A Study of Hybrid Similarity Measures for Semantic Relation Extraction

Alexander Panchenko and Olga Morozova
Center for Natural Language Processing (CENTAL)
Université catholique de Louvain, Belgium
{alexander.panchenko, olga.morozova}@uclouvain.be

Abstract

This paper describes several novel hybrid semantic similarity measures. We study various combinations of 16 baseline measures based on WordNet, Web as a corpus, corpora, dictionaries, and encyclopedia. The hybrid measures rely on 8 combination methods and 3 measure selection techniques and are evaluated on (a) the task of predicting semantic similarity scores and (b) the task of predicting semantic relation between two terms. Our results show that hybrid measures outperform single measures by a wide margin, achieving a correlation up to 0.890 and MAP(20) up to 0.995.

1 Introduction

Semantic similarity measures and relations are proven to be valuable for various NLP and IR applications, such as word sense disambiguation, query expansion, and question answering.

Let R be a set of synonyms, hypernyms, and co-hyponyms of terms C, established by a lexicographer. A *semantic relation extraction* method aims at discovering a set of relations $\hat{R}$ approximating R. The quality of the relations provided by existing extractors is still lower than the quality of the manually constructed relations. This motivates the development of new relation extraction methods.

A well-established approach to relation extraction is based on lexico-syntactic patterns (Auger and Barrière, 2008). In this paper, we study an alternative approach based on *similarity measures*. These methods do not return a type of the relation between words ($\hat{R} \subseteq C \times C$). However, we assume that the methods should retrieve *a mix* of synonyms, hypernyms, and co-hyponyms for practical use in text processing applications and evaluate them accordingly.

A multitude of measures was used in the previous research to extract synonyms, hypernyms, and co-hyponyms. Five key approaches are those based on a distributional analysis (Lin, 1998b), Web as a corpus (Cilibrasi and Vitanyi, 2007), lexico-syntactic patterns (Bollegala et al., 2007), semantic networks (Resnik, 1995), and definitions of dictionaries or encyclopedias (Zesch et al., 2008a). Still, the existing approaches based on these single measures are far from being perfect. For instance, Curran and Moens (2002) compared distributional measures and reported Precision@1 of 76% for the best one. For improving the performance, some attempts were made to combine single measures, such as (Curran, 2002; Cederberg and Widdows, 2003; Mihalcea et al., 2006; Agirre et al., 2009; Yang and Callan, 2009). However, most studies are still not taking into account the whole range of existing measures, combining mostly sporadically different methods.

The main contribution of the paper is a systematic analysis of 16 baseline measures, and their combinations with 8 fusion methods and 3 techniques for the combination set selection. We are first to propose hybrid similarity measures based on all five extraction approaches listed above; our combined techniques are original as they exploit all key types of resources usable for semantic relation extraction – corpus, web corpus, semantic networks, dictionaries, and encyclopedias. Our experiments confirm that the combined measures are more precise than the single ones. The best found hybrid measure combines 15 baseline measures with the supervised learning. It outperforms

Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data (Hybrid2012), EACL 2012, pages 10–18,
Avignon, France, April 23 2012.

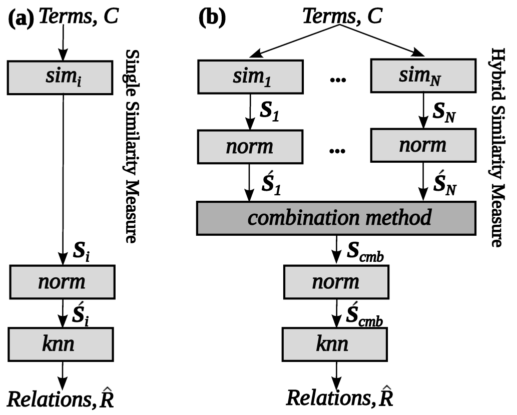

Figure 1: (a) Single and (b) hybrid relation extractors based on similarity measures.

all tested single and combined measures by a large margin, achieving a correlation of 0.870 with human judgements and MAP(20) of 0.995 on the relation recognition task.

2 Similarity-based Relation Extraction

In this paper a similarity-based relation extraction method is used. In contrast to the traditional approaches, relying on a single measure, our method relies on a hybrid measure (see Figure 1). A *hybrid similarity measure* combines several *single similarity measures* with a *combination method* to achieve better extraction results. To extract relations $\hat{R}$ between terms C, the method calculates pairwise similarities between them with the help of a similarity measure. The relations are established between each term $c \in C$ and the terms most similar to c (its nearest neighbors). First, a term-term ($C \times C$) similarity matrix $\mathbf{S}$ is calculated with a similarity measure sim, as depicted in Figure 1 (a). Then, these similarity scores are mapped to the interval $[0;1]$ with a $norm$ function as follows: $\acute{\mathbf{S}} = \frac{\mathbf{S} - min(\mathbf{S})}{max(\mathbf{S})}$. Dissimilarity scores are transformed into similarity scores: $\acute{\mathbf{S}} = \mathbf{1} - norm(\mathbf{S})$. Finally, the knn function calculates semantic relations between terms with a k-NN thresholding: $\hat{R} = \bigcup_{i=1}^{|C|} \{\langle c_i, c_j \rangle : (c_j \in \text{ top } k\% \text{ of } c_i) \wedge (s_{ij} > 0)\}$. Here, k is a percent of top similar terms to a term c_i. Thus, the method links each term c_i with $k\%$ of its nearest neighbours.

3 Single Similarity Measures

A similarity measure *extracts* or *recalls* a similarity score $s_{ij} \in \mathbf{S}$ between a pair of terms $c_i, c_j \in C$. In this section we list 16 baseline measures exploited by hybrid measures. The measures were selected as (a) the previous research suggests that they are able to capture synonyms, hypernyms, and co-hyponyms; (b) they rely on all main resources used to derive semantic similarity – semantic networks, Web as a corpus, traditional corpora, dictionaries, and encyclopedia.

3.1 Measures Based on a Semantic Network

We test 5 measures relying on WORDNET semantic network (Miller, 1995) to calculate the similarities: Wu and Palmer (1994) (1), Leacock and Chodorow (1998) (2), Resnik (1995) (3), Jiang and Conrath (1997) (4), and Lin (1998a) (5). These measures exploit the lengths of the shortest paths between terms in a network and probability of terms derived from a corpus. We use implementation of the measures available in WORDNET::SIMILARITY (Pedersen et al., 2004).

A limitation of these measures is that similarities can only be calculated upon 155.287 English terms from WordNet 3.0. In other words, these measures *recall* rather than *extract* similarities. Therefore, they should be considered as a source of common lexico-semantic knowledge for a hybrid semantic similarity measure.

3.2 Web-based Measures

Web-based metrics use Web search engines for calculation of similarities. They rely on the number of times the terms co-occur in the documents as indexed by an information retrieval system. We use 3 baseline web measures based on index of YAHOO! (6), BING (7), and GOOGLE over the domain `wikipedia.org` (8). These three measures exploit Normalized Google Distance (NGD) formula (Cilibrasi and Vitanyi, 2007) for transforming the number of hits into a similarity score. Our own system implements BING measure, while Measures of Semantic Relatedness (MSR) web service[1] calculates similarities with YAHOO! and GOOGLE.

The coverage of languages and vocabularies by web-based measures is huge. Therefore, it is assumed that they are able to *extract* new lexico-semantic knowledge. Web-based measures are limited by constraints of a search engine API (hundreds of thousands of queries are needed).

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

3.3 Corpus-based Measures

We tested 5 measures relying on corpora to calculate similarity of terms: two baseline distributional measures, one novel measure based on lexico-syntactic patterns, and two other baseline measures. Each of them uses a different corpus.

Corpus-based measures are able to *extract* similarity between unknown terms. Extraction capabilities of these measures are limited by a corpus. If terms do not occur in a text, then it would be impossible to calculate similarities between them.

Distributional Measures

These measures are based on a distributional analysis of a 800M tokens corpus WaCypedia (Baroni et al., 2009) tagged with TreeTagger and dependency-parsed with MaltParser. We rely on our own implementation of two distributional measures. The distributional measure (9) performs Bag-of-words Distributional Analysis (BDA) (Sahlgren, 2006). We use as features the 5000 most frequent lemmas (nouns, adjectives, and verbs) from a context window of 3 words, excluding stopwords. The distributional measure (10) performs Syntactic Distributional Analysis (SDA) (Lin, 1998b). For this one, we use as features the 100.000 most frequent dependency-lemma pairs. In our implementation of SDA a term c_i is represented with a feature $\langle dt_j, w_k \rangle$, if w_k is not in a stoplist and dt_j has one of the following dependency types: NMOD, P, PMOD, ADV, SBJ, OBJ, VMOD, COORD, CC, VC, DEP, PRD, AMOD, PRN, PRT, LGS, IOBJ, EXP, CLF, GAP . For both BDA and SDA: the feature matrix is normalized with Pointwise Mutual Information; similarities between terms are calculated with a cosine between their respective feature vectors.

Pattern-based Measure

We developed a novel similarity measure PatternWiki (13), which relies on 10 lexico-syntactic patterns. [2] First, we apply the patterns to the WaCypedia corpus and get as a result a list of concordances (see below). Next, we select the concordances which contain at least two terms from the input vocabulary C. The semantic similarity s_{ij} between each two terms $c_i, c_j \in C$ is equal to the number of their co-occurences in the same concordance.

The set of the patterns we used is a compilation of the 6 classical Hearst (1992) patterns, aiming at the extraction of hypernymic relations, as well as 3 patterns retrieving some other hypernyms and co-hyponyms and 1 synonym extraction pattern, which we found in accordance with Hearst's pattern discovery algorithm. The patterns are encoded in a form of finite-state transducers with the help of a corpus processing tool Unitex [3] (Paumier, 2003). The main graph is a cascade of the subgraphs, each of which encodes one of the patterns. For example, Figure 2 presents the graph which extracts, e. g.:

- `such diverse {[occupations]} as {[doctors]}, {[engineers]} and {[scientists]}[PATTERN=1]`

Figure brackets mark the noun phrases, which are in the semantic relation, nouns and compound nouns stand between the square brackets. Unitex enables the exclusion of meaningless adjectives and determiners out of the tagging, while the patterns containing them are still being recognized. So, the notion of a pattern has more general sense with respect to other works such as (Bollegala et al., 2007), where each construction with a different lexical item, a word form or even a punctuation mark is regarded as a unique pattern. The nouns extracted from the square brackets are lemmatized with the help of DELA dictionary[4], which consists of around 300,000 simple and 130,000 compound words. If the noun to extract is a plural form of a noun in the dictionary, then it is re-written into the respective singular form. Semantic similarity score is equal to the number of co-occurences of terms in the square brackets within the same concordance (the number of extractions between the terms).

Other Corpus-based Measures

In addition to the three measures presented above, we use two other corpus-based measures available via the MSR web service. The measure (11) relies on the Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997) trained on the TASA corpus (Veksler et al., 2008). LSA calculates similarity of terms with a cosine between their respective vectors in the "concept space". The measure (12) relies on the NGD formula (see Section 3.2), where counts are derived from the Factiva corpus (Veksler et al., 2008).

[2] Available at http://http://cental.fltr.ucl.ac.be/team/~morozova/pattern-wiki.tar.gz

[3] http://igm.univ-mlv.fr/~unitex/

[4] Available at http://infolingu.univ-mlv.fr/

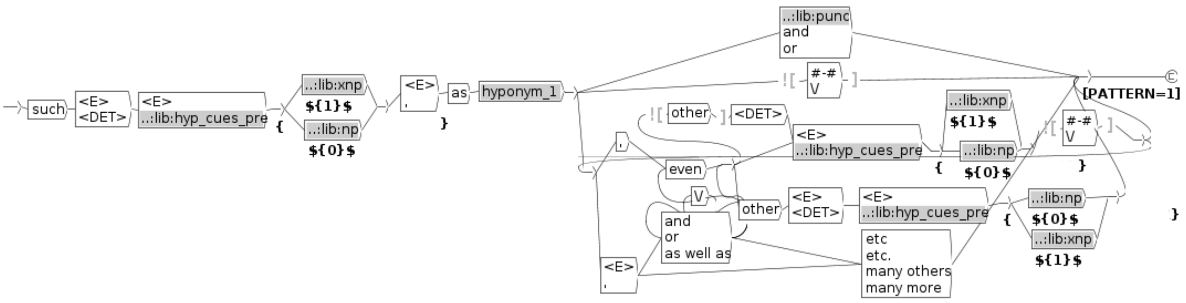

Figure 2: An example of a UNITEX graph for hypernym extraction (subgraphs are marked with gray; <E> defines zero; <DET> defines determiners; bold symbols and letters outside of the boxes are annotation tags)

3.4 Definition-based Measures

We test 3 measures which rely on explicit definitions of terms specified in dictionaries. The first metric WktWiki (14) is a novel similarity measure that stems from the Lesk algorithm (Pedersen et al., 2004) and the work of Zesch et al. (2008a). WktWiki operates on Wiktionary definitions and relations and Wikipedia abstracts. WktWiki calculates similarity as follows. First, definitions for each input term $c \in C$ are built. A "definition" is a union of all available glosses, examples, quotations, related words, and categories from Wiktionary and a short abstract of the corresponding Wikipedia article (a name of the article must exactly match the term c). We use all senses corresponding to a surface form of term c. Then, each term $c \in C$ of the 1000 most frequent lemmas is represented as a bag-of-lemma vector, derived from its "definition". Feature vectors are normalized with Pointwise Mutual Information and similarities between terms are calculated with a cosine between them. Finally, the pairwise similarities between terms $\mathbf{S}$ are corrected. The highest similarity score is assigned to the pairs of terms which are directly related in Wiktionary. [5]

WktWiki is different to the work of Zesch et al. (2008b) in three aspects: (a) terms are represented in a word space, and not in a document space; (b) both texts from Wiktionary and Wikipedia are used; (c) relations of Wiktionary are used to update similarity scores.

In addition to WktWiki, we operate with 2 baseline measures relying on WordNet glosses available in a WORDNET::SIMILARITY package: Gloss Vectors (Patwardhan and Pedersen, 2006) (15) and Extended Lesk (Banerjee and Pedersen, 2003) (16). The key difference between WktWiki and WordNet-based measures is that the latter uses definitions of related terms.

Extraction capabilities of definition-based measures are 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 in all languages; Wikipedia has 3.866.773 English articles and around 20.8 millons of articles in all languages.

4 Hybrid Similarity Measures

A hybrid similarity measure combines several single similarity measures described above with one of the combination methods described below.

4.1 Combination Methods

A goal of a combination method is to produce similarity scores which perform better than the scores of input single measures. A combination method takes as an input a set of similarity matrices $\{\mathbf{S}_1, \ldots, \mathbf{S}_K\}$ produced by K single measures and outputs a combined similarity matrix $\mathbf{S}_{cmb}$. We denote as s_{ij}^k a *pairwise similarity score* of terms c_i and c_j produced by k-th measure. We test the 8 following combination methods:

Mean. A mean of K pairwise similarity scores:

$$\mathbf{S}_{cmb} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{S}_k \Leftrightarrow s_{ij}^{cmb} = \frac{1}{K} \sum_{k=1,K} s_{ij}^k.$$

Mean-Nnz. A mean of those pairwise similarity scores which have a non-zero value:

$$s_{ij}^{cmb} = \frac{1}{|k : s_{ij}^k > 0, k = 1, K|} \sum_{k=1,K} s_{ij}^k.$$

[5] We used JWKTL library (Zesch et al., 2008a), as an API to Wiktionary and DBpedia.org as a source of Wikipedia short abstracts (dumps were downloaded in October 2011).

Mean-Zscore. A mean of K similarity scores transformed into Z-scores:

$$s_{ij}^{cmb} = \frac{1}{K}\sum_{k=1}^{K}\frac{s_{ij}^{k}-\mu_k}{\sigma_k},$$

where μ_k is a mean and σ_k is a standard deviation of similarity scores of k-th measure ($\mathbf{S}_k$).

Median. A median of K pairwise similarities:

$$s_{ij}^{cmb} = median(s_{ij}^{1},\ldots,s_{ij}^{K}).$$

Max. A maximum of K pairwise similarities:

$$s_{ij}^{cmb} = max(s_{ij}^{1},\ldots,s_{ij}^{K}).$$

Rank Fusion. First, this combination method converts each pairwise similarity score s_{ij}^{k} to a rank r_{ij}^{k}. Here, $r_{ij}^{k}=5$ means that term c_j is the 5-th nearest neighbor of the term c_i, according to the k-th measure. Then, it calculates a combined similarity score as a mean of these pairwise ranks: $s_{ij}^{cmb} = \frac{1}{K}\sum_{k=1,K} r_{ij}^{k}$.

Relation Fusion. This combination method gathers and unites the best relations provided by each measure. First, the method retrieves relations extracted by single measures with the function knn described in Section 2. We have empirically chosen an "internal" kNN threshold of 20% for this combination method. Then, a set of extracted relations R_k, obtained from the k-th measure, is encoded as an adjacency matrix $\mathbf{R}_k$. An element of this matrix indicates whether terms c_i and c_j are related:

$$r_{ij}^{k} = \begin{cases} 1 & \text{if semantic relation } \langle c_i, c_j\rangle \in R_k \\ 0 & \text{else} \end{cases}$$

The final similarity score is a mean of adjacency matrices: $\mathbf{S}_{cmb} = \frac{1}{K}\sum_{i=1}^{K}\mathbf{R}_i$. Thus, if two measures are combined and the first extracted the relation between c_i and c_j, while the second did not, then the similarity s_{ij} will be equal to 0.5.

Logit. This combination method is based on logistic regression (Agresti, 2002). We train a binary classifier on a set of manually constructed semantic relations R (we use BLESS and SN datasets described in Section 5). Positive training examples are "meaningful" relations (synonyms, hyponyms, etc.), while negative training examples are pairs of semantically unrelated words (generated randomly and verified manually). A semantic relation $\langle c_i, c_j\rangle \in R$ is represented with a vector of pairwise similarities between terms c_i, c_j calculated with K measures $(s_{ij}^{1},\ldots,s_{ij}^{K})$ and a binary variable r_{ij} (category):

$$r_{ij} = \begin{cases} 0 & \text{if } \langle c_i, c_j\rangle \text{ is a random relation} \\ 1 & \text{otherwise} \end{cases}$$

For evaluation purposes, we use a special 10-fold cross validation ensuring that all relations of one term c are always in the same training/test fold. The results of the training are $K+1$ coefficients of regression $(w_0, w_1, \ldots, w_K)$. We apply the model to combine similarity measures as follows:

$$s_{ij}^{cmb} = \frac{1}{1+e^{-z}}, z = w_0 + \sum_{k=1}^{K} w_k s_{ij}^{k}.$$

4.2 Combination Sets

Any of the 8 combination methods presented above may combine from 2 to 16 single measures. Thus, there are $\sum_{m=2}^{16} C_{16}^{m} = \sum_{m=2}^{16}\frac{16!}{m!(16-m)!} = 65535$ ways to choose which single measures to use in *a* combination method. We apply three methods to find an efficient combination of measures in this search space: expert choice of measures, forward stepwise procedure, and analysis of a logistic regression model.

Expert choice of measures is based on the analytical and empirical properties of the measures. We chose 5 or 9 measures which perform well and rely on complimentary resources: corpus, Web, WordNet, etc. Additionally, we selected a group of all measures except for one which has shown the worst results on all datasets. Thus, according to this selection method we have chosen three groups of measures (see Section 3 and Table 1 for notation):

- $E5 = \{3, 9, 10, 13, 14\}$
- $E9 = \{1, 3, 9-11, 13-16\}$
- $E15 = \{1, 2, 3, 4, 5, 6, 8-16\}$

Forward stepwise procedure is a greedy algorithm which works as follows. It takes as an input all measures, a method of their combination such as *Mean*, and a criterion such as Precision at $k=50$. It starts with a void set of measures. Then, at each iteration it adds to the combination one measure which brings the biggest improvement to the criterion. The algorithm stops when no measure can improve the criteria. According to this method, we have chosen four groups of the measures [6]:

- $S7 = \{9-11, 13-16\}$
- $S8a = \{9-16\}$
- $S8b = \{1, 9-11, 13-16\}$
- $S10 = \{1, 6, 9-16\}$

The last measure selection technique is based on analysis of *logistic regression* trained on all 16 measures as features. Only measures with positive coefficients are selected. According to this method, 12 measures were chosen:

- $R12 = \{3, 5, 6, 8-16\}$

We test combination methods on the 8 sets of measures specified above. Remarkably, all three selection techniques constantly choose six following measures – $9, 10, 11, 14, 15, 16$, i. e., C-BowDA, C-SynDA, C-LSA-Tasa, D-WktWiki, N-GlossVectors, and N-ExtendedLesk.

5 Evaluation

Evaluation relies on human judgements about semantic similarity and on manually constructed semantic relations. [7]

Human Judgements Datasets. This kind of ground truth enables *direct* assessment of measure performance and *indirect* assessment of extraction quality with this measure. Each of these datasets consists 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. We use three standard human judgements datasets – MC (Miller and Charles, 1991), RG (Rubenstein and Goodenough, 1965) and WordSim353 (Finkelstein et al., 2001), composed of 30, 65, and 353 pairs of terms respectively. Let $\mathrm{s} = (s_{i1}, s_{i2}, \ldots, s_{iN})$ be a vector of ground truth scores, and $\hat{\mathrm{s}} = (\hat{s}_{i1}, \hat{s}_{i2}, \ldots, \hat{s}_{iN})$ be a vector of similarity scores calculated with a similarity measure. Then, the quality of this measure is assessed with Spearman's correlation between s and ŝ.

Semantic Relations Datasets. This kind of ground truth enables *indirect* assessment of measure performance and *direct* assessment of extraction quality with the measure. Each of these datasets consists of a set of semantic relations R, such as ⟨agitator, syn, activist⟩, ⟨hawk , hyper, predator⟩, ⟨gun, syn,weapon⟩, and ⟨dishwasher, cohypo, reezer⟩. Each "target" term has roughly the same number of meaningful and random relations. We use two semantic relation datasets: BLESS (Baroni and Lenci, 2011) and SN. The first is used to assess *hypernyms* and *co-hyponyms* extraction. BLESS relates 200 target terms (100 animate and 100 inanimate nouns) to 8625 relatum terms with 26554 semantic relations (14440 are meaningful and 12154 are random). Every relation has one of the following types: hypernym, co-hyponym, meronym, attribute, event, or random. We use the second dataset to evaluate synonymy extraction. SN relates 462 target terms (nouns) to 5910 relatum terms with 14682 semantic relations (7341 are meaningful and 7341 are random). We built SN from WordNet, Roget's thesaurus, and a synonyms database [8].

This kind of evaluation is based on the number of correctly extracted relations with the method described in Section 2. Let $\hat{R}_k$ be a set of extracted semantic relations at a certain level of the kNN threshold k. Then, precision, recall, and mean average precision (MAP) at k are calculated correspondingly as follows: $P(k) = \frac{|R \cap \hat{R}_k|}{|\hat{R}_k|}, R(k) = \frac{|R \cap \hat{R}_k|}{|R|}, M(k) = \frac{1}{k} \sum_{i=1}^{k} P(i)$. The quality of a similarity measure is assessed with the six following statistics: $P(10)$, $P(20)$, $P(50)$, $R(50)$, $M(20)$, and $M(50)$.

6 Results

Table 1 and Figure 3 present performance of the single and hybrid measures on the five ground truth datasets listed above. The first three columns of the table contain correlations with human judgements, while the other columns present performance on the relation extraction task.

The first part of the table reports on scores of 16 single measures. Our results show that the measures are indeed complimentary – there is no measure which performs best on all datasets. For instance, the measure based on a syntactic distributional analysis C-SynDA performed best on the MC dataset achieving a correlation of 0.790; the WordNet measure WN-LeacockChodorow achieved the top score of 0.789 on the RG dataset;

[6] We used Mean as a hybrid measure and the following criteria: MAP(20), MAP(50), P(10), P(20) and P(50). We kept measures which were selected by most of the criteria.

[7] An evaluation script is available at `http://cental.fltr.ucl.ac.be/team/~panchenko/sre-eval/`

[8] `http://synonyms-database.downloadaces.com`

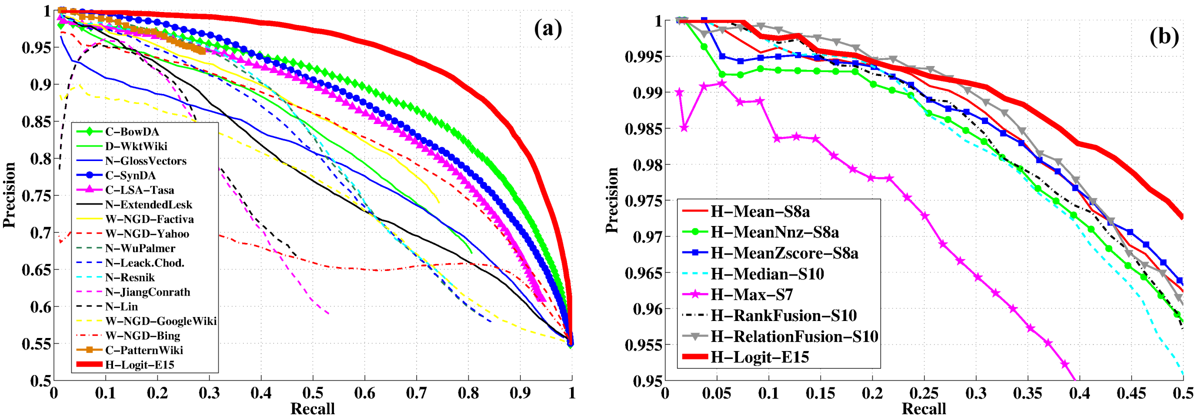

Figure 3: Precision-Recall graphs calculated on the BLESS dataset of (a) 16 single measures and the best hybrid measure H-Logit-E15; (b) 8 hybrid measures.

the corpus based measure C-NGD-Factiva was best on the WordSim353 dataset, achieving a correlation of 0.600. On the BLESS dataset, syntactic distributional analysis C-SynDA performed best for high precision among single measures achieving MAP(20) of 0.984, while the bag-of-words distributional measure C-BowDA was the best for high recall with R(50) of 0.772. On the SN dataset, the WordNet-based measure N-WuPalmer was best both for precision and recall.

The second part of Table 1 presents performance of the hybrid measures. Our results show that if signals from complimentary resources are used, then the retrieval of semantically similar words is significantly improved. Most of the hybrid measures outperform the single measures on all the datasets. We tested each of the 8 combination methods presented in Section 4.1 with each of the 8 sets of measures specified in Section 4.2. We report on the best metrics among all 64 hybrid measures. A notion H-Mean-S8a means that the *Mean* combination method provides the best results with the set of measures *S8a*.

Measures based on the mean of non-zero similarities H-MeanNnz-S8a and H-MeanNnz-E5 performed best on MC and WordSim353 datasets respectively. They achieved correlations of 0.878 and 0.740, which is higher than scores of any other measure. At the same time, measure H-MeanZscore-S8b provided the best scores on the RG dataset among all single and hybrid measures, achieving correlation of 0.890. Supervised measure H-Logit-E15 based on Logistic Regression provided the very best results on both semantic relations datasets BLESS and SN. Furthermore, it outperformed all single and hybrid measures on that task, in terms of both precision and recall, achieving MAP(20) of 0.995 and R(50) of 0.818 on BLESS and MAP(20) of 0.993 and R(50) of 0.819 on SN. H-Logit-E15 makes use of 15 similarity measures and disregards only the worst single measure W-NGD-Bing.

As we can see in Figure 3 (b), combining similarity scores with a *Max* function appears to be the worst solution. Combination methods based on an average and a median, including Rank and Relation Fusion, perform much better. These methods provide quite similar results: in the high precision range, they perform nearly as well as a supervised combination. Relation Fusion even manages to slightly outperform Logit on the first 10-15 k-NN (see Figure 3). However, all unsupervised combination methods are significantly worse if higher recall is needed.

We conclude that the H-Logit-E15 is the best hybrid similarity measure for semantic relation extraction and in terms of plausibility with human judgements among all single and hybrid measures examined in this paper.

7 Discussion

Hybrid measures achieve higher precision and recall than single measures. First, it is due to the reuse of common lexico-semantic information (such as that a “car” is a synonym of a “vehicle”) 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,

Similarity Measure	**MC**	**RG**	**WS**	**BLESS**						**SN**					
	ρ	*ρ*	*ρ*	*P(10)*	*P (20)*	*M(20)*	*P(50)*	*M(50)*	*R(50)*	*P(10)*	*P(20)*	*M(20)*	*P(50)*	*M(50)*	*R(50)*
Random	*0.056*	*-0.047*	*-0.122*	0.546	0.542	0.549	0.544	0.546	0.522	0.504	0.502	0.507	0.499	0.502	0.498
1. N-WuPalmer	0.742	0.775	0.331	0.974	0.929	0.972	0.702	0.879	0.674	0.982	0.959	**0.981**	0.766	0.917	**0.763**
2. N-Leack.Chod.	0.724	**0.789**	0.295	0.953	0.901	0.954	0.702	0.863	0.648	0.984	0.953	**0.981**	0.757	0.913	0.755
3. N-Resnik	0.784	0.757	0.331	0.970	0.933	0.970	0.700	0.879	0.647	0.948	0.908	0.948	0.724	0.874	0.722
4. N-JiangConrath	0.719	0.588	0.175	0.956	0.872	0.920	0.645	0.817	0.458	0.931	0.857	0.911	0.625	0.808	0.570
5. N-Lin	0.754	0.619	0.204	0.949	0.884	0.918	0.682	0.822	0.451	0.939	0.877	0.920	0.611	0.827	0.566
6. W-NGD-Yahoo	*0.330*	0.445	0.254	0.940	0.907	0.941	0.783	0.885	0.648	—	—	—	—	—	—
7. W-NGD-Bing	*0.063*	*0.181*	*0.060*	0.724	0.706	0.713	0.650	0.690	0.600	0.659	0.619	0.671	0.633	0.648	0.633
8. W-NGD-GoogleWiki	*0.334*	0.502	0.251	0.874	0.837	0.872	0.703	0.814	0.649	—	—	—	—	—	—
9. C-BowDA	0.693	0.782	0.466	0.971	0.947	0.969	0.836	0.928	**0.772**	0.974	0.932	0.968	0.742	0.896	0.740
10. C-SynDA	**0.790**	0.786	0.491	0.985	0.953	**0.984**	0.811	0.925	0.749	0.978	0.945	0.972	0.751	0.907	0.743
11. C-LSA-Tasa	0.694	0.605	0.566	0.968	0.937	0.967	0.802	0.912	0.740	0.903	0.846	0.895	0.641	0.803	0.609
12. C-NGD-Factiva	0.603	0.599	**0.600**	0.959	0.916	0.959	0.786	0.894	0.681	0.906	0.857	0.904	0.731	0.835	0.543
13. C-PatternWiki	0.461	0.542	0.357	0.972	0.951	0.976	0.944	0.957	0.287	0.920	0.904	0.907	0.891	0.900	0.295
14. D-WktWiki	0.759	0.754	0.521	0.943	0.905	0.946	0.750	0.876	0.679	0.922	0.887	0.918	0.725	0.854	0.656
15. D-GlossVectors	0.653	0.738	0.322	0.894	0.860	0.901	0.742	0.843	0.686	0.932	0.899	0.933	0.722	0.864	0.709
16. D-ExtenedLesk	0.792	0.718	0.409	0.937	0.866	0.939	0.711	0.843	0.657	0.952	0.873	0.943	0.655	0.832	0.654
H-Mean-S8a	0.834	0.864	0.734	0.994	0.980	0.994	0.870	0.960	0.804	0.985	0.965	0.985	0.788	0.928	0.787
H-MeanZscore-S8a	0.830	0.864	0.728	0.994	0.981	0.993	0.874	0.961	0.808	0.986	0.967	0.986	0.793	0.932	0.792
H-MeanNnz-S8a	**0.843**	0.847	**0.740**	0.993	0.977	0.991	0.865	0.956	0.799	0.986	0.967	0.985	0.803	0.933	0.802
H-Median-S10	0.821	0.842	0.647	0.995	0.976	0.992	0.843	0.950	0.779	0.975	0.934	0.970	0.724	0.892	0.721
H-Max-S7	0.802	0.816	0.654	0.979	0.957	0.979	0.839	0.936	0.775	0.980	0.957	0.979	0.786	0.922	0.785
H-RankFusion-S10	—	—	—	0.994	0.978	0.993	0.864	0.956	0.798	0.976	0.929	0.971	0.745	0.896	0.744
H-RelationFusion-S10	—	—	—	0.996	0.982	0.995	0.840	0.952	0.758	0.986	0.963	0.981	0.781	0.920	0.749
H-Logit-E15	0.793	**0.870**	0.690	0.995	0.987	**0.995**	0.885	0.968	**0.818**	0.995	0.984	**0.993**	0.821	0.951	**0.819**
H-MeanNnz-E5	**0.878**	0.878	0.482	0.986	0.956	0.984	0.784	0.922	0.725	0.975	0.938	0.969	0.768	0.906	0.766
H-MeanZscore-S8b	0.844	**0.890**	0.616	0.992	0.977	0.991	0.844	0.953	0.780	0.995	0.985	0.995	0.815	0.950	0.814

Table 1: Performance of 16 single and 8 hybrid similarity measures on human judgements datasets (MC, RG, WordSim353) and semantic relation datasets (BLESS and SN). The best scores in a group (single/hybrid) are in bold; the very best scores are in grey. Correlations *in italics* mean $p > 0.05$, otherwise $p \leq 0.05$.

they only can 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 enables keeping high precision for frequent terms (e. g., "disease") present in WordNet and dictionaries, and empowers calculation of relations between rare terms unlisted in the handcrafted resources (e. g., "bronchocele") with web and corpus measures.

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

Finally, the supervised combination method works better than unsupervised ones because of two reasons. First, the measures generate scores which have quite different distributions on the range $[0; 1]$. The averaging of such scores may be suboptimal. Logistic Regression overcomes this issue by assigning appropriate weights $(w_1, \ldots, w_k)$ to the measures in the linear combination z. Second, training procedure enables the model to assign higher weights to the measures which provide better results, while for the methods based on averaging all weight are equal.

8 Conclusion

In this work, we designed and studied several hybrid similarity measures in the context of semantic relation extraction. We have undertaken a systematic analysis of 16 baseline measures, 8 combination methods, and 3 measure selection techniques. The combined measures were thoroughly evaluated on five ground truth datasets: MC, RG, WordSim353, BLESS, and SN. Our results have shown that the hybrid measures outperform the single measures on all datasets. In particular, a combination of 15 baseline corpus-, web-, network-, and dictionary-based measures with Logistic Regression provided the best results. This method achieved a correlation of 0.870 with human judgements and MAP(20) of 0.995 and Recall(50) of 0.818 at predicting semantic relation between terms.

This paper also sketched two novel single similarity measures performing comparably with the baselines – WktWiki, based on definitions of Wikipedia and Wiktionary; and PatternWiki, based on patterns applied on Wikipedia abstracts. In the future research, we are going to apply the developed methods to query expansion.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of NAACL-HLT 2009*, pages 19–27.

Alan Agresti. 2002. *Categorical Data Analysis (Wiley Series in Probability and Statistics)*. 2 edition.

Alain Auger and Caroline Barrière. 2008. Pattern-based approaches to semantic relation extraction: A state-of-the-art. *Terminology Journal*, 14(1):1–19.

Satanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *IJCAI*, volume 18, pages 805–810.

Marco Baroni and Alexandro Lenci. 2011. How we blessed distributional semantic evaluation. *GEMS (EMNLP), 2011*, pages 1–11.

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

D. Bollegala, Y. Matsuo, and M. Ishizuka. 2007. Measuring semantic similarity between words using web search engines. In *WWW*, volume 766.

S. Cederberg and D. Widdows. 2003. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Proceedings HLT-NAACL*, page 111118.

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

James R. Curran and Marc Moens. 2002. Improvements in automatic thesaurus extraction. In *Proceedings of the ACL-02 workshop on Unsupervised Lexical Acquisition*, pages 59–66.

James R. Curran. 2002. Ensemble methods for automatic thesaurus extraction. In *Proceedings of the EMNLP-02*, pages 222–229. ACL.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *WWW 2001*, pages 406–414.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *ACL*, pages 539–545.

Kris Heylen, Yves Peirsman, Dirk Geeraerts, and Dirk Speelman. 2008. Modelling word similarity: an evaluation of automatic synonymy extraction algorithms. *LREC'08*, pages 3243–3249.

Jay J. Jiang and David W. Conrath. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *ROCLING X*, pages 19–33.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psych. review*, 104(2):211.

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

Dekang Lin. 1998a. An Information-Theoretic Definition of Similarity. In *ICML*, pages 296–304.

Dekang Lin. 1998b. Automatic retrieval and clustering of similar words. In *ACL*, pages 768–774.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI'06*, pages 775–780.

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

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

Siddharth Patwardhan and Ted Pedersen. 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.

Sébastien Paumier. 2003. *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Ph.D. thesis, Université de Marne-la-Vallée.

Ted Pedersen, Siddaharth Patwardhan, and Jason Michelizzi. 2004. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, pages 38–41. ACL.

Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *IJCAI*, volume 1, pages 448–453.

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

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

Vladislav D. Veksler, Ryan Z. Govostes, and Wayne D. Gray. 2008. Defining the dimensions of the human semantic space. In *30th Annual Meeting of the Cognitive Science Society*, pages 1282–1287.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of ACL'1994*, pages 133–138.

Hui Yang and Jamie Callan. 2009. A metric-based framework for automatic taxonomy induction. In *ACL-IJCNLP*, page 271279.

Torsen Zesch, Christof Müller, and Irina Gurevych. 2008a. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of LREC'08*, pages 1646–1652.

Torsen Zesch, Christof Müller, and Irina Gurevych. 2008b. Using wiktionary for computing semantic relatedness. In *Proceedings of AAAI*, page 45.