Journal of Artificial Intelligence Research 30 (2007) 181-212 Submitted 03/07; published 10/07

Knowledge Derived From Wikipedia For Computing Semantic Relatedness

Simone Paolo Ponzetto PONZETTO@EML-RESEARCH.DE
Michael Strube STRUBE@EML-RESEARCH.DE
EML Research gGmbH, Natural Language Processing Group
Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg, Germany
http://www.eml-research.de/nlp

Abstract

Wikipedia provides a semantic network for computing semantic relatedness in a more structured fashion than a search engine and with more coverage than WordNet. We present experiments on using Wikipedia for computing semantic relatedness and compare it to WordNet on various benchmarking datasets. Existing relatedness measures perform better using Wikipedia than a baseline given by Google counts, and we show that Wikipedia outperforms WordNet on some datasets. We also address the question whether and how Wikipedia can be integrated into NLP applications as a knowledge base. Including Wikipedia improves the performance of a machine learning based coreference resolution system, indicating that it represents a valuable resource for NLP applications. Finally, we show that our method can be easily used for languages other than English by computing semantic relatedness for a German dataset.

1. Introduction

While most advances in Natural Language Processing (NLP) have been made recently by investigating data-driven methods, namely statistical techniques, we believe that further advances crucially depend on the availability of world and domain knowledge. This is essential for high-level linguistic tasks which require language understanding capabilities such as question answering (e.g., Hovy, Gerber, Hermjakob, Junk, & Lin, 2001) and recognizing textual entailment (Bos & Markert, 2005; Tatu, Iles, Slavick, Novischi, & Moldovan, 2006, inter alia). However, there are not many domain-independent knowledge bases available which provide a large amount of information on named entities (the leaves of the taxonomy) and contain continuously updated knowledge for processing current information.

In this article we approach the problem from a novel[1] perspective by making use of a wide coverage online encyclopedia, namely Wikipedia. We use the "encyclopedia that anyone can edit" to compute semantic relatedness by taking the system of categories in Wikipedia as a semantic network. That way we overcome the well known knowledge acquisition bottleneck by deriving a knowledge resource from a very large, collaboratively created encyclopedia. Then the question is whether the quality of the resource is high enough to be used successfully in NLP applications. By performing two different evaluations we provide an answer to that question. We do not only show that Wikipedia derived semantic relatedness correlates well with human judgments, but also that such information can be used to include lexical semantic information in a NLP application, namely coreference resolution, where world knowledge has been considered important since early

1. This article builds upon and extends Ponzetto and Strube (2006a) and Strube and Ponzetto (2006).

research (Charniak, 1973; Hobbs, 1978), but has been integrated only recently by means of WordNet (Harabagiu, Bunescu, & Maiorano, 2001; Poesio, Ishikawa, Schulte im Walde, & Vieira, 2002).

We begin by introducing Wikipedia and measures of semantic relatedness in Section 2. In Section 3 we show how semantic relatedness measures can be ported to Wikipedia. We then evaluate our approach using datasets designed for evaluating such measures in Section 4. Because all available datasets are small and seem to be assembled rather arbitrarily we perform an additional extrinsic evaluation by means of a coreference resolution system in Section 5. In Section 6 we show that relatedness measures computed using Wikipedia can be easily ported to a language other than English, i.e. German. We give details of our implementation in Section 7, present related work in Section 8 and conclude with future work directions in Section 9.

2. Wikipedia and Semantic Relatedness Measures

In this section we describe the structure of Wikipedia and present the measures we use for computing semantic relatedness within its categorization network.

2.1 Wikipedia

Wikipedia is a multilingual web based encyclopedia. Being a collaborative open source medium, it is edited by volunteers. Wikipedia provides a very large domain-independent encyclopedic repository. The English version, as of 14 February 2006, contains 971,518 articles with 18.4 million internal hyperlinks[2].

The text in Wikipedia is highly structured. Apart from article pages being formatted in terms of sections and paragraphs, various relations exists between the pages themselves. These include:

Redirect pages: These pages are used to redirect the query to the actual article page containing information about the entity denoted by the query. This is used to point alternative expressions for an entity to the same article, and accordingly models *synonymy*. Examples include CAR and SICKNESS[3] redirecting to the AUTOMOBILE and DISEASE pages respectively, as well as U.S.A., U.S., USA, US, ESTADOS_UNIDOS and YANKEE_LAND all redirecting to the UNITED_STATES page.

Disambiguation pages: These pages collect links for a number of possible entities the original query could be pointed to. This models *homonymy*. For instance, the page BUSH contains links to the pages SHRUB, BUSH_LOUISIANA, GEORGE_H.W._BUSH and GEORGE_W._BUSH.

Internal links: Articles mentioning other encyclopedic entries point to them through *internal hyperlinks*. This models *article cross-reference*. For instance, the page 'PATAPHYSICS contains links to the term inventor, ALFRED_JARRY, followers such as RAYMOND_QUENEAU, as well as distinctive elements of the philosophy such as NONSENSICAL and LANGUAGE.

Since May 2004 Wikipedia provides also a semantic network by means of its *categories*: articles can be assigned one or more categories, which are further categorized to provide a so-called

2. Wikipedia can be downloaded at `http://download.wikimedia.org`. In our experiments we use the English and German Wikipedia database dump from 19 and 20 February 2006, except where otherwise stated.
3. In the following we use Sans Serif for words and queries, CAPITALS for Wikipedia pages and SMALL CAPS for concepts and Wikipedia categories.

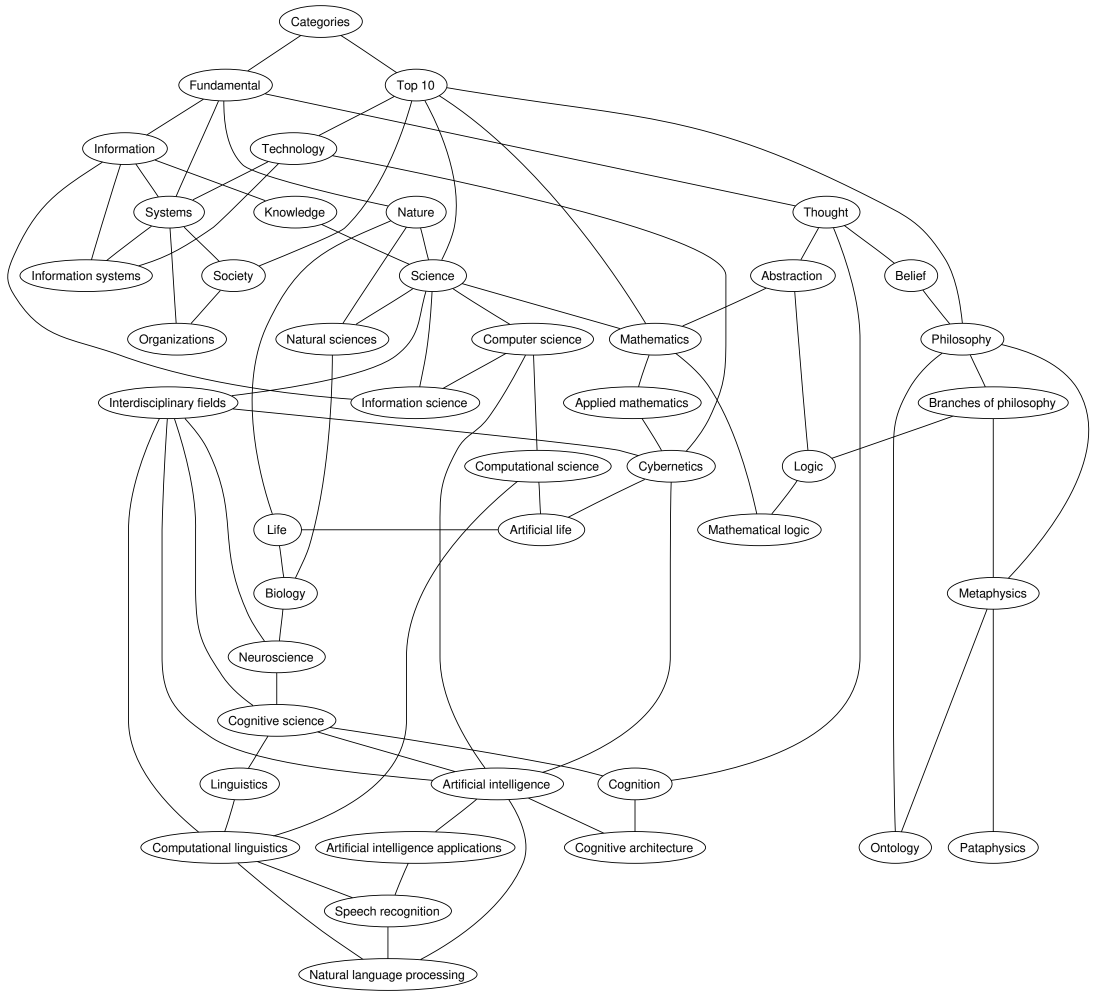


Figure 1: Wikipedia category network. The top nodes in the network (CATEGORIES, FUNDAMENTAL, TOP 10) are structurally identical to the more content bearing categories.

"category tree". In practice, this "tree" is not designed as a strict hierarchy, but allows multiple categorization schemes to coexist simultaneously. The category system is considered a directed acyclic graph, though the encyclopedia editing software does not prevent the users to create cycles in the graph (which nevertheless should be avoided according to the Wikipedia categorization guidelines). Due to this flexible nature, we refer to the Wikipedia "category tree" as the *category network*. As of February 2006, 94% of the articles have been categorized into 103,759 categories. An illustration of some of the higher regions of the hierarchy is given in Figure 1.

The strength of Wikipedia lies in its size, which could be used to overcome the limited coverage and scalability issues of current knowledge bases. But the large size represents also a challenge: the search space in the Wikipedia category graph is very large in terms of depth, branching factor and multiple inheritance relations. Problems arise also in finding robust methods for retrieving relevant information. For instance, the large amount of disambiguation pages requires an efficient algorithm for disambiguating queries, in order to be able to return the desired articles.

Since Wikipedia exists only since 2001 and has been considered a reliable source of information for an even shorter amount of time (Giles, 2005), researchers in NLP have only begun recently to work with its content or use it as a resource. Wikipedia has been used successfully for applications such as question answering (Ahn, Jijkoun, Mishne, Müller, de Rijke, & Schlobach, 2004; Ahn, Bos, Curran, Kor, Nissim, & Webber, 2005; Lo & Lam, 2006, inter alia), named entity disambiguation (Bunescu & Paşca, 2006), text categorization (Gabrilovich & Markovitch, 2006) and computing document similarity (Gabrilovich & Markovitch, 2007).

2.2 Taxonomy Based Semantic Relatedness Measures

Approaches to measuring semantic relatedness that use lexical resources transform that resource into a network or graph and compute relatedness using paths in it. An extensive overview of lexical resource-based approaches to measuring semantic relatedness is presented in Budanitsky and Hirst (2006).

2.2.1 TERMINOLOGY

Semantic relatedness indicates how much two concepts are semantically distant in a network or taxonomy by using all relations between them (i.e. hyponymic/hypernymic, antonymic, meronymic and any kind of functional relations including *is-made-of*, *is-an-attribute-of*, etc.). When limited to hyponymy/hyperonymy (i.e. *isa*) relations, the measure quantifies *semantic similarity* instead (see Budanitsky & Hirst, 2006, for a discussion of *semantic relatedness* vs. *semantic similarity*). In fact, two concepts can be related but are not necessarily similar (e.g. cars and gasoline, see Resnik, 1999). While the distinction holds for a lexical database such as WordNet, where the relations between concepts are semantically typed, it cannot be applied when computing metrics in Wikipedia. This is because the category relations in Wikipedia are neither typed nor show a uniform semantics. The Wikipedia categorization guidelines state that "categories are mainly used to browse through similar articles". Therefore users assign categories rather liberally without having to make the underlying semantics of the relations explicit.

In the following, we use the more generic term of *semantic relatedness*, as it encompasses both WordNet and Wikipedia measures. However, it should be noted that when applied to WordNet, the measures below indicate semantic similarity, as they make use only of the subsumption hierarchy.

2.2.2 PATH BASED MEASURES

These measures compute relatedness as a function of the number of edges in the path between two nodes c_1 and c_2 the words w_1 and w_2 are mapped to. Rada, Mili, Bicknell, and Blettner (1989) traverse MeSH, a term hierarchy for indexing articles in Medline, and compute semantic distance straightforwardly in terms of the number of edges between terms in the hierarchy. Accordingly, semantic relatedness is defined as the inverse score of the semantic distance (pl henceforth).

Since the edge counting approach relies on a uniform modeling of the hierarchy, researchers started to develop measures for computing semantic relatedness which abstract from this problem. Leacock and Chodorow (1998) propose a normalized path-length measure which takes into account the depth of the taxonomy in which the concepts are found (lch). Wu and Palmer (1994) present instead a scaled measure which takes into account the depth of the nodes together with the depth of their least common subsumer (wup).

2.2.3 INFORMATION CONTENT BASED MEASURES

The measure of Resnik (1995) computes the relatedness between the concepts as a function of their information content, given by their probability of occurrence in a corpus (res). Relatedness is modeled as "the extent to which they [the concepts] share information", and is given by the information content of their least common subsumer. Similarly to the path-length based measures, more elaborate measure definitions based on information content have been later developed. This includes the measures from Jiang and Conrath (1997) and Lin (1998), hereafter referred to respectively as jcn and lin, which have been both shown to correlate better with human judgments than Resnik's measure.

2.2.4 TEXT OVERLAP BASED MEASURES

Lesk (1986) defines the relatedness between two words as a function of text (i.e. gloss) overlap. The *extended gloss overlap* (*lesk*) measure of Banerjee and Pedersen (2003) computes the overlap score by extending the glosses of the concepts under consideration to include the glosses of related concepts in a hierarchy. Given two glosses g_1 and g_2 taken as definitions for the words w_1 and w_2, the overlap score $overlap(g_1, g_2)$ is computed as $\sum_n m^2$ for n phrasal m-word overlaps (Banerjee & Pedersen, 2003). The overlap score is computed using a non-linear function, as the occurrences of words in a text collection are known to approximate a Zipfian distribution.

3. Computing Semantic Relatedness with Wikipedia

Wikipedia based semantic relatedness computation is described in the following Subsections:

1. Retrieve two unambiguous Wikipedia pages which a pair of words, w_1, w_2 (e.g. king and rook) refer to, namely $pages = \{p_1, p_2\}$ (Section 3.1).

2. Connect to the category network by parsing the pages and extracting the two sets of categories $C_1 = \{c_1 | \, c_1 \text{ is\_category\_of } p_1\}$ and $C_2 = \{c_2 | \, c_2 \text{ is\_category\_of } p_2\}$ the pages are assigned to (Section 3.2).

3. Compute the set of paths between all pairs of categories of the two pages, namely $paths = \{path_{c_1,c_2} | \, c_1 \in C_1, c_2 \in C_2\}$ (Section 3.2).

4. Compute semantic relatedness based on the two pages extracted (for text overlap based measures) and the paths found along the category network (for path length and information content based measures) (Section 3.3).

3.1 Page Retrieval and Disambiguation

Given a pair of words, w_1 and w_2, page retrieval for page p is accomplished by

1. querying the page titled as the word w,

2. following all *redirects* (e.g. CAR redirecting to AUTOMOBILE),

3. resolving *ambiguous* page queries. This is due to many queries in Wikipedia returning a *disambiguation page*. For instance, querying king returns the Wikipedia disambiguation page KING, which points to other pages including MONARCH, KING (CHESS), KING KONG, KING-FM (a broadcasting station), B.B. KING (the blues guitarist) and MARTIN LUTHER KING.

We choose an approach to disambiguation which maximizes relatedness, namely *we let the page queries disambiguate each other* (see Figure 2). If a disambiguation page p_1 for querying word w_1 is hit, we first get all the hyperlinks in page p_2 obtained by querying the other word w_2 without disambiguating. This is to bootstrap the disambiguation process, since it could be the case that both queries are ambiguous, e.g. king and rook. We then take the other word w_2 and all the Wikipedia internal links of page p_2 as a *lexical association list* $L_2 = \{w_2\} \cup \{l_2 | \, l_2 \text{ is\_a\_link\_in } p_2\}$ to be used for disambiguation – i.e., we use the term list {rook, rook (chess), rook (bird), rook (rocket), ...} for disambiguating the page KING. Links such as rook (chess) are split to extract the label between parentheses – i.e., rook (chess) splits into rook and chess. If a link in p_1 contains any occurrence of a disambiguating term $l_2 \in L_2$ (i.e. the link to KING (CHESS) in the KING page containing the term chess extracted from the ROOK page), the linked page is returned (KING (CHESS)), else we return the first article linked in the disambiguation page (MONARCH).

This disambiguation strategy provides a less accurate solution than following all disambiguation page links. Nevertheless it realizes a more practical solution as many of those pages contain a large number of links (e.g. 34 and 13 for the KING and ROOK pages respectively).

3.2 Category Network Search

Given the pages p_1 and p_2, we extract the lists of categories C_1 and C_2 they belong to (i.e. both KING (CHESS) and ROOK (CHESS) belong to the CHESS PIECES category). Given the category sets C_1 and C_2, for each category pair $\langle c_1, c_2 \rangle, c_1 \in C_1, c_2 \in C_2$ we look for all paths connecting the two categories c_1 and c_2. We perform a depth-limited search of maximum depth of 4 for a least common subsumer. We additionally limit the search to any category of a level greater than 2, i.e. we do not consider the levels between 0 and 2 (where level 0 is represented by the top node CATEGORIES of Figure 1). We noticed that limiting the search improves the results. This is probably due to the upper regions of the Wikipedia category network being too strongly connected (see Figure 1). Accordingly, the value of the search depth was established during system prototyping by finding the depth search value which maximizes the correlation between the relatedness scores of the best performing Wikipedia measure and the human judgments given in the datasets from Miller and Charles (1991) and Rubenstein and Goodenough (1965).

3.3 Relatedness Measure Computation

Finally, given the set of paths found between all category pairs, we compute the network based measures by selecting the paths satisfying the measure definitions, namely the shortest path for path-based measures and the path with the most informative least common subsumer for information content based measures.

In order to apply Resnik's measure to Wikipedia we couple it with an intrinsic information content measure relying on the hierarchical structure of the category network (Seco, Veale, & Hayes, 2004), rather than computing the information content from the probabilities of occurrence of the

function GET-PAGES(w_1, w_2) **returns** $pages$

1: $pages \leftarrow \{\emptyset\}$
2: $pages \leftarrow pages \cup$ GET-UNAMBIGUOUS-PAGE(w_1, w_2)
3: $pages \leftarrow pages \cup$ GET-UNAMBIGUOUS-PAGE(w_2, w_1)
4: **return** $pages$

function GET-UNAMBIGUOUS-PAGE(w_1, w_2) **returns** $page$

1: $page \leftarrow getArticleTitled(w_1)$
2: **while** $page$ is a redirection page **do**
3:     $page \leftarrow followRedirect(page)$
4: **end while**
5: **while** $page$ is a disambiguation page **do**
6:     $l_0 \leftarrow$ first_link_in $page$
       $otherPage \leftarrow getArticleTitled(w_2)$,
       $L_1 = \{l_1 |\, l_1$ is_a_link_in $page\}$
       $L_2 = \{w_2\} \cup \{l_2 |\, l_2$ is_a_link_in $otherPage\}$
7:     **for each** $l_i \in L_1$
8:         **for each** $l_j \in L_2$
9:             **if** MATCHES?(l_i,l_j) **then**
10:                 $page \leftarrow getArticleTitled(l_i)$, **goto** (5)
11:             **end if**
12:         **end for**
13:     **end for**
14:     $page \leftarrow getArticleTitled(l_0)$
15: **end while**
16: **return** $page$

function MATCHES?(l_1, l_2) **returns** true or false

1: $T_1 \leftarrow$ SPLIT-BY-PARENTHESIS(l_1)
    $T_2 \leftarrow$ SPLIT-BY-PARENTHESIS(l_2)
2: **for each** $t_i \in T_1$
3:     **for each** $t_j \in T_2$
4:         **if** ORTHOGRAPHICALLY-MATCHES(t_i, t_j) **then**
5:             **return** true
6:         **end if**
7:     **end for**
8: **end for**
9: **return** false

Figure 2: Algorithm for Wikipedia page retrieval and disambiguation

concepts in a corpus. Seco et al. (2004) show that this method correlates better with human judgments than the original approach from Resnik (1995). The intrinsic information content of a category node n in the hierarchy is given as a function of its child nodes, namely

$$ic(n) = 1 - \frac{\log(hypo(n) + 1)}{\log(C)} \tag{1}$$

where $hypo(n)$ is the number of hyponyms of node n and C equals the total number of conceptual nodes in the hierarchy.

Gloss overlap measures are computed from article pages, since no relevant text is given in the category pages. In order to adapt the Lesk measure to Wikipedia (Equation 2), gloss overlap measures (*gloss*) are computed from the first paragraph of the pages. The relatedness score is given by applying a double normalization step to the overlap score. We first normalize by the sum of text lengths and then take the output as the value of the hyperbolic tangent function in order to minimize the role of outliers skewing the score distribution.

$$lesk\_wikipedia(t_1, t_2) = tanh\left(\frac{overlap(t_1, t_2)}{length(t_1) + length(t_2)}\right) \tag{2}$$

4. Experiments

This section describes an evaluation of our methodology based on experiments with word pair lists. We compare the performance of WordNet and Wikipedia based relatedness measures on datasets which have been extensively used in the literature as standard benchmark tests. In addition, we evaluate the performance of the relatedness measures derived from Wikipedia using different versions of the online encyclopedia between February and May 2007.

4.1 Experiments for English

We evaluate the relatedness measures on four standard datasets, namely Miller and Charles' (1991) list of 30 noun pairs (hereafter referred to as M&C), Rubenstein and Goodenough's (1965) 65 word pair synonymity list (R&G) of which M&C is a subset, the WordSimilarity-353 Test Collection (353-TC) from Finkelstein, Gabrilovich, Matias, Rivlin, Solan, Wolfman, and Ruppin (2002)[4], and finally the 2,682 pairs from the nominal only subset (KLEB) of the reader based lexical cohesion dataset from Beigman Klebanov and Shamir (2006). As the 353-TC dataset is partitioned into training and testing subsets, we experiment both with the full list (353 word pairs) and its test data subset (153 pairs). Similarly, as the KLEB dataset contains a relatively large amount of noun pairs, it was split into two 50-50% partitions for performing machine learning based experiments on learning the relatedness of words.

4.1.1 EVALUATION

Following the literature on semantic relatedness, we evaluate performance by taking the Pearson product-moment correlation coefficient r between the relatedness measure scores and the corresponding human judgments. For each dataset we report the correlation computed on all pairs (*all*). In the case of word pairs where at least one of the words could not be found in the lexical resource

4. Available at `http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/wordsim353.html`.

Dataset		Google	WordNet 2.1					Wikipedia (February 2006)					SVM
		jaccard	*pl*	*wup*	*lch*	*res*	*lesk*	*pl*	*wup*	*lch*	*res*	*gloss*	
M&C	all	0.33	0.72	0.77	**0.82**	0.78	0.37	0.47	0.43	0.45	0.23	0.46	
	non-miss	0.33	0.72	0.77	**0.82**	0.78	0.37	0.50	0.49	0.49	0.29	0.47	
R&G	all	0.22	0.78	0.82	**0.86**	0.81	0.35	0.51	0.50	0.52	0.30	0.46	
	non-miss	0.22	0.78	0.82	**0.86**	0.81	0.35	0.53	0.54	0.55	0.34	0.47	
353-TC	all	0.08	0.28	0.30	0.34	0.34	0.21	0.45	0.48	**0.49**	0.38	0.21	
full	non-miss	0.08	0.27	0.32	0.36	0.35	0.21	0.45	0.48	**0.49**	0.39	0.21	
353-TC	all	0.10	0.29	0.28	0.35	0.38	0.21	0.50	0.55	**0.57**	0.46	0.23	**0.62**
test	non-miss	0.10	0.28	0.30	0.38	0.39	0.21	0.50	0.55	**0.57**	0.46	0.23	
KLEB	all	0.02	0.07	0.15	0.15	0.18	0.14	0.29	0.28	**0.30**	0.18	0.13	
full	non-miss	0.02	0.10	0.15	0.15	0.18	0.14	0.30	0.29	**0.31**	0.19	0.13	
KLEB	all	0.03	0.06	0.15	0.14	0.18	0.16	0.31	0.30	**0.32**	0.19	0.15	**0.38**
test	non-miss	0.03	0.09	0.15	0.14	0.18	0.16	0.31	0.31	**0.32**	0.18	0.15	

Table 1: Results on correlation with human judgments of relatedness measures

(i.e. WordNet or Wikipedia) the relatedness score is set to 0. In addition, we report the correlation score obtained by disregarding such pairs containing missing words (*non-miss*). As a baseline, we compute for each word pair w_1 and w_2 the Google correlation coefficient by taking the Jaccard similarity coefficient (Salton & McGill, 1983) on page hits.

$$jaccard = \frac{Hits(w_1 \text{ AND } w_2)}{Hits(w_1) + Hits(w_2) - Hits(w_1 \text{ AND } w_2)}$$

This co-occurrence distributional similarity measure serves as baseline. We choose the Jaccard similarity coefficient because it is a combinatorial measure which does not take into account the actual word distributions (Lee, 1999) — which we do not have here, as we take only Google hits. This models also the usage of other similarity coefficients, e.g. van Rijsbergen (1979) shows that Dice and Jaccard's coefficients are monotonic in each other.

4.1.2 EXPERIMENTAL SETTING

Experiments were performed for each measure on all datasets. For the 353-TC and the KLEB, we experiment on integrating different measures by performing regression using a Support Vector Machine (Vapnik, 1995) to estimate the functional dependence of the human relatedness judgments on multiple relatedness scores. The learner was trained and tested using all available Google, WordNet and Wikipedia scores. We used an RBF kernel with degree 3. Feature selection was performed to find the optimal feature space using a genetic algorithm (Mierswa, Wurst, Klinkenberg, Scholz, & Euler, 2006) through cross-validation on the training data. In addition, we performed model selection for optimal parameter estimation as a grid search (Hsu, Chang, & Lin, 2006).

4.1.3 DISCUSSION

Table 1 shows the correlation coefficients of the different measures with human judgments. Best performance per dataset is highlighted in bold[5]. Both WordNet and Wikipedia perform better than

5. Differences in performance are statistically significant at 95% significance level ($p = 0.05$). For computing statistical significance we performed a paired *t*-test on each dataset for pairs of corresponding relatedness measures (e.g. between the WordNet and Wikipedia path measures). Additionally, we performed the test between each WordNet and Wikipedia measure and the Google baseline, and between the SVM combined measure and the best performing measure on the 353-TC and KLEB test datasets. The only statistically non-significant differences in performance were found between the *lesk* and the Wikipedia *gloss* measure on the M&C dataset.

the Google baseline. While WordNet performs extremely well on the M&C and R&G datasets, its performance drastically decreases when applied to the 353-TC and KLEB datasets. Wikipedia however does not perform as well on the M&C and R&G datasets but outperforms WordNet on 353-TC and KLEB. In the case of the KLEB full dataset, we report a performance competitive with a state-of-the-art WordNet based measure (information content induced from taxonomy and gloss information, Beigman Klebanov, 2006). This is not due to coverage, because in the 353-TC dataset there are only 2 pairs containing at least one word not present in WordNet, where these amount to 13 for Wikipedia. In the KLEB dataset 114 pairs are missing in WordNet while 150 are missing in Wikipedia. The problems seem to be caused rather by *sense proliferation* in WordNet. The measures are in fact computed by looking at all possible sense pairs for the given words (as no word senses are given), and taking the best scoring (e.g. shortest, most informative) path. This allows for unplausible paths to be returned. It should be noted however that this is not caused by WordNet itself, as it has to provide sense coverage, but rather by the relatedness measures. In fact, no sense disambiguation apart from the one performed by the measures themselves is possible. Using Wikipedia pages as entry points, we have access to the page texts and hyperlinks, which can be used to disambiguate and subsequently limit and focus the search. As an example, using fertility to disambiguate egg, we correctly return the Wikipedia page OVUM, whereas the shortest path in WordNet makes use of the second sense for egg, namely 'oval reproductive body of a fowl (especially a hen) used as food'.

In addition to the problem of sense proliferation, WordNet seems to suffer in principle of a *link proliferation* problem, e.g., the shortest path between egg and fertility traverses the hierarchy through one of the root nodes (i.e. ENTITY). One could suggest to limit the search in WordNet as we did in Wikipedia, though it should be noted that this is supposed to be taken care by the measures themselves, e.g. by scaling by the depth of the path nodes.

Besides, Wikipedia performs better than WordNet in the present experimental setting because the 353-TC and the KLEB datasets model semantic relatedness, rather than similarity. The WordNet measures we used (all except for *lesk*) are instead designed to quantify similarity, thus yielding a poor performance. This is supported by the 353-TC annotation guidelines, which define similarity as 'belonging to the same domain or representing features of the same concept', as well as by Beigman Klebanov (2006) reporting competitive results (Pearson correlation coefficient $r = 0.47$) with her WordNet-based relatedness measure. 353-TC contains also highly rated word pairs such as cell and phone or energy and crisis which are closely related, since they tend to occur frequently together, but not similar, as they share few or no properties at all. Finally, additional support is given by the fact that the most competitive results given by Wikipedia are on the KLEB dataset, which is specifically designed with *relatedness* in mind (Beigman Klebanov & Shamir, 2006).

Finkelstein et al. (2002) suggest that integrating a word-vector based relatedness measure with a WordNet based one is useful, as it accounts for word co-occurrences and helps recovering from cases in which the words cannot be found in the available resources, e.g. dictionary or ontology. Accordingly, on the 353-TC and KLEB test sets we report the best performance by *integrating* all available measures and performing feature selection. On the 353-TC data, the score of $r = 0.62$ outperforms the combined WordNet–word-vector measure of Finkelstein et al. (2002) ($r = 0.55$), as well as the score of $r = 0.38$ on the KLEB test data outperforming the score of $r = 0.32$ obtained by using the best performing (Wikipedia-based) relatedness measure (lch). Instead of integrating a word-vector based relatedness measure with a WordNet based one, our results indicate that a competitive performance can be achieved also by using a different knowledge base such as Wikipedia.

From the analysis above, one could conclude that Wikipedia yields better relatedness scores than WordNet when applied to datasets designed to evaluate the relatedness of words. In practice, we believe that it is extremely difficult to perform a fair comparison of the two knowledge sources when limiting the application to such small datasets. In addition, it is not always clear which linguistic notion (i.e. similarity vs. relatedness) underlies the datasets (cf. 353-TC). This is the reason why we do not perform additional experiments making use of other datasets from synonymity tests such as the 80 TOEFL (Landauer & Dumais, 1997), 50 ESL (Turney, 2001) or 300 Reader's Digest Word Power Game (Jarmasz & Szpakowicz, 2003) questions. These datasets pose also a problem since they contain verbs, which are unlikely to be found in an encyclopedic resource such as Wikipedia. These are all reasons why we evaluate in Section 5 our approach by applying it to a real-world NLP task, namely coreference resolution, where the relatedness between hundreds of thousands of word pairs has to be computed, thus providing a more reliable evaluation.

4.2 Evaluation of Wikipedia Throughout Time

One of the most appealing features of Wikipedia is not only that it provides a large coverage knowledge base, but also that it shows a quasi-exponential growth with respect to the number of articles (Table 2)[6]. We evaluated whether such growth rate affects our methodology for computing the relatedness of words. The experiments with the word pairs datasets, using the Wikipedia English database dump from 19 February 2006, were repeated using the dumps from 25 September 2006 and 27 May 2007. The performance of our Wikipedia-based relatedness measures on the M&C, R&G, 353-TC and KLEB datasets are presented in Tables 3 and 4. As highlighted in Figure 3, the only notable differences in performance between different Wikipedia versions are on the M&C and R&G dataset. Nevertheless a simple one-tailed paired sample *t*-test at the 0.05 level reveals that none of the variations between different Wikipedia versions are statistically significant. Qualitative analysis reveals that the improvements are due to few queries getting correctly disambiguated – i.e. lad correctly disambiguates to BOY in the September 2006 and May 2007 Wikipedia, rather than SECT, as in the February version where the LAD disambiguation page contains SECT as the first link used for disambiguation occurring in "Lad is a sub-sect of the Jainist Digambara sect". These differences are accidental and can be easily spotted by a significance test.

	WordNet 2.1	English Wikipedia		
		Feb. 06	Sep. 06	May 07
#word-sense pairs/#articles	207,016	971,518	1,403,207	1,803,843
#synsets/#categories	117,597	103,759	165,744	244,376

Table 2: Statistics on WordNet and Wikipedia

These results show that our method is robust (thus replicable) disregarding the Wikipedia version. However, we did not observe any improvement despite the quasi-exponential growth of Wikipedia, because the articles added to Wikipedia did not provide crucial information with respect to our experiments.

6. See for instance `http://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia's_growth`.

Dataset		Wikipedia					SVM
		pl	*wup*	*lch*	*res*	*gloss*	
M&C	all	**0.59**	0.54	0.58	0.31	0.58	
	non-miss	0.63	0.61	**0.64**	0.41	0.60	
R&G	all	0.57	0.59	**0.60**	0.30	0.52	
	non-miss	0.60	0.65	**0.66**	0.37	0.54	
353-TC full	all	0.47	0.49	**0.51**	0.35	0.25	
	non-miss	0.48	0.50	**0.53**	0.36	0.26	
353-TC test	all	0.52	0.57	**0.60**	0.49	0.26	**0.62**
	non-miss	0.52	0.57	**0.61**	0.49	0.26	
KLEB full	all	0.26	0.26	**0.27**	0.13	0.11	
	non-miss	0.28	0.28	**0.29**	0.15	0.11	
KLEB test	all	0.27	0.26	**0.28**	0.14	0.12	**0.37**
	non-miss	0.29	0.29	**0.31**	0.17	0.13	

Table 3: Correlation of Wikipedia scores (September 2006) with human judgments

Dataset		Wikipedia					SVM
		pl	*wup*	*lch*	*res*	*gloss*	
M&C	all	**0.57**	0.54	**0.57**	0.31	0.55	
	non-miss	**0.57**	0.54	**0.57**	0.31	0.55	
R&G	all	0.58	0.58	**0.61**	0.38	0.53	
	non-miss	0.58	0.58	**0.61**	0.38	0.53	
353-TC full	all	0.48	0.52	**0.53**	0.41	0.20	
	non-miss	0.48	0.52	**0.54**	0.41	0.20	
353-TC test	all	0.54	0.63	**0.64**	0.60	0.22	**0.66**
	non-miss	0.54	0.63	**0.64**	0.60	0.22	
KLEB full	all	0.31	0.31	**0.33**	0.19	0.10	
	non-miss	0.32	0.33	**0.34**	0.20	0.10	
KLEB test	all	0.28	0.29	**0.30**	0.18	0.11	**0.37**
	non-miss	0.29	0.30	**0.31**	0.18	0.11	

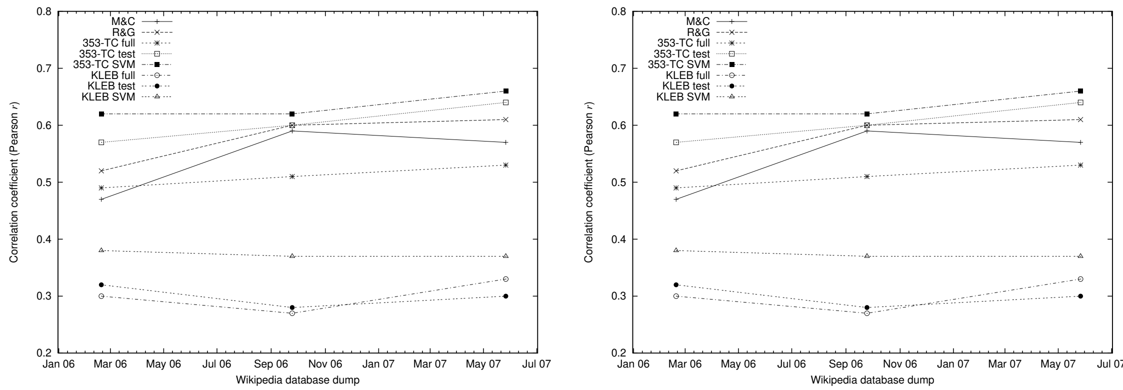
Table 4: Correlation of Wikipedia scores (May 2007) with human judgments

5. Case Study: Coreference Resolution

We extend a machine learning based coreference resolver with features capturing different semantic knowledge sources. These features represent relatedness scores mined from WordNet and Wikipedia. Coreference resolution provides an application to evaluate the performance of the relatedness measures we previously evaluated using only datasets of limited size. This *extrinsic* evaluation provides a better insight on the usefulness of Wikipedia relatedness measures for NLP applications than the intrinsic evaluation described in Section 4.

5.1 Machine Learning Based Coreference Resolution and Semantic Knowledge

The last years have seen a boost of work devoted to the development of machine learning based coreference resolution systems (Soon, Ng, & Lim, 2001; Ng & Cardie, 2002; Yang, Zhou, Su, & Tan, 2003; Luo, Ittycheriah, Jing, Kambhatla, & Roukos, 2004, inter alia). While machine learning has proved to yield performance rates fully competitive with rule based systems, current coreference resolution systems are mostly relying on rather shallow features, such as the distance between the coreferent expressions, string matching, and linguistic form. These shallow features are not sufficient for correctly identifying many of the coreferential relations between expressions



(a) All pairs (b) Non-missing pairs

Figure 3: Performance variation of Wikipedia-based relatedness measures throughout time

in a text. As an example, consider a fragment from the Automatic Content Extraction (ACE) 2003 data.

> But frequent visitors say that given the sheer weight of **the country**'s totalitarian ideology and generations of mass indoctrination, changing **this country**'s course will be something akin to turning a huge ship at sea. Opening **North Korea** up, even modestly, and exposing **people** to the idea that Westerners – and South Koreans – are not devils, alone represents an extraordinary change. [...] as **his people** begin to get a clearer idea of the deprivation **they** have suffered, especially relative to **their** neighbors. "**This** is **a society** that has been focused most of all on stability, [...]".

In order to correctly resolve the coreferent expressions highlighted in bold (which are all annotated as coreferent in the ACE data), lexical semantic and encyclopedic knowledge is required, i.e., that North Korea is a country, that countries consist of people and are societies. The resolution requires a knowledge base (e.g. generated from Wikipedia) look-up and reasoning on the content relatedness holding between the different expressions (e.g. as a path measure along the links of the WordNet and Wikipedia semantic networks). In the following we explore the scenario of including knowledge mined from WordNet and Wikipedia for coreference resolution. We start with a machine learning based baseline system taken from Ponzetto and Strube (2006b), which includes the set of shallow linguistic features from Soon et al. (2001) as well as semantic parsing information in terms of *semantic role labeling* (Gildea & Jurafsky, 2002; Carreras & Màrquez, 2005, SRL henceforth), and analyze the performance variations given by including the previously discussed relatedness measures in the feature set. An overview of the system we present in the remainder of the section is given in Figure 4.

5.2 Coreference Resolution Using Semantic Knowledge Sources

This subsection presents our coreference resolution system which uses semantic relatedness features induced from WordNet and Wikipedia to capture information from these knowledge sources.

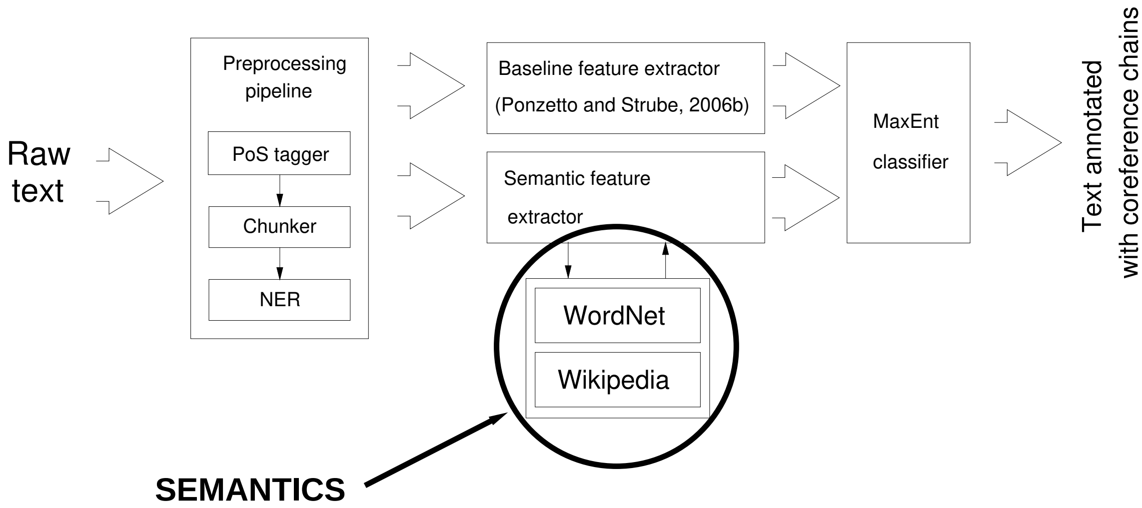


Figure 4: Overview of the coreference resolution system for extrinsic evaluation of WordNet and Wikipedia relatedness measures. We start with a baseline system from Ponzetto and Strube (2006b) which includes the features from Soon et al. (2001) and semantic role information. We then include at different times features from WordNet and Wikipedia and register performance variations.

5.2.1 CORPORA USED

To establish a competitive coreference resolver, the system was initially prototyped using the MUC-6 and MUC-7 data sets (Chinchor & Sundheim, 2003; Chinchor, 2001), using the standard partitioning of 30 texts for training and 20-30 texts for testing. Then, we developed and tested the system with the ACE 2003 Training Data corpus (Mitchell, Strassel, Przybocki, Davis, Doddington, Grishman, Meyers, Brunstain, Ferro, & Sundheim, 2003)[7]. Both the Newswire (NWIRE) and Broadcast News (BNEWS) sections where split into 60-20-20% document-based partitions for training, development, and testing, and later per-partition merged (MERGED) for system evaluation. The distribution of coreference chains and referring expressions is given in Table 5.

5.2.2 LEARNING ALGORITHM

For learning coreference decisions, we used a Maximum Entropy (Berger, Della Pietra, & Della Pietra, 1996) model, implemented using the MALLET library[8]. Coreference resolution is viewed as a binary classification task: given a pair of REs, the classifier has to decide whether they are coreferent or not. The MaxEnt model produces a probability for each category y (coreferent or not) of a candidate pair, conditioned on the context x in which the candidate occurs. The conditional probability is calculated by:

7. We used the training data corpus only, as the availability of the test data is restricted to ACE participants. Therefore, the results we report cannot be compared directly with those using the official test data.
8. `http://mallet.cs.umass.edu`

	BNEWS (147 docs – 33,479 tokens)			
	#coref chains	#pronouns	#common nouns	#proper names
TRAIN.	587	876 (36.1%)	572 (23.6%)	980 (40.3%)
DEVEL	201	315 (33.4%)	163 (17.3%)	465 (49.3%)
TEST	228	291 (30.7%)	238 (25.1%)	420 (44.2%)
TOTAL	1,016	1,482	973	1,865
TOTAL (%)		34.3%	22.5%	43.2%

	NWIRE (105 docs – 57,205 tokens)			
	#coref chains	#pronouns	#common nouns	#proper names
TRAIN.	904	1037 (24.3%)	1210 (28.3%)	2023 (47.4%)
DEVEL	399	358 (20.3%)	485 (27.5%)	923 (52.2%)
TEST	354	329 (21.6%)	484 (31.7%)	712 (46.7%)
TOTAL	1,657	1,724	2,179	3,658
TOTAL (%)		22.8%	28.8%	48.4%

Table 5: Partitions of the ACE 2003 training data corpus

$$p(y|x) = \frac{1}{Z_x}\left[\sum_i \lambda_i f_i(x,y)\right]$$

where $f_i(x, y)$ is the value of feature i on outcome y in context x, and λ_i is the weight associated with i in the model. Z_x is a normalization constant. The features used in our model are all binary-valued feature functions (or indicator functions), e.g.

$$f_{\text{WIKI\_PL}}(\text{WIKI\_PL} = 0.1, \text{COREF}) = \begin{cases} 1 & \text{if candidate pair is coreferent and} \\ & \text{semantic relatedness of their lexical} \\ & \text{heads is 0.1 using the } pl \text{ measure} \\ & \\ 0 & \text{otherwise} \end{cases}$$

We use the L-BFGS algorithm (Malouf, 2002) to estimate the parameters of the Maximum Entropy model. To prevent the model from overfitting, we employ a tunable Gaussian prior as a smoothing method. Training is performed using multi-conditional learning (McCallum, Pal, Druck, & Wang, 2006), a state-of-the-art hybrid method combining generative and discriminative methods for model parameter estimation.

We apply a set of preprocessing components including a POS tagger (Giménez & Màrquez, 2004), NP chunker (Kudoh & Matsumoto, 2000) and the *Alias-I LingPipe* Named Entity Recognizer[9] to the text in order to identify the noun phrases, which are further taken as referring expressions (REs) to be used for instance generation. Therefore, we use automatically extracted noun phrases, rather than assuming perfect NP chunking. This is in contrast to other related works in coreference resolution (e.g., Luo et al., 2004; Kehler, Appelt, Taylor, & Simma, 2004).

Instances are created following Soon et al. (2001). We create a positive training instance from each pair of adjacent coreferent REs. Negative instances are obtained by pairing the anaphoric

9. `http://alias-i.com/lingpipe`

REs with any RE occurring between the anaphor and the antecedent. During testing each text is processed from left to right: each RE is paired with any preceding RE from right to left, until a pair labeled as coreferent is output, or the beginning of the document is reached. The classifier imposes a partitioning on the available REs by clustering each set of expressions labeled as coreferent into the same coreference chain.

5.2.3 BASELINE SYSTEM FEATURES

Following Ng and Cardie (2002), our core baseline system reimplements the Soon et al. (2001) system. The system uses twelve features. Given a potential antecedent RE_i and a potential anaphor RE_j the features are computed as follows[10].

(a) Lexical features

1. **STRING_MATCH** T if RE_i and RE_j have the same spelling, else F.
2. **ALIAS** T if one RE is an alias of the other; else F.

(b) Grammatical features

3. **I_PRONOUN** T if RE_i is a pronoun; else F.
4. **J_PRONOUN** T if RE_j is a pronoun; else F.
5. **J_DEF** T if RE_j starts with *the*; else F.
6. **J_DEM** T if RE_j starts with *this*, *that*, *these*, or *those*; else F.
7. **NUMBER** T if both RE_i and RE_j agree in number; else F.
8. **GENDER** U if either RE_i or RE_j have an undefined gender. Else if they are both defined and agree T; else F.
9. **PROPER_NAME** T if both RE_i and RE_j are proper names; else F.
10. **APPOSITIVE** T if RE_j is in apposition with RE_i; else F.

(c) Semantic features

11. **WN_CLASS** U if either RE_i or RE_j have an undefined WordNet semantic class. Else if they both have a defined one and it is the same T; else F.

(d) Distance features

12. **DISTANCE** how many sentences RE_i and RE_j are apart.

In addition to the 12 features from Soon et al. (2001), we employ SRL features taken from Ponzetto and Strube (2006b). Semantic roles are taken from the output of the ASSERT parser (Pradhan, Ward, Hacioglu, Martin, & Jurafsky, 2004), an SVM based semantic role tagger which uses a full syntactic analysis to automatically identify all verb predicates in a sentence together with their semantic arguments, and output them as PropBank arguments (Palmer, Gildea, & Kingsbury, 2005). It

10. Possible values are U(nknown), T(rue) and F(alse). Note that in contrast to Ng and Cardie (2002) we interpret ALIAS as a lexical feature, as it solely relies on string comparison and acronym string matching.

is often the case that the semantic arguments output by the parser do not align with any of the previously identified noun phrases. In this case, we pass a semantic role label to a RE only when the two phrases share the same head. Labels have the form "$\text{ARG}_1\text{\_pred}_1 \ldots \text{ARG}_n\text{\_pred}_n$" for n semantic roles filled by a constituent, where each semantic argument label is always defined with respect to a predicate. Given such level of semantic information available at the RE level, we introduce two new features.

(e) SRL features

13. **I_SEMROLE** the semantic role argument-predicate pairs of RE_i.
14. **J_SEMROLE** the semantic role argument-predicate pairs of RE_j.

For the ACE 2003 data, 11,406 of 32,502 automatically extracted noun phrases were tagged with 2,801 different argument-predicate pairs. Our baseline feature set is obtained by starting with all the features from Soon et al. (2001), plus the SRL features, and removing those selected using a backward feature selection (see Subsection 5.3.2).

5.2.4 WordNet Features

The WN_CLASS feature from the baseline system is very noisy, because of the lack of coverage, sense proliferation and ambiguity[11]. We accordingly enrich the semantic information available to the classifier by using semantic similarity measures based on the WordNet taxonomy (Pedersen, Patwardhan, & Michelizzi, 2004). The measures we use include path length based measures (Rada et al., 1989; Wu & Palmer, 1994; Leacock & Chodorow, 1998), as well as ones based on information content (Resnik, 1995; Jiang & Conrath, 1997; Lin, 1998).

In our case, the measures are obtained by computing the similarity scores between the head lemmata (for common nouns, e.g. house) or full NPs (for named entities, e.g. George W. Bush) of each potential antecedent-anaphor pair. In order to deal with the sense disambiguation problem, we factorize over all possible sense pairs: given a candidate pair, we take the cross product of each antecedent and anaphor sense to form pairs of synsets. For each measure WN_SIMILARITY, we compute the similarity score for all synset pairs, and create the following features.

15. **WN_SIMILARITY_BEST** the *highest* similarity score from all $\langle \text{SENSE}_{RE_i,n}, \text{SENSE}_{RE_j,m} \rangle$ synset pairs.

16. **WN_SIMILARITY_AVG** the *average* similarity score from all $\langle \text{SENSE}_{RE_i,n}, \text{SENSE}_{RE_j,m} \rangle$ synset pairs.

Pairs containing REs which cannot be mapped to WordNet synsets are assumed to be maximally dissimilar, i.e. their similarity score is set to 0.

5.2.5 Wikipedia Features

We include features derived from Wikipedia. Pages and paths are retrieved following the procedure described in Section 3. As in the case of WordNet, we query the head lemmata of common nouns or the full NPs for named entities. Given a candidate coreference pair $\text{RE}_{i/j}$ and the disambiguated

11. Following Soon et al. (2001) we mapped each RE to the first WordNet sense of the head noun.

Wikipedia pages $p_{RE_{i/j}}$ they point to, obtained by querying pages titled as $t_{RE_{i/j}}$, we extract the following features:

17. **I/J_GLOSS_CONTAINS** U if no Wikipedia page titled $t_{RE_{i/j}}$ is available. Else T if the first paragraph of text of $p_{RE_{i/j}}$ contains $t_{RE_{j/i}}$; else F.

18. **I/J_RELATED_CONTAINS** U if no Wikipedia page titled as $t_{RE_{i/j}}$ is available. Else T if at least one Wikipedia hyperlink of $p_{RE_{i/j}}$ contains $t_{RE_{j/i}}$; else F.

19. **I/J_CATEGORIES_CONTAINS** U if no Wikipedia page titled as $t_{RE_{i/j}}$ is available. Else T if the list of categories $p_{RE_{i/j}}$ belongs to contains $t_{RE_{j/i}}$; else F.

20. **GLOSS_OVERLAP** the overlap score between the first paragraph of text of p_{RE_i} and p_{RE_j} computed using Equation 2.

Additionally, we use the Wikipedia category graph. We use the relatedness measures described in Subsections 2.2.2 and 2.2.3. Given $p_{RE_{i/j}}$ and the lists of categories $C_{RE_{i/j}}$ they belong to, we factorize over all possible category pairs. That is, we take the cross product of each antecedent and anaphor category to form pairs of 'Wikipedia synsets'. For each measure WIKI_RELATEDNESS, we compute the relatedness score for all category pairs, and create the following features.

21. **WIKI_RELATEDNESS_BEST** the *highest* relatedness score from all $\langle C_{RE_i,n}, C_{RE_j,m}\rangle$ category pairs.

22. **WIKI_RELATEDNESS_AVG** the *average* relatedness score from all $\langle C_{RE_i,n}, C_{RE_j,m}\rangle$ category pairs.

5.3 Experiments

5.3.1 PERFORMANCE METRICS

We report in the following tables the MUC score (Vilain, Burger, Aberdeen, Connolly, & Hirschman, 1995). Scores in Table 6 are computed for all noun phrases appearing in either the key or the system response, whereas Tables 7 and 8 refer to scoring only those phrases which appear in both the key and the response. We therefore discard those responses not present in the key, as the only referring expressions annotated in the ACE 2003 data belong to the categories *Person* (i.e. humans), *Organization* (e.g. corporations, agencies), *Facility* (i.e. buildings), *Location* (e.g. geographical areas and landmasses) and *Geo-political entities* (i.e. nations, regions, their government and people, cf. the 'North Korea' example above)[12]. This makes it impossible to perform 'full' coreference resolution including, e.g. the identification of referential expressions, and implies that the results establish the upper limit of the improvements given by our semantic features.

We also report the accuracy score for all three types of ACE mentions, namely pronouns, common nouns and proper names. Accuracy is the percentage of REs of a given mention type correctly resolved divided by the total number of REs of the same type having a direct antecedent given in the key. A RE is said to be correctly resolved when both it and its direct antecedent identify mentions which belong to the same coreference class in the key.

12. Cf. the ACE 2003 Entity Detection and Tracking (EDT) Annotation Guidelines available at `http://projects.ldc.upenn.edu/ace/docs/EDT-Guidelines-V2-5.pdf`: "We do not identify mentions of animals or most inanimate objects at this time".

	MUC-6			MUC-7		
original	R	P	F_1	R	P	F_1
Soon et al.	58.6	67.3	62.3	56.1	65.5	60.4
duplicated baseline	64.9	65.6	65.3	55.1	68.5	61.1

Table 6: Results on MUC

5.3.2 FEATURE SELECTION

For determining the relevant feature sets we follow an iterative procedure similar to the wrapper approach for feature selection (Kohavi & John, 1997) using the development data. The feature selection algorithm performs a hill-climbing search along the feature space. In the case of the baseline system we perform backward feature selection, since we are interested in obtaining a minimal feature set which provides the most competitive baseline[13]. We start with a model based on all available features (Ponzetto & Strube, 2006b). Then we train models obtained by removing one feature at a time. We choose the worst performing feature, namely the one whose removal gives the largest improvement based on the MUC score F-measure on the development data, and remove it from the model. We then train classifiers removing each of the remaining features separately from the enhanced model. The process is iteratively run as long as significant improvement is observed.

For evaluating WordNet and Wikipedia features, we perform instead forward greedy feature selection: we start with the minimal set of previously kept baseline features and iteratively add those features from WordNet or Wikipedia which give the best MUC score F-measure improvement on the development data for each selection step. This is because we are interested in evaluating the additional contribution of such information, and avoid external factors such as improvements due to further removal of the baseline features. A summary of the features selected by the backward and forward feature selections is given in Table 9.

5.3.3 RESULTS

Table 6 compares the results between our duplicated Soon et al. (2001) baseline and the original system on the MUC data. We assume that the slight improvements of our system are due to the use of current preprocessing components and another classifier. Tables 7 and 8 show a comparison of the performance between our baseline system (Ponzetto & Strube, 2006b) and the ones incremented with semantic features mined from WordNet and Wikipedia on the ACE data. Statistically significant performance improvements are highlighted in bold[14].

13. This is under the assumption that adding the SRL features to the core baseline features from Soon et al. (2001) could not yield the best performance. In practice, analysis of feature selection on the development data highlights that Soon et al.'s (2001) features such as J_DEM, PROPER_NAME and WN_CLASS (BNEWS), J_DEM (NWIRE) and DISTANCE (MERGED) are indeed removed from the baseline feature set when SRL information is included. None of the best performing features of Soon et al. (2001) (STRING_MATCH, ALIAS and APPOSITIVE) is removed, thus supporting their feature relevance analysis (pp. 534–535).

14. We take performance variations between different experimental runs to be statistically significant in case the changes in the MUC F-measure are statistically significant at the 0.05 level or higher. We follow Soon et al. (2001) in performing a simple one-tailed, paired sample *t*-test between the baseline system's MUC score F-measure and each of the other systems' F-measure scores on the test documents.

	BNEWS						NWIRE					
	R	P	F_1	A_p	A_{cn}	A_{pn}	R	P	F_1	A_p	A_{cn}	A_{pn}
baseline	50.5	82.0	62.5	44.2	17.4	58.0	56.3	86.7	68.3	43.8	35.0	71.6
+WordNet	**59.1**	**82.4**	**68.8**	43.1	**40.9**	**64.6**	**62.4**	81.4	**70.7**	**45.4**	**43.0**	68.7
+Wikipedia	**58.3**	81.9	**68.1**	41.2	**38.9**	**62.3**	**60.7**	81.8	**69.7**	**44.1**	**40.1**	71.6

Table 7: Results on the ACE 2003 data (BNEWS and NWIRE sections)

	R	P	F_1	A_p	A_{cn}	A_{pn}
baseline	54.5	85.4	66.5	40.5	30.1	73.0
+WordNet	**60.6**	79.4	**68.7**	**42.4**	**43.2**	66.0
+Wikipedia	**59.4**	82.2	**68.9**	38.9	**41.4**	**74.5**

Table 8: Results ACE (merged BNEWS/NWIRE)

5.3.4 DISCUSSION

The tables show that *semantic features improve system recall*, rather than acting as a 'semantic filter' improving precision. Semantics therefore seems to trigger a response in cases where more shallow features do not seem to suffice. A one-tailed, paired sample *t*-test reveals that on the BNEWS and MERGED sections *the difference in performance between WordNet and Wikipedia are not statistically significant* ($p < 0.05$), thus proving that Wikipedia is indeed competitive with WordNet in the coreference resolution scenario.

WordNet and Wikipedia features tend to consistently increase performance on common nouns on all dataset partitions. WordNet features are able to improve by 23.5%, 8% and 13.1% the accuracy rate for common nouns on the BNEWS, NWIRE and MERGED datasets (+35, +28 and +65 correctly resolved common nouns out of 149 and 349 and 498 respectively), whereas employing Wikipedia yields slightly smaller improvements (+21.5%, +5.1% and +11.3% accuracy increase on the same datasets). The accuracy on common nouns shows that features derived from Wikipedia are competitive with the ones from WordNet. The performance gap on all three datasets is relatively small, which indicates the usefulness of using an encyclopedic knowledge base as a replacement for a lexical taxonomy.

If semantic relatedness clearly helps for common nouns, it does not always improve the performance on proper names, where features such as string matching and alias suffice, cf. the performance degradation induced by WordNet on the NWIRE and MERGED datasets. This suggests that the semantic information we use tends to play a role mostly for common noun resolution, where surface features cannot account for complex preferences and semantic knowledge is required. Nevertheless, Wikipedia exhibits in general a better performance for the resolution of proper names than WordNet, as it yields results which are always at least as good as the baseline. This is not surprising, as Wikipedia contains a larger amount of information about named entities than WordNet. In particular, qualitative analysis on the development data shows that Wikipedia is useful for instance for identifying cases of REs coreferent with the same geo-political discourse entity, e.g. Yemeni and Yemen or American and United States, thanks to the feature of redirection, i.e. YEMENI redirects to YEMEN. While from a linguistic point of view it is far from clear whether such cases represent genuine cases of coreference, redirection helps to cover meronymy.

		BNEWS	NWIRE	MERGED
backward feature selection	starting features removed	WN_CLASS PROPER_NAME J_DEM	J_DEM	DISTANCE
forward feature selection	WordNet features added	jcn average jcn best pl best wup average	jcn best lin average pl best	lch best pl best pl average
	Wikipedia features added	wup average pl best pl average $gloss$	wup average lch best	pl best pl average

Table 9: Feature selection

Feature selection improves the results[15]. This is due to the fact that our full feature set is extremely redundant: in order to explore the usefulness of the knowledge sources we included overlapping features (i.e. using *best* and *average* similarity/relatedness measures at the same time), as well as features capturing the same phenomenon from different points of view (i.e. using *multiple* measures at the same time). In order to yield the desired performance improvements, it turns out to be essential to filter out irrelevant features. For instance, in the case of Wikipedia none of the I/J_GLOSS, RELATED or CATEGORIES_CONTAINS features survives the feature selection process (see Table 9). On the other hand, in all cases for both WordNet and Wikipedia at least one best and one average measure is always included among the selected features. This suggests that including information about all available senses (in terms of average relatedness measures) provides sensible information to handle ambiguity. Finally, multiple measures are included in most of the selected feature sets, e.g. the selected features for WordNet on the BNEWS data include both the measure from Wu and Palmer (1994) and the one from Jiang and Conrath (1997), indicating that rather than having a best overall measure, competitive results can be obtained by integrating different ones.

6. Experiments for German

Except for Gurevych (2005) who ported semantic relatedness measures to German using GermaNet (Lemnitzer & Kunze, 2002), the topic of semantic relatedness has been explored almost exclusively for the English language using WordNet. Research about semantic relatedness in languages other than English has been hindered by differences in the structure and organization of the respective wordnets and by a large variation in coverage. For instance, Gurevych and Niederlich (2005a) had to implement an API specifically designed for GermaNet access. Furthermore, GermaNet is much smaller than WordNet and it did not even cover all word pairs in the relatively small dataset provided by Rubenstein and Goodenough (1965) which Gurevych and Niederlich (2005b) translated into German. In contrast, the structure and organization of Wikipedia is the same across all languages, so that semantic relatedness measures developed for English can be applied to other languages

15. We experienced during system prototyping that simply including the full feature set without performing feature selection gave worst or no significant performance variations at all.

	GermaNet 4.0	German Wikipedia		
		Feb. 06	Jun. 06	Sep. 06
#word-sense pairs/#articles	60,646	387,586	410,586	471,065
#synsets/#categories	41,777	25,035	28,656	33,130

Table 10: Statistics on GermaNet and Wikipedia

without changing the methods for accessing them. The coverage of the German Wikipedia is also considerably large as shown in Table 10.

To our knowledge there exist no datasets for evaluating semantic relatedness measures in languages other than English. Only Gurevych translated the dataset by Rubenstein and Goodenough (1965) into German and supplied it with judgments by native speakers of German. Gurevych (2005) evaluated several semantic relatedness measures on the R&G dataset using GermaNet as knowledge source. Because GermaNet did not cover all words in the 65 word pairs they report only results for 57 word pairs with the best results obtained by *res* followed by *lin*. In Table 11 we adopt the results reported by Gurevych and supply them with numbers computed for all 65 word pairs including the ones which are not covered by GermaNet. As in the case of the results given in Table 1, word pairs where at least one of the words could not be found in GermaNet are assigned a relatedness score of 0[16].

The numbers we obtain in our experiments using Wikipedia compare well with the numbers they reported. In contrast to our results for the English R&G dataset, Wikipedia performs as good as GermaNet when considering all pairs – viz., no statistically significant difference using a one-tailed paired sample *t*-test ($p < 0.05$) between the respective best performing measures, GermaNet lin and Wikipedia wup – and shows a performance only slightly below when considering available pairs only. The German R&G dataset is the one on which Wikipedia shows the best overall performance, and the only one on which we have a larger coverage than the corresponding wordnet (4 missing pairs versus 8).

As previously pointed out in the analysis of the performance on the 353-TC English dataset, Wikipedia is able to yield a competitive performance on datasets specifically designed to capture semantic relatedness, rather than the stricter notion of similarity. This seems to be supported in the present scenario as well, as the data in Gurevych and Niederlich (2005b) are rated explicitly for semantic relatedness, using the definition from Budanitsky and Hirst (2006) that we summarized in Section 2.2.1.

7. Implementation Details

Wikipedia is freely available for download, and can be accessed using robust Open Source applications, e.g. the MediaWiki software[17], integrated within a Linux, Apache, MySQL and PHP (LAMP) software bundle. We briefly present in the following the main components of the Application Programming Interface (API) we developed as part of the present work[18]. The architecture of the WikiRelate! API (Ponzetto & Strube, 2007a) consists of the following modules:

16. The complete list of German word pairs with the corresponding human judgments and automatically computed measures can be obtained from Gurevych and Niederlich (2005b).
17. `http://www.mediawiki.org`
18. The WikiRelate! software can be downloaded from `http://www.eml-research.de/nlp`.

Dataset		Google	GermaNet 4.0			Wikipedia (February 2006)				
		jaccard	*lin*	*res*	*lesk*	*pl*	*wup*	*lch*	*res*	*gloss*
R&G	all	0.26	0.66	0.64	0.49	0.58	0.65	0.64	0.62	0.33
	non-miss	0.26	0.73	0.76	0.53	0.62	0.70	0.69	0.67	0.34
						Wikipedia (June 2006)				
	all					0.58	0.65	0.64	0.58	0.33
	non-miss					0.61	0.69	0.69	0.62	0.35
						Wikipedia (September 2006)				
	all					0.59	0.65	0.64	0.50	0.38
	non-miss					0.63	0.70	0.69	0.55	0.40

Table 11: Results for German R&G Dataset

1. **RDBMS**: at the lowest level, the encyclopedia content is stored into a relational database management system (e.g. MySQL).

2. **MediaWiki**: a suite of PHP routines for interacting with the RDBMS.

3. **WWW-Wikipedia Perl library**[19]: responsible for querying MediaWiki, parsing and structuring the returned encyclopedia pages.

4. **XML-RPC server**: an intermediate communication layer between Java and the Perl routines.

5. **Java wrapper library**: provides a simple façade to create and access the encyclopedia page objects and compute the relatedness scores.

The information flow of the API is summarized by the sequence diagram in Figure 5. The higher input/output layer the user interacts with is provided by a Java API from which Wikipedia can be queried. The API provides factory classes for querying Wikipedia, in order to retrieve the encyclopedia entries as well the relatedness scores for word pairs. In practice, the Java library works as a wrapper in order to provide a simple user access in terms of a façade. The library is responsible for issuing HTTP requests to an XML-RPC daemon which provides a layer for calling Perl routines from the Java API. Perl routines take care of the bulk of querying encyclopedia entries to the MediaWiki software (which in turn queries the database) and efficiently parsing the text responses into structured objects.

8. Related Work

In this section we relate our work to the existing body of literature on computing semantic relatedness and its application to various NLP tasks.

19. `http://search.cpan.org/dist/WWW-Wikipedia/`

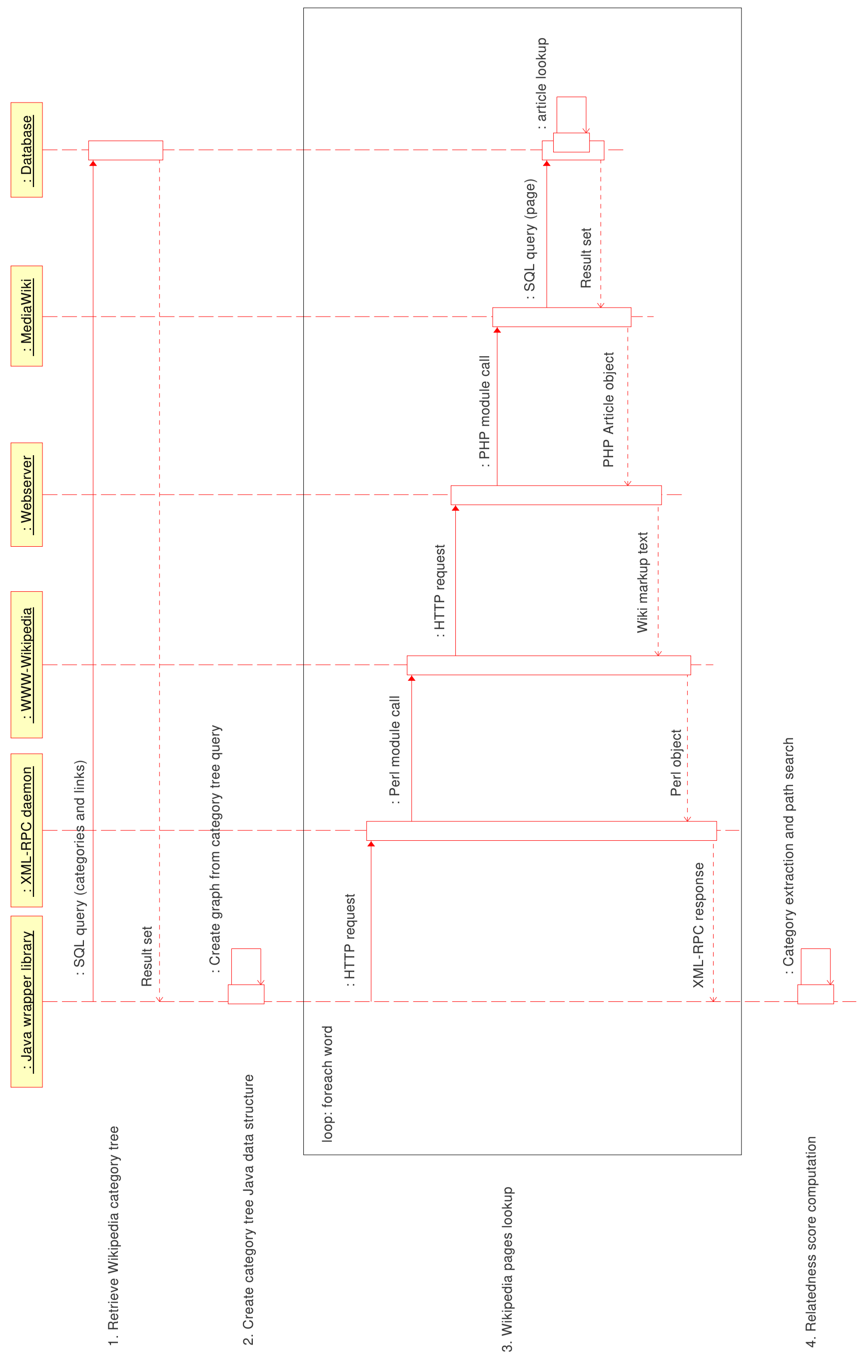


Figure 5: WikiRelate! API processing sequence diagram. Wikipedia pages and relatedness measures are accessed through a Java API façade. The wrapper communicates with a Perl library designed for Wikipedia access and parsing through an XML-RPC server. WWW-Wikipedia in turn accesses the database where the encyclopedia is stored by means of appropriate queries to MediaWiki.

8.1 Computing Semantic Relatedness

Research in the area of computing semantic relatedness and similarity can be generally divided into two categories – firstly measures of distributional similarity using largely unstructured information such as text and secondly approaches using structured lexical databases such as WordNet.

Measures of distributional similarity (Landauer & Dumais, 1997; Lee, 1999; Dagan, 2000; Turney, 2001; Weeds & Weir, 2005, inter alia) are based on the *distributional* hypothesis, i.e. on the hypothesis that similar words appear in similar contexts and hence have similar meaning. For reasons discussed in detail by Budanitsky and Hirst (2006, pp.41–44), measures of distributional similarity and measures of semantic relatedness are distinct, because (1) measures of semantic relatedness cover relations between concepts while measures of distributional similarity capture relations between words; (2) semantic relatedness is a symmetric relation while distributional similarity is a potentially asymmetric relationship; (3) measures of semantic relatedness depend on a predefined knowledge source which is created by humans and may be presumed "true, unbiased and complete" (Budanitsky & Hirst, 2006, p.43); measures of distributional similarity depend entirely on corpora causing problems of imbalance and data sparseness; this problem may only be overcome by using representative, very large corpora; however, computing distributional similarity does not scale well (Gorman & Curran, 2006). Budanitsky and Hirst (2006) conclude that measures of distributional similarity cannot replace measures of semantic relatedness and similarity.

Approaches using structured lexical databases can be traced back to work by Rada et al. (1989) who measured semantic similarity in MeSH, a term hierarchy for indexing articles in Medline. They compute semantic similarity straightforwardly in terms of the numbers of edges between terms in the hierarchy. Research in this area proceeded in two directions. Firstly, different knowledge sources were proposed. Early on, WordNet was used to provide a broad coverage lexical database (Resnik, 1993; Wu & Palmer, 1994; Resnik, 1995). Later Jarmasz and Szpakowicz (2003) explored the use of Roget's Thesaurus for computing semantic similarity. Secondly, major advances were achieved by developing more sophisticated measures of semantic similarity and relatedness.

In this article we propose a new knowledge source for computing semantic relatedness, i.e. Wikipedia and its associated categorization network. We believe that many NLP applications will benefit from using Wikipedia and evaluate this hypothesis by including Wikipedia based semantic relatedness measures as features in a state-of-the-art coreference resolution system.

8.2 Using Semantic Relatedness in Coreference and Other NLP Applications

Vieira and Poesio (2000), Harabagiu et al. (2001), and Markert and Nissim (2005) explore the use of WordNet for different coreference resolution subtasks, such as resolving bridging references, *other*- and definite NP anaphora, and MUC-style coreference resolution. All of them present systems which infer coreference relations from a set of potential antecedents by means of a WordNet search. Our approach to WordNet here is to cast the search results in terms of semantic similarity measures. Their output can be used as features for a learner. These measures are not specifically developed for coreference resolution but simply taken 'off-the-shelf' and applied to our task without any specific tuning — e.g. in contrast to Harabagiu et al. (2001), who weight WordNet relations differently in order to compute the confidence measure of the path.

Semantic relatedness measures have been proven to be useful in many applications in Natural Language Processing such as word sense disambiguation (Kohomban & Lee, 2005; Patwardhan, Banerjee, & Pedersen, 2005), information retrieval (Finkelstein et al., 2002), information extraction pattern induction (Stevenson & Greenwood, 2005), interpretation of noun compounds (Kim & Baldwin, 2005), paraphrase detection (Mihalcea, Corley, & Strapparava, 2006) and spelling correction (Budanitsky & Hirst, 2006).

9. Conclusions

In this article we investigated the use of Wikipedia for computing semantic relatedness and its application to a real-world NLP task, coreference resolution. We assumed the Wikipedia category graph to represent a semantic network modeling relations between concepts, and we computed their relatedness from it. Even if the categorization feature has been introduced into Wikipedia only three years ago, our results indicate that semantic relatedness computed using the Wikipedia category network consistently correlates better with human judgments than a simple baseline based on Google counts. It is also competitive with WordNet for datasets specifically modeling semantic relatedness human judgments. Because all available dataset are small and seem to be assembled rather arbitrarily we perform an extrinsic evaluation with an NLP application, i.e. a coreference resolution system, where we register for some datasets no statistically significant differences between the improvements given by features induced from WordNet and the ones from Wikipedia.

Wikipedia provides a large amount of information as encyclopedic entries at the leaves of the category network, e.g. named entities. The encyclopedia gets continuously updated and the derived knowledge can be used to analyze current information. The text and the category network both provide semi-structured information and can be mined with more precision than unstructured data gathered from the web. Unfortunately, the Wikipedia categorization still suffers from some limitations, i.e., it cannot be considered a fully-fledged ontology, as the relations between categories are not semantically-typed. In the near future we will concentrate on making the semantic relations between concepts explicit in the Wikipedia category network (Ponzetto & Strube, 2007b). The availability of explicit semantic relations will allow for inducing *semantic similarity* rather than *semantic relatedness* measures, which may be more suitable for coreference resolution. What is most interesting about our results is that they indicate that a collaboratively created folksonomy can actually be used in NLP applications with the same benefit as hand-crafted taxonomies or ontologies.

Acknowledgments. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first author has been supported by a KTF grant (09.003.2004). We thank three anonymous JAIR reviewers for their extensive reviews and our colleague Vivi Nastase for useful feedback.

References

Ahn, D., Jijkoun, V., Mishne, G., Müller, K., de Rijke, M., & Schlobach, S. (2004). Using Wikipedia at the TREC QA track. In *Proceedings of the Thirteenth Text REtrieval Conference,* Gaithersburg, Md., 16–19 November 2004.

Ahn, K., Bos, J., Curran, J. R., Kor, D., Nissim, M., & Webber, B. (2005). Question answering with QED at TREC-2005. In *Proceedings of the Fourteenth Text REtrieval Conference,* Gaithersburg, Md., 15–18 November 2005.

Banerjee, S., & Pedersen, T. (2003). Extended gloss overlap as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence,* Acapulco, Mexico, 9–15 August 2003, pp. 805–810.

Beigman Klebanov, B. (2006). Semantic relatedness: Computational investigation of human data. In *Proceedings of the 3rd Midwest Computational Linguistics Colloquium,* Urbana-Champaign, Ill., 20-21 May 2006.

Beigman Klebanov, B., & Shamir, E. (2006). Reader-based exploration of lexical cohesion. *Language Resources and Evaluation*, *40*(2), 109–126.

Berger, A., Della Pietra, S. A., & Della Pietra, V. J. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, *22*(1), 39–71.

Bos, J., & Markert, K. (2005). Recognising textual entailment with logical inference. In *Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processing,* Vancouver, B.C., Canada, 6–8 October 2005, pp. 628–635.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, *32*(1), 13–47.

Bunescu, R., & Paşca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics,* Trento, Italy, 3–7 April 2006, pp. 9–16.

Carreras, X., & Màrquez, L. (2005). Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the 9th Conference on Computational Natural Language Learning,* Ann Arbor, Mich., USA, 29–30 June 2005, pp. 152–164.

Charniak, E. (1973). Jack and Janet in search of a theory of knowledge. In *Advance Papers from the Third International Joint Conference on Artificial Intelligence, Stanford, Cal.*, pp. 337–343, Los Altos, Cal. W. Kaufmann.

Chinchor, N. (2001). Message Understanding Conference (MUC) 7. LDC2001T02, Philadelphia, Penn: Linguistic Data Consortium.

Chinchor, N., & Sundheim, B. (2003). Message Understanding Conference (MUC) 6. LDC2003T13, Philadelphia, Penn: Linguistic Data Consortium.

Dagan, I. (2000). Contextual word similarity. In Dale, R., Moisl, H., & H., S. (Eds.), *Handbook of Natural Language Processing*, pp. 459–476. New York, N.Y.: Marcel Dekker Inc.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, *20*(1), 116–131.

Gabrilovich, E., & Markovitch, S. (2006). Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the 21st National Conference on Artificial Intelligence,* Boston, Mass., 16–20 July 2006, pp. 1301–1306.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence,* Hyderabad, India, 6–12 January 2007, pp. 1606–1611.

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, *28*(3), 245–288.

Giles, J. (2005). Internet encyclopedias go head to head. *Nature*, *438*, 900–901.

Giménez, J., & Màrquez, L. (2004). SVMTool: A general POS tagger generator based on support vector machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation,* Lisbon, Portugal, 26–28 May 2004, pp. 43–46.

Gorman, J., & Curran, J. R. (2006). Scaling distributional similarity to large corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics,* Sydney, Australia, 17–21 July 2006, pp. 361–368.

Gurevych, I. (2005). Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing,* Jeju Island, South Korea, 11–13 October 2005, pp. 767–778.

Gurevych, I., & Niederlich, H. (2005a). Accessing GermaNet data and computing semantic relatedness. In *Companion Volume to the Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,* Ann Arbor, Mich., 25–30 June 2005, pp. 5–8.

Gurevych, I., & Niederlich, H. (2005b). Measuring semantic relatedness of GermaNet word senses. Tech. rep., EML Research gGmbH.

Harabagiu, S. M., Bunescu, R. C., & Maiorano, S. J. (2001). Text and knowledge mining for coreference resolution. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics,* Pittsburgh, Penn., 2–7 June 2001, pp. 55–62.

Hobbs, J. R. (1978). Resolving pronominal references. *Lingua*, *44*, 311–338.

Hovy, E., Gerber, L., Hermjakob, U., Junk, M., & Lin, C.-Y. (2001). Question answering in Webclopedia. In *Proceedings of the Thirteenth Text REtrieval Conference,* Gaithersburg, Md., 13–16 November 2001.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2006). *A Practical Guide to Support Vector Classification*. http://www.csie.ntu.edu.tw/∼cjlin/papers/guide/guide.pdf.

Jarmasz, M., & Szpakowicz, S. (2003). Roget's Thesaurus and semantic similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing,* Borovets, Bulgaria, 10–12 September 2003, pp. 212–219.

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics (ROCLING).*

Kehler, A., Appelt, D., Taylor, L., & Simma, A. (2004). The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* Boston, Mass., 2–7 May 2004, pp. 289–296.

Kim, S. N., & Baldwin, T. (2005). Automatic interpretation of noun compounds using WordNet similarity. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing,* Jeju Island, South Korea, 11–13 October 2005, pp. 945–956.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence Journal*, *97*(1-2), 273–324.

Kohomban, U. S., & Lee, W. S. (2005). Learning semantic classes for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,* Ann Arbor, Mich., 25–30 June 2005, pp. 34–41.

Kudoh, T., & Matsumoto, Y. (2000). Use of Support Vector Machines for chunk identification. In *Proceedings of the 4th Conference on Computational Natural Language Learning,* Lisbon, Portugal, 13–14 September 2000, pp. 142–144.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In Fellbaum, C. (Ed.), *WordNet. An Electronic Lexical Database*, chap. 11, pp. 265–283. Cambridge, Mass.: MIT Press.

Lee, L. (1999). Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics,* College Park, Md., 20–26 June 1999, pp. 25–31.

Lemnitzer, L., & Kunze, C. (2002). GermaNet – representation, visualization, application. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation,* Las Palmas, Canary Islands, Spain, 29–31 May 2002, pp. 1485–1491.

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 Conference on Systems Documentation,* Toronto, Ontario, Canada, pp. 24–26.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning,* Madison, Wisc., 24–27 July 1998, pp. 296–304.

Lo, K. K., & Lam, W. (2006). Using semantic relations with world knowledge for question answering. In *Proceedings of the Fifteenth Text REtrieval Conference,* Gaithersburg, Md., 14–17 November 2006.

Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., & Roukos, S. (2004). A mention-synchronous coreference resolution algorithm based on the Bell Tree. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics,* Barcelona, Spain, 21–26 July 2004, pp. 136–143.

Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the 6th Conference on Computational Natural Language Learning,* Taipei, Taiwan, 31 August – 1 September 2002, pp. 49–55.

Markert, K., & Nissim, M. (2005). Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics*, *31*(3), 367–401.

McCallum, A., Pal, C., Druck, G., & Wang, X. (2006). Multi-conditional learning: Generative/discriminative training for clustering and classification. In *Proceedings of the 21st National Conference on Artificial Intelligence,* Boston, Mass., 16–20 July 2006, pp. 433–439.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). YALE: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, Philadelphia, Penn., 20–23 August 2006, pp. 935–940.

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence,* Boston, Mass., 16–20 July 2006, pp. 775–780.

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

Mitchell, A., Strassel, S., Przybocki, M., Davis, J., Doddington, G., Grishman, R., Meyers, A., Brunstain, A., Ferro, L., & Sundheim, B. (2003). TIDES extraction (ACE) 2003 multilingual training data. LDC2004T09, Philadelphia, Penn.: Linguistic Data Consortium.

Ng, V., & Cardie, C. (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* Philadelphia, Penn., 7–12 July 2002, pp. 104–111.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, *31*(1), 71–105.

Patwardhan, S., Banerjee, S., & Pedersen, T. (2005). SenseRelate::TargetWord – A generalized framework for word sense disambiguation. In *Proceedings of the 20th National Conference on Artificial Intelligence,* Pittsburgh, Penn., 9–13 July 2005.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet::Similarity – Measuring the relatedness of concepts. In *Companion Volume to the Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* Boston, Mass., 2–7 May 2004, pp. 267–270.

Poesio, M., Ishikawa, T., Schulte im Walde, S., & Vieira, R. (2002). Acquiring lexical knowledge for anaphora resolution. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation,* Las Palmas, Canary Islands, Spain, 29–31 May 2002, pp. 1220–1225.

Ponzetto, S. P., & Strube, M. (2006a). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* New York, N.Y., 4–9 June 2006, pp. 192–199.

Ponzetto, S. P., & Strube, M. (2006b). Semantic role labeling for coreference resolution. In *Companion Volume to the Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics,* Trento, Italy, 3–7 April 2006, pp. 143–146.

Ponzetto, S. P., & Strube, M. (2007a). An API for measuring the relatedness of words in Wikipedia. In *Companion Volume to the Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics,* Prague, Czech Republic, 23–30 June 2007, pp. 49–52.

Ponzetto, S. P., & Strube, M. (2007b). Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence,* Vancouver, B.C., Canada, 22–26 July 2007, pp. 1440–1447.

Pradhan, S., Ward, W., Hacioglu, K., Martin, J. H., & Jurafsky, D. (2004). Shallow semantic parsing using Support Vector Machines. In *Proceedings of the Human Language Technology*

Conference of the North American Chapter of the Association for Computational Linguistics, Boston, Mass., 2–7 May 2004, pp. 233–240.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric to semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, *19*(1), 17–30.

Resnik, P. (1993). *Selection and Information: A Class-based Approach to Lexical Relationships*. Ph.D. thesis, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, Penn.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence,* Montréal, Canada, 20–25 August 1995, Vol. 1, pp. 448–453.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, *11*, 95–130.

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

Salton, G., & McGill, M. (1983). *Introduction to Modern Information Retrieval*. New York, N.Y.: McGraw-Hill.

Seco, N., Veale, T., & Hayes, J. (2004). An intrinsic information content metric for semantic similarity in WordNet. In *Proceedings of the 16th European Conference on Artificial Intelligence,* Valencia, Spain, 23–27 August 2004, pp. 1089–1090.

Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, *27*(4), 521–544.

Stevenson, M., & Greenwood, M. (2005). A semantic approach to IE pattern induction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,* Ann Arbor, Mich., 25–30 June 2005, pp. 379–386.

Strube, M., & Ponzetto, S. P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence,* Boston, Mass., 16–20 July 2006, pp. 1419–1424.

Tatu, M., Iles, B., Slavick, J., Novischi, A., & Moldovan, D. (2006). COGEX at the Second Recognizing Textual Entailment Challenge. In *Proceedings of the Second PASCAL Recognising Textual Entailment Challenge Workshop,* Venice, Italy, 10 April 2006, pp. 104–109.

Turney, P. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning,* Freiburg, Germany, 3–7 September, 2001, pp. 491–502.

van Rijsbergen, C. (1979). *Information Retrieval*. London, U.K.: Butterworths.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, Germany.

Vieira, R., & Poesio, M. (2000). An empirically-based system for processing definite descriptions. *Computational Linguistics*, *26*(4), 539–593.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., & Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pp. 45–52, San Mateo, Cal. Morgan Kaufmann.

Weeds, J., & Weir, D. (2005). Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, *31*(4), 439–475.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics,* Las Cruces, N.M., 27–30 June 1994, pp. 133–138.

Yang, X., Zhou, G., Su, J., & Tan, C. L. (2003). Coreference resolution using competition learning approach. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* Sapporo, Japan, 7–12 July 2003, pp. 176–183.