\left(c_i, c_j\right) = -\log p\left(lso\left(c_i, c_j\right)\right). \tag{5}$$

The Jiang and Conrath (1997) distance measure:

$$dist_{JC}\left(c_i, c_j\right) = 2\log p\left(lso\left(c_i, c_j\right)\right) - \log p\left(c_i\right) - \log p\left(c_j\right). \tag{6}$$

The Lin (1998) similarity measure:

$$sim_L\left(c_i, c_j\right) = \frac{2\log p\left(lso\left(c_i, c_j\right)\right)}{\log p\left(c_i\right) + \log p\left(c_j\right)}. \tag{7}$$

In recent years, some new similarity or relatedness measures appeared, however, to the best of the author's knowledge, none of them is entirely knowledge-based, and they are usually topic dominated methods. For example, Zugang, Jia and Yaping (2018) developed an interesting semantic relatedness measure for geographical applications and McInnes et al. (2014) proposed a measure to be applied in medicine.

Table 1. The Rubenstein-Goodenough set of word pairs with human ratings of semantic relatedness

1	cord	smile	0.02	34	car	journey	1.55
2	rooster	voyage	0.04	35	cemetery	mound	1.69
3	noon	string	0.04	36	glass	jewel	1.78
4	fruit	furnace	0.05	37	magician	oracle	1.82
5	autograph	shore	0.06	38	crane	implement	2.37
6	automobile	wizard	0.11	39	brother	lad	2.41
7	mound	stove	0.14	40	sage	wizard	2.46
8	grin	implement	0.18	41	oracle	sage	2.61
9	asylum	fruit	0.19	42	bird	crane	2.63
10	asylum	monk	0.39	43	bird	cock	2.63
11	graveyard	madhouse	0.42	44	food	fruit	2.69
12	glass	magician	0.44	45	brother	monk	2.74
13	boy	rooster	0.44	46	asylum	madhouse	3.04
14	cushion	jewel	0.45	47	furnace	stove	3.11
15	monk	slave	0.57	48	magician	wizard	3.21
16	asylum	cemetery	0.79	49	hill	mound	3.29
17	coast	forest	0.85	50	cord	string	3.41
18	grin	lad	0.88	51	glass	tumbler	3.45

19	shore	woodland	0.90	52	grin	smile	3.46
20	monk	oracle	0.91	53	serf	slave	3.46
21	boy	sage	0.96	54	journey	voyage	3.58
22	automobile	cushion	0.97	55	autograph	signature	3.59
23	mound	shore	0.97	56	coast	shore	3.60
24	lad	wizard	0.99	57	forest	woodland	3.65
25	forest	graveyard	1.00	58	implement	tool	3.66
26	food	rooster	1.09	59	cock	rooster	3.68
27	cemetery	woodland	1.18	60	boy	lad	3.82
28	shore	voyage	1.22	61	cushion	pillow	3.84
29	bird	woodland	1.24	62	cemetery	graveyard	3.88
30	coast	hill	1.26	63	automobile	car	3.92
31	furnace	implement	1.37	64	midday	noon	3.94
32	crane	rooster	1.41	65	gem	jewel	3.94
33	hill	woodland	1.48				

Source: Budanitsky, Hirst, 2006

4. A Modification of the Leacock-Chodorow measure

We used two popular benchmark data sets in order to analyse the quality of the Leacock-Chodorow measure and to find possibilities of improving it. The first data set is the Rubenstein-Goodenough (RG65) data set of 65 pairs of nouns (see Table 1) meant rather for assessing similarity than relatedness. The second data set is the Fickelstein (F353) (avail. at http://alfonseca.org/eng/research/wordsim353.html) set of 353 pairs of terms, meant rather for assessing relatedness. The RG65 dataset was analysed in the research carried out by Budanitsky and Hirst (2006) for some of the mentioned methods. One has to keep in mind that different values given in various studies are of different nature, some are distances (dissimilarities) and some are similarities. Therefore, in order to achieve some kind of comparison basis, one has to make them uniform, e.g. transform the values to the relatedness measure on the interval [0; 1]. The same goes for judgements provided by humans, they are usually given on different scales, e.g. in the RG65 set, the scale was from 0 to 4, and in the F353 set, the scale was from 0 to 10. After standardising the results, it turned out that the Leacock-Chodorow measure proved to be the best (in the case of both RG65 and F353 sets), both in terms of the medium arithmetic absolute deviation from the human judgement and in terms of the correlation measured by the Spearman rank correlation coefficient.

Taking a closer look at some particular pairs of words and at formula (2), it is easy to observe that the reason the Leacock-Chodorow measure has proven wrong is probably too deep normalisation. If both compared words are in the middle of the WordNet taxonomy, or even at the top, but do not have much in common (which occurs very often), formula (2) tends to assign relatedness values of medium size while human values are close to zero. This observation is illustrated in two graphs in Figure 3. To the right of the number 0.5, on the horizontal axis, the values of the mean absolute deviation between the Leacock-Chodorow measure and human judgements stabilise, the worst departures from human judgement occur at the beginning. Therefore, we suggest to modify the Leacock-Chodorow measure in the following way. Up to a certain threshold *T*, say *T* = 0.5, calculate the measure in the form of a linear combination of global normalisation (as in the Leacock-Chodorow measure) with coefficient α and local normalisation with coefficient 1 – α, see formula (8). The local depth is the sum of the taxonomy depths of both concepts c_i, c_j.

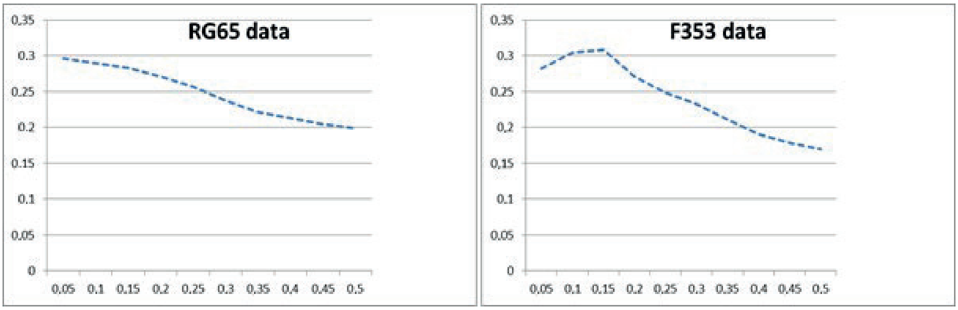

Figure 3. Arithmetic mean absolute deviation between the Leacock-Chodorow method and human judgements for the pairs of concepts for which the human judgement is below the value given on the horizontal axis.

Source: own elaboration

$$sim\left(c_i,c_j\right) = -\alpha \cdot \log_2 \frac{len\left(c_i,c_j\right)}{2 \cdot \max\_depth} - \left(1-\alpha\right) \cdot \log_2 \frac{len\left(c_i,c_j\right)}{2 \cdot local\_depth}. \qquad (8)$$

Above the threshold *T*, calculate the measure as in the original Leacock-Chodorow formula. As far as the choice of α is concerned, we propose the value of the Leacock-Chodorow measure for α, albeit other options, e.g. the squared measure, might also be attractive.

5. Experimental evaluation

We evaluated our proposal using both RG65 and F353 datasets. We applied two criteria. The first one was the arithmetic mean absolute deviation between human judgements and those resulting from the methods for the pairs of words for which the human judgement did not exceed 0.05; 0.1; 0.15; 0.2; 0.25; 0.3; 0.35; 0.4; 0.45; 0.5. The second one was the Spearman rank correlation coefficient between the measures' rank values and the rank values of human judgements with the correction for tied ranks.

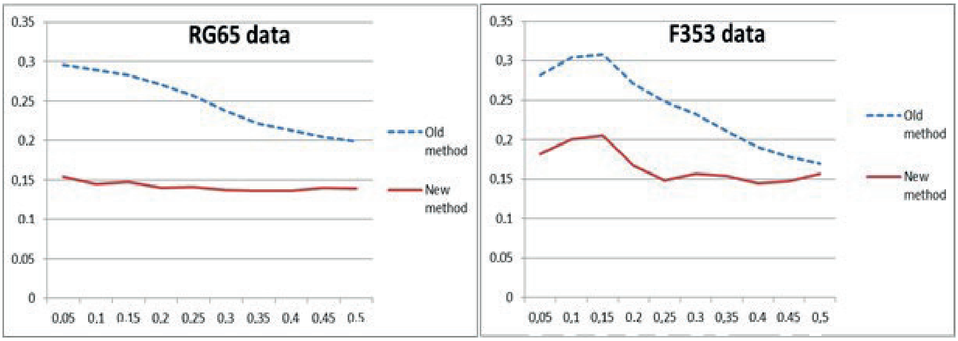

Figure 4. Absolute deviation between both methods and human judgements for the pairs of concepts for which the human judgement is below the value given on the horizontal axis

Source: own elaboration

The results of the first criterion are presented in Figure 3. It follows clearly that the modification achieved much smaller deviations from human judgements than the original Leacock-Chodorow formula. The results of the second criterion are as follows. For the RG65 dataset, the original Leacock-Chodorow formula achieved 0.782 correlation and our modification achieved 0.783. For the F353 dataset, the original Leacock-Chodorow formula achieved 0.317 correlation while our modification had 0.333 correlation.

6. Conclusions

In our opinion, the proposed modification achieves better results because it does not have the drawback of relating the distance between two terms to the whole depth of the WordNet taxonomy. The modification creates perspectives for further developments in mimicking human judgements more closely, e.g. one can use different α in formula (8) or one can take into account terms or concepts closely related to the concepts analysed with a view to smoothing 'discrepancies' resulting from single word analysis. One can also try to determine α by means of optimisation techniques with respect to, e.g. topic-oriented measures or with respect to WordNet searching techniques aimed at analysing closely related terms.

References

Bird S., Loper E., Klein E. (2009), *Natural Language Processing with Python*, O'Reilly Media Inc., Sebastopol.

Budanitsky A., Hirst G. (2006), *Evaluating WordNet-based Measures of Lexical Semantic Relatedness*, "Computational Linguistics", vol. 32, issue 1, pp. 13–47.

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

Hirst G., St-Onge D. (1998), *Lexical chains as representations of context for the detection and correction of malapropisms*, [in:] Ch. Fellbaum (ed.), *WordNet: An Electronic Lexical Database*, The MIT Press, Cambridge, pp. 305–332.

Jiang J., Conrath D. (1997), *Semantic similarity based on corpus statistics and lexical taxonomy*, Proceedings of International Conference on Research in Computational Linguistics, Taiwan, pp. 19–33.

Leacock C., Chodorow M. (1998), *Combining local context and WordNet similarity for word sense identification*, [in:] Ch. Fellbaum (ed.), *WordNet: An Electronic Lexical Database*, The MIT Press, Cambridge, pp. 265–283.

Lin D. (1998), *Automatic retrieval and clustering of similar words*, Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING–ACL '98), Montreal, pp. 296–304.

McInnes B., Pedersen T., Liu Y., Melton G., Pakhomov S. (2014), *U-path: An undirected path-based measure of semantic similarity*, Proceedings of the Annual Symposium of the American Medical Informatics Association, Washington, pp. 882–891.

Resnick P. (1995), *Using information content to evaluate semantic similarity*, Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, pp. 448–453.

Wu Z., Palmer M. (1994), *Verbs semantics and lexical selection*, Proceedings of the 32nd annual meeting on Association for Computational Linguistics, ACL '94, Association for Computational Linguistics, Stroudsburg, pp. 133–138.

Zugang C., Jia S., Yaping Y. (2018), *An Approach to Measuring Semantic Relatedness of Geographic Terminologies Using a Thesaurus and Lexical Database Sources*, "International Journal of Geo-Information", vol. 7(3), pp. 98–12.

Modyfikacja miary semantycznego podobieństwa pojęć Leacock-Chodorowa

Streszczenie: Miary semantycznego podobieństwa pojęć można podzielić na dwa rodzaje: metody oparte na wiedzy i metody oparte na bazie tekstów. Techniki oparte na wiedzy stosują stworzone przez człowieka słowniki oraz inne opracowania. Techniki oparte na bazie tekstów oceniają podobieństwo semantyczne dwóch pojęć, odwołując się do obszernych baz dokumentów tekstowych. Niektórzy badacze twierdzą, że miary oparte na wiedzy są lepsze jakościowo od tych opartych na bazie tekstów, ale o wiele istotniejsze jest to, że te drugie zależą bardzo mocno od użytej bazy tekstów. W niniejszym artykule przedstawiono propozycję modyfikacji najlepszej metody pomiaru semantycznego podobieństwa pojęć, opartej na sieci WordNet, a mianowicie miary Leacock-Chodorowa. Ta miara była najlepsza w kilku eksperymentach badawczych oraz można zapisać ją za pomocą prostej formuły. Nową propozycję oceniono na podstawie dwóch popularnych benchmarkowych zbiorów par pojęć, tj. zbioru 65 par pojęć Rubensteina-Goodenougha oraz zbioru 353 par pojęć Fickelsteina. Wyniki pokazują, że przedstawiona propozycja spisała się lepiej od tradycyjnej miary Leacock-Chodorowa.

Słowa kluczowe: badanie tekstu, sieć WordNet, podobieństwo semantyczne słów,
miara Leacock-Chodorowa

JEL: C39, C65, Z13

OPEN ACCESS	© by the author, licensee Lodz University – Lodz University Press, Łódź, Poland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license CC-BY (https://creativecommons.org/licenses/by/4.0/)
	Received: 2020-10-18; verified: 2020-12-15. Accepted: 2021-02-26
COPE Member since 2018 JM13703	This journal adheres to the COPE's Core Practices https://publicationethics.org/core-practices