Comparison of the Baseline Knowledge-, Corpus-, and Web-based Similarity Measures for Semantic Relations Extraction

Alexander Panchenko
Center for Natural Language Processing (CENTAL)
Université catholique de Louvain, Belgium
alexander.panchenko@student.uclouvain.be

Abstract

Unsupervised methods of semantic relations extraction rely on a similarity measure between lexical units. Similarity measures differ both in kinds of information they use and in the ways how this information is transformed into a similarity score. This paper is making a step further in the evaluation of the available similarity measures within the context of semantic relation extraction. We compare 21 baseline measures – 8 knowledge-based, 4 corpus-based, and 9 web-based metrics with the BLESS dataset. Our results show that existing similarity measures provide significantly different results, both in general performances and in relation distributions. We conclude that the results suggest developing a combined similarity measure.

1 Introduction

Semantic relations extraction aims to discover meaningful lexico-semantic relations such as synonyms and hyponyms between a given set of lexically expressed concepts. Automatic relations discovery is a subtask of automatic thesaurus construction (see Grefenstette (1994), and Panchenko (2010)).

A set of semantic relations R between a set of concepts C is a binary relation $R \subseteq C \times T \times C$, where T is a set of semantic relation types. A relation $r \in R$ is a triple $\langle c_i, t, c_j \rangle$ linking two concepts $c_i, c_j \in C$ with a semantic relation of type $t \in T$. We are dealing with six types of semantic relations: hyperonymy, co-hyponymy, meronymy, event (associative), attributes, and random: $T = \{hyper, coord, mero, event, attri, random\}$. We describe analytically and compare experimentally methods, which discover set of semantic relations $\hat{R}$ for a given set of concepts C. A semantic relation extraction algorithm aims to discover $\hat{R} \sim R$.

One approach for semantic relations extraction is based on the lexico-syntactic patterns which are constructed either manually (Hearst, 1992) or semi-automatically (Snow et al., 2004). The alternative approach, adopted in this paper, is unsupervised (see e.g. Lin (1998a) or Sahlgren (2006)). It relies on a *similarity measure* between lexical units. Various measures are available. We compare 21 baseline measures: 8 knowledge-based, 4 corpus-based, and 9 web-based. We would like to answer on two questions: "What metric is most suitable for the unsupervised relation extraction?", and "Does various metrics capture the same semantic relations?". The second question is particularly interesting for developing of a meta-measure combining several metrics. This information may also help us choose a measure well-suited for a concrete application.

We extend existing surveys in three ways. First, we ground our comparison on the BLESS dataset[1], which is open, general, and was never used before for comparing all the considered metrics. Secondly, we face corpus-, knowledge-, and web-based, which was never done before. Thirdly, we go further than most of the comparisons and thoroughly compare the metrics with respect to relation types they provide. We report empirical relation distributions for

[1] http://sites.google.com/site/geometricalmodels/sharedevaluation

Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP 2011, pages 11–21,
Edinburgh, Scotland, UK, July 31, 2011.

each measure and check if they are significantly different. Next, we propose a way to find the measures with the most and the least similar relation distributions. Finally, we report information about redundant measures in an original way – in a form of an undirected graph.

2 Methodology

2.1 Similarity-based Semantic Relations Discovery

We use an unsupervised approach to calculate set of semantic relations R between a given set of concepts C (see algorithm 1). The *method* uses one of 21 similarity *measures* described in sections 2.2 to 2.4. First, it calculates the concept×concept similarity matrix $\mathbf{S}$ with a measure sim. Since some similarity measures output scores outside the interval $[0;1]$ we transform them with the function $normalize$ as following: $\mathbf{S} \leftarrow \frac{(\mathbf{S}-min(\mathbf{S}))}{max(\mathbf{S})}$. If we deal with a dissimilarity measure, we additionally transform its score $\mathbf{S}$ to similarity as following: $\mathbf{S} \leftarrow 1 - normalize(\mathbf{S})$. Finally, the function $threshold$ calculates semantic relations R between concepts C with the k-NN thresholding: $\bigcup_{i=1}^{|C|} \{\langle c_i, t, c_j\rangle : c_j \in \text{ top } k\% \text{ concepts } \wedge s_{ij} \geq \gamma\}$. Here k is the percent of the top similar concepts to a concept c_i, and γ is a small value which ensures than nearly-zero pairwise similarities s_{ij} will be ignored. Thus, the method links each concept c_i with $k\%$ of its nearest neighbours.

Algorithm 1: Computing semantic relations

Input: Concepts C, Sim.parameters P, Threshold k, Min.similarity value γ
Output: Unlabeled semantic relations $\hat{R}$

$\mathbf{S} \leftarrow sim(C, P)$;
$\mathbf{S} \leftarrow normalize(\mathbf{S})$;
$\hat{R} \leftarrow threshold(\mathbf{S}, k, \gamma)$;
return $\hat{R}$;

Below we list the pairwise similarity measures sim used in our experiments with references to the original papers, where all details can be found.

2.2 Knowledge-based Measures

The knowledge-based metrics use a hierarchical semantic network in order to calculate similarities. Some of the metrics also use counts derived from a corpus. We evaluate eight knowledge-based measures listed below. Let us describe them in the following notations: c_r is the root concept of the network; h is the height of the network; $len(c_i, c_j)$ is the length of the shortest path in the network between concepts; c_{ij} is a lowest common subsumer of concepts c_i and c_j; $P(c)$ is the probability of the concept, estimated from a corpus (see below). Then, the Inverted Edge Count measure (Jurafsky and Martin, 2009, p. 687) is

$$s_{ij} = len(c_i, c_j)^{-1}; \tag{1}$$

Leacock and Chodorow (1998) measure is

$$s_{ij} = -log\frac{len(c_i, c_j)}{2h}; \tag{2}$$

Resnik (1995) measure is

$$s_{ij} = -log(P(c_{ij})); \tag{3}$$

Jiang and Conrath (1997) measure is

$$s_{ij} = (2log(P(c_{ij})) - (log(P(c_i)) + log(P(c_j))))^{-1}; \tag{4}$$

Lin (1998b) measure is

$$s_{ij} = (\frac{2log(P(c_{ij}))}{log(P(c_i) + log(P(c_j))}; \tag{5}$$

Wu and Palmer (1994) measure is

$$s_{ij} = \frac{2len(c_r, c_{ij})}{len(c_i, c_{ij}) + len(c_j, c_{ij}) + 2len(c_r, c_{ij})}. \tag{6}$$

Extended Lesk (Banerjee and Pedersen, 2003) measure is

$$s_{ij} = \sum_{c_i \in C_i} \sum_{c_j \in C_j} sim_g(c_i, c_j), \tag{7}$$

where sim_g is a gloss-based similarity measure, and set C_i includes concept c_i and all concepts which are directly related to it.

Gloss Vectors measure (Patwardhan and Pedersen, 2006) is calculated as a cosine (9) between context vectors $\mathbf{v}_i$ and $\mathbf{v}_j$ of concepts c_i and c_j. A context vector calculated as following:

$$\mathbf{v}_i = \sum_{\forall j : c_j \in G_i} \mathbf{f}_j. \tag{8}$$

Here $\mathbf{f}_j$ is a first-order co-occurrence vector, derived from the corpus of all glosses, and G_i is concatenation of glosses of the concept c_i and all concepts which are directly related to it.

We experiment with measures relying on the WORDNET 3.0 (Miller, 1995) as a semantic network and SEMCOR as a corpus (Miller et al., 1993).

2.3 Corpus-based measures

We use four measures, which rely on the bag-of-word distributional analysis (BDA) (Sahlgren, 2006). They calculate similarity of concepts c_i, c_j as similarity of their feature vectors $\mathbf{f}_i, \mathbf{f}_j$ with the following formulas (Jurafsky and Martin, 2009, p. 699): cosine

$$s_{ij} = \frac{\mathbf{f}_i \cdot \mathbf{f}_j}{\|\mathbf{f}_i\| \, \|\mathbf{f}_j\|}, \tag{9}$$

Jaccard

$$s_{ij} = \frac{\sum_{k=1}^{N} min(f_{ik}, f_{jk})}{\sum_{k=1}^{N} max(f_{ik}, f_{jk})}, \tag{10}$$

Manhattan

$$s_{ij} = \sum_{k=1}^{N} |f_{ik} - f_{jk}|, \tag{11}$$

Euclidian

$$s_{ij} = \sqrt{\sum_{k=1}^{N} (f_{ik} - f_{jk})^2}. \tag{12}$$

The feature vector $\mathbf{f}_i$ is a first-order co-occurrence vector. The context of a concept includes all words from a sentence where it occurred, which pass a stop-word filter (around 900 words) and a stop part-of-speech filter (nouns, adjectives, and verbs are kept). The frequencies f_{ij} are normalized with Poinwise Mutual Information (PMI): $f_{ij} = log(f_{ij}/(count(c_i)count(f_j)))$. In our experiments we use two general English corpora (Baroni et al., 2009): WACYPEDIA (800M tokens), and PUKWAC (2000M tokens). These corpora are POS-tagged with the TreeTagger (Schmid, 1994).

2.4 Web-based measures

The web-based metrics use the Web text search engines in order to calculate the similarities. They rely on the number of times words co-occur in the documents indexed by an information retrieval system. Let us describe these measures in the following notation: h_i is the number of documents (hits) returned by the system by the query "c_i"; h_{ij} is the number of hits returned by the query "c_i AND c_j"; and M is number of documents indexed by the system. We use two web-based measures: Normalized Google Distance (NGD) (Cilibrasi and Vitanyi, 2007):

$$s_{ij} = \frac{max(log(h_i, h_j)) - log(h_{ij})}{log(M) - min(log(h_i), log(h_j))}, \tag{13}$$

and PMI-IR similarity (Turney, 2001) :

$$s_{ij} = log\left(\frac{h_{ij} \sum_i \sum_j h_i h_j}{h_i h_j \sum_i h_{ij}}\right). \tag{14}$$

We experiment with 5 NGD measures based on Yahoo, YahooBoss [2], Google, Google over Wikipedia, and Factiva [3]; and with 4 PMI-IR measures based on YahooBoss, Google, Google over Wikipedia, and Factiva. We perform search among all indexed documents or within the domain `wikipedia.org` (we denote the latter measures with the postfix -W).

2.5 Classification of the measures

It might help to understand the results if we mention that (1) - (6) are measures of *semantic similarity*, while (7) and (8) are measures of *semantic relatedness*. Semantic relatedness is a more general notion than semantic similarity (Budanitsky and Hirst, 2001). A measure of semantic similarity uses only hierarchical and equivalence relations of the semantic network, while a measure of semantic relatedness also use relations of other types. Furthermore, measures (1), (2), (3), are "pure" semantic similarity measures since they use only semantic network, while (3), (4), and (5) combine information from a semantic network and a corpus.

The corpus-based and web-based measures are calculated differently, but they are both clearly *distributional* in nature. In that respect, the web-based measures use the Web as a corpus. Figure 1 contains

[2] `http://developer.yahoo.com/search/boss/`
[3] `http://www.factiva.com/`

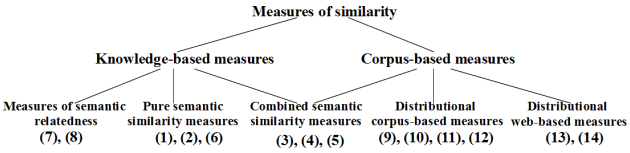

Figure 1: Classification of the measures used in the paper.

a more precise classification of the considered measures, according to their properties. Finally, both (8) and (9)-(12), rely on the vector space model.

2.6 Experimental Setup

We experiment with the knowledge-based measures implemented in the WORDNET::SIMILARITY package (Pedersen et al., 2004). Our own implementation is used in the experiments with the corpus-based measures and the web-based measures relying on the YAHOO BOSS search engine API. We use the MEASURES OF SEMANTIC RELATEDNESS web service [4] to assess the other web measures.

The evaluation was done with the BLESS set of semantic relations. It relates 200 target concepts to some 8625 relatum concepts with 26554 semantic relations (14440 are correct and 12154 are random). Every relation has one of the following six types: hyponymy, co-hyponymy, meronymy, attribute, event, and random. The distribution of relations among those types is given in table 1. Each concept is a single English word.

3 Results

3.1 Comparing General Performance of the Similarity Measures

In our evaluation semantic relations extraction was viewed as a retrieval task. Therefore, for every metric we calculated precision, recall, and F1-measure with respect to the golden standard. Let $\hat{R}$ be set of extracted semantic relations, and R be set of semantic relations in the BLESS. Then

$$Precision = \frac{|R \cap \hat{R}|}{|\hat{R}|}, Recall = \frac{|R \cap \hat{R}|}{|R|}.$$

An extracted relation $\langle c_i, t_?, c_j \rangle \in \hat{R}$ matches a relation from the evaluation dataset $\langle c_i, t, c_j \rangle \in R$ if

[4] http://cwl-projects.cogsci.rpi.edu/msr/

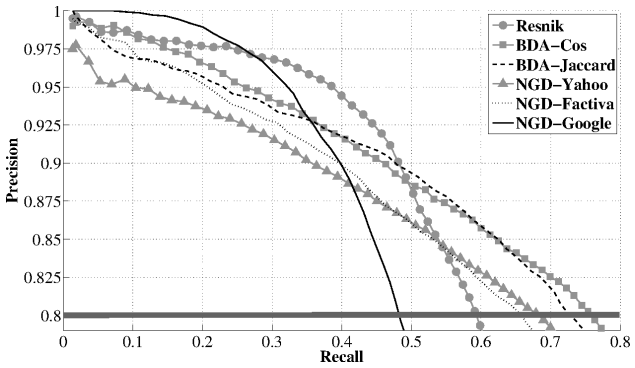

Figure 2: Precision-recall graph of the six similarity measures (kNN threshold value $k = 0 - 52\%$).

$t \neq random$. Thus, an extracted relation is correct if it has any type in BLESS, but random.

General performance of the measures is presented in table 1 (columns 2-4). The Resnik measure (3) is the best among the knowledge-based measures; the NGD (13) measure relying on the Yahoo search engine is the best results among the web-based measures. Finally, the cosine measure (9) (BDA-Cos) is the best among all the measures. The table 2 demonstrate some extracted relations discovered with the BDA-Cos measure.

In table 1 we ranked the measures based on their F-measure when precision is fixed at 80% (see figure 2). We have chosen this precision level, because it is a point when automatically extracted relations start to be useful. It is clear from the precision-recall graph (figure 2) that if another precision level is fixed then ranking of the metrics will change. Analysis of this and similar plots for other measures shows us that: (1) the best knowledge-based metric is Resnik; (2) the BDA-Cos is the best among the corpus-based measures, but BDA-Jaccard is very close to it; (3) the three best web-based measures are NGD-Google (within the precision range 100-90%), NGD-Factiva (within the precision range 90%-87%), and NGD-Yahoo (starting from the precision level 87%). In these settings, choose of the most suitable metric may depend on the application. For instance, if just a few precise relations are needed then NGD-Google is a good choice. On the other hand, if we tolerate a slightly less precision, and if we need many relations then the BDA-Cos is the best choice.

Figure 3 depicts learning curve of the BDA-Cos

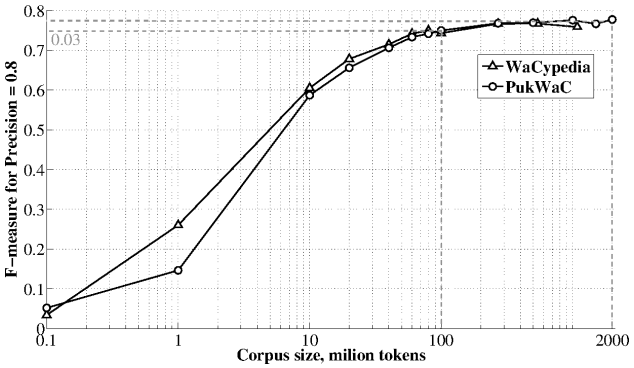

Figure 3: Learning curves of the BDA-Cos on the WaCypedia and PukWaC corpora (0.1M–2000M tokens).

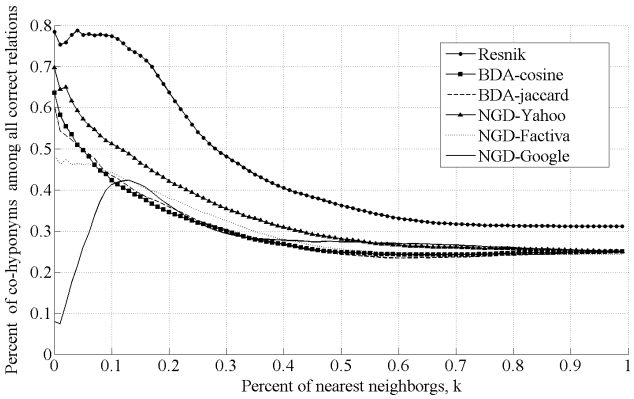

Figure 4: Percent of co-hyponyms among all correctly extracted relations for the six best measures.

measure.Dependence of the F-measure at the precision level of 80% from the corpus size is not linear. F-measure improves up to 44% when we increase corpus size from 1M to 10M tokens; increasing corpus from 10M to 100M tokens gives the improvement of 16%; finally, increasing corpus from 100M to 2000M tokens gives the improvement of only 3%.

3.2 Comparing Relation Distributions of the Similarity Measures

In this section, we are trying to figure out what types of semantic relations the measures find. We compare distributions of semantic relations against the BLESS dataset. Generally, if two measures have equal general performances, one may want to choose a metric which provides more relations of a certain type, depending on the application. This information may be also valuable in order to decide which metrics to combine in a meta-metric.

Distribution of Relation Types. In this section, we estimate empirical relation distribution of the metrics over five relation types: hyponymy, co-hyponymy, meronymy, attribute, and event. To do so we calculate percents of correctly extacted relations of type t for a each measure:

$$Percent = \frac{\hat{R}_t}{|R \cap \hat{R}|}, \text{ where } \bigcup_{t \in T} \hat{R}_t = |R \cap \hat{R}|.$$

Here $|R \cap \hat{R}|$ is a set of all correctly extracted relations, and $\hat{R}_t$ is a set of extracted relations of type t. Figure 4 demonstrates that percent of extracted relations of certain type depends on the value of k (c.f. section 2.1). For instance, if $k = 10\%$ then 77% of extracted relations by Resnik are co-hyponyms, but if $k = 40\%$ then the same measure outputs 40% of co-hyponyms. We report relations distribution at two levels of the threshold k – 10% and 40%.

The empirical distributions are reported in columns 5-9 of the table 1. Each of those columns correspond to one semantic relation type t, and contains two numbers: p_{10} – percent of relations of type t when $k = 10\%$, and p_{40} – percent of relations of type t when $k = 40\%$. We represent those two values in the following format: $p_{10}|p_{40}$. For instance, $77|40$ behind the Resnik measure means that when $k = 10\%$ it extracts 77% of co-hypernyms, and when $k = 40\%$ it extracts 40% of co-hypernyms.

If the threshold k is 10% then the biggest fraction of extracted relations are co-hyponyms – from 35% for BDA-Manhattan to 77% for Resnik measure. At this threshold level, the knowledge-based measures mostly return co-hyponyms (60% in average) and hyperonyms (23% in average). The corpus-based metrics mostly return co-hyponyms (38% in average) and event relations (26% in average). The web-based measures return many (48% in average) co-hyponymy relations.

If the threshold k is 40% then relation distribution of all the measures significantly changes. Most of the relations returned by the knowledge-based measures are co-hyponyms (36%) and meronyms (24%). The majority of relations discovered by the corpus-based metrics are co-hyponyms (33%), event relations (26%), and meronyms (20.33%). The web-based measures at this threshold value return many event relations (32%).

	General Performance			Semantic Relations Distribution				
Measure	k	**Recall**	**F1**	**hyper**, %	**coord**, %	**attri**, %	**mero**, %	**event**, %
Resnik	**40%**	**0.59**	**0.68**	**9 | 14**	**77 | 40**	**4 | 8**	**6 | 22**	**4 | 15**
Inv.Edge-Counts	38%	0.56	0.66	22 | 15	61 | 40	4 | 8	7 | 22	6 | 15
Leacock-Chodorow	38%	0.56	0.66	22 | 15	61 | 40	4 | 8	7 | 22	6 | 15
Wu Palmer	37%	0.54	0.65	20 | 15	64 | 42	3 | 8	7 | 22	5 | 13
Lin	36%	0.53	0.64	30 | 16	52 | 31	4 | 7	8 | 29	5 | 16
Gloss Overlap	36%	0.53	0.63	5 | 6	52 | 34	7 | 12	18 | 21	18 | 27
Jiang-Conrath	35%	0.52	0.63	38 | 16	45 | 30	4 | 6	8 | 29	5 | 18
Extended Lesk	30%	0.45	0.57	21 | 14	39 | 30	1 | 9	29 | 28	9 | 19
BDA-Cos	**52%**	**0.76**	**0.78**	**9 | 7**	**42 | 27**	**11 | 20**	**15 | 17**	**23 | 30**
BDA-Jaccard	51%	0.75	0.77	10 | 7	45 | 27	8 | 16	16 | 20	20 | 27
BDA-Manhattan	37%	0.54	0.65	7 | 6	35 | 24	17 | 22	10 | 15	31 | 34
BDA-Euclidian	21%	0.30	0.44	7 | 7	31 | 18	20 | 26	12 | 13	30 | 37
NGD-Yahoo	**46%**	**0.68**	**0.74**	**7 | 6**	**51 | 30**	**9 | 18**	**17 | 20**	**15 | 25**
NGD-Factiva	47%	0.66	0.72	10 | 8	44 | 28	8 | 19	23 | 22	16 | 25
NGD-YahooBOSS	35%	0.51	0.63	13 | 10	54 | 36	4 | 10	14 | 20	15 | 22
NGD-Google	33%	0.48	0.60	1 | 7	41 | 28	45 | 19	2 | 19	11 | 28
NGD-Google-W	29%	0.43	0.56	8 | 9	45 | 31	8 | 14	20 | 21	19 | 25
PMI-YahooBOSS	29%	0.43	0.56	15 | 12	53 | 38	3 | 9	15 | 20	13 | 20
PMI-Factiva	25%	0.28	0.44	8 | 8	42 | 30	10 | 17	21 | 20	18 | 24
PMI-Google	12%	0.18	0.29	8 | 8	55 | 35	7 | 15	17 | 21	12 | 22
PMI-Google-W	9%	0.13	0.23	12 | 11	47 | 38	7 | 11	20 | 20	13 | 19
Random measure				8 | 9	24 | 25	20 | 19	22 | 20	26 | 27
BLESS dataset				9	25	20	19	27

Table 1: Columns 2-4: Recall and F-measure when Precision$= 0.8$ (correct relations of all types vs random relations). Columns 5-9: percent of extracted relations of a certain type with respect to all correctly extracted relations, when threshold k equal 10% or 40%. The best measure are sorted by F-measure; the best measures are in bold.

ant	**banana**	**fork**	**missile**	**salmon**
cockroach (coord)	mango (coord)	prong (mero)	warhead (mero)	trout (coord)
grasshopper (coord)	pineapple (coord)	spoon (coord)	weapon (hyper)	mackerel (coord)
silverfish (coord)	papaya (coord)	knife (coord)	deploy (event)	herring (coord)
wasp (coord)	pear (coord)	lift (event)	nuclear (attri)	fish (event)
insect (hyper)	ripe (attri)	**fender (random)**	bomb (coord)	tuna (coord)
arthropod (hyper)	peach (coord)	plate (coord)	destroy (event)	oily (attri)
industrious (attri)	coconut (coord)	rake (coord)	rocket (coord)	poach (event)
ladybug (coord)	fruit (hyper)	shovel (coord)	arm (hyper)	catfish (coord)
bee (coord)	apple (coord)	handle (mero)	propellant (mero)	catch (event)
beetle (coord)	apricot (coord)	sharp (attri)	**bolster (random)**	fresh (attri)
locust (coord)	strawberry (coord)	spade (coord)	launch (event)	cook (event)
dragonfly (coord)	ripen (event)	napkin (coord)	deadly (attri)	cod (coord)
hornet (coord)	plum (coord)	cutlery (hyper)	**country (random)**	smoke (event)
creature (hyper)	grapefruit (coord)	head (mero)	strike (event)	seafood (hyper)
crawl (event)	cherry (coord)	scissors (coord)	defuse (event)	eat (event)

Table 2: Examples of the discovered semantic relations with the bag-of-words distributional analysis (BDA-Cos).

Interestingly, for the most of the measures, percent of extracted hyponyms and co-hyponyms decreases as the value of k increase, while the percent of other relations increases. In order to make it clear, we grayed cells of the table 1 when $p_{10} \geq p_{40}$.

Similarity to the BLESS Distribution. In this section, we check if relation distributions (see table 1) are completely biased by the distribution in the evaluation dataset. We compare relation distributions of the metrics with the distribution in the BLESS on the basis of the χ^2 goodness of fit test [5] (Agresti, 2002) with $df = 4$. A random similarity measure is completely biased by the distribution in the evaluation dataset: $\chi^2 = 5.36, p = 0.252$ for $k = 10\%$ and $\chi^2 = 3.17, p = 0.53$ for $k = 40\%$. On the other hand, distributions of all the 21 measures are significantly different from the distribution in the BLESS ($p < 0.001$). The value of chi-square statistic varies from $\chi^2 = 89.94$ (NGD-Factiva, $k = 10\%$) to $\chi^2 = 4000$ (Resnik, $k = 10\%$).

Independence of Relation Distributions. In this section, we check whether relation distributions of the various measures are significantly different. In order to do so, we perform the chi-square independence test on the table 1. Our experiments shown that there is a significant interaction between the type of the metric and the relations distribution: $\chi^2 = 10487, p < 0.001, df = 80$ for all the metrics; $\chi^2 = 2529, df = 28, p < 0.001$ for the knowledge-based metrics; $\chi^2 = 245, df = 12, p < 0.001$ for the corpus-based metrics; and $\chi^2 = 3158, df = 32, p < 0.001$ for the web-based metrics. Thus, there is a clear dependence between the type of measure and the type of relation it extracts.

Most Similar and Dissimilar Measures. In this section, we would like to find the most similar and disimilar measures. This information is particularly useful for the combination of the metrics. In order to find redundant measures, we calculate distance x_{ij} beween measures sim_i and sim_j, based on the χ^2-statistic:

$$x_{ij} = x_{ji} = \sum_{t \in T} \frac{(|\hat{R_t^i}| - |\hat{R_t^j}|)^2}{|\hat{R_t^j}|}, \qquad (15)$$

where $\hat{R_t^i}$ is ensemble of correctly extracted relations of type t with measure sim_i. We calculate these distances for all pairs of measures and then rank the pairs according to the value of x_{ij}. Table 3 present list of the most similar and dissimilar metrics obtained this way. Figure 7 reports in a compact way all the pairwise similarities $(x_{ij})_{21\times21}$ between the 21 metrics. In this graph, an edge links two measures, which have the distance value $x_{ij} < 220$. The graph was drawn with the Fruchterman and Reingold (1991) force-directed layout algorithm. One can see that relation distributions of the web- and corpus-based measures are quite similar. The knowledge-based measures are much different from them, but similar among themselves.

Distribution of Similarity Scores. In this section, we compare distributions of similarity scores across relation types with the following procedure: (1) Pick a closest relatum concept c_j per relation type t for each target concept c_i. (2) Convert similarity scores associated to each target concept to z-scores. (3) Summarize the distribution of similarities across relations by plotting the z-scores grouped by relations in a box plot. (4) Verify the statistical significance of the differences in similarity scores across relations by performing the Tukey's HSD test.

Figure 6 presents the distributions of similarities across various relation types for Resnik, BDA-Cos, and NGD-Yahoo. First, meaningful relation types for these three measures are significantly different ($p < 0.001$) from random relations. The only exception is the Resnnik measure – its similarity scores for the attribute relations are not significantly different ($p = 0.178$) from random relations. Thus, the best three measures provide scores which let us separate incorrect relations from the correct ones if an appropriate threshold k is set. Second, the similarity scores have highest values for the co-hyponymy relations. Third, BDA-Cos, BDA-Jaccard, NGD-Yahoo, NGD-Factiva, and PMI-YahooBoss provide the best scores. They let us clearly ($p < 0.001$) separate meaningful relations from the random ones. From the other hand, the poorest scores were provided by BDA-Manhattan, BDA-Euclidian, NGD-YahooBoss, and NGD-Google, because their scores let us clearly separate only co-hyponyms from the random relations.

Corpus Size. Table 1 presented relation distribution of the BDA-Cos trained on the 2000M token

[5] Here and below, we calculate the χ^2 statistic from the table 1 (columns 5-9), where percents are replaced with frequencies.

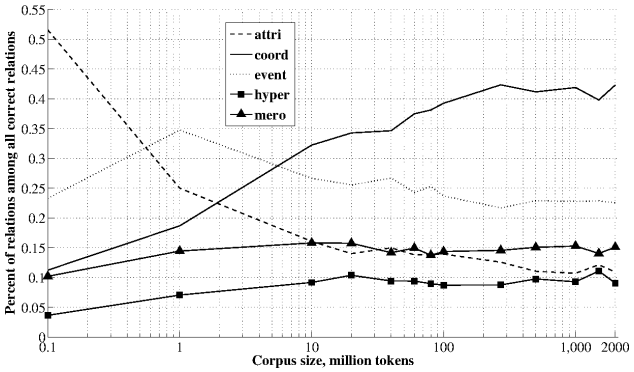

Figure 5: Semantic relations distribution function of corpus size (BDA-Cos measure, PukWaC corpus).

corpus UKWAC. Figure 5 shows the relation distribution function of the corpus size. First, if corpus size increases then percent of attribute relations decreases, while percent of co-hyponyms increases. Second, corpus size does not drastically influence the distribution for big corpora. For instance, if we increase corpus size from 100M to 2000M tokens then the percent of relations change on 3% for attributes, on 3% co-hyponyms, on 1% events, on 0.7% hyperonyms, and on 0.4% meronyms.

4 Related Work

Prior research provide us information about general performances of the measures considered in this paper, but not necessarily on the task of semantic relations extraction. For instance, Mihalcea et al. (2006) compare two corpus-based (PMI-IR and LSA) and six knowledge-based measures on the task of text similarity computation. The authors report that PMI-IR is the best measure; that, similarly to our results, Resnik is the best knowledge-based measure; and that simple average over all 8 measures is even better than PMI-IR. Budanitsky and Hirst (2001) report that Jiang-Conrath is the best knowledge-based measure for the task of spelling correction. Patwardhan and Pedersen (2006) evaluate six knowledge-based measures on the task of word sense disambiguation and report the same result. This contradicts our results, since we found Resnik to be the best knowledge-based measure.

Peirsman et al. (2008) compared general performances and relation distributions of distributional methods using a lexical database. Sahlgren (2006) evaluated syntagmatic and paradigmatic bag-of-word models. Our findings mostly fits well these and other (e.g. Curran (2003) or Bullinaria and Levy (2007)) results on the distributional analysis. Lindsey et al. (2007) compared web-based measures. Authors suggest that a small search domain is better than the whole Internet. Our results partially confirm this observation (NGD-Factiva outperforms NGD-Google), and partially contradicts it (NGD-Yahoo outperforms NGD-Factiva).

Van de Cruys (2010) evaluates syntactic, and bag-of-words distributional methods and suggests that the syntactic models are the best for the extraction of tight synonym-like similarity. Wandmacher (2005) reports that LSA produces 46.4% of associative relations, 15.2% of synonyms, antonyms, hyperonyms, co-hyponyms, and meronyms, 5.6% of syntactic relations, and 32.8% of erroneous relations. We cannot compare these results to ours, since we did not evaluate neither LSA nor syntactic models.

A common alternative to our evaluation methodology is to use the Spearman's rank correlation coefficient (Agresti, 2002) to compare the results with the human judgments, such as those obtained by Rubenstein and Goodenough (1965) or Miller and Charles (1991).

5 Conclusion and Future Work

This paper has compared 21 similarity measures between lexical units on the task of semantic relation extraction. We compared their general performances and figured out that Resnik, BDA-Cos, and NGD-Yahoo provide the best results among knowledge-, corpus-, and web-based measures, correspondingly. We also found that (1) semantic relation distributions of the considered measures are significantly different; (2) all measures extract many co-hyponyms; (3) the best measures provide the scores which let us clearly separate correct relations from the random ones.

The analyzed measures provide complimentary types of semantic information. This suggests developing a combined measure of semantic similarity. A combined measure is not presented here since designing an integration technique is a complex research goal on its own right. We will address this problem in our future research.

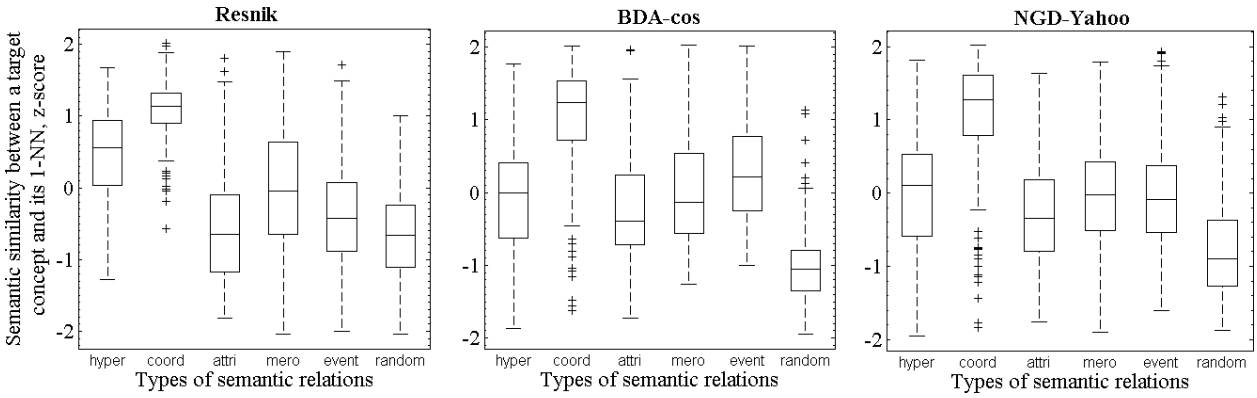

Figure 6: Distribution of similarities accross relation types for Resnik, BDA-Cos, and NGD-Yahoo measures.

Most Similar Measures			**Most Disimilar Measures**		
sim_i	sim_j	x_{ij}	sim_i	sim_j	x_{ij}
Leacock-Chodorow	Inv.Edge-Counts	0	NGD-Google	Extended Lesk	39935.16
BDA-Jaccard	BDA-Cos	7.17	Jiang-Conrath	NGD-Google	27478.90
NGD-YahooBOSS	PMI-YahooBOSS	19.58	Lin	NGD-Google	17527.22
Wu-Palmer	Inv.Edge-Counts	24.00	NGD-Google	Wu-Palmer	17416.95
Wu-Palmer	Leacock-Chodorow	24.00	NGD-Google	PMI-YahooBOSS	13390.66
BDA-Manhattan	BDA-Euclidian	25.37	Inv.Edge-Counts	NGD-Google	12012.79
PMI-Google-W	NGD-Factiva	27.65	Leacock-Chodorow	NGD-Google	12012.79
PMI-Google	NGD-Yahoo	33.42	NGD-Google	Resnik	11750.41
NGD-Google-W	NGD-Factiva	40.03	NGD-Google	NGD-YahooBOSS	11556.69
NGD-W	PMI-Factiva	42.17	BDA-Euclidian	Extended Lesk	8411.66
Gloss Overlap	NGD-Yahoo	53.64	NGD-Factiva	NGD-Google	8066.75
NGD-Factiva	PMI-Factiva	58.13	BDA-Euclidian	Resnik	6829.71
Lin	Jiang-Conrath	58.42	PMI-Google-W	NGD-Google	6574.62
Gloss Overlap	NGD-Google-W	62.46	BDA-Manhattan	Extended Lesk	6428.47

Table 3: List of the most and least similar measures ($k = 10\%$).

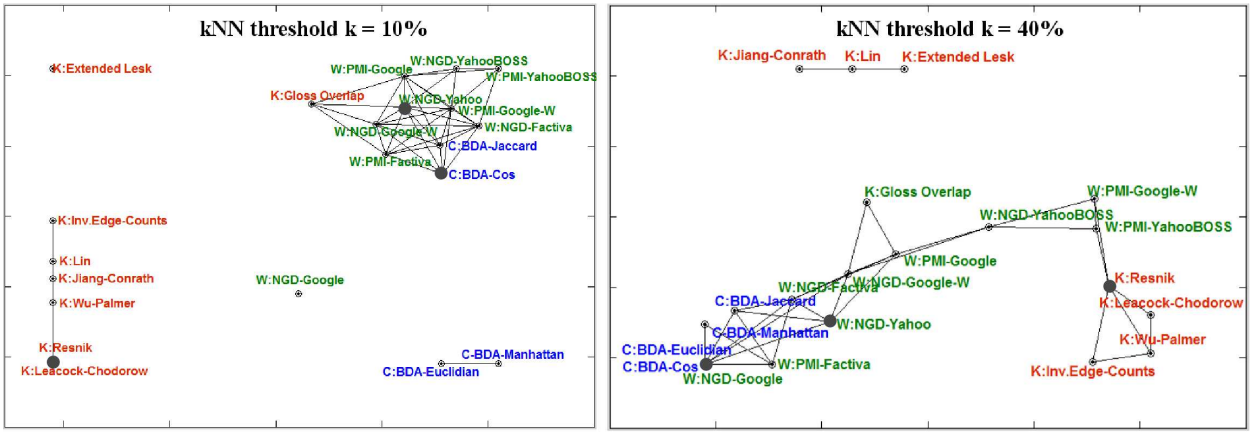

Figure 7: Measures grouped according to similarity of their relation distributions with (15). An edge links measures sim_i and sim_j if $x_{ij} < 220$. The knowledge-, corpus-, and web-based measures are marked in red, blue, and green correspondingly and with the prefixes 'K','C',and 'W'. The best measures are marked with a big circle.

6 Acknowledgments

I would like to thank Thomas François who kindly helped with the evaluation methodology, and my supervisor Dr. Cédrick Fairon. The two anonymous reviewers, Cédrick Fairon, Thomas François, Jean-Leon Bouraoui, and Andew Phillipovich provided comments and remarks, which considerably improved quality of the paper. This research is supported by Wallonie-Bruxelles International.

References

Alan Agresti. *Categorical Data Analysis (Wiley Series in Probability and Statistics)*. Wiley series in probability and statistics. Wiley Interscience, Hoboken, NJ, 2 edition, 2002.

Satanjeev Banerjee and Ted Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *International Joint Conference on Artificial Intelligence*, volume 18, pages 805–810, 2003.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009.

Alexander Budanitsky and Graeme Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources*, volume 2, 2001.

John A. Bullinaria and Joseph P. Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510, 2007.

Rudi L. Cilibrasi and Paul M. B. Vitanyi. The Google Similarity Distance. *IEEE Trans. on Knowl. and Data Eng.*, 19(3):370–383, 2007.

James R. Curran. *From distributional to semantic similarity*. PhD thesis, University of Edinburgh, 2003.

Thomas M. J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991.

Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery (The Springer International Series in Engineering and Computer Science)*. Springer, 1 edition, 1994. ISBN 0792394682.

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

Jay J. Jiang and David W. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *International Conference Research on Computational Linguistics (ROCLING X)*, pages 19–33, 1997.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2009.

Claudia Leacock and Martin Chodorow. Combining Local Context and WordNet Similarity for Word Sense Identification. *An Electronic Lexical Database*, pages 265–283, 1998.

Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 768–774. Association for Computational Linguistics, 1998a.

Dekang Lin. An Information-Theoretic Definition of Similarity. In *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, 1998b.

Robert Lindsey, Vladislav D. Veksler, Alex Grintsvayg, and Wayne D. Gray. Be wary of what your computer reads: the effects of corpus selection on measuring semantic relatedness. In *8th International Conference of Cognitive Modeling, ICCM*, 2007.

Rado Mihalcea, Corley Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 775. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press, 2006.

George A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

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

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics, 1993.

Alexander Panchenko. Can we automatically reproduce semantic relations of an information retrieval thesaurus? In *4th Russian Summer School in Information Retrieval*, pages 13–18. Voronezh State University, 2010.

Siddharth Patwardhan and Ted Pedersen. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. *Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, page 1, 2006.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. WordNet:: Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004 on XX*, pages 38–41. Association for Computational Linguistics, 2004.

Yves Peirsman, Kris Heylen, and Dirk Speelman. Putting things in order. First and second order context models for the calculation of semantic similarity. *Proceedings of the 9th Journées internationales d'Analyse statistique des Données Textuelles (JADT 2008)*, pages 907–916, 2008.

Philip Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence.*, volume 1, pages 448–453, 1995.

H. Rubenstein and J.B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

Magnus Sahlgren. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Stockholm University, 2006.

Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. pages 44–49, 1994.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems (NIPS)*, 17:1297–1304, 2004.

Peter Turney. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the twelfth european conference on machine learning (ecml-2001)*, 2001.

Tim Van de Cruys. *Mining for Meaning: The Extraction of Lexicosemantic Knowledge from Text*. PhD thesis, University of Groningen, 2010.

Tonio Wandmacher. How semantic is Latent Semantic Analysis? *Proceedings of TALN/RECITAL*, 2005.

Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.