COGNITIVE STUDIES | ÉTUDES COGNITIVES, 20
Warsaw 2020

Article No.: 2325
DOI: 10.11649/cs.2325

Citation: Maziarz, M., & Rudnicka. E. (2020). Expanding WordNet with gloss and polysemy links for evocation strength recognition. *Cognitive Studies | Études cognitives*, *2020*(20), Article 2325. https://doi.org/10.11649/cs.2325

MAREK MAZIARZ
Wroclaw University of Science and Technology, Wroclaw, Poland
https://orcid.org/0000-0003-0318-2869
marek.maziarz@pwr.edu.pl

EWA RUDNICKA
Wroclaw University of Science and Technology, Wroclaw, Poland
https://orcid.org/0000-0002-8738-2739
ewa.rudnicka@pwr.edu.pl

EXPANDING WORDNET WITH GLOSS AND POLYSEMY LINKS FOR EVOCATION STRENGTH RECOGNITION

Abstract

Evocation — a phenomenon of sense associations going beyond standard (lexico)-semantic relations — is difficult to recognise for natural language processing systems. Machine learning models give predictions which are only moderately correlated with the evocation strength. It is believed that ordinary graph measures are not as good at this task as methods based on vector representations. The paper proposes a new method of enriching the WordNet structure with weighted polysemy and gloss links, and proves that Dijkstra's algorithm performs equally as well as other more sophisticated measures when set together with such expanded structures.
Keywords: evocation; WordNet; glosses; polysemy; evocation strength; semantic relations

1 Introduction

Evocation is a psycho-linguistic phenomenon of associations arising between specific word *senses* that go beyond standard (lexico)-semantic relations. For example, *ankle-n-1*, in the sense of 'a gliding joint between the distal ends of the tibia and fibula and the proximal end of the talus', evokes *swell-v-3*, in the sense of 'expand abnormally' (Boyd-Graber et al., 2006; Nikolova et al., 2009). As such, they resemble simple free *word associations* that immediately come to the mind of a native speaker when presented with a stimulus word, e.g. *girl* and *boy*, or *harm* and *bad*. Both evocations and word associations are asymmetric and weighted relations, yet word associations do not specify which particular word senses are involved (Ma, 2013).

From a theoretical perspective, the phenomenon of evocation should also be distinguished from word/sense similarity and semantic relatedness. Relatedness is usually recognised as more general than similarity (Agirre et al., 2009; Ballatore et al., 2014; Faruqui et al., 2016). Following Cramer (2008), we treat the three concepts, i.e. similarity – relatedness – evocation, as forming a chain of subsumed senses, with similarity as the narrowest and evocation as the broadest one. Semantic similarity signifies close semantic resemblance (such as synonymy, near-synonymy or hyponymy). Relatedness covers both close semantic relations as well as more distant relationships

(such as topic generalisation, remote superordinate chains or meronymy/holonymy). Evocation allows for even weaker semantic associations. In particular, the association relationship includes co-hyponyms, opposites, collocates, superordinates and synonyms (Cruse, 2006, p. 191). The three semantic relationships may link word senses or words themselves.[1]

Evocation recognition can facilitate such natural language processing (henceforth, NLP) tasks as measuring textual similarity and relatedness or detecting coreference (Hayashi, 2016). Recognising evocations and assessing their strength remains a very difficult task for NLP (Cramer, 2008). This is due to the lack of large scale language resources that cover this specific type of relations. Currently, there exists a gold standard evocation data set, namely The Princeton Evocation Data.[2] It consists of 120,000 sense pairs annotated by 20 trained undergraduates from Princeton University (Boyd-Graber et al., 2006). The experiment was conducted as follows. Pairs of word senses were randomly selected from Princeton WordNet (Fellbaum, 1998; Miller & Fellbaum, 2007).[3] Every pair of word senses was annotated by at least three annotators with respect to the degree of its association strength ranging from 0 to 100. Due to the random design of the experiment, two-thirds of the word sense pairs selected were judged to be completely unrelated (receiving a score of 0).

This paper presents a novel method of optimising the WordNet structure for the purposes of evocation strength recognition. Graph distance has been used as a measure of semantic dissimilarity between concepts. Following the suggestions of Hayashi (2016, p. 1666), an effort has been made to improve the graph topology in order to achieve more adequate distance measurements. The WordNet graph has been extended with new types of edges, namely gloss links and polysemy links (Sec. 3.1). The links have been directed and their weights have been optimised in concurrent experiments (Sec. 3.2). A simple evocation strength measure is proposed, being the inverse of Dijkstra's distance (Sec. 3.3). It works equally as well as the best individual measures reported in the literature, despite the fact that it is much simpler (Sec. 3.4). Including gloss and polysemy links significantly increases the efficiency of evocation recognition when compared to the bare WordNet graph. The proof that such a way of enriching the WordNet structure is possible is the main contribution of the present paper.

2 Related work

The task of evocation strength recognition is a challenge for NLP. Sense associations spread in various directions and are asymmetrical (Cramer, 2008). There are no language resources designed

[1] We assume the following set of definitions: (Def. 1) *Orthographic word* or *token* is a string of letters (and other symbols) of the English alphabet delimited in writing usually by spaces (Lyons, 1977, pp. 49–50; Saeed, 2003, pp. 55–56). (Def. 2) *Word-form* is either an orthographic word (in the case of one-word lexical items), or a sequence of orthographic words (in the case of multi-word lexical items, treated as 'words with spaces') as seen from the perspective of inflection, see Saeed (2003, p. 56), Lyons (1977, p. 50), Sag et al. (2002). (Def. 3) *Lemma* is the canonical word-form that was chosen to represent other inflectional forms in a dictionary as an entry term (Lyons, 1977, p. 50; Saeed, 2003, p. 56; Svensen, 2009, p. 93). (Def. 4) *Word* is the class of all word-forms equivalent — according to English inflectional patterns — to the same lemma and representing one sense or several related senses. (Ex.) For instance, we treat the word [*go*] as the class of all semantically related word-forms, including *go*, *goes*, *going*, *went*, *gone*, equivalent to each other, since they all might be equated with the lemma *go* according to inflectional rules (like adding the affix -(*e*)*s* to a stem and irregular verb alternations codified in grammars of English). (Remark 1) We may operationalise the notions of *lemma*, *word* and *equivalence relation* by taking the output of existing English lemmatizers, ascribing to a given word-form its lemma. (Remark 2) For WordNet words the definition of the term *lemma* could be narrowed to basic word-forms of nouns, adjectives, verbs, and adverbs. Hence, we regarded nominal, adjectival, verbal or adverbial word-forms as representing the same word if only they shared the same lemma in WordNet and were semantically related. (Remark 3) Because of the isomorphism between lemma and its word/equivalence class, we talked about lemmas in such a manner as if we would describe words themselves. (Def. 5) *Sense* or *meaning* is the triple <lemma, POS, sense number>, where parts of speech and sense numbers (called *variants*) are taken from Princeton WordNet.

[2] `https://wordnet.cs.princeton.edu/downloads.html`

[3] 1,000 synsets — sets of synonymous lexical units — classified as denoting the so called core concepts were involved in the experiment.

to aid evocation recognition. WordNets mainly focus on paradigmatic relations, while various thesauri and ontologies capture special vocabulary and taxonomic dependencies. Valence lexicons cover predicate-argument relations. Therefore, the Princeton evocation data set was intended to complement the taxonomy of WordNet.

Ordinary semantic similarity measures have proven to be completely inefficient in capturing evocation strength. When Boyd-Graber et al. (2006) implemented them, they achieved only 0.131 of Spearman's ρ.[4] Hayashi (2016) confirmed this finding — his individual WordNet-based measures achieved Pearson's correlation $r = 0.15$ at most. His results were much better for complex vector-space-based measures, with $max(r) = 0.30$ (cf. Figure 8 at the end of this paper). His final model, performing at $r = 0.4391$, was a neural network combining a dozen individual measures, with no feature playing the central role. According to Hayashi, further advancement in the field of evocation recognition should proceed in two complementary directions: (i) applying more sophisticated machine learning frameworks and (ii) gathering and merging new and better features. He suggested making use of high quality word/sense vector representations and relational features, as well as more adequate semantic networks (in which distance measures could be applied). Most of Hayashi's best individual measures rely on calculating distances in different semantic spaces. Two out of four of his best measures reaching r values higher than 0.2 are cosine functions (for word2vec and AutoExtend vectors), and one is the AutoExtend difference of two vectors.

Cattle and Ma (2017) focused on predicting *word* association strength in the Princeton evocation data set. Although the task was different (instead of concept evocations, they were seeking word associations), the results were strikingly similar to those obtained by Hayashi. Again, cosine vector similarities (such as w2v, GloVe and w2g embeddings) proved to be the best. More recently, Kacmajor and Kelleher (2019) tested several individual measures of *word* similarity on the evocation set. The authors divided their measures into four broad groups: (i) knowledge-based distance measures, (ii) measures utilizing vector space models constructed out of existing lexical resources, (iii) distributional vector space measures based on large corpora, and (iv) hybrid approaches mixing knowledge-based and distributional approaches. The main claim of their paper is that measures based on WordNet and other lexical resources are inadequate in the evocation task, because most WordNet/lexical resource relation instances are taxonomic in nature. On the other hand, distributional and hybrid models perform well with intuitive evocation associations. It seems that the WordNet structure itself is unfit for the task of evocation recognition.

3 Experiments

This paper will show that it is possible to construct a WordNet-based distance measure which performs better than other knowledge-based features, and no worse than vector space-based measures. We made use of the implementation of Dijkstra's algorithm in the `igraph` library in R (Csardi & Nepusz, 2006).

The experiment design was as follows:

- Firstly, four different versions of the WordNet graph were constructed and the most successful one (achieving the best Pearson's r correlations) was selected. The graphs consisted of three different types of semantic relations: (i) pure WordNet links, (ii) gloss links, and (iii) relations between different senses of the same polysemous lemma (Sec. 3.1). Next, Dijkstra's distance measuring algorithm was applied to the obtained structures in order to obtain the best predictions of evocation strength.

- Secondly, local optima were identified in parameter spaces (the axes represented the costs of edges in the algorithm, Sec. 3.2).

[4]Boyd-Graber et al. (2006) checked simple path measures, as well as the Lesk measure. Finally, the cosine between Latent Semantic Analysis vectors turned out to be the best measure.

- Having found the minimum points, in the third step several similarity measures in the form of functions of Dijkstra's distance: $Sim = f(Dist)$ were evaluated. Both graph structures and similarity functions were compared. Based on several quantitative-qualitative criteria, one measure was chosen (Sec. 3.3).
- Finally, the efficiency of the measure in the evocation recognition task was tested on the validation data set. Different graph topologies were compared together with evocation measures from the literature (Sec. 3.4).

Since the proposed association strength function was strikingly simple (cf. Sec. 3.3), the main emphasis was placed on the optimization of the WordNet structure. The idea was to give it a shape that would facilitate evocation recognition. The optimisation procedure and the final evaluation were run on three independent subsets of the evocation data set:[5]

- S_1 — 2,000 evocation pairs used for checking the efficiency of the WordNet graph and its extensions, and for tuning the weights of edges,
- S_2 — 10,000 sense pairs used to determine the choice of the best similarity measure,
- S_3 — the final testing data set of 108,000 evocation pairs for choosing the best graph topology.

The experiment proceeded in the following manner: (i) four differently structured WordNet graphs were constructed, (ii) various combinations of relation weights and fitted response surfaces were tested on two testing samples (10% of the total number of instances, samples S_1 and S_2).

During the preparatory phase, the efficiency of Dijkstra's algorithm was tested on differently structured WordNet graphs. Evocation strength recognition was performed on the smaller set of 2,000 evocation pairs (S_1). Networks were unweighted. Technically, it was obtained by applying the weight (cost) of 1 to all edge types (cf. Table 1).

Table 1: Edge type combinations tested on unweighted graphs and on the testing set S_1. Symbols: wn — WordNet relations, g — gloss relations, polyWN — the set of all pairs of polysemous lemma senses taken from WordNet, polySC — the set of all pairs of polysemous lemma senses co-occurring in SemCor altogether with polysemy patterns and top level noun and verb synsets, N — number of relation instances, d — directed edges, u — undirected links. Please note that the calculated correlations do not contain the cases of disconnected graph edges (`NA` and `Inf` values were excluded).

graph configuration	r Dist	directionality of links	vector of costs
`wn`	-0.167	(d)	(1)
`wn+g`	-0.189	(d, d)	(1, 1)
`wn+g+polyWN`	-0.184	(d, d, u)	(1, 1, 1)
`wn+g+polySC`	-0.214	(d, d, d/u)	(1, 1, 1)

The four graph structures will be inspected in the forthcoming sections.

[5] All subsets were randomly chosen from the original data set.

3.1 Relation types

The graphs were constructed out of the following types of edges:

- directed WordNet edges (380,000 links in total, symbol `wn`);
- directed gloss relation instances (820,000 links, marked with `g`);
- bidirectional polysemy links between different WordNet senses (400,000 links in total; $\frac{(n-1)n}{2}$ links for each n-sense lemma, symbol `polyWN`);
- a heterogenous set of edges made out of SemCor[6] polysemy links and upper synsets of nominal and verbal WordNet hierarchies (marked jointly as `polySC`), including:
  - directed polysemy links collected from the SemCor corpus; the links were established every time a sense pair appeared in a very similar context;
  - polysemy patterns extracted from the previous set via a generalization from a given polysemy pair to a pair of corresponding semantic domains ('lexicographer files' of the considered synsets);
  - top level noun and verb WordNet synsets linked to their semantic domains via undirected edges to facilitate linking synsets with the top level of polysemy patterns.

It is important to distinguish between two different types of WordNet relations: paradigmatic relations (hyponymy, meronymy, antonymy etc.) and gloss relations, emerging from the WordNet gloss annotation process (cf. Suderman & Ide, 2006).[7] WordNet was first tested without glosses (symbol: `wn`), and was then later tested with the addition of glosses (symbol: `wn+g`), see Table 2 below.

Table 2: Testing graph structures. WordNet relations were given the weight of "1", weights $X1$ and $X2$ were tested in the range $[0, 12]$ in order to find optimal values. Symbols are identical as those used in the previous table. Testing set S_1.

graph configuration	N 10^6	directionality of links	vector of costs
`wn`	0.38	(d)	(1)
`wn+g`	0.82	(d, d)	(1, $X1$)
`wn+g+polyWN`	1.20	(d, d, u)	(1, $X1$, $X2$)
`wn+g+polySC`	0.83	(d, d, d/u)	(1, $X1$, $X2$)

Next, the WordNet graphs were extended further by adding polysemy links. Polysemy is a phenomenon in which the same word-form constitutes a sign for different related meanings (Cruse, 2006, pp. 133–134).[8] Polysemous lemmas are to be understood as those WordNet lemmas which

[6] WordNet sense annotated SemCor corpus (Chklovski & Mihalcea, 2002), v. 3., can be obtained from `https://web.eecs.umich.edu/~mihalcea/downloads.html`.

[7] The data set can be obtained from the WordNet Gloss Project site, v. 1.0, `https://wordnetcode.princeton.edu/glosstag.shtml`.

[8] The most frequent words possess the high level of ambiguity (cf. Zipf's law; Zipf, 1945). For instance, Merriam Webster Pocket Dictionary gives 63 meanings for the word go (Small et al., 1988). In Oxford Lexico (`https://www.lexico.com/definition/go`) the same word received 47 distinct senses: among others 'move from one place to another; travel', 'leave; depart', 'proceed or turn out in a specified way', '(of a machine or device) function' or 'an attempt or trial at something' (a noun sense). For Natural Language Processing this ambiguity is very

possess two or more semantically related senses.[9] Polysemy should be carefully discerned from homonymy, that is from accidental relationships between word meanings (Allen, 2014, p. 150n).[10] In contrast to real polysemy cases, homonymous sense pairs are usually not related at all (Lyons, 1995, p. 59). Establishing links between semantically related WordNet senses is important, because the lexical net does not possess explicit information on the sense relationship.

In our experiments we tested the result of adding combinations of all lemma senses.[11] The `polyWN` set counted 200,000 undirected links. We also took the SemCor and inspected the corpus characteristics of polysemy patterns. It was assumed that those lemma senses and their semantic domains which co-occurred in the very same text were semantically related.[12] The relation was kept directed and forced to go from the preceding sense/semantic domain to the succeeding one. 7,100 sense pairs were collected from SemCor. Below are examples of polysemy:

> **buffer-n-1** `[n:substance]` '(chemistry) an ionic compound that resists changes in its pH'
> → **buffer-v-1** `[v:change]` 'add a buffer (a solution)'
>
> **time-n-2** `[n:time]` 'a period of time considered as a resource under your control and sufficient to accomplish something'
> → **time-n-3** `[n:time]` 'an indefinite period (usually marked by specific attributes or activities)'

Taking into account the semantic domains of all senses (the so called *lexicographer files*), a small directed net of 800 relation instances was obtained. Below we give the subset of the relations for the node `[n:animal]`:

`[n:animal]` ↔ `[n:act]`
`[n:animal]` ↔ `[n:body]`
`[n:animal]` ↔ `[n:food]`
`[n:animal]` ↔ `[n:person]`
`[n:animal]` ↔ `[n:substance]`
`[n:animal]` → `[v:cognition]`
`[n:animal]` → `[v:competition]`
`[n:animal]` ↔ `[v:motion]`

This net reflects the hidden relationships between different semantic categories. We call them polysemy *patterns* and use them to pin up the upper levels of nominal and verbal WordNet

problematic, since in usage it is very difficult to discern all different senses or shades of meaning, and the choice of proper dictionary is very important (Agirre & Edmonds, 2007). Some dictionaries have very coarse-grained sense distinctions, while others possess very fined lists of meanings. WordNet, because of its vast computational applications, is often used as a source of word senses, however, at the same time, it is criticised for its overly detailed sense distinctions (Edmonds, 2004). Clustering WordNet senses seems a reasonable solution to the problem (Agirre & Lopez de Lacalle, 2003). Yet another solution is to seek for the so called *polysemy patterns*, that is regular polysemy types (Vicente & Falkum, 2017).

[9] Please note that we consider meanings belonging to distinct parts of speech as cases of polysemy of *the same* lemma, not the result of word formation. Hence, we treat the unmarked change of word category as *semantic* derivation. In theoretical linguistics the status of the conversion (zero-derivation) sense pairs is not entirely clear (Schmid, 2007). Although many researchers treat zero-derivation as a purely word-formational process, some portray it as mere syntactic change, cf. the description of theoretical positions in Schönefeld (2005, pp. 135–138). For instance, Baker (2003) ascribes a syntactic category (POS) not lexemes but syntactic phrases (pp. 266n). Lexicographers often put related meanings characterised by different parts of speech into the same lexical entry, differentiating it from homonymy cases (Saeed, 2003, p. 80). They are treated as polysemous senses representing the same lemma (Svensen, 2009, pp. 95, 97).

[10] These accidental relationships can be due either to the converging evolution of native vocabulary, or to loans from other languages (Jackson, 2002, pp. 2–3).

[11] Hence also with homonymy.

[12] In such a manner we overcame the challenging homonymy problem.

hierarchies, i.e. those synsets that do not possess any hypernym.[13] We made the latter links[14] undirected in order to enable free movement up and down — from the polysemy patterns net and back again. The net consisted of 3,400 directed links. In comparison to WordNet paradigmatic relations, gloss links and the huge polysemy set derived from WordNet, the size of the *polySC* set is rather modest.

3.2 Setting weights

Each of the four relation types described above was given a weight (the cost of Dijkstra's algorithm). WordNet links were equipped with a weight of "1", all other link types were tested for the optimal values in the range $[0, 12]$. Again, the tests were conducted on the same testing sample of 2,000 evocation pairs (S_1). For the original WordNet graph (wn), the correlation values of Pearson's r and Spearman's ρ measures were calculated. Then, we merged the base graph with glosses (`wn+g`) and checked correlations in a systematic one-factor design (Fig. 1). The response curve $r(g)$ is a polynomial of the fifth degree (adjusted $R^2 = 0.8956$, p-value of the F-test $= 1.219 \times 10^{-9}$). It suggests that finding an adequate optimization solution is a demanding task. The long right tail of the response curve monotonously climbing up to correlations close to $r \sim -0.10$ — proves the importance of using the gloss relation set. We estimated the global minimum location at ($g = 1.4$).

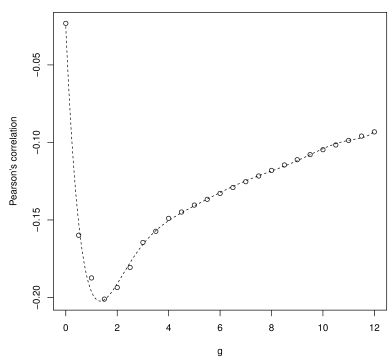

Figure 1: Directed WordNet graph expanded with gloss edges (`wn+g`): Pearson's correlation r between the evocation strength and Dijkstra distance. The global minimum is clearly visible in the range $[1, 2]$.

The two remaining polysemy graphs were extensions of the `wn+g` graph, hence we expected response surfaces of a higher degree. For the 2-factor ($X1 - X2$) problem we conducted a quadruple third order design (CCD Augmented by I-optimal design; cf. Yang, 2008, Tab. 20), because ordinary second order designs were inefficient. The response surfaces of the fourth and fifth order were fitted to the experimental points with acceptable adjusted R-squared values.

Figures 2 and 3 present the response surface for the `wn+g+polyWN` graph. Generally speaking, the plot reproduces the shape of the 1-factor $r(g)$ response curve with a long tail stretching out to high g weights and a valley of minima extending meridionally for the wide range of `polyWN` values. Comparing the `wn+g` minimum and the ravine of the `wn+g+polyWN` graph, one may notice that the optimal area is shifted to higher g values. From the two deepest minima we chose the one deposited in the point $g = 2.5$, $polyWN = 3.2$.

The shape of the SemCor polysemy response surface (Fig. 4 and 5) is a twin to the WordNet polysemy graph, but the localisation of the optimum area is closer to the original wn+g value.

[13] In the case of nouns we took all nouns that did have a hypernym marked with the `[n:Tops]` domain).
[14] That is from synsets to domains and vice versa.

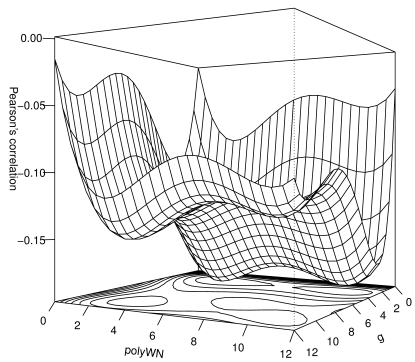

Figure 2: Response surface for the WordNet polysemy graph: the 4^{th} order polynomial, adj. $R^2 = 0.81$, p-value of F-test $= 1.329 \times 10^{-13}$, sample S_1.

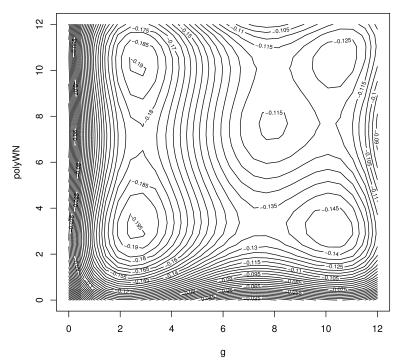

Figure 3: Contour plot for the WordNet polysemy complete graph, the local optimum at (2.5, 3.2) gives the smallest r value.

It is probably caused by different cardinalities of the two polysemy sets. The `polyWN` is a very large matrix consisting of 200,000 bi-directional edges. Its mass of edges linking different parts of graph, especially different parts of speech, greatly affects the length of paths within the graph. The `polySC` set, though 200 times smaller, is much more accurate; it does not contain many homonymy pairs or too distant relationships. The fitted models suggest the better performance of the `polySC` network.

We evaluated how sufficient all these WordNet topologies are on another sample set S_2.

3.3 Similarities

We treated weights as a cost of every step and aimed to establish the optimal set of these parameters. We identified the local optima of the discussed network models using the S_1 set. We will now turn our attention to the evaluation of the efficacy of all these structures. The second testing set, consisting of 10,000 evocation pairs, was used here (the S_2 set). Table 3 collates the data,

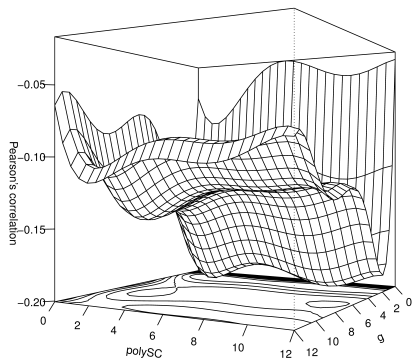

Figure 4: Response surface for the SemCor polysemy graph: 5^{th} order polynomial, adj. $R^2 = 0.93$, p-value of F-test $< 2.2 \times 10^{-16}$, sample S_1.

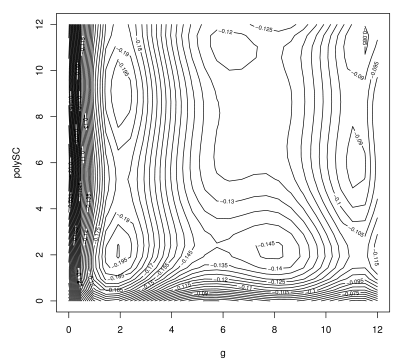

Figure 5: Contour plot for the SemCor polysemy graph, the local optimum is at (1.75, 2).

while Figure 6 presents differences in the correlations between Dijkstra's distance and evocation strength. All three expanded WordNet graphs proved to be better at predicting evocation strength based on the bare Dijkstra's *Dist* measure.

Following Ge & Qiu (2008, p. 382), for this task we employed three measures of semantic similarity, where *Dist* is calculated with Dijkstra's algorithm (Cormen et al., 2001, Sec. 24.3):

$$Sim1 = \frac{1}{Dist + 1}, \tag{1}$$

$$Sim2 = \frac{1}{Dist^2 + 1}, \tag{2}$$

$$Sim3 = \frac{1}{e^{Dist}}, \tag{3}$$

If no path could be established, we ascribed to a sense pair the distance of maximum shortest path length in a graph plus one. To this set of similarity functions we also added a natural, although theoretically not so plausible, measure — the inverse of Dijkstra's distance:

Table 3: Local optima for different expansions of the WordNet graph and different weights (the cost of Dijkstra's algorithm) tested on the set of 10,000 evocation pairs randomly sampled from the whole set (the set S_2). The correlation is given for pairs of Dijkstra's distances and evocation strength values.

relation types	vector of costs	r Dist	ρ Dist
wn	(1)	-0.128	-0.147
wn+g	(1,1.4)	-0.197	-0.191
wn+g+polyWN	(1,2.5,3.2)	-0.190	-0.184
wn+g+polySC	(1,1.75,2)	-0.204	-0.198

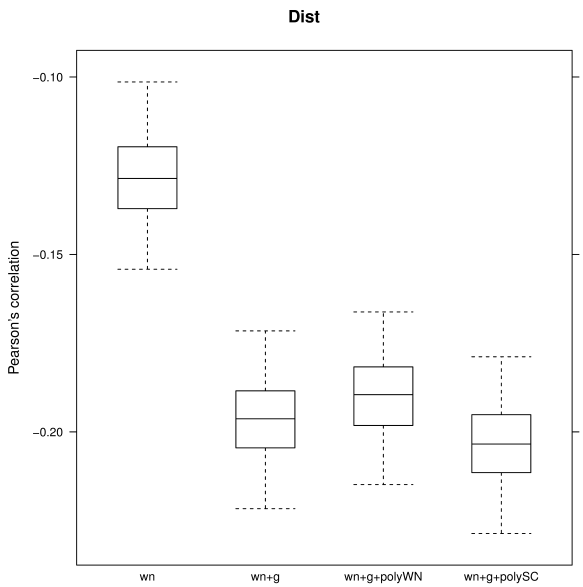

Figure 6: 95% bootstrap confidence intervals (B=1,000 iterations) of the Dijkstra's *Dist* measure for four inspected graph structures and optimized weights values.

$$Sim0 = \begin{cases} \frac{1}{Dist} & Dist > 1 \\ 1 & Dist \leq 1 \end{cases} \tag{4}$$

After the transformation of the *Dist* measure into the measures of similarity *Sim0*–*Sim4*, we obtained higher correlation ratios. Table 4 presents the values of Pearson's correlation for all *Sim* functions and the S_2 data set. For all measures and graphs, we also prepared bootstrap samples to assess their distributional properties (the number of iterations $B = 1000$). We checked standard deviations and calculated mean deviations as well as ranges.[15] *Sim2* gives the highest correlations, while Sim1 the lowest. *Sim2* and *Sim3* measures are characterised by relatively large variances, while *Sim0* and especially Sim1 functions preserve smaller variances. The graph of Cullen and

[15]We define *range* as a difference between a maximum and a minimum value.

Frey proves that all measures gave values close to the normal distribution, with *Sim2* and *Sim0* being the closest to the point (0,3).

Table 4: Local optima for different expansions of the WordNet graph and different weights (the cost of Dijkstra's algorithm) tested on the set of 10,000 evocation pairs randomly sampled from the whole set (the set S_2). The correlation is given for pairs of similarity and evocation scores. The symbol "sd" stands for standard deviation.

graph	r			
configuration	Sim0	Sim1	Sim2	Sim3
wn	0.213	0.199	0.229	0.223
wn+g	0.251	0.241	0.259	0.255
wn+g+polyWN	0.246	0.236	0.251	0.235
wn+g+polySC	0.258	0.249	0.265	0.254
mean r	0.242	0.231	0.251	0.242
mean sd	0.025	0.021	0.030	0.034
range	0.045	0.050	0.036	0.031

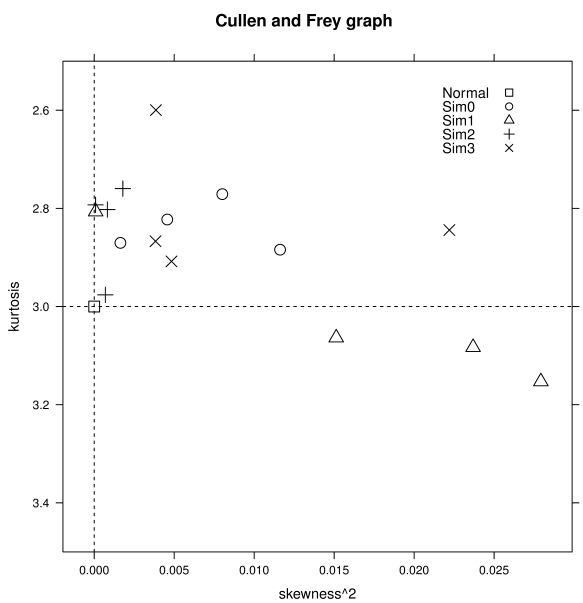

Figure 7: The Cullen and Frey graph for bootstrap distributions of all similarity measures.

For further experiments we have chosen the *Sim0* measure, taking into account the qualitative-quantitative criteria mentioned above. This measure gave high correlation values and was characterised by small variance. It also ensured a relatively large range; its distribution was very close to the normal distribution. Obviously, this decision is not fully objective and is a matter of a researcher's choice.

3.4 Similarity measure efficiency

The optimizing procedure conducted on the S_1 set (2,000 evocation pairs) led to the choice of the most promising parameter values for gloss edges (`g`) and polysemy links (`polyWN` and `polySC`). Correlations for Dijkstra's distances and evocation strength improved after the optimalisation of weights on the gloss-enlarged graph and still remained high after adding new polysemy links. The set (10 pairs) was used to support the choice of the best similarity measure. The final evaluation was performed on a large set of the remaining 90% of evocation pairs (S_3) with the use of a compromise measure *Sim0*. Despite a sharp disproportion in size between the testing sets and the validation set (1 : 10 ratio), final experiments conducted on 107,000 evocation examples confirmed the following findings:

- The optimizing procedure gave *Sim0* the highest r values within the group of Hayashi's knowledge-based individual measures (Fig. 8, the group "K").
- The simple similarity measure performed comparable to Hayashi's best distributional individual measures (Fig. 8, groups "KV" and "DV").
- The most optimal WordNet configuration included glosses and SemCor polysemy links (`wn+g+polySC`).

We performed a paired bootstrap test on six pairs of graphs at the 95% confidence level, using the Bonferroni correction (the significance level of each paired test was adjusted to the $\frac{5}{6}$% level). Due to the large size of the $S3$ data set, most tests led to significant differences. It transpired that the *wn+g+polySC* network model is the most suitable for evocation recognition (the mean correlation of the *Sim0* measure is 0.251). Sole WordNet relations (the graph `wn`) were much worse than any of the extended structures (for details see Table 5).

Table 5: One-sided paired bootstrap percentile test for the difference between r values for *Sim0* at the 95% confidence level. Symbols: "=" marks insignificant differences, while ">" and "<" designate differences which are significant, wherein the former symbol is read as 'a row is greater than a column', and the latter has the opposite meaning. The direction of the test was established on the basis of the previous experiment (on S_1).

graph configuration	wn	wn+g	wn+g+ polyWN	r Sim0
wn	=			.209
wn+g	>	=		.245
wn+g+polyWN	>	=	=	.246
wn+g+polySC	>	>	>	.251

4 Conclusions

The main contribution of this paper is the presentation of a novel method of optimizing lexical net structure for the needs of evocation recognition. We started with a bare WordNet graph and expanded it with glosses and polysemy links. In order to achieve better agreement with gold standard evocation strength values, the graph relations were differentiated through different weights (the cost of each edge) and optimized on a very small subset of the original data set. The optimization process led to new network structures which gave distances better correlated with

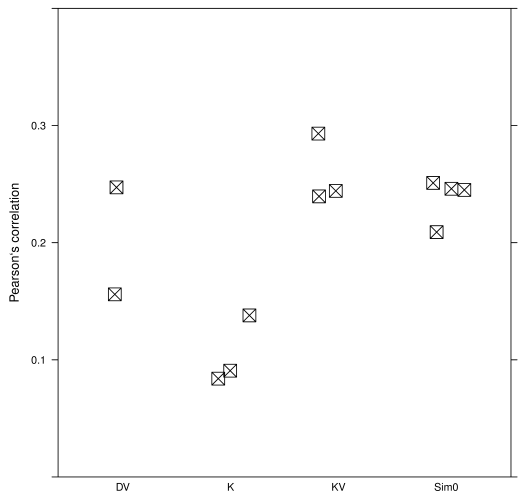

Figure 8: Efficiency of several evocation measures taken from Hayashi (2016). With "Sim0" we mark our similarity measure (4) run on different WordNet structures. Symbols: DV — corpus-based distributional vectors (*ldaSim*, *w2vSim*), K — knowledge-based measures (*wupSim*, *lexNW*, *dirRel*), KV — WordNet-based vector spaces (*posSim*, *autoexSim*, *relVec*).

the evocation strength. The WordNet structure most optimal for the task was extended with gloss relations and a small set of polysemy patterns and instances derived from the SemCor corpus.[16] This fact may be evidence for the importance of polysemy links in the lexico-semantic system.

We also proposed an evocation measure in the form of inverse Dijkstra's distance which performed well not only when compared with other graph measures, but also in comparison with sophisticated vector space models, of which some were highly-dimensioned ($100 - 300$D vectors). Contrary to the well-established opinion of the superiority of vector representations, it is probable that the common denominator of all the successful individual measures is adequate distance measuring methodology. Our simple knowledge-based measure has recently been reused — together with some other credible individual measures — in a neural-network framework, with the overall NN efficiency $r = 0.4415$, and has turned out to be the best feature of all those used (Janz & Maziarz, in press).

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on — NAACL '09*, 19. Association for Computational Linguistics. `https://doi.org/10.3115/1620754.1620758`

Agirre, E., & Edmonds, P. (2007). *Word sense disambiguation: Algorithms and applications*. Springer. `https://doi.org/10.1007/1-4020-4809-2`

Agirre, E., & Lopez de Lacalle, O. (2003). Clustering WordNet word senses. In N. Nicolov, K. Bontcheva, G. Angelova, & R. Mitkov (Eds.), *Recent Advances in Natural Language Processing III: Selected papers from RANLP 2003* (pp. 121–130). `https://doi.org/10.1075/cilt.260.13agi`

[16] We publish the whole data set on Git-Hub repository `https://github.com/MarekMaziarz/SemCor-polysemy-patterns` on open license.

Allen, K. (2014). *Linguistic meaning*. Routledge.

Baker, M. C. (2003). *Lexical categories: Verbs, nouns and adjectives*. Cambridge University Press. https://doi.org/10.1017/CBO9780511615047

Ballatore, A., Bertolotto, M., & Wilson, D. C. (2014). An evaluative baseline for geo-semantic relatedness and similarity. *GeoInformatica*, *18*, 747–767. https://doi.org/10.1007/s10707-013-0197-8

Boyd-Graber, J., Fellbaum, C., Osherson, D., & Schapire, R. (2006). Adding dense, weighted, connections to WordNet. In *Proceedings of the Global WordNet Conference*. http://umiacs.umd.edu/~jbg/docs/jbg-jeju.pdf

Cattle, A., & Ma, X. (2017). Predicting word association strengths. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1283–1288). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1132

Chklovski, T., & Mihalcea, R. (2002). Building a sense tagged corpus with open mind word expert. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: Recent successes and future directions* (Vol. 8, pp. 116–122). Association for Computational Linguistics. https://doi.org/10.3115/1118675.1118692

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). *Introduction to algorithms*. MIT Press.

Cramer, I. (2008). How well do semantic relatedness measures perform? A meta-study. In *Proceedings of the 2008 Conference on Semantics in Text Processing* (pp. 59–70). Association for Computational Linguistics. https://doi.org/10.3115/1626481.1626487

Cruse, A. (2006). *Glossary of semantics and pragmatics*. Edinburgh University Press.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal: Complex Systems*, Article 1695. http://igraph.org

Edmonds, P. (2004). Lexical disambiguation. In *Elsevier encyclopedia of language & linguistics* (pp. 43–62). Elsevier.

Faruqui, M., Tsvetkov, Y., Rastogi, P., & Dyer, C. (2016). Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP* (pp. 30–35). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2506

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. MIT Press. https://doi.org/10.7551/mitpress/7287.001.0001

Ge, J., & Qiu, Y. (2008). Concept similarity matching based on semantic distance. In *2008 Fourth International Conference on Semantics, Knowledge and Grid* (pp. 380–383). IEEE. https://doi.org/10.1109/SKG.2008.24

Hayashi, Y. (2016). Predicting the evocation relation between lexicalized concepts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 1657–1668). Association for Computational Linguistics.

Jackson, H. (2002). *Lexicography: An introduction*. Routledge.

Janz, A., & Maziarz, M. (in press). Chaining polysemous senses for evocation recognition. In *Proceedings of the 12th International Conference on Computational Collective Intelligence*. Springer.

Kacmajor, M., & Kelleher, J. D. (2019). Capturing and measuring thematic relatedness. *Language Resources and Evaluation*, *54*(3), 645–682. https://doi.org/10.1007/s10579-019-09452-w

Lyons, J. (1977). *Semantics* (Vol. 1). Cambridge University Press. https://doi.org/10.1017/CBO9781139165693

Lyons, J. (1995). *Linguistic semantics: An introduction*. Cambridge University Press. https://doi.org/10.1017/CBO9780511810213

Ma, X. (2013). Evocation: Analyzing and propagating a semantic link based on free word association. *Language Resources and Evaluation*, *47*(3), 819–837. https://doi.org/10.1007/s10579-013-9219-2

Miller, G. A., & Fellbaum, C. (2007). WordNet then and now. *Language Resources and Evaluation*, *41*(2), 209–214. https://doi.org/10.1007/s10579-007-9044-6

Nikolova, S. S., Boyd-Graber, J., Fellbaum, C., & Cook, P. (2009). Better vocabularies for assistive communication aids: Connecting terms using semantic networks and untrained annotators. *In ACM Conference on Computers and Accessibility*. ACM Press. https://doi.org/10.1145/1639642.1639673

Saeed, J. (2003). *Semantics* (2nd ed.). Blackwell Publishing.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing: CICLing 2002* (pp. 1–15). Springer. https://doi.org/10.1007/3-540-45715-1_1

Schmid, H.-J. (2007). Laurie Bauer and Salvador Valera (eds.), Approaches to conversion/zero-derivation. Münster, New York, Munich, and Berlin: Waxmann, 2005. 175 pp., £19.90 (pb.), ISBN 3-8309-1456-3 [Review]. *English Language & Linguistics*, *11*(3), 587–590. https://doi.org/10.1017/S1360674307002407

Schönefeld, D. (2005). Zero-derivation-functional change-metonymy. In L. Bauer & S. Valera (Eds.), *Approaches to conversion/zero-derivation* (pp. 131–160). Waxmann.

Small, S. L., Cottrell, G. W., & Tanenhaus, M. K. (1988). Preface. In S. L. Small, G. W. Cottrell, & M. K. Tanenhaus (Eds.), *Lexical ambiguity resolution: Perspectives from psycholinguistics, neuropsychology, and artificial intelligence*. Morgan Kaufmann. https://doi.org/10.1016/B978-0-08-051013-2.50004-5

Suderman, K., & Ide, N. (2006). Layering and merging linguistic annotations. In *Proceedings of the 5th Workshop on NLP and XML (NLPXML-2006): Multi-Dimensional Markup in Natural Language Processing* (pp. 89–92). https://doi.org/10.3115/1621034.1621052

Svensen, B. (2009). *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge University Press.

Vicente, A., & Falkum, I. L. (2017). Polysemy. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.325

Yang, Y. (2008). *Multiple criteria third-order response surface design and comparison*. https://www.researchgate.net/publication/254671895_Multiple_Criteria_Third-Order_Response_Surface_Design_and_Comparison

Zipf, G. K. (1945). The meaning-frequency relationship of words. *The Journal of General Psychology*, *33*(2), 251–256. https://doi.org/10.1080/00221309.1945.10544509

This research was financed by the National Science Centre, Poland, grant number 2018/29/B/HS2/02919, and supported by the CLARIN-PL[17] research infrastructure.

The authors declare that they have no competing interests.

Both the authors participated equally in preparing conception and academic editing of this article.

Publisher: Institute of Slavic Studies, Polish Academy of Sciences

Publishing History: Received 2020-07-02; Accepted 2020-09-14; Published 2020-12-23.

[17] http://clarin-pl.eu