УДК 811.111'37'374-115:004

EVALUATING SEMANTIC SIMILARITY MEASURES FOR ENGLISH VERBS IN WORDNET AND THESAURI

Iryna Dilay, Mykhailo Bilynskyi

Ivan Franko National University of Lviv,
1, Universytetska St., Lviv, 79000, Ukraine irynadilay@gmail.com, bislo@ukrpost.ua

Semantic similarity measures the distance between concepts and is based on their likeness. WordNet-based similarity metrics summarized by Pedersen can serve to compare both the distances between separate concepts and the metrics themselves. Establishing and comparing WordNet-based verb similarity can be applicable for a number of NLP tasks. However, the measures, being predominantly of a non-linear character, fail to account for synonymy of words, as well as convey the principles of mental lexicon structuring, in particular the assymetry of associations. The combination of concept-based and word-based similarity measures can be leveraged to solve this problem. Attention will be paid to the applicability of the synonymous distance metric to the reversal of multiple (over twenty) thesauri of English verbs and the study of versatile issues of the geometry of semantic spaces based on vast numbers of semantic proximity values in the (near-)synonymy of verbs.

Keywords: semantic similarity measure, asymmetry of associations, taxonomic relations, synset, corpus, WordNet, thesauri of English verbs, reverse synonymous strings of verbs, uneven distance-from-the-dominant scales, (non-)euclidean geometry, vector analysis, angular geometry in a thesaurus.

Introduction

Measures of semantic similarity and relatedness between concepts are widely used in Natural Language Processing. Their applications are versatile, starting with lexicography, translation, spelling correction, information retrieval (IR), document retrieval (DR), language teaching and ending up with plagiarism detection and ontologies comparison.

The majority of the metrics are WordNet-based, though, the new ones also appeared, such as corpus-based LSA, Wikipedia-based WikiRelate and ESA. A new trend is to integrate the measures in order to enhance their performance. The measures are typically compared with the gold standard, i.e., human judgments obtained through associative experiments. The most attested reported experiments are scaling 30 noun pairs by Miller &Charles (1991), 65 word pairs by Rubenstein et al. (1965), and the 353 Test Collection by Finkelstein et al. (2002). However, they mostly deal with the semantic similarity of nouns, assuming that nouns give the best representation of concepts. The standards for verbs, as well as the evaluation of verb similarity measures, to the best of our knowledge, have not

© I. Dilay, M. Bilynskyi, 2016

received enough attention in literature (see the discussion of it in Yang and Powers (2006)).

Thus**, the objective** of this study is to assess the semantic similarity measures of English verbs exemplified by the verbs belonging to the same semantic class – verbs of cognition. Also we will enumerate the applications of the positional string metric for the reversal of a number of dictionaries of synonyms.

Previous research

Semantic similarity presupposes only synonymic and homonymic relations that build taxonomies. All other relations, such as meronymy, are treated as semantic relatedness and, as a rule, accompany the study of semantic similarity. Semantic relatedness, thus, incorporates semantic similarity and is a broader term. The majority of the measures focus on semantic similarity. Normally, the measures are ontology-based, particularly WordNet-based, and fall into several groups: path finding measures, information content measures, context vector, extended gloss overlap and their enhancements.

Path length counts the edges between concepts. The relatedness score is inversely proportional to the number of nodes along the shortest path between the synsets. The shortest possible path occurs when the two synsets are the same, in which case the length is 1. Thus, the maximum relatedness value is 1. It is simple but requires a rich hierarchy and only uses 'is-a' relation. Another pass finding metric by Wu and Palmer (1994) calculates the path length to the least common subsumer and is scaled by a subsumer's path to root. The formula of wup is score = 2*depth(lcs) / (depth(s1) + depth(s2)) . This means that 0 < score <= 1. The score can never be zero because the depth of the LCS is never zero (the depth of the root of a taxonomy is one). Leacock and Chodorow (lch) is - log (length / (2 * D)), where length is the length of the shortest path between the two synsets (using node-counting) and D is the maximum depth of the taxonomy (Leacock & Chodorow, 1998). It takes into account the depth of the taxonomy as well as the presence or absence of a unique root node. Besides, Hirst and St-Onge (hso) works by finding lexical chains linking the two word senses (Hirst & St-Onge, 1998). There are three classes of relations that are considered: extra-strong, strong, and medium-strong. The maximum relatedness score is 16 and the measure is highly WordNet specific.

Resnik (1998) calculates the Info Content (IC) of the least common subsume (LCS). Unlike the path edge-based measures, the approach takes into account corpora. T*he* value will be greater than 0 and will depend on the size of the corpus. It combines taxonomic lexical knowledge (WordNet) with probabilistic models. Jiang and Conrath (jcn) is an extension of Resnik (Jiang & Conrath, 1997). It scales LCS by IC of the concepts whereby accounting for IC of individual concepts. The relatedness value returned by the jcn measure is equal to 1 / jcn_distance, where jcn_distance is equal to IC (synset1) IC(synset2) - 2 * IC(lcs). Lin is one more extension of Resnik IC. The relatedness value returned by the lin measure is a number equal to 2 * IC (lcs) / (IC(synset1) + IC(synset2)) where IC(x) is the information content of x. One can observe, then, that the relatedness value will be greater-than or equal-to zero and less-than or equal-to one.

Context vector (CV) measure by Patwardhan and Pedersen (2003) is derived from the co-occurrence statistics of corpora. It is computationally intensive, adapts the word sense disambiguation method and LSI and builds co-occurrence vectors that represent the contextual profile of concepts. The cosine of the angles determines relatedness.

The extended gloss overlap measure (lesk) finds overlaps in the glosses of the two synsets (Banerjee, 2003). The extended gloss overlap measure (lesk). The relatedness score is the sum of the squares of the overlap lengths. For example, a single word overlap results in a score of 1. Two single words overlap results in a score of 2. A two word overlap (i.e., two consecutive words) results in a score of 4. A three word overlap results in a score of 9.

More recently, cross-ontology measures, enhancements and hybrid measures have been suggested. The applications of the measures have been widely discussed in the literature below. Yang, Bhavsar and Boley (2008) provide an overview of the existing algorithms and propose their own approach to semantic matching that offers a better concept granularity measure. The basic assumption is that the concepts at upper layers have more general semantics, while the concepts at the lower layers have more concrete semantics and higher similarity, consequently granularity is measured based on upward and downward path lengths. One of the major conclusions drawn by the authors is that "the human judgment of information sources was demonstrated to be a non-linear process toward their similarity" (Yang, Bhavsar, Booley, 2008, p. 10) and is affected by the granularity of the taxonomy. Thus, the shortest path length and concept depth appear to be the most efficient structural characteristics.

Wang and Hirst (2011) reinvestigated the rationale for and the effectiveness of adopting the notions of depth and density in WordNet-based semantic similarity measures. The human judgment of lexical semantic similarity imposes limitations on relying on these notions. New definitions of depth and density have been offered by the authors, which yield better performance of the two features. The authors start with the critical examination of the existing approaches to semantic similarity measures, namely, the distributional ones (based on text corpora) and lexical resource-based ones (here exemplified by using WordNet and within lexical groups theory suggested also for an arbitrary thesaurus and/or all the availably thesauri). Acknowledging the high cost of WordNet compilation and its benefit for the study, the authors mention, in particular, the off-the-shelf toolkit of Pederson et al. (2004) which allows comparing the existing algorithms. Density is defined as the number of edges leaving the concept or its parent node(s), (Wang & Hirst, 2011, p. 1004]. The basic assumption is that everything else being equal, two nodes are semantically closer if (a) they reside deeper in the WordNet hierarchy, or (b) they are more densely connected locally. The new measures were compared with human judgments and showed significant improvement in performance.

It was much earlier when G. Hirst and D. St-Onge discovered a method of detecting malapropisms which heavily relies on the semantic relatedness of the words in a chain (Hirst & St- Onge, 1998). Though the initial performance of their approach is not very high, they have managed to thoroughly discuss the notion of semantic relatedness in their work. A lexical chain is viewed as a cohesive chain in which one word is in a cohesive relation of some kind to the other word. Such chains are claimed to provide enough context to resolve lexical ambiguity and, consequently, detect malapropisms. The initial idea was to apply Roget's thesaurus as a source of lexical chains. However, due to the fact that at that time it was not available electronically and imposed some other limitations, the authors adapted their method to WordNet. Their definition of semantic relatedness centers upon the notion of the synset. There have been defined three kinds of relations going upward, downward and horizontally: extra-strong having the highest weight (hold between a word and its literal repetition), strong (also see, antonymy, attribute, cause, entailment,

holonymy, hyperonymy, hyponymy, meronymy, pertinence and similarity), medium-strong (when there is allowable path connecting two synsets). A path is a sequence between two and five links between synsets and is allowable if it corresponds to one of the possible patterns. The longer the path and the more changes of direction, the lower the weight. On the path each lexical relation is taken into consideration: an upward direction corresponds to generalization, a downward relation corresponds to specification. Horizontal links are less typical due to the synset structure of WordNet (Fellbaum, 1998). The rules for the allowable path have been set: no other direction may precede an upward link; at the most one change of direction is allowed. The crucial observation the authors make is that WordNet, unlike Roget's, is restricted to formal relations rather than connections by general association and has a varying conceptual density.

There were the authors who did not rely exclusively on WordNet. Stube (2007) argues that Wikipedia provides a better knowledge base than a search engine and has more coverage (proper names, etc.) than WordNet. However, the best performance can be achieved when integrating Google, Wiki and WordNet. The major problem is to find efficient mining methods. Wikipedia is big in size, characterized by variable depth, branching, multiple category relations including meronymy, and rather than being a well-structured taxonomy, it is a category tree of folksonomy, a collaborative tagging system. It imposes its limitations, since folksonomy does not strive for correct conceptualization. The main idea was to apply similarity measures developed for WordNet to Wikipedia and account for their efficacy. The authors used edge-counts, info content, and text overlap based measures. Then, the relatedness measures were evaluated on three standard datasets which are the results of human judgments tests: a) Miller and Charles (1991) – the list of 30 noun pairs; b) Rubenstein and Goodenough (1965) – the list of 65 noun pairs; c) WordSimilarity – 353-Test Collection by Finkelstein et al. (2002). The performance between the relatedness measures and human judgments was evaluated by taking the Pearson product- moment correlation. While WordNet performs better on small datasets (a) and (b), its performance decreases with the 353-Test Collection.

Pedersen et al. (2007) focused on domain independent measures and attempted to adapt them to a biomedical domain. On applying different metrics, the authors came to the conclusion that the highest correlation is achieved by the Context vector measure based on cosine similarity. They advocated the use of corpora to enhance the performance and proved that the metrics can be applied to specific domains.

Discussion

The works on semantic similarity have grown exponentially and can hardly be exhaustively summarized and compared. The underlying enhancement idea is that WordNet-based similarity measures can be improved with corpus data, other ontologies (Wikipedia, other thesauri, medical taxonomies inter alia), as well as search engines. There is also a need for the human judgments standards, which can be used to evaluate the performance of the automatic techniques. They should not be limited to nouns exclusively.

Having analyzed the most widely used and attested measures of semantic relatedness and similarity, the shortcomings of the purely computational methodology can be observed. First, according to J. Aitchison, "Different words require different levels of activation in order to be born: very frequently used words require relatively little to trigger them, while uncommon words are harder to arouse". Hence, frequency affects the semantic distance.

Leveraging corpus frequencies data can help solve this problem. Second, the same applies to abstract and specific vocabulary. Some metrics acknowledge this problem and come up with feasible solutions. The most important reservation, however, is that "we cannot assume that going upstairs uses identical muscles to going down but in a reverse sequence" (Aitchison, 2012, p. 207). We cannot assume

that the distance from w_i to w_j is the same as the distance from w_j to w_i. A vast majority of measures fail to reflect the asymmetric structure of the mental lexicon relations.

The asymmetry of the mental lexicon relations is not reflected in the WordNet as a conceptual thesaurus linking concepts rather than words. It is the processing of form, looking for the right word, which requires more effort than searching for the right concept (Aitchison, 2012). Thus, the important implication is that the distance between the concepts is more measurable as they fill in a certain template in the human mind, whereas the distance between word forms that embody the concepts is less tangible, oftentimes asymmetric and idiosyncratic. Speakers may be lost for words but not for the meaning they want to convey. Nonetheless, it is improper to claim that WordNet is insufficient as a resource or the measures are faulty. It focuses on the conceptual similarity rather than word similarity. Partially, it is one of the reasons why the measures cannot reach the gold standard. It is not clear what humans take as a starting point of their similarity judgment: a form (word) or a concept (meaning).

Based on the previous research in disparate fields (NLP, lexical semantics, corpus linguistics, mental lexicon, and psycholinguistics) and the reservations mentioned above, the possible extension and improvement of the similarity metrics can be achieved through the integration of word-based and concept-based metrics. To this end, the verbs (we limited ourselves here to the study to the cognitive verbs) have been drawn from WordNet. The total number of the cognitive verbs synsets found in WordNet is 376. On analyzing verb similarity in the WordNet, Yang and Powers (2006) observed that verb hierarchy in WordNet is sparse and somewhat limited and drew a conclusion that in comparison with nouns verb hierarchy exists in a very shallow way in human. The verb similarity matrices have been extracted for the principle measures from WordNet::Similarity Web Interface created by Pedersen (2015). It has been used as a user-friendly comprehensive resource to calculate the semantic similarity of the verbs in WordNet. This package consists of Perl modules along with supporting Perl programs that implement the semantic relatedness measures described by Leacock & Chodorow (1998), Jiang & Conrath (1997), Resnik (1998), Lin (1998), Hirst & St-Onge (1998), Wu & Palmer (1994), the extended gloss overlap measure by Banerjee and Pedersen (2003), and two measures based on context vectors by Patwardhan (2003). The same results can be obtained from WS4J Demo (WordNet Similarity for Java at http://ws4jdemo.appspot.com/) maintained by Hideki Shima.

Based on the literature reviewed, it was observed that the best measures for verbs can be Hirst and St-Onge, as it can calculate all parts of speech, Context Vector (or Gloss Vector) by Patwardhan and Pedersen which is derived from the co-occurrence statistics in corpora and extended Gloss Overlap. On running the verbs through the metrics, it was noticed that only the Hirst & St-Onge measure can yield different results depending on the direction of similarity, *i. e.*, from $v1$ to $v2$ and from $v2$ to $v1$. For example, the similarity score for *notice* vs. *know* is 4, whereas for *know* vs. *notice* is 0. No other WordNet::Similarity measure

reflected this asymmetry. This observation called for a more thorough analysis. The peculiarity of hso lies in the fact that they use a recursive algorithm and impose limitations on the allowable paths to be measured. Certain path configurations do not work. As a result, the path in the opposite direction can be different from the initial one. It is in accord with the aforementioned psycholinguistic principles. However, hso is still a non-linear measure that cannot account for the linear distances between synonymous words.

The next step was to use dictionaries and thesauri in order to build similarity matrices for the verbal forms based on their linear similarity. A simple formula was applied: $w = \frac{(n+1)-r}{n}$ (Levickij, 1989), where *w* stands for the weight of the component, *n* for the number of components, and *r* for the rank of the component in the definition or a synonymic string. Weights are below 1 and reflect the distance of the component as well as its presence. The matrices reflecting the asymmetry of associations have been built (See Table 1).

Table 1

A fragment of a dictionary/thesaurus-based know matrix

components, Synonyms / verbs	know	understand	see	perceive	comprehend	apprehend	recognize	catch	discern	grasp	appreciate	tell	fathom
1. know	+	+	+	+	+	+	+	+	+	+	+	+	+
2. understand	+	+	+	+	+	+	+		+	+	+	+	+
3. see	+	+	+	+	+	+	+		+	+	+	+	+
4. perceive	+	+	+	+	+	+	+		+	+		+	
5. comprehend	+	+	+	+	+	+		+	+	+	+		+
6. apprehend	+	+	+	+	+	+	+	+	+	+			
7. recognize	+	+	+	+	+	+	+				+	+	
8. catch	+	+	+	+	+	+	+	+	+	+			
9. discern	+	+	+	+	+		+		+				
10. grasp		+	+	+	+	+	+	+	+	+			
11. appreciate	+	+		+	+	+	+			+	+		
12. fathom		+	+		+					+	+		+

Finally, the fundamental question needs to be answered. How far do the computed findings agree with human judgments? To evaluate the efficiency of the above metrics, the experiment has been set. Human labelers, who were English native speakers, have been asked to assign a similarity score from 0 to 10 for the suggested verb pairs. In a way, the experiment replicated the 353-Test Collection by Finkelstein et al. (2002), though the 353-Test was developed for nouns without taking into consideration asymmetry. Our experiment included 30 verb pairs, in fact, 15 pairs, where *v*1 is a stimulus and *v*2 is a reaction, and 15 verb pairs where *v*2 is a stimulus and *v*1 is a reaction. The pairs have been retrieved from the matrices of the verbs reporting asymmetry. The objective of the experiment was to find out whether the labelers can capture asymmetry and, of course, reach agreement on assigning semantic similarity scores. This can be used as a benchmark to draw a comparison with concept based and word form based similarity measures.

The mean value for the fourteen labelers was calculated, as well as standard deviation. The mean standard deviation was 1.8 on a scale of 10. The labelers did notice the asymmetry of similarity direction and assigned the respective values. The agreement is the highest with

the most 'similar' or the least 'similar' verbs, whereas the verbs in-between show diverse results. The labelers more or less agreed on the pairs *know* vs. *notice* (the mean of 5.75, StDev of 0.8), *notice* vs. *know* (mean = 5.0; StDev = 1.0), *know* vs. *remember* (mean = 8.5; StDev = 1.0), *remember* vs. *know* (mean = 8.0; StDev = 1.7), *think* vs. *know* (mean = 4.75; StDev = 0.4), *realize* vs. *know* (mean = 6.25; StDev = 0.4), *think* vs. *perceive* (mean = 8.25; StDev = 0.8), and *perceive* vs. *think* (mean = 7.5; StDev = 1.1). Surprisingly, for these verbs they agreed on the scores in the reversed verb pairs too. This observation requires further study.

The comparison of the hso values with human judgments revealed little agreement. Hso was normalized from the scale of 16 to the scale of 10 to make a compatible scale. The average deviation from the human mean was 3.58. In general, the scores in hso were considerably lower than in the benchmark, though the general asymmetry tendency was retained. When compared against human judgments, other measures, which do not reflect asymmetry, produced varied results. Lesk, res, vector_paired, lch and jcn had to be normalized to the scale with max = 10. Lin did not consider the selected items similar, similarity = 0. Jcn produced a very low similarity result, almost negligible. A little bit better, but still low similarity score was observed with lesk. Out of the remaining measures, the deviation from the gold standard was the lowest with wup (- 0.29) and vector_paired (1.23). The largest deviation was observed with hso (3.58), though it took into account verb asymmetry and nearsynonyms. See the results in Table 2. The highlighted lines in Table 2 reveal the opposite direction of associations.

The compared dictionaries and thesauri showed the deviation of 2.81 from the benchmark. Still the choice of the dictionaries and their focus on synonyms only is crucial. Some thesauri do not consider the analyzed items synonymous at all.

The integration of WordNet-based and dictionary-based measures proves to be promising. The mean between wup and dictionary -based scores was calculated. The deviation from the human judgment is significantly smaller: 1.02 as compared to 1.8 (general WordNet-based mean). For some pairs it was 0: *know – interpret*: wup = 4, w=5, compared against human = 4.5. More research should be done to validate these findings.

The Levitsky formalism mentioned above was being applied to the study of limited areas of vocabulary throughout the 1980s. It is still applicable in the same manual procedure in the course of reconstructing extents of relatedness between the words in the strings belonging to identifiable lexical- semantic groups. It was it the only known metric (cf. those suggested by Czekanowski, perhaps the most suitable, Jaccard and Levin, not to be mixed with Levine) fit for the study of similarity quantification in use in earlier decades. As it is based on the number of synonyms in the string and the ordinal position of a constituent in the series its applicability increases in the thesauri that opt for the non-alphabetical placement of synonyms after string headword (dominant). In most of the processed thesauri this is the principle. The LUMPE: Lviv University Metric Project of English begun in 1995 was triggered by the usability of personal computers The hardware gave an impetus to the development of usable software possibilities and most importantly at-hand facilities of reliable data storage and parameters resettable data processing. The factor that thesauri at the time were unavailable electronically, and most of them still are now, taken as a hindrance in some of the NLP synonymy frameworks (see above), was not an issue as query-sensitive digitalization of such paper objects had to be developed in order to make the metric runnable and diversely applicable for NLP tasks.

Owing to the application of the said formalism to each of the strings of the thesaurus the latter is getting turned into a metric object. The metric weight scale being evenly distributed between the constituents of the string, the step of distancing applicable to each subsequent constituent from the dominant equals the weight factor value of the last constituent. As the density of the string rests on the number of (near-)synonyms involved, the grading thresholds in the proximity would be responsible for numerically divergent inventories of variant means of signification. The reading of the thesaurus at its metric processing is taken to be direct as the left hand-side, or headword, i.e. its signified element, is juxtaposed with the right hand-side signifying elements within the string. The string is then reflected in an evenly descending metric. It is set with no possibility for equally distanced elements or gaps in the filling of the string. Most of thesauri ascribe separate sets of synonyms to multiple meanings of a polysemous word. The list of dominants is superseded by the complete list of lexical items, hence a proportion of elements from the right hand-side part of the thesaurus are not ascribed explicit synonymous stringing.

So as to overcome these counter-intuitive limitations of thesauri their reversal was suggested to develop a gradually extended multi-step framework which was being refined for a number of years. With this purpose all the verbs from the right hand-side of the thesaurus are shifted to the left hand- side (string dominants) position. The dominants of the respective strings within which they occur in the direct thesaurus with the corresponding proximity (which is the inverse of distance) measures would constitute the composition of the reverse string. The consequences of such a procedure imply that reverse strings modeled on a given thesaurus are characterized by an (un-)even distancing of the constituents from the dominant. Also, they admit of identically positioned, i.e. equally distanced, synonyms within the string. The string is not strictly linear then, or same-distance constituents could be placed at certain values along its descending linear contour. In contrast with the direct thesaurus, in the reverse thesaurus stretches of proximity can remain unfilled with any synonyms at all.

The dominant of the string from the direct thesaurus is ascribed the maximum (1.0) proximity value. In cases of polysemy the reversal brings about a cluster of strings with the repeated maximum values, although a dismantling of such clusters into the primary, secondary and subsequent reverse epidigmatic sub-thesauri is doable. The verbs that initiate (a) string(s) in the direct thesaurus, but do not figure within the constituents make-up of its strings, are attributed the dominant's (1.00) value at the thesaurus reversal. More importantly, the reversal brings to light latent, uneven as to their constituents' distancing, synonymous strings. They are initiated by the verbs missing from the list of dominants in the direct thesaurus that nonetheless occur within one or more string(s) initiated by other dominants. These become just members of the 'unearthed' reverse strings and such strings are being modeled solely due to the reversal procedure of the thesaurus.

The perception of the semantic distance in the reverse string rests on the perception of the distancing of the dominant from its respective string members (trick of the "right-to-left" reading in the usual thesaurus). In this way the headword is diversely and unevenly distanced from its reverse string members.

Table 2

Evaluation of WordNet-Based verb similarity measures performance against human judgements

#	Verb 1	Verb 2	Humanmea n	HumanStDe v	hsonorm.	Human-hso	path	Human-path	jsn	resnorm	Human- res	vector	human-rector	lin	lchnorm	Human-lch	LeskNorm	vector_ pairsno rm	Human-vp	wup	Human- wup
1	know	notice	**5.75**	0.82	**0**	5.7	**3.3**	2.45	**0**	**4.3**	1.4	**2.2**	3.5	**0**	**6.0**	-0.2	**0.3**	**2.2**	3.5	**6.6**	-0.8
2	know	interpret	**4.5**	2.87	**3.1**	1.4	**2.5**	2	**0**	**0**	4.5	**4.3**	0.2	**0**	**5.2**	-0.7	**0.4**	**2.0**	2.5	**4.0**	0.5
3	know	remember	**8.5**	1	**3.1**	5.4	**5.0**	-4.5	**0**	**5.5**	3	**9.0**	-0.5	**0**	**2.7**	5.8	**1.5**	**6.0**	1.5	**8.0**	0.5
4	know	Think	**4.5**	0.86	**3.1**	1.4	**5.0**	-0.5	**0**	**5.6**	-1.1	**9.0**	-4.5	**0.6**	**7.1**	-2.6	**1.6**	**7.5**	-3	**8.0**	-3.5
5	know	memorize	**7.75**	1.29	**2.5**	5.2	**2.0**	5.7	**0**	**0**	7.75	**0.9**	6.8	**0**	**4.6**	3.15	**0.1**	**3.0**	4.5	**3.3**	3.9
6	know	intuit	**7.25**	1.78	**2.5**	4.7	**2.0**	5.2	**0**	**0**	7.25	**1.1**	6.1	**0**	**4.6**	2.65	**0.1**	**2.5**	5	**3.3**	3.9
7	know	fathom	**4.25**	3.19	**3.1**	1.1	**2.5**	1.7	**0**	**0.7**	3.5	**0.9**	3.3	**0**	**5.2**	-95	**0.1**	**3.7**	0.5	**4.0**	0.2
8	know	realize	**5.75**	1.08	**3.1**	2.6	**5.0**	0.7	**0**	**6.4**	-0.6	**9.0**	-3.2	**0.8**	**7.1**	-1.3	**6.4**	**2.6**	3.1	**8.5**	-2.7
9	know	Grasp	**6.0**	2.82	**3.1**	2.9	**2.5**	3.5	**0**	**6.2**	-0.2	**3.9**	2.1	**0**	**6.2**	-0.2	**0.5**	**6.2**	-0.2	**8.4**	-2.4
10	notice	Know	**5.0**	0.70	**2.5**	2.5	**3.3**	1.7	**0**	**4.3**	0.7	**2.2**	2.8	**0**	**6.0**	-1	**0.3**	**2.2**	2.8	**6.6**	-1.6
11	interpret	Know	**6.5**	1.65	**0**	6.5	**2.5**	4	**0**	**0**	6.5	**4.3**	2.2	**0**	**5.2**	1.3	**0.4**	**2.0**	4.5	**4.0**	2.5
12	remembe	Know	**8.0**	1.73	**3.7**	4.3	**5.0**	3	**0**	**5.5**	2.5	**9.0**	-1	**0**	**2.7**	5.3	**1.5**	**6.0**	2	**8.0**	0
13	think	Know	**4.75**	0.43	**3.7**	1	**5.0**	-0.2	**0**	**5.6**	-1.7	**9.0**	-4.2	**0.6**	**7.1**	-3.1	**1.6**	**7.5**	-2.7	**8.0**	-3.2
14	memoriz	Know	**8.0**	2.34	**0**	8	**2.0**	6	**0**	**0**	8	**0.9**	7.1	**0**	**4.6**	3.4	**0.1**	**3.0**	5	**3.3**	4.7
15	intuit	Know	**7.5**	1.62	**0**	7.5	**2.0**	5.5	**0**	**0**	7.5	**1.1**	6.4	**0**	**4.6**	2.9	**0.1**	**2.5**	5	**3.3**	4.2
16	fathom	Know	**4.25**	3.26	**0**	4.2	**2.5**	2	**0**	**0.7**	3.5	**0.9**	3.3	**0**	**5.2**	-0.9	**0.1**	**3.7**	0.5	**4.0**	0.2
17	realize	Know	**6.25**	0.43	**2.5**	3.7	**5.0**	1.2	**0**	**6.4**	-0.1	**9.0**	-2.7	**0.8**	**7.1**	-0.8	**6.4**	**2.6**	3.6	**8.5**	-2.2
18	grasp	Know	**5.0**	3.08	**2.5**	2.5	**2.5**	2.5	**0**	**6.2**	1.2	**3.9**	1.1	**0**	**6.2**	-1.2	**0.5**	**6.2**	-1.2	**8.4**	-3.4

The entire "0-to-1" string can be arbitrarily split into arbitrary or inventory justifiable parts, and the respective synonyms would be placed within each of these, which creates a visualization curve for each string, producing an onomasiological atlas of synonymous relatedness for each thesaurus. The curve is imputed parametrization in the form of the intactness (attributed the value of (0), increase, respectively, (+1), and decrease assigned (-1) put consecutively in the inverse Shaumjanian notation starting with the concluding lag of the scale.

The first such reversal of the thesaurus of English verbs within LUMPE was accomplished using *Webster's New World Thesaurus* (1986) in the late 1990s (Bilynsky, 1999). It included over 31,000 values of the semantic distance between verbs. By docking the object with the derivation reflection (where obliging, on condition of no category, or through the rewriting procedures even suffix, constraint(s)) of each of the synonyms of the verbal string in the respective slots within the stringing of the deverbal word families engine and reversing the obtained derivational thesaurus seventeen derivationally reflected usually much smaller than the one for the verbs (also atlased as described above in the complete configuration 'prompted' by the derivational productivity of the dominant, or, for visualization purposes, configured separately) thesauri were compiled.

Metrically reversed in the course of the last almost twenty years and applicable to the developed upgradable multiple queries framework have been *Collins Thesaurus*, *Collins Thesaurus: The Ultimate Word Finder* (alphabetical listing inside strings), *Chamber's Study Thesaurus*, *Encarta Pocket Thesaurus*, *Concise Oxford Thesaurus*, *Penguin Thesaurus*, *The Penguin Dictionary of English Synonyms and Antonyms*, *Longman Pocket Thesaurus*, *Roget's Thesaurus* (separately for the aggregate and semantic fields applications), *Roget's College Thesaurus in Dictionary Form, Roget's II New Thesaurus*, *Roget's 21st c. Thesaurus* (alphabetical listing inside strings), *Webster's New World Thesaurus for the 1990s*, *Webster's New Thesaurus* (alphabetical listing inside strings, no polysemy markings), *Merriam Webster Thesaurus*, *The Wordsworth Thesaurus* (alphabetical listing inside string, no polysemy markings). Alongside the application of the developed framework to the processed present-day English paper thesauri the databases are being also compiled on the available electronic resources: *Synonymy.com*, *WordNet*, *Visual Thesaurus* as well as *Thesaurus.com.* with synonymous and antonymous alphabetical central and peripheral strings based on the *21st c. Roget's Thesaurus* with a particular emphasis on the inclusion of phrases. Added to the present-day English synonymous lexicography have been some of the available 'shallow history' thesauri, like the turn of the 20th c. *Ordway's Thesaurus* and still to process, possibly, some compatibles.

In principle, the entire aggregate megabase (in the 'Big Data' spirit) is attainable with the overwriting of fully coincident strings, accompanied with sources labeling, as one and separate listings of all strings to the dominant with a minimally divergent composition.

All the developed queries aim at the processing of the characteristics of the synonymous strings reversal in a chosen database. As the basic unit of analysis is the string it is characterized by a given number of constituents. All the queries could be run on the string's lengths indiscriminately, an arbitrary string length or a range of lengths. The obtained distribution numbers can be illustrated by individual or listed (if need may be, exhaustively) examples and (re-)set visualization.

The number of strings in the thesaurus, the lengths recurrence patterns as well as extends of polysemy in the dominants prove to be of relevance for revealing the metric of verbal synonymy. It certainly rests on the precise nature of the processed thesaurus. But, obviously, there are some special contour characteristics. For instance, at the tripartite stacking of the "0-to-1" scale shorter strings tend to reveal a better representation of the last ('remote') third of the scale whereas longer strings show a better representation of the first ('close') third of the scale. More layered stacking is justifiable for longer strings or for the thesauri that typically provide bigger extents of near- synonymy rather than merely stricter synsets.

The distribution of the distance values of the onomasiological flanks of the string is an additional factor of its contour in the onomasiological atlas. In provides a search engine (series length(s) sensitive) for discovering the strings that both begin or/and end at a specific determined stretch in fact, arbitrary, but more practically, three to five partite discreteness of the scale. The values of the semantic distance between the reverse string dominant and string members are distributed unevenly, thus it is of interest to generalize on the location of the

filled in vs. empty scale stretches (and their precise/approximate contouring) as well the ensuing statistics relevant for the processed thesauri.

A string of synonyms in a thesaurus (and this is applicable both to the input (direct) and reverse versions) may be free of any instance of reflexivity of the bilateral bond. A segment of such strings in a thesaurus, which could be tentatively termed as referential islands in its make-up, appears rather substantial, and may call for some amount of analysis with reference to thesauri individually or collectively. At the same time there are constituents in the thesaurus string that reveal a reflexive bond with an amount of asymmetry of the two-way relatedness. The respective sum total values of proximity ascribed to the constituents of this physical and 'mirrored' versions of the string will be responsible for two vectors. They characterized the aggregate two-way proximity between the dominant of the string and string constituents as the respective differential and/or angle. The angular geometry of each string established according to these principles is measurable in degrees or radians. Also, the quota of 'two-traffic' distance measures against the total number of constituents in such non-island strings is representable owing to a separate formalism.

The ongoing LUMPE project originally suggested and developed for English verbs provides a possibility of students' research work and will serve as a vast on-line reference resource in the future. There are also plans, in part already materialized, to extend it over adjectives and nouns.

Conclusion

Though for some applications, the asymmetry appears to be irrelevant and can be neglected, it does matter which word or concept is taken as a stimulus and which one as a reaction for pairwise semantic similarity measures. Intelligent systems cannot ignore this universal peculiarity of human mind associations. Consequently, the similarity and relatedness-based measures need to be reviewed taking into consideration the direction of the path between the concepts (words). It can be done by referring to other sources than ontologies only and by taking into account the linear semantic distance between the words that belong to the same synset.

The integration of corpus-based, ontology-based and dictionary/ (thesaurus)-based (multiple querried many-layered resettable data-driven exemplifiable and evidence visaulizable) measures can solve this problem and enhance the performance of automated measures for versatile NLP applications, which is the promising direction for future work.

References

Aitchison, J. (2012). *Words in the Mind: An Introduction to the Mental Lexicon.* John Wiley & Sons.

Banerjee, S. (2003). Extended gloss overlapProceedings of the Eighteenth International Joint Conference on Artificial Intelligence. *Extended gloss overlaps as a measure of semantic relatedness*, (pp. 805–810). Acapulco, Mexico.

Bilynsky, M. (1999). *English Verbal Synonyms: A Dictionary of Semantic Distances.* Lviv: Lviv University Publishers.

Butterfield, J. (2003). *Collins English Dictionary: Complete and Unabridged* (6 ed.). Glasgow: HarperCollins Publishers.

Davidson, G. (Ed.). (2003). *Roget's Thesaurus of English Words and Phrases.* Penguin Books. Fellbaum, C. (1998). *WordNet. An Electronic Lexical Database.* Cambridge, MA: The MIT Press.

Finkelstein, L. G. (2002, January). Placing search in context: the concept revisited. *ACM Transactions on Information Systems, 20*(1), pp. 116-131.

Hirst, G. & D. St-Onge. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum (Ed.), *WordNet. An Electronic Lexical Database* (pp. 305–332). Cambridge, MA: The MIT Press.

Jellis, S. (Ed.). (2002). *Microsoft Encarta College Thesaurus.* St. Martin's Press.

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

Leacock, C. & M. Chodorow . (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database* (pp. 265–283). Cambridge, MA.

Levickij, V. V. (1989). *Eksperimental'nye Metody v Semasiologii (Experimental Methods in Semasiology)*. Voronezh: Izd-vo Voronezhskogo universiteta.

Lin, D. (1998). An information-theoretic definition of similarity. *15th International Conference of Machine Learning*, (pp. 296–304). Madison, WI.

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

New Webster's Dictionary and Thesaurus of the English Language. (1993). Dansbury, CT: Lexicon Publications, INC.

Patwardhan, S. (2003). *Incorporating dictionary and corpus information into a context vector measure of semantic relatedness.*

Pedersen, T. (2015). *WordNet::Similarity Web Interface*. Retrieved from http://maraca.d.umn.edu/.

Pedersen, T. et. al. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Jounal of Biomedical Informatics*, 288–299.

Pedersen, T., Patwardhan S., Michelizzi J. (2004). WordNet::Similarity–measuring the relatedness of concepts. *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI)*, (pp. 144–152). San Jose, CA.

Resnik, P. (1998). WordNet and class-based probabilities. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database* (pp. 239–263). Cambridge, MA: MIT Press.

Rooney, K. (Ed.). (2001). *Longman Dictionary of Contemporary English* (3 ed.). Harlow: Longman.

Rubenstein, H. & H. Goodenough. (1965). Contextual correlates of synonymy. *Communications of the ACM, 8*(10), pp. 627–633.

Simpson, J. W. (Ed.). (1989). *The Oxford English Dictionary on Historical Principles (OED)* (2 ed.). Oxford: OUP.

Stube, M. S. (2007). WikiRelate! Computing semantic relatedness using Wikipedia. *Proceedings of IJCAI-07*, (pp. 1419–1429).

Wang, T. & G. Hirst. (2011). Refining the notions of depth and density in WordNet-based semantic similarity measures. *EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (pp. 1003-1011). Stroudsburg, PA, USA.

Wu, Z. & M. Palmer. (1994). Verbs semantics and lexical selection. *ACL '94 Proceedings of the 32nd annual meeting on Association for Computational Linguistics (June 27–30)*, (pp. 133–138). Las Cruces, New Mexico.

Yang, D. &. M. Powers. (2006). Verb similarity in the taxonomy of Word-Net. *Proceedings of the Third International WordNet Conference (GWC-06)*, (pp. 121–128). Jeju Island, Korea.

Yang, Yu., Bhavsar V., Booley H. (2008). On semantic concept similarity methods. *Proceedings of the 4th International Conference on Information & Communication Technology and System* , (pp. 4-11).

ОЦІНЮВАННЯ МІР СЕМАНТИЧНОЇ СХОЖОСТІ АНГЛІЙСЬКИХ ДІЄСЛІВ У WORDNET ТА ТЕЗАУРУСАХ

Ірина Ділай, Михайло Білинський

Львівський національний університет імені Івана Франка вул. Університетська 1, Львів 79000, Україна irynadilay@gmail.com, bislo@ukrpost.ua

У статті розглянуто міри семантичної подібності між концептами у WordNet. Виявлено, що семантична подібність у WordNet ґрунтується на таксономічних зв'язках: гіпонімії та синонімії. У результаті порівняння, помічено, що міри не враховують асиметрію ментальних асоціацій, тобто відстань між двома концептами є майже завжди сталою. Запропоновано інтегрувати досліджувані міри з такими, в основі яких лежить словоформа, а також контек- стуальна корпусна інформація, задля наближення до «золотого стандарту». Також подано апробовані параметри лексикометричного аналізу дієслівних тезаурусів англійської мови.

Ключові слова: міра семантичної подібності, асиметрія асоціацій, таксономічні зв'язки, си-нонімічний ряд, корпус, WordNet, зворотна синонімія, тезауруси, (не-)Евклідова геометрія, векторний аналіз, кутова геометрія у тезаурусі.