A Study of Heterogeneous Similarity Measures for Semantic Relation Extraction

Alexander Panchenko
Center for Natural Language Processing (CENTAL), Université catholique de Louvain
College Erasme, 1 place Blaise Pascal, B-1348 Louvain-la-Neuve (Belgium)
alexander.panchenko@student.uclouvain.be

RÉSUMÉ

Etude des mesures de similarité hétérogènes pour l'extraction de relations sémantiques

L'article évalue un éventail de mesures de similarité qui ont pour but de prédire les scores de similarité sémantique et les relations sémantiques qui s'établissent entre deux termes, et étudie les moyens de combiner ces mesures. Nous présentons une analyse comparative à grande échelle de 34 mesures basées sur des réseaux sémantiques, le Web, des corpus, ainsi que des définitions. L'article met en évidence les forces et les faiblesses de chaque approche en contexte de l'extraction de relations. Enfin, deux techniques de combinaison de mesures sont décrites et testées. Les résultats montrent que les mesures combinées sont plus performantes que toutes les mesures simples et aboutissent à une corrélation de 0,887 et une Precision(20) de 0,979.

ABSTRACT

This paper evaluates a wide range of heterogeneous semantic similarity measures on the task of predicting semantic similarity scores and the task of predicting semantic relations that hold between two terms, and investigates ways to combine these measures. We present a large-scale benchmarking of 34 knowledge-, web-, corpus-, and definition-based similarity measures. The strengths and weaknesses of each approach regarding relation extraction are discussed. Finally, we describe and test two techniques for measure combination. These combined measures outperform all single measures, achieving a correlation of 0.887 and Precision(20) of 0.979.

MOTS-CLÉS : Similarité sémantique, Relations sémantiques, Similarité distributionnelle.

KEYWORDS: Semantic Similarity, Semantic Relations, Distributional Similarity.

1 Introduction

Semantic relations provide information about terms which have similar or related *meanings*. This kind of knowledge about language has proven to be valuable for various *NLP applications*, such as word sense disambiguation (Patwardhan *et al.*, 2003), query expansion (Hsu *et al.*, 2006), document categorization (Tikk *et al.*, 2003), or question answering (Sun *et al.*, 2005).

Let R be a set of synonymy, hypernymy, co-hypernymy, and associative relations between a set of terms C, established manually. A semantic relation extraction aims at discovering relations

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 29–42,
Grenoble, 4 au 8 juin 2012.

$\hat{R} \subseteq C \times C$ which would be as close to R as possible in terms of precision and recall :

$$\hat{R}^* = \arg\max_{\hat{R}} \frac{Precision(R,\hat{R}) \cdot Recall(R,\hat{R})}{Precision(R,\hat{R}) + Recall(R,\hat{R})}, Precision(R,\hat{R}) = \frac{|R \cap \hat{R}|}{|\hat{R}|}, Recall(R,\hat{R}) = \frac{|R \cap \hat{R}|}{|R|}.$$

The quality of the relations provided by existing extraction methods is still lower than the quality of manually constructed relations (see Section 5). This motivates the development of new relation extraction techniques.

One common approach to relation extraction is based on lexico-syntactic patterns such as those proposed by Hearst (1992). We use another extraction principle based on a *semantic similarity measure* between terms. The studied methods extract or recall pairs of semantically similar terms $\langle c_i, c_j \rangle$, but do not return the type of the relationship between them. Nonetheless, we suppose that the extractors must retrieve a mix of synonyms, hypernyms, co-hypernyms, and associations for practical use in NLP systems.

Existing similarity measures rely on one of these four sources of information – semantic networks (Resnik, 1995), Web corpus (Cilibrasi et Vitanyi, 2007), traditional corpora (Lin, 1998b), definitions of dictionaries (Lesk, 1986) or encyclopedia (Zesch *et al.*, 2008a). Prior research (Sahlgren, 2006; Heylen *et al.*, 2008; Panchenko, 2011) suggests that measures based on these sources of information are complementary. The goals of this work is to compare measures based on these four sources of information, and meta-measures combining information from different sources.

The main contributions of this paper are twofold. First, we present a comparative study of the heterogeneous baseline similarity measures. Several authors compared existing measures (see Section 5), but we do it on a large scale. We are the first to compare as many as 34 similarity measures based on the four sources of information listed above. Second, we present two combined metrics which use all the four information sources to calculate similarity (semantic networks, Web corpora, corpora, and definitions). Our experiments show that the measures based on complementary sources of information outperform all baseline measures by a wide margin achieving a correlation with human judgements up to 0.887 and Precision(20) up to 0.979 for the relation extraction task from a closed number of word pairs.

2 Similarity Measures

This section describes 34 knowledge-, web-, corpus-, and definition-based similarity measures, studied in this paper, as well as two combined measures.

Knowledge-based Measures We tested 6 knowledge-based measures based on WordNet (Miller, 1995) and SemCor corpus (Miller *et al.*, 1993) [1] : Inverted Edge Count (Jurafsky et Martin, 2009, p. 687), Leacock et Chodorow (1998), Resnik (1995), Jiang et Conrath (1997), Lin (1998a), and Wu et Palmer (1994). These measures use the following variables to compute the similarities : length of the shortest path in the network between terms c_i and c_j ; length of the shortest path from c_i to the lowest common subsumer (LCS) of c_i and c_j ; length of the shortest path from the root term to the LCS of c_i and c_j ; probability of c, estimated from a corpus ; probability of the LCS of c_i and c_j.

1. We used the implementation available in the package WordNet::Similarity (Pedersen *et al.*, 2004).

The complexity of the knowledge-based measures is mainly bounded by the computation time of the shortest paths between the nodes of the network. A limitation of these measures is that similarities can only be calculated between the 155.287 English terms encoded in the WordNet 3.0. For instance, since the named entity "TALN" is not present in WordNet, no relations between "TALN" and other words can be retrieved. Therefore, these measures are only able to *recall* provided beforehand lexico-semantic knowledge.

Web-based Measures Web-based metrics use the Web as a corpus in order to calculate similarities. They rely on the number of times terms co-occur in documents indexed by a Web search engine. In particular, web-based measures rely on the number of documents (hits) returned by the system by the query "c_i" and the number of hits returned by the query "c_i AND c_j".

We tested 9 measures relying either on Normalized Google Distance (NGD) (Cilibrasi et Vitanyi, 2007) or on Pointwise Mutual Information (PMI-IR) formula (Turney, 2001). We experimented with 5 NGD measures based respectively on BING, YAHOO, YAHOOBOSS, GOOGLE, and GOOGLE over the domain `wikipedia.org`, and with 4 PMI-IR measures based respectively on BING, YAHOOBOSS, GOOGLE, and GOOGLE over the domain `wikipedia.org`. [2]

The complexity of the web-based measures is mainly bounded by the maximum number of queries per second. For instance, BING allows not more than 7 queries per second for free ; GOOGLE allows 100 queries per day for free or 1000 queries for 5$; YAHOO asks 0.80$ for 1000 queries [3]. Web-based measures provide huge coverage of vocabulary in tens of languages. Therefore they are able to extract *new* lexico-semantic knowledge.

Corpus-based Measures We experimented with 13 measures which calculate the similarity between terms based on statistics derived from a corpus. Ten of them are based on the Distributional Analysis (Sahlgren, 2006; Curran, 2003). These distributional measures use 800M token corpus WACYPEDIA (Baroni *et al.*, 2009) tagged with TREETAGGER (Schmid, 1994) and dependency-parsed with MALTPARSER (Hall *et al.*, 2011). The distributional measures use context window or syntactic context techniques to calculate the similarities.

Our implementation of the distributional measures builds a feature matrix $\mathbf{F}$ from a corpus D, such that each term $c_i \in C$ is represented with a row-vector $\mathbf{f}_i$. The feature matrix is then normalized with Pointwise Mutual Information :

$$f_{ij} = log \frac{P(c_i, f_j)}{P(c_i)P(f_j)} = log \frac{f_{ij}}{n(c_i)\sum_i f_{ij}}. \quad (1)$$

Here, f_{ij} is an element of $\mathbf{F}$ is the number of times term c_i was represented with the feature f_j, $n(c_i)$ is the frequency of term c_i in the corpus. Finally, the similarity between the terms c_i and c_j is computed as the cosine between their respective feature vectors $\mathbf{f}_i, \mathbf{f}_j$:

$$s_{ij} = sim(c_i, c_j) = \frac{\mathbf{f}_i \cdot \mathbf{f}_j}{\left\|\mathbf{f}_i\right\| \left\|\mathbf{f}_j\right\|}. \quad (2)$$

Our choice of cosine among other metrics is in line with previous findings (Curran, 2003; Panchenko, 2011). The different distributional measures only vary in the way they build feature

2. Our own system is used in the experiments with measures based on BING (`http://www.bing.com/toolbox/bingdeveloper/`) and YAHOOBOSS (`http://developer.yahoo.com/search/boss/`), and Measures of Semantic Relatedness (MSR) web service (`http://cwl-projects.cogsci.rpi.edu/msr/`) is used for the measures based on GOOGLE and YAHOO !.

3. These rates were up-to date on April 2012. It is likely that the Bing API will be commercialized in future similarly to the YahooBoss.

vectors. The first seven measures perform a Bag-of-words Distributional Analysis (BDA). So, they construct the feature matrix **F** with the context window technique (Van de Cruys, 2010). We tested seven context window sizes – 1, 2, 3, 5, 8, 10 words, and a sentence. A term is represented with a bag of lemmas from a context window, passing a stop-word filter (around 900 words) and a stop part-of-speech filter (nouns, adjectives and verbs are kept).

The other three measures perform Syntactic Distributional Analysis (SDA). So, they construct the feature matrix **F** with the syntactic context technique (Lin, 1998b; Van de Cruys, 2010). Let the term $c_i =$ "cat" be linked with syntactic dependency $dt_j =$ OBJ with the word $w_k =$ "catch". Syntactic context of the term c_i is a bag of dependency-word pairs linked to it $\{\langle dt_j, w_k\rangle : w_k \notin Stoplist \wedge dt_k \in DT\}$, where DT is a set of dependency types used by a measure. [4]

In addition to these 10 distributional measures, we test 3 corpus-based measures available via the MSR web service. Two of them are based on the Factiva corpus (Veksler *et al.*, 2008), and use NGD and PMI-IR similarity functions (see above). The third measure rely on the Latent Semantic Analysis (Landauer et Dumais, 1997), trained on the TASA corpus (Veksler *et al.*, 2008). LSA calculates the similarity of terms with cosine (2) between term vectors in the "concept space".

The complexity of the corpus-based measures is mainly bounded by the time required to pre-process a corpus. In that respect, NGD and PMI-IR are the fastest methods, since they only require a corpus to be indexed in a standard way. BDA require more computational resources since pairwise similarities should be calculated between high-dimensional term vectors. Finally, LSA and SDA are the least scalable methods since the former performs a computationally heavy singular value decomposition of the term-document matrix, and the latter requires dependency parsing of the corpus. Similarly to web-based methods, corpus-based measures are able to extract relations between unknown terms. However, extraction capability of such measures is limited by the corpus – if "TALN" does not occur in the text then it would be impossible to obtain its relations.

Definition-based Measures We experimented with 6 measures which rely on explicit definitions of terms. The first four measures use definitions and relations of Wiktionary and abstracts of Wikipedia. [5] Our implementation of these four measures is similar to the techniques proposed by Zesch *et al.* (2008b). Our measures are different from the previously proposed in three aspects : (a) they represent each term c_i as a bag-of-words vector, while the measures of Zesch *et al.* (2008b) represent terms as concept vectors [6] ; (b) we use both texts from Wiktionary and Wikipedia in order to represent a term, which is not the case in the original work ; (c) we use semantic relations listed in Wiktionary to update similarity scores.

Algorithm 1 depicts pseudocode of these measures. First, it builds the definitions D for input terms C from the information available in Wiktionary and Wikipedia. The function $get\_wiktionary\_def$ returns for each term $c \in C$ a text composed of glosses, examples, quotations, related words, and categories found in Wiktionary (all meanings corresponding to a surface form of c are used). We remove syntax- and etymology-related categories such as "English nouns" or "Japanese proper names" with a stoplist of 94 words, such as "noun" or "esperanto". Next, the function $get\_wikipedia\_def$ returns for each term c a short abstract from the corresponding

4. We tested three models which use 6, 9, or 21 types of syntactic dependencies : $DT_6 =$ { NMOD, SBJ, OBJ, COORD, AMOD, IOBJ} ; $DT_9 =$ { NMOD, ADV, SBJ, OBJ, VMOD, COORD, AMOD, PRN, IOBJ } ; $DT_{21} =$ { NMOD, P, PMOD, ADV, SBJ, OBJ, VMOD, COORD, CC, VC, DEP, PRD, AMOD, PRN, PRT, LGS, IOBJ, EXP, CLF, GAP }.

5. We experimented with data downloaded on October 2011 from `www.wiktionary.org` and `www.dbpedia.org`.

6. An element f_{ij} of a *concept vector* equals to tf.idf score of term c_i in the definition d_j, while an element of *bag-of-words vector* f_{ij} equals to normalized frequency of word c_j in the definition d_i of term c_i.

Algorithm 1: Wiktionary-based sim.measure

Input: Terms C, $UseWikipedia$, Number of features β
Output: Similarity matrix, $\mathbf{S}$ $[C \times C]$

$D \leftarrow get\_wiktionary\_def(C)$;
if $UseWikipedia$ **then**
  $D \leftarrow D \cup get\_wikipedia\_def(C)$
$\mathbf{F} \leftarrow construct\_fmatrix(C, D, \beta)$;
$\mathbf{F} \leftarrow pmi(\mathbf{F})$;
$\mathbf{S} \leftarrow cos(\mathbf{F})$;
$\mathbf{S} \leftarrow update\_similarity(\mathbf{S})$;
return $\mathbf{S}$;

Algorithm 2: Relation fusion sim.measure

Input: Sim.matrices produced by N measures $\{\mathbf{S}_1, \ldots, \mathbf{S}_N\}$, kNN threshold k
Output: Similarity matrix, $\mathbf{S}_{cmb}[C \times C]$

for i=1,N **do**
  $R_i \leftarrow threshold(\mathbf{S}_i, k)$;
  $\mathbf{R}_i \leftarrow relation\_matrix(R_i)$
$\mathbf{S}_{cmb} \leftarrow \frac{1}{N} \sum_{i=1}^{N} \mathbf{R}_i$;
return $\mathbf{S}_{cmb}$;

Wikipedia article (the name of the article must *exactly* match the term c). Next, the feature matrix $\mathbf{F}$ is constructed : each term $c_i \in C$ is represented as a bag-of-words vector $\mathbf{f}_i$, derived from its definition. These feature vectors are normalized with Pointwise Mutual Information (1). Pairwise similarities of terms are calculated with cosine (2). Finally, the pairwise similarities are corrected with the function $update\_similarity$. It assigns the highest similarity score to the pairs of terms which are directly related in Wiktionary :

$$s_{ij}^{updated} = \begin{cases} 1 & \text{if semantic relation } \langle c_i, c_j \rangle \text{ is listed in Wiktionary} \\ s_{ij} & \text{otherwise} \end{cases} \quad (3)$$

We tested four variations of this measure : two of them use only Wiktionary (1000 and 2500 features β), while the others use both Wiktionary and Wikipedia (1000 and 2500 features β). [7]

In addition to these four measures, we tested two measures based on WordNet glosses available in the package WordNet::Similarity : Extended Lesk (Banerjee et Pedersen, 2003) and Gloss Vectors (Patwardhan et Pedersen, 2006). The key difference between Wiktionary- and WordNet-based measures is that the latter uses definitions of related terms.

The complexity of the definition-based measures is mainly bounded by the time required to preprocess definitions and calculate pairwise similarities between them. In that respect, measures based on Wiktionary and WordNet are similar since they use the bag-of-word model to represent terms. The extraction capability of definition-based measures is limited by the number of available definitions. As of October 2011 WordNet contains 117.659 definitions (glosses) ; Wiktionary contains 536.594 definitions in English and 4.272.902 definitions on all languages ; Wikipedia has 3.866.773 English articles and 20.8 million of articles for all languages.

Combined Similarity Measures We tested two combination techniques – similarity and relation fusion. These methods take as input a set of similarity matrices $\{\mathbf{S}_1, \ldots, \mathbf{S}_N\}$ produced by N combined measures. The output of a combination is a similarity matrix $\mathbf{S}_{cmb}$.

Similarity fusion combines N similarity measures with a simple mean over their respective pairwise similarity scores : $\mathbf{S}_{cmb} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{S}_i$.

Relation fusion keeps only the best relations provided by each measure ; then all these relations are merged. First, the algorithm retrieves the relations extracted by single measures with function

7. We used the JWKTL library (Zesch *et al.*, 2008a) as an API to Wiktionary, and `DBpedia.org` as a source of Wikipedia abstracts. In particular, we used this version of abstracts : `http://downloads.dbpedia.org/3.7/en/long_abstracts_en.nt.bz2`

$threshold$ (a kNN technique described in Section 3). Then each set of relations R_i is encoded in an adjacency matrix $\mathbf{R}_i$. An element of this matrix indicates if the terms c_i and c_j are related :

$$r_{ij} = \begin{cases} 1 & \text{if } \langle c_i, c_j \rangle \in R_k \\ 0 & \text{else} \end{cases} \quad (4)$$

The final similarity score is an average over adjacency matrices (line 4). In our experiments we empirically chose an internal kNN threshold k of 20%.

Expert approach was used to compose three groups of measures of the 34 measures. These groups of measures are combined with two techniques described above. The first group contains 4 measures (see Tables 1 and 2) : WN-Resnik, BDA-3-5000, SDA-21-100000, Def-WktWiki-1000. The second group contains 8 measures – the 4 previous ones plus WN-WuPalmer, LSA-Tasa, Def-GlossVec., and Def-Ext.Lesk. The third group contains 14 measures – the 8 previous ones plus WN-LeacockChodorow, WN-Lin, WN-JiangConrath, NGD-Factiva, NGD-Yahoo, and NGD-GoogleWiki. The running time required to calculate a similarity with a combined measure is close to the sum of times required by the measures used in a combination.

3 Evaluation

Our comparison of similarity measures is based on human judgments about semantic similarity and on semantic relations fixed manually by lexicographers [8].

Human Judgements This kind of evaluation is a standard and simple way to assess a semantic similarity measure. We used three classical human judgement datasets – MC (Miller et Charles, 1991), RG (Rubenstein et Goodenough, 1965) and WordSim353 (Finkelstein *et al.*, 2001) composed of 30, 65, and 353 pairs of terms respectively. Each of these datasets is composed of N tuples $\langle c_i, c_j, s_{ij} \rangle$, where c_i, c_j are terms, and s_{ij} is their similarity obtained by human judgement. Let $\mathbf{s} = (s_{i1}, s_{i2}, \ldots, s_{iN})$ be a vector of ground truth scores, and $\hat{\mathbf{s}} = (\hat{s}_{i1}, \hat{s}_{i2}, \ldots, \hat{s}_{iN})$ be a vector of similarity scores calculated by a measure. Then, the quality of the measure is assessed with Pearson and Spearman's correlation between $\mathbf{s}$ and $\hat{\mathbf{s}}$.

Semantic Relations This ground truth is composed of semantic relations $\langle c_i, type, c_j \rangle$, such as ⟨agitator, synonym, activist⟩, ⟨dishwasher, co-hyponym, freezer⟩, ⟨hawk , hypernym, predator⟩, and ⟨gun, synonym,weapon⟩. The dataset contains both meaningful and random relations. The evaluation is based on the number of correctly ranked relations. In order to extract relations R between a set of terms C, we follow a standard procedure. First, pairwise similarities between terms are calculated and saved in a $[C \times C]$ similarity matrix $\mathbf{S}$. The similarity scores are mapped to the interval $[0; 1]$. Second, each term c_i is linked with k% of its nearest neighbours : $\hat{R} = \bigcup_{i=1}^{|C|} \left\{ \left\langle c_i, c_j \right\rangle : (c_j \in \text{ top } k\% \text{ terms of } c_i) \wedge (s_{ij} \geq 0) \right\}, s_{ij} \in \mathbf{S}$.

Let $\hat{R}_k$ be a set containing top k % semantic relations for each target word c_i, and R be a set of all correct semantic relations. Then, Precision, Recall, F1-measure at k are calculated as follows : $P(k) = \frac{|R \cap \hat{R}_k|}{|\hat{R}_k|}, R(k) = \frac{|R \cap \hat{R}_k|}{|R|}, F(k) = \frac{P(k) \cdot R(k)}{P(k)+R(k)}$. Each "target" term c_i has roughly the same number of meaningful and random relations. That is why for a random measure $P(50) \approx 0.5$ and not $\frac{|\hat{R}|}{|C^2|} \approx 0$ as in the case of an open vocabulary relation extraction. We argue that this kind

8. Evaluation datasets and scripts are available at : `http://cental.fltr.ucl.ac.be/team/~panchenko/sre-eval/`

of evaluation should give a good idea about the relative performances of different measures. However, the performance scores in this evaluation should not be confused with the performance scores in an open-vocabulary relation extraction task. In this work, the quality of a similarity measure is assessed with the four statistics : $P(10)$, $P(20)$, $P(50)$, $F(50)$.

We used two semantic relation datasets : BLESS (Baroni et Lenci, 2011), and SN. The first one relates 200 target terms (100 animate and 100 inanimate nouns) to 8625 relatum terms with 26.554 semantic relations (14.440 are meaningful and 12.154 are random). Every relation has one of the following types : hypernymy, co-hypernymy, meronymy, attribute, event, or random. We built the SN (Semantic Neighbors) dataset in order to complement the BLESS, because it contains no synonyms. [9] SN relates 462 target terms (nouns) to 5910 relatum terms with 14.682 semantic relations (7341 are meaningful and 7341 are random). The SN contains synonyms coming from three sources : WordNet 3.0 (Miller, 1995), Roget's thesaurus (Kennedy et Szpakowicz, 2008), and a synonyms database [10].

4 Results

Human Judgements Table 1 presents correlations of the 34 single and the 3 combined measures with human judgements. We ranked the measures according to their Spearman's correlation. The best measures in each group (knowledge-, web-based etc.) are in bold. We observed that correlations of most web-based measures with human judgements are low and not significant in most of the cases. PMI-IR and NGD over Wikipedia are two exceptions. They provided the best results among the web measures. However, generally, knowledge-, corpus-, and definition-based measures perform far better than those relying on the Web as a corpus. Particularly high correlations with human judgements were observed for the following single similarity measures : *WN-Resnik, SDA-21-100000, Def-WktWiki-1000, BDA-3-5000*, and *WN-LeacockChodorow*. However, the similarity fusion of 14 measures *Cmb-Avg-14* outperformed all single measures on MC and RG datasets. In the same time, similarity fusion of 8 measures (*Cmb-Avg-8*) was better that any single measure on the WordSim353 pairs.

Semantic Relations Table 2 presents performance of the measures at relation extraction. We ranked the measures according to P(20) and P(50) statistics. We would like to recall that our evaluation procedure is different from an open vocabulary extraction and a random measure would achieve $P(50) \approx 0.5$ (see the first line of Table 2). The knowledge-,web-, corpus-,and definition-based measures are grouped and the best metrics in each group are in bold. Figure 1(c) depicts Precision-Recall graph of four variations of the definition-based measures. The following single measures provided the best scores in this evaluation : *WN-Resnik, SDA-21-100000, BDA-3-5000, Def-WktWiki-1000*, and *WN-WuPalmer*.

Our experiments showed that measures which use both Wiktionary and Wikipedia (denoted as *Def-WktWiki-**) are better on most of the datasets than measures relying only on Wiktionary (*Def-Wkt-**). In particular, *Def-WktWiki-1000* outperformed all definition-based measures, including those based on WordNet. On the BLESS dataset, the syntactic distributional analysis *SDA-21-10000* achieved the best precision among the single measures (0.953), while bag-of-words distributional analysis *BDA-3-5000* achieved the highest recall (0.835). On the SN dataset, the

9. SN dataset is available at `http://cental.fltr.ucl.ac.be/team/~panchenko/sre-eval/sn.csv`
10. `http://synonyms-database.downloadaces.com/`

WordNet-based measure WN-WuPalmer performed best achieving P(20) of 0.959 and P(50) of 0.764. However, the relation fusion of 8 measures (*Cmb-Rel-8*) outperformed all single measures on both datasets achieving P(20) of 0.975 and P(50) of 0.802 on the BLESS and P(20) of 0.971 and P(50) of 0.760 on the SN dataset.

Summary Results obtained on the human judgements and semantic relation datasets are overlapping but not identical. We used the following criterion in order to decide which measures are the best : a measure should be the best in its group (e. g. among corpus-based measures) in both types of evaluations. According to this criterion, the best single metrics are the WordNet measure *WN-Resnik*, the bag-of-words distributional measure *BDA-3-5000*, the syntactic distributional measure *SDA-21-100000*, and the measure *Def-WktWiki-1000* based on Wiktionary and Wikipedia. Figure 1 depicts distributions of similarity scores for these four most successful metrics. Our experiments showed that, for these measures there is a significant difference in distributions of scores of meaningful and random relations. This means that an appropriate kNN threshold level k clearly separates meaningful relations from the random ones. The best combined measure and the best measure overall is *Cmb-Rel-8*. It is based on the eight following measures : *WN-Resnik, BDA-3-5000, SDA-21-100000, Def-WktWiki-1000, WN-WuPalmer, LSA-Tasa, Def-GlossVec.*, and *Def-Ext.Lesk*. This result is interesting as combination of the four strongest measures (those listed in Figure 1 and denoted as *Cmb-*-4* in Tables 1 and 2) can benefit of redundancy provided by the additional weaker measures. Our results suggest that performance of the combinations based on 14 measures is very close to the performance of *Cmb-Rel-8* (see Figure 1(b) and Table 3). Thus, redundancy provided by the additional 6 measures does not improve the results with respect to the set of 8 measures.

Sim.Measure	MC Dataset		RG Dataset		WordSim353 Dataset	
	Pearson	*Spearman*	*Pearson*	*Spearman*	*Pearson*	*Spearman*
Random	0.172 ***	0.056 ***	-0.060 ***	-0.047 ***	-0.158 ***	-0.122 ***
WN-Resnik	0.823	**0.784**	0.823	0.757	0.350	**0.330**
WN-Short.Path	0.755	0.724	0.782	0.788	0.366	0.290
WN-Leack.Chod.	0.779	0.724	0.841	**0.789**	0.313	0.295
WN-WuPalmer	0.768	0.742	0.800	0.775	0.270	0.330
WN-Lin	0.769	0.754	0.737	0.619	0.287	0.203
WN-JiangConrath	0.473 *	0.719	0.575	0.587	0.227	0.175
NGD-Bing	0.035 ***	0.063 ***	0.174 ***	0.181 ***	0.042 ***	0.058 ***
NGD-Yahoo	0.387 **	0.330 ***	0.448	0.445	0.290	0.254
NGD-Google	0.085 ***	0.019 ***	-0.013 ***	-0.012 ***	0.120 **	0.150 *
NGD-GoogleWiki	0.306 ***	0.334 ***	0.452	**0.501**	0.205	0.250
PMI-IR-Bing	0.079 ***	0.120 ***	0.116 ***	0.149 ***	0.000 ***	0.003 ***
PMI-IR-Google	0.046 ***	-0.107 ***	-0.061 ***	-0.039 ***	0.097 ***	0.113 **
PMI-IR-GoogleWiki	0.508 *	**0.498** *	0.401	0.411	0.254	**0.279**
BDA-sent-10000	0.642	0.638	0.694	0.703	0.383	0.362
BDA-1-5000	0.658	0.676	0.704	0.758	0.448	0.438
BDA-2-5000	0.667	0.638	0.698	0.734	0.441	0.439
BDA-3-5000	0.722	0.692	0.752	0.782	0.467	**0.465**
BDA-5-5000	0.710	0.683	0.755	**0.787**	0.467	0.455
BDA-8-5000	0.707	0.697	0.746	0.764	0.455	0.440
BDA-10-5000	0.710	0.718	0.746	0.764	0.443	0.425
SDA-6-100000	0.759	0.790	0.741	0.792	0.380	0.496
SDA-9-100000	0.756	0.790	0.732	0.787	0.384	0.491
SDA-21-100000	0.756	**0.790**	0.731	0.785	0.384	0.490
LSA-Tasa	0.737	0.694	0.645	0.604	0.527	0.565
NGD-Factiva	0.602	0.602	0.618	0.599	0.565	0.599
PMI-Factiva	0.312 ***	0.442 **	0.436	0.517	0.314	0.559
Def-WN-GlossVec.	0.566	0.653	0.647	0.738	0.383	0.322
Def-WN-Ext.Lesk	0.355 ***	0.792	0.340 *	0.717	0.209	0.409
Def-Wkt-1000	0.625	0.687	0.655	0.760	0.416	0.492
Def-Wkt-2500	0.625	0.687	0.655	0.760	0.382	0.527
Def-WktWiki-1000	0.704	**0.759**	0.701	**0.754**	0.453	**0.545**
Def-WktWiki-2500	0.704	0.759	0.701	0.754	0.416	0.520
Cmb-Avg-4	0.847	0.859	0.867	0.887	0.500	0.508
Cmb-Avg-8	0.858	0.858	0.867	0.883	0.537	**0.555**
Cmb-Avg-14	0.847	**0.859**	0.867	**0.887**	0.500	0.508

TABLE 1 – Evaluation on the human judgement datasets (MC, RG, and WordSim353). Here (*) means $p \leq 0.01$, (**) means $p \leq 0.05$, (***) means $p > 0.05$, otherwise $p \leq 0.001$. The best results for each group of measures are in bold. The very best results are in grey.

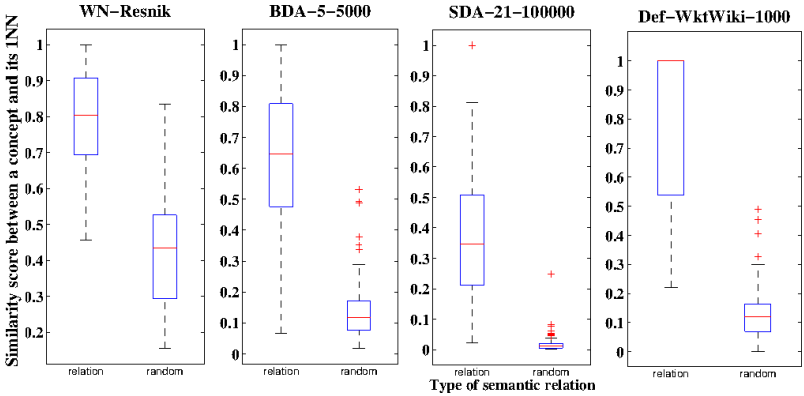

FIGURE 1 – Distribution of 1-NN similarity scores of the four best single measures on the BLESS dataset. Here "random" and "relation" are distributions of scores between random and meaningful relations. The distributions were calculated as suggested in (Baroni et Lenci, 2011).

Sim.Measure	BLESS Dataset				SN Dataset			
	P(10)	*P (20)*	*P(50)*	*F(50)*	*P(10)*	*P(20)*	*P(50)*	*F(50)*
Random	0.546	0.541	0.543	0.522	0.504	0.501	0.498	0.498
WN-Resnik	0.977	**0.958**	**0.718**	0.690	0.948	0.908	0.725	0.725
WN-Short.Path	0.967	0.925	0.722	0.693	0.981	0.947	0.752	0.752
WN-Leack.Chod.	0.967	0.925	0.722	0.693	0.982	0.951	0.756	0.756
WN-WuPalmer	0.978	0.938	0.706	0.678	0.979	**0.959**	**0.764**	0.764
WN-Lin	0.975	0.919	0.776	0.745	0.924	0.853	0.637	0.637
WN-JiangConrath	0.981	0.909	0.732	0.703	0.916	0.835	0.615	0.615
NGD-Bing	0.725	0.692	0.695	0.670	0.676	0.682	0.639	0.639
NGD-Yahoo	0.940	0.907	**0.782**	0.751	—	—	—	—
NGD-YahooBoss	0.847	0.843	0.747	0.718	—	—	—	—
NGD-Google	0.991	**0.934**	0.651	0.625	—	—	—	—
NGD-GoogleWiki	0.874	0.836	0.702	0.674	—	—	—	—
PMI-IR-Bing	0.675	0.650	0.692	0.667	0.610	0.608	0.647	0.647
PMI-IR-YahooBOSS	0.823	0.822	0.724	0.696	—	—	—	—
PMI-IR-Google	0.822	0.749	0.660	0.634	—	—	—	—
PMI-IR-GoogleWiki	0.791	0.761	0.676	0.649	—	—	—	—
BDA-sent-10000	0.962	0.920	0.799	0.767	0.941	0.898	0.724	0.724
BDA-1-5000	0.971	0.940	0.826	0.793	0.969	0.926	0.737	0.737
BDA-2-5000	0.966	0.939	0.829	0.796	0.970	0.929	0.738	0.738
BDA-3-5000	0.970	0.947	**0.835**	0.802	0.974	0.932	0.743	0.743
BDA-5-5000	0.975	0.946	0.833	0.800	0.971	0.929	0.744	0.744
BDA-8-5000	0.974	0.943	0.827	0.794	0.968	0.924	0.741	0.741
BDA-10-5000	0.972	0.941	0.821	0.789	0.962	0.922	0.737	0.737
SDA-6-100000	0.984	0.948	0.810	0.778	0.978	0.945	0.749	0.749
SDA-9-100000	0.984	0.951	0.809	0.777	0.977	0.945	0.753	0.753
SDA-21-100000	0.985	**0.953**	0.810	0.778	0.978	**0.946**	**0.753**	0.753
LSA-Tasa	0.967	0.936	0.801	0.769	0.901	0.839	0.637	0.637
NGD-Factiva	0.959	0.916	0.800	0.768	0.900	0.832	0.651	0.651
PMI-Factiva	0.903	0.860	0.816	0.784	0.826	0.768	0.606	0.606
Def-WN-GlossVec.	0.894	0.860	0.742	0.712	0.930	0.872	0.719	0.719
Def-WN-Ext.Lesk	0.940	0.870	0.716	0.687	0.950	**0.895**	0.653	0.653
Def-Wkt-1000	0.926	0.885	0.783	0.752	0.907	0.868	0.678	0.678
Def-Wkt-2500	0.915	0.882	0.754	0.754	0.928	0.898	0.704	0.704
Def-WktWiki-1000	0.942	**0.905**	**0.785**	0.725	0.917	0.878	0.696	0.696
Def-WktWiki-2500	0.931	0.891	0.765	0.734	0.937	0.912	**0.726**	0.726
Cmb-Avg-4	0.992	0.969	0.787	0.756	0.980	0.952	0.768	0.768
Cmb-Rel-4	0.989	0.970	0.737	0.708	0.975	0.943	0.696	0.696
Cmb-Avg-8	0.994	0.974	0.774	0.743	0.955	0.875	0.660	0.660
Cmb-Rel-8	0.994	0.975	**0.802**	0.770	0.989	**0.971**	**0.760**	0.760
Cmb-Avg-14	0.994	**0.979**	0.792	0.760	0.957	0.880	0.663	0.663
Cmb-Rel-14	0.994	0.973	0.811	0.779	0.987	0.966	0.759	0.759

TABLE 2 – Evaluation of the measures on the semantic relation datasets (BLESS and SN). Here *P(x)*, and *F(x)* are Precision, and F-measure as specified in Section 3. The best results for each group of measures are in bold. The very best results are in grey.

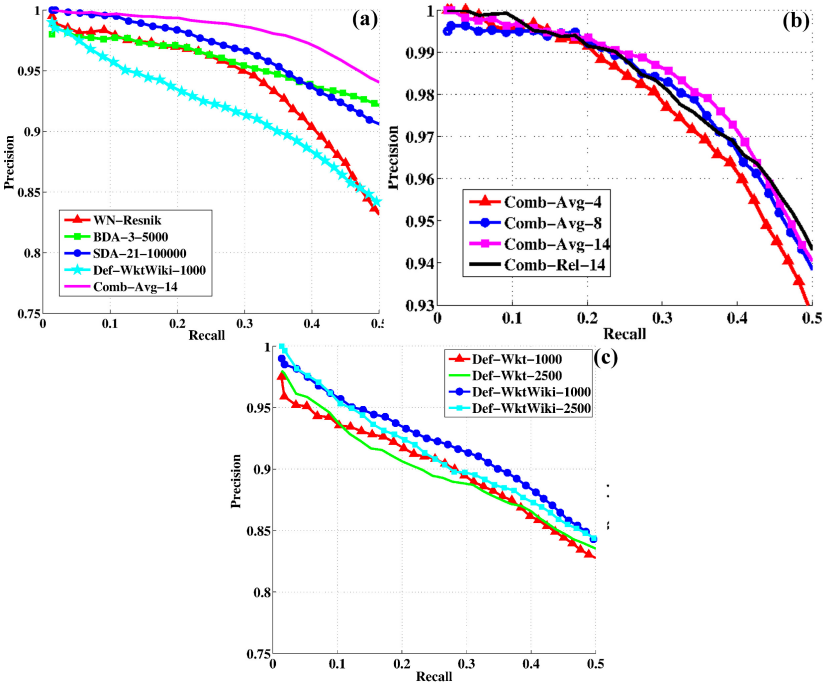

FIGURE 2 – Precision-Recall graphs of (a) the best single and combined measures ; (b) four combined measures ; (c) measures based on Wiktionary and Wikipedia.

Discussion There is a huge difference in performance between web-based and corpus-based measures. This is likely to be due to the noisy nature of the web documents (BDA/SDA use a more precise and linguistically motivated representation of a term) and the fact that the counts of a search engine API are rough approximations of the real counts. Similarly, the higher performance of the knowledge- and definition-based methods is likely due to the more linguistically precise representation of the terms. Some web measures yield significantly worst results than others. Following (Veksler *et al.*, 2008), we suggest that the variance in the results are due to differences in the corpora indexed by different search engines. For instance, Web measures over Wikipedia or Factiva provide better results since this corpora contain less noisy documents than the heterogeneous Web collection indexed by Bing.

Combined measures achieve higher precision and recall with respect to the single measures. First, this is due to the reuse of common lexico-semantic information (such as “car” being a synonym of “vechicle”) via knowledge- and definition-based measures. Measures based on WordNet and dictionary definitions achieve high precision as they rely on fine-grained manually constructed

resources. However, due to limited coverage of these resources they can only determine relations between a limited number of terms. On the other hand, measures based on web and corpora are nearly unlimited in their coverage, but provide less precise results. Combination of the measures let us keep high precision for frequent terms present in WordNet and dictionaries and at the same time calculate relations between rare terms unlisted in the handcrafted resources with web and corpus measures.

Second, combinations work well because, as it was found in previous research (Sahlgren, 2006; Heylen *et al.*, 2008; Panchenko, 2011), different measures provide complementary types of semantic relations. For instance, WordNet-based measures score high hypernyms, distributional analysis score high co-hypernymy and synonyms, etc. In that respect, a combination helps to recall more diverse relations. For example, a WordNet-based measure may return the hyponym ⟨salmon, seafood⟩, while a corpus-based measure would extract the co-hypernym ⟨salmon, mackerel⟩.

5 Related Work

There exists a significant body of literature about single measures discussed in this paper. However, just a few works compared different measures and their combinations. Furthermore, even less people evaluated the performance of these measures on the relation extraction task. One notable exception is the work of Curran et Moens (2002). The authors evaluated nine BDA measures and 14 weight functions and reported $Precision(5)$ of 0.52, and $Precision(10)$ of 0.45 for the best measure – Jaccard similarity with t-test weight function. Van de Cruys (2010) studied distributional measures and reported that : the optimal context window sizes for BDA is 2-5 words ; SDA is the best distributional measure. Budiu *et al.* (2007) compared LSA, PMI-IR, and GLSA. The authors found that GLSA performs better on the synonymy tests, while PMI-IR works better on the human judgement datasets. Agirre *et al.* (2009) compared 3 WordNet-based and 20 distributional measures (BDA and SDA) as well as their combinations. The authors found that a supervised combination of distributional and WordNet measures outperforms all measures on all datasets. Similarity measures which rely on Wikipedia, Wiktionary, WordNet and their combinations are described in the work of Zesch *et al.* (2007, 2008b). Navarro *et al.* (2009) described another method for extraction of synonyms from Wiktionary. Two promising measures which rely on Wikipedia were proposed by Strube et Ponzetto (2006) and Gabrilovich et Markovitch (2007).

Some studies compare the measures in context of NLP applications. For instance, Mihalcea *et al.* (2006) studied PMI-IR, LSA, and six WordNet-based measures on the text similarity task. The authors found that PMI-IR and Resnik are best corpus- and knowledge-based measures correspondingly ; and that an average over eight measures outperforms single measures. Budanitsky et Hirst (2006) found that the WN-JiangConrath is the best knowledge-based measure for the spelling correction application. Patwardhan et Pedersen (2006) report the same result for the task of word sense disambiguation. SDA was used by Grefenstette (1994) to induce a thesaurus.

In prior research, some attempts were made to combine baseline measures, including (Curran, 2002; Cederberg et Widdows, 2003; Mihalcea *et al.*, 2006; Agirre *et al.*, 2009). However, those studies did not take into account the whole range of existing information sources.

6 Conclusion

In this paper we compared 34 knowledge-, corpus-, web-, and definition-based measures on the task of predicting semantic similarity scores and semantic relations that hold between two terms. We also described and tested two techniques for measure combination. Our results show that the combined measures outperform all single measures achieving a correlation of 0.887 on RG dataset and $Precision(20)$ of 0.979 on the BLESS dataset. In the future research, we are going to estimate the precision of the relation extraction on the whole vocabulary C. The obtained relations will be applied in context of text classification and query expansion applications.

Références

AGIRRE, E., ALFONSECA, E., HALL, K., KRAVALOVA, J., PAŞCA, M. et SOROA, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. *In Proceedings of NAACL-HLT 2009*, pages 19–27.

BANERJEE, S. et PEDERSEN, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. *In IJCAI*, volume 18, pages 805–810.

BARONI, M., BERNARDINI, S., FERRARESI, A. et ZANCHETTA, E. (2009). The wacky wide web : A collection of very large linguistically processed web-crawled corpora. *LREC*, 43(3):209–226.

BARONI, M. et LENCI, A. (2011). How we blessed distributional semantic evaluation. *Workshop on Geometrical Models of Natural Language Semantics, EMNLP 2011*, pages 1–11.

BUDANITSKY, A. et HIRST, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

BUDIU, R., ROYER, C. et PIROLLI, P. (2007). Modeling information scent : A comparison of lsa, pmi and glsa similarity measures on common tests and corpora. pages 314–332. In RIAO.

CEDERBERG, S. et WIDDOWS, D. (2003). Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. *In Proceedings HLT-NAACL*, pages 111–118.

CILIBRASI, R. L. et VITANYI, P. M. B. (2007). The Google Similarity Distance. *IEEE Trans. on Knowl. and Data Eng.*, 19(3):370–383.

CURRAN, J. R. (2002). Ensemble methods for automatic thesaurus extraction. *In Proceedings of the EMNLP-02*, pages 222–229. ACL.

CURRAN, J. R. (2003). *From distributional to semantic similarity*. Thèse de doctorat, University of Edinburgh.

CURRAN, J. R. et MOENS, M. (2002). Improvements in automatic thesaurus extraction. *In Proceedings of the ACL-02 workshop on Unsupervised Lexical Acquisition*, pages 59–66.

FINKELSTEIN, L., GABRILOVICH, E., MATIAS, Y., RIVLIN, E., SOLAN, Z., WOLFMAN, G. et RUPPIN, E. (2001). Placing search in context : The concept revisited. *In WWW 2001*, pages 406–414. ACM.

GABRILOVICH, E. et MARKOVITCH, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. *In IJCAI*, volume 6, page 12.

GREFENSTETTE, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Springer.

HALL, J., NILSSON, J. et NIVRE, J. (2011). Single malt or blended ? a study in multilingual parser optimization. volume 43 de *Text, Speech and Language Technology*, pages 19–33. Springer.

HEARST, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *In Proceedings of the 14th conference on Computational linguistics*, pages 539–545. ACL.

HEYLEN, K., PEIRSMAN, Y., GEERAERTS, D. et SPEELMAN, D. (2008). Modelling word similarity : an evaluation of automatic synonymy extraction algorithms. *LREC'08*, pages 3243–3249.

HSU, M.-H., TSAI, M.-F. et CHEN, H.-H. (2006). Query expansion with conceptnet and wordnet : An intrinsic comparison. *Information Retrieval Technology*, pages 1–13.

JIANG, J. J. et CONRATH, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *In ROCLING X*, pages 19–33.

JURAFSKY, D. et MARTIN, J. H. (2009). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

KENNEDY, A. et SZPAKOWICZ, S. (2008). Evaluating rogets thesauri. *ACL-08 HLT*, pages 416–424.

LANDAUER, T. K. et DUMAIS, S. T. (1997). A solution to plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

LEACOCK, C. et CHODOROW, M. (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. *An Electronic Lexical Database*, pages 265–283.

LESK, M. (1986). Automatic sense disambiguation using machine readable dictionaries : how to tell a pine cone from an ice cream cone. *In Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM.

LIN, D. (1998a). An Information-Theoretic Definition of Similarity. *In In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304.

LIN, D. (1998b). Automatic retrieval and clustering of similar words. *In Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 768–774. ACL.

MIHALCEA, R., CORLEY, C. et STRAPPARAVA, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *In AAAI'06*, pages 775–780.

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

MILLER, G. A. et CHARLES, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

MILLER, G. A., LEACOCK, C., TENGI, R. et BUNKER, R. T. (1993). A semantic concordance. *In Proceedings of the workshop on Human Language Technology*, pages 303–308. ACL.

NAVARRO, E., SAJOUS, F., BRUNO, G., PRÉVOT, L., SHUKAI, H., TZU-YI, K., MAGISTRY, P. et CHU-REN, H. (2009). Wiktionary and nlp : improving synonymy networks. *In Proceedings of the 2009 Workshop on The People's Web Meets NLP : Collaboratively Constructed Semantic Resources*, People's Web '09, pages 19–27. Association for Computational Linguistics.

PANCHENKO, A. (2011). Comparison of the baseline knowledge-, corpus-, and web-based similarity measures for semantic relations extraction. *GEMS Workshop (EMNLP)*, pages 11–21.

PATWARDHAN, S., BANERJEE, S. et PEDERSEN, T. (2003). Using measures of semantic relatedness for word sense disambiguation. *In* GELBUKH, A., éditeur : *Computational Linguistics and Intelligent Text Processing*, volume 2588 de *LNCS*, pages 241–257. Springer Berlin.

PATWARDHAN, S. et PEDERSEN, T. (2006). Using WordNet-based context vectors to estimate the semantic relatedness of concepts. *Making Sense of Sense : Bringing Psycholinguistics and Computational Linguistics Together*, page 1.

PEDERSEN, T., PATWARDHAN, S. et MICHELIZZI, J. (2004). Wordnet : : Similarity : measuring the relatedness of concepts. *In Demonstration Papers at HLT-NAACL 2004*, pages 38–41. ACL.

RESNIK, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *In Proceedings of the 14th IJCAI conference.*, volume 1, pages 448–453.

RUBENSTEIN, H. et GOODENOUGH, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

SAHLGREN, M. (2006). *The Word-Space Model : Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Thèse de doctorat, Stockholm University.

SCHMID, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. pages 44–49.

STRUBE, M. et PONZETTO, S. P. (2006). Wikirelate ! computing semantic relatedness using wikipedia. *In Proceedings of the AAAI*, volume 21, pages 14–19.

SUN, R., JIANG, J., FAN, Y., HANG, T., TAT-SENG, C. et yen KAN, C. M. (2005). Using syntactic and semantic relation analysis in question answering. *In Proceedings of TREC*.

TIKK, D., YANG, J. D. et BANG, S. L. (2003). Hierarchical text categorization using fuzzy relational thesaurus. *KYBERNETIKA-PRAHA*, 39(5):583–600.

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

Van de CRUYS, T. (2010). *Mining for Meaning : The Extraction of Lexicosemantic Knowledge from Text*. Thèse de doctorat, University of Groningen.

VEKSLER, V. D., GOVOSTES, R. Z. et GRAY, W. D. (2008). Defining the dimensions of the human semantic space. *In 30th Annual Meeting of the Cognitive Science Society*, pages 1282–1287.

WU, Z. et PALMER, M. (1994). Verbs semantics and lexical selection. *In Proceedings of the 32nd meeting on Association for Computational Linguistics*, pages 133–138.

ZESCH, T., GUREVYCH, I. et MÜHLHÄUSER, M. (2007). Comparing wikipedia and german wordnet by evaluating semantic relatedness on multiple datasets. *In HLT-NAACL 2007*, pages 205–208.

ZESCH, T., MÜLLER, C. et GUREVYCH, I. (2008a). Extracting lexical semantic knowledge from wikipedia and wiktionary. *In Proceedings of LREC'08*, pages 1646–1652.

ZESCH, T., MÜLLER, C. et GUREVYCH, I. (2008b). Using wiktionary for computing semantic relatedness. *In Proceedings of AAAI*, volume 2008, page 45.