WikiRelate! Computing Semantic Relatedness Using Wikipedia

Michael Strube and **Simone Paolo Ponzetto**
EML Research gGmbH
Schloss-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
http://www.eml-research.de/nlp

Abstract

Wikipedia provides a knowledge base for computing word relatedness in a more structured fashion than a search engine and with more coverage than WordNet. In this work 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 when applied to the largest available dataset designed for that purpose. The best results on this dataset are obtained by integrating Google, WordNet and Wikipedia based measures. We also show that including Wikipedia improves the performance of an NLP application processing naturally occurring texts.

Introduction

Semantic relatedness indicates how much two concepts are related in a taxonomy by using all relations between them (i.e. hyponymic/hypernymic, meronymic and any kind of functional relations including *has-part*, *is-made-of*, *is-an-attribute-of*, etc.). When limited to hyponymy/hyperonymy (i.e. *is-a*) relations, the measure quantifies *semantic similarity* instead. Semantic relatedness measures are used in many applications in Natural Language Processing (NLP) such as word sense disambiguation (Patwardhan et al., 2005), information retrieval (Finkelstein et al., 2002), interpretation of noun compounds (Kim & Baldwin, 2005) and spelling correction (Budanitsky & Hirst, 2006).

Most of the work dealing with relatedness and similarity measures has been developed using WordNet (Fellbaum, 1998). While WordNet represents a well structured taxonomy organized in a meaningful way, questions arise about the need for a larger coverage. E.g., WordNet 2.1 does not include information about named entities such as *Condoleezza Rice*, *Salvador Allende* or *The Rolling Stones* as well as specialized concepts such as *exocytosis* or *P450*.

In contrast, Wikipedia provides entries on a vast number of named entities and very specialized concepts. The English version, as of 14 February 2006, contains 971,518 articles with 18.4 million internal hyperlinks, thus providing a large coverage knowledge resource developed by a large community, which is very attractive for information extraction applications[1]. In addition, since May 2004 it provides also a taxonomy by means of its *categories*: articles can be assigned one or more categories, which are further categorized to provide a category tree. In practice, the taxonomy is not designed as a strict hierarchy or tree of categories, but allows multiple categorization schemes to co-exist simultaneously. As of January 2006, 94% of the articles have been categorized into 91,502 categories.

The strength of Wikipedia lies in its size, which could be used to overcome current knowledge bases' limited coverage and scalability issues. Such size represents on the other hand a challenge: the search space in the Wikipedia category graph is very large in terms of depth, branching factor and multiple inheritance relations, which creates problems related to finding efficient mining methods. In addition, the category relations in Wikipedia cannot only be interpreted as corresponding to *is-a* links in a taxonomy since they denote meronymic relations as well. As an example, the Wikipedia page for the Nigerian musician *Fela Kuti* belongs not only to the categories MUSICAL ACTIVISTS and SAXOPHONISTS (*is-a*) but also to the 1938 BIRTHS (*has-property*)[2]. This is due to the fact that, rather than being a well-structured taxonomy, the Wikipedia category tree is an example of a *folksonomy*, namely a collaborative tagging system that enables the users to categorize the content of the encyclopedic entries. Folksonomies as such do not strive for correct conceptualization in contrast to systematically engineered ontologies. They rather achieve it by collaborative approximation.

In this paper we explore the idea of using Wikipedia for computing semantic relatedness. We make use of the online encyclopedia and its folksonomy for computing the relatedness of words and evaluate the performance on standard datasets designed for that purpose. Since the datasets are limited in size, we additionally apply these measures to a real-world NLP application, using semantic relatedness as a feature for a machine learning based coreference resolution system (Ponzetto & Strube, 2006).

[1] Wikipedia can be downloaded at http://download.wikimedia.org. In our experiments we use the English Wikipedia database dump from 19 February 2006.

[2] In the following we use *Italics* for words and queries, CAPITALS for Wikipedia pages and SMALL CAPS for concepts and Wikipedia categories.

Related Work

Approaches to measuring semantic relatedness that use lexical resources (instead of distributional similarity of words, e.g. Landauer & Dumais (1997) and Turney (2001)) transform that resource into a network or graph and compute relatedness using paths in it. Rada et al. (1989) traverse MeSH, a term hierarchy for indexing articles in Medline, and compute semantic relatedness straightforwardly in terms of the number of edges between terms in the hierarchy. Jarmasz & Szpakowicz (2003) use the same approach with *Roget's Thesaurus* while Hirst & St-Onge (1998) apply a similar strategy to WordNet. 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 (Wu & Palmer, 1994; Resnik, 1995; Leacock & Chodorow, 1998; Finkelstein et al., 2002; Banerjee & Pedersen, 2003, inter alia). Those researchers, however, focused on developing appropriate measures while keeping WordNet as the *de facto* primary knowledge source.

Since Wikipedia exists only since 2001 and has been considered a reliable source of information for an even shorter amount of time, not many researchers in NLP have worked with its content, and even less have used it as resource. Few researchers have explored the use of Wikipedia for applications such as question answering (Ahn et al., 2004) and named entity disambiguation (Bunescu & Paşca, 2006) and showed promising results.

In our work we combine these two lines of research. We apply well established semantic relatedness measures originally developed for WordNet to the open domain encyclopedia Wikipedia. This way we hope to encourage further research taking advantage of the resources provided by Wikipedia.

Semantic Relatedness Measures

The measures we use for computing semantic relatedness fall into three broad categories.

Path based measures. These measures compute relatedness as a function of the number of edges in the taxonomy along the path between two conceptual nodes c_1 and c_2 the words w_1 and w_2 are mapped to (e.g. via disambiguation and sense assignment). The simplest path-based measure is the straightforward edge counting method of Rada et al. (1989, pl, henceforth), which defines semantic distance as the number of nodes in the taxonomy along the shortest path between two conceptual nodes. Accordingly, semantic relatedness is defined as the inverse score of the semantic distance. Leacock & Chodorow (1998, lch) propose a normalized path-length measure which takes into account the depth of the taxonomy in which the concepts are found

$$lch(c_1, c_2) = -\log \frac{length(c_1, c_2)}{2D}$$

where $length(c_1, c_2)$ is the number of nodes along the shortest path between the two nodes (as given by the edge counting method), and D is the maximum depth of the taxonomy. Wu & Palmer (1994, wup) present instead a scaled measure which takes into account the depth of the nodes together with the depth of their least common subsumer, *lcs*.

$$wup(c_1, c_2) = \frac{depth(lcs_{c_1,c_2})}{depth(c_1) + depth(c_2)}$$

Information content based measures. The measure of Resnik (1995, res) computes the relatedness between the concepts as a function of their information content, given by their probability of occurrence in a corpus. Relatedness is modeled as "the extent to which they [the concepts] share information", and is given by the information content, *ic*, of their least common subsumer.

$$res(c_1, c_2) = ic(lcs_{c_1,c_2})$$

In the case of Wikipedia we couple the Resnik measure with an intrinsic information content measure relying on the hierarchical structure of the category tree (Seco et al., 2004), rather than computing the information content from the probabilities of occurrence of the concepts in a corpus. This method has been proven to correlate better with human judgements. The intrinsic information content of a category node n in the hierarchy is given as a function of its hyponyms, namely

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

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

Text overlap based measures. We finally use measures based on the relatedness between two words defined as a function of text (i.e. gloss) overlap (Lesk, 1986). An example of such measure is the *extended gloss overlap* ($lesk$) measure of Banerjee & Pedersen (2003). This measure computes the overlap score by extending the glosses of the concepts under consideration to include the glosses of related concepts in a hierarchy. In the case of Wikipedia, since no relevant text is given in the category pages, text overlap measures are computed from article pages only. Given two texts t_1 and t_2 taken as definitions for the words w_1 and w_2, the overlap score $overlap(t_1, t_2)$ is computed as $\sum_n m^2$ for n phrasal m-word overlaps (Banerjee & Pedersen, 2003). In order to adapt the Lesk measure to Wikipedia, text overlap measures were computed from Wikipedia 'glosses' (viz., the first paragraph of text of the pages, $gloss$) and full page texts ($text$). 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.

$$relate_{gloss/text}(t_1, t_2) = tanh\left(\frac{overlap(t_1, t_2)}{length(t_1) + length(t_2)}\right)$$

Computing Semantic Relatedness with Wikipedia

Wikipedia mining works in our system as follows: given the word pairs i, j (*king* and *rook* for instance) we first retrieve the Wikipedia pages which they refer to. We then hook to the category tree by extracting the categories the pages belong to. Finally, we compute relatedness based on the pages extracted and the paths found along the category taxonomy.

Page retrieval and disambiguation. Page retrieval for page $p_{i/j}$ is accomplished by first querying the page titled as the word i/j. Next, we follow all redirects (i.e. CAR redirecting to AUTOMOBILE) and resolve *ambiguous* page queries, as many queries to Wikipedia return a *disambiguation page*, namely pages hyperlinking to entries which are candidate targets for the given original query. 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. As we are interested here in relatedness, we opt for an approach to disambiguation which maximizes relatedness, namely we let the page queries disambiguate each other. If a disambiguation page $p_{i/j}$ for querying word i/j is hit, we first get all the hyperlinks in the page $p_{j/i}$ obtained by querying the other word j/i *without disambiguating*. This is to bootstrap the disambiguation process, as well as it could be the case that *both* queries are ambiguous, e.g. *king* and *rook*. We take the other word j/i and all the Wikipedia internal links of the page $p_{j/i}$ as a *lexical association list* L 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 $p_{i/j}$ contains any occurrence of a disambiguating term $l \in L$ (i.e. the link *king (chess)* in the KING page containing the term *chess* extracted from the ROOK page), the linked page is returned, else we return the first article linked in the disambiguation page. – This disambiguation strategy probably offers a less accurate solution than following all disambiguation page links. Nevertheless it offers a more practical solution as many of those pages contain a large number of links.

Category tree search. The Wikipedia pages suffice only to compute the text overlap measures. Additionally, paths along the category tree are needed for computing path and information based measures. Given the pages p_i and p_j, we extract the lists of categories C_i and C_j they belong to (i.e. KING (CHESS) belongs to the CHESS PIECES category). That is, we assume that category links in the pages are the primitive concepts in the taxonomy which the words denote. Given the category lists, for each category pair $\langle c_i, c_j \rangle, c_i \in C_i, c_j \in C_j$ we perform a depth-limited search of maximum depth of 4 for a least common subsumer. We noticed that limiting the search improves the results. This is probably due to the upper regions of the Wikipedia category tree being too strongly connected. Accordingly, the value of the search depth was established during system prototyping on the datasets from Miller & Charles (1991) and Rubenstein & Goodenough (1965).

Relatedness measure computation. Finally, given the set of paths found between the category pairs, we compute the taxonomy based measures by selecting the paths satisfying the measure definitions, namely the shortest path for path-based measures and the path which maximizes information content for information content based measures.

Experiments

We evaluated the relatedness measures on three standard datasets, namely Miller & Charles' (1991) list of 30 noun pairs (M&C), the 65 word synonymity list from Rubenstein & Goodenough (1965, R&G) of which M&C is a subset, and finally the WordSimilarity-353 Test Collection (Finkelstein et al., 2002, 353-TC)[3]. As the 353-TC dataset includes two sets, we experiment both with the full list (353 word pairs), and its test data subset (153 pairs).

Following the literature on semantic relatedness, we evaluated performance by taking the Pearson product-moment correlation coefficient r between the relatedness measure scores and the corresponding human judgements. For each dataset we report the correlation computed on all pairs, as well as the one obtained by disregarding missing pairs which could not be found. As a baseline, we compute for each word pair i and j the Google correlation coefficient by taking the Jaccard similarity coefficient on page hits.

$$jaccard = \frac{Hits(i \text{ AND } j)}{Hits(i) + Hits(j) - Hits(i \text{ AND } j)}$$

Experiments were performed for each measure on all datasets. Additionally, since the 353-TC dataset is large enough to be partioned into training and testing, 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 judgements 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. Model selection for optimal parameter estimation was performed as a grid search through cross-validation on the training data (Hsu et al., 2006).

Table 1 shows the correlation coefficients of the different measures with human judgements. Best performance per dataset is highlighted in bold[4]. Both WordNet and

[3] Available at http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/wordsim353.html

[4] 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 test dataset, namely the Wikipedia

Dataset		Google	WordNet					Wikipedia						SVM
		jaccard	*pl*	*wup*	*lch*	*res*	*lesk*	*pl*	*wup*	*lch*	*res*	*gloss*	*text*	
M&C	all	0.26	0.71	0.77	**0.82**	0.78	0.37	0.45	0.40	0.41	0.23	0.46	0.46	
	non-missing	0.26	0.71	0.77	**0.82**	0.78	0.37	0.49	0.45	0.46	0.29	0.47	0.47	
R&G	all	0.41	0.78	0.82	**0.86**	0.81	0.34	0.53	0.49	0.50	0.31	0.46	0.46	
	non-missing	0.41	0.78	0.82	**0.86**	0.81	0.34	0.56	0.52	0.54	0.34	0.47	0.47	
353-TC full	all	0.18	0.28	0.30	0.34	0.34	0.21	0.45	0.47	**0.48**	0.37	0.20	0.19	
	non-missing	0.18	0.27	0.32	0.36	0.35	0.21	0.46	**0.48**	**0.48**	0.38	0.20	0.20	
353-TC test	all	0.27	0.29	0.28	0.35	0.38	0.21	0.50	0.54	**0.55**	0.45	0.22	0.22	**0.59**
	non-missing	0.27	0.29	0.30	0.38	0.39	0.21	0.50	0.54	**0.55**	0.45	0.22	0.22	

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

Wikipedia perform better than the Google baseline, which seems to suggest that using structured knowledge sources for relatedness computation yields more accurate results. While WordNet performs extremely well on the small datasets (M&C and R&G), its performance drastically decreases when applied to a larger dataset such as 353-TC. Wikipedia however does not perform as well on the smaller datasets but outperforms WordNet on 353-TC by a large margin. This is not due to coverage, as 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. The problems seem to be caused rather by *sense proliferation*. 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, more informative) path. This allows for unplausible paths to be returned. As an example, the shortest path returned for the pair *stock* and *jaguar* uses an infrequent sense of *stock* ('not used technically; any animals kept for use or profit'), which was not the one intended by the human judges as they assigned a low correlation score to the pair. It should be noted however that this does not constitute a problem for WordNet itself, as it has to provide coverage, but rather for the relatedness measures. Additionally no sense disambiguation is possible, as the input consists only of two, possibly unrelated, words. On the contrary, using Wikipedia pages as taxonomy 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'[5]. In addition to this, 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. In general, it seems that comparing WordNet and Wikipedia on the 353-TC dataset reveals a classic coverage/precision dilemma. While Wikipedia still suffers a more limited coverage than WordNet, by using it we can direct the path search via disambiguation using resources such as text and links.

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 test set we report the best performance by *integrating* all available knowledge sources. The score of $r = 0.59$ outperforms the combined WordNet–word-vector measure of Finkelstein et al. (2002) ($r = 0.55$), with the correlation score dropping minimally when leaving the Google scores out ($r = 0.58$). 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 simply using a different knowledge base such as Wikipedia.

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. 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. Besides, the only available 'not-so-small' dataset for evaluating relatedness measures, namely the 353-TC dataset, has been criticized in the literature for having been built in a methodologically unsolid way and accordingly for not being able to provide a suitable benchmarking dataset (Jarmasz & Szpakowicz, 2003). These are all reasons why we turn in the next section to the application of such measures 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.

lch. The only statistically non-significant differences in performance were found between the *lesk* and both the Wikipedia *gloss* and *text* measures on the M&C dataset.

[5]This is not to deny that sometimes we return paths which make no sense at all. For instance, given the pair *Arafat-terror*, the word *terror* gets disambiguated by returning the UNITED_STATES page, such that the shortest path is *Arafat* → ARAFAT → REBELS → REBELLION → POLITICS ← COUNTRIES ← NORTH_AMERICAN_COUNTRIES ← UNITED_STATES ← *terror*.

	BNEWS						NWIRE					
	R	P	F_1	A_p	A_{cn}	A_{pn}	R	P	F_1	A_p	A_{cn}	A_{pn}
baseline	46.7	86.2	60.6	36.4	10.5	44.0	56.7	88.2	69.0	37.6	23.1	55.6
+WordNet	**54.8**	86.1	**66.9**	**36.8**	**24.8**	**47.6**	**61.3**	84.9	**71.2**	**38.9**	**30.8**	55.5
+Wiki	**52.7**	**86.8**	**65.6**	36.1	**23.5**	**46.2**	**60.6**	83.6	**70.3**	**38.0**	**29.7**	55.2
+SRL	**53.3**	85.1	**65.5**	**37.1**	**13.9**	**46.2**	**58.0**	**89.0**	**70.2**	**38.3**	**25.0**	**56.0**
all features	**59.1**	84.4	**69.5**	**37.5**	**27.3**	**48.1**	**63.1**	83.0	**71.7**	**39.8**	**31.8**	52.8

Table 2: Results on the ACE 2003 data (BNEWS and NWIRE sections)

Case Study: Coreference Resolution

We present in this section an extension of a machine learning based coreference resolver which uses relatedness scores as features for classifying referring expressions (REs) as denoting the same discourse entities (see Ponzetto & Strube (2006) for an in-depth description of the system).

To establish a competitive baseline system, we reimplemented the machine learning based coreference resolver of Soon et al. (2001). Coreference resolution is viewed as a binary classification task: given a pair of referring expressions (REs), the classifier has to decide whether they are coreferent or not. For learning coreference decisions, we used a Maximum Entropy (Berger et al., 1996) model. Instances are created following Soon et al. (2001).

In order to test the effects of including semantic relatedness information within a coreference learner, the system is first run using the 12 features of the baseline model to be replicated, viz., shallow surface features, such as the distance between the potentially coreferent expressions, string matching and linguistic form (i.e. pronoun, demonstrative). We then explore the contribution of features capturing semantic relatedness. These are computed by taking the relatedness score of the RE pairs, obtained by querying the head lemma of the REs (i.e. *diver* for *the Russian divers*) or, in the case of named entities, the full linguistic expression. No relatedness score is computed for pairs including pronouns. We evaluate four expanded feature sets, namely adding (1) WordNet features; (2) Wikipedia features; (3) semantic role label features (Gildea & Jurafsky, 2002, SRL) (4) all available features. For all feature sets we determine the relevant features following an iterative procedure similar to the wrapper approach for feature selection (Kohavi & John, 1997).

The system was developed and tested with the ACE 2003 Training Data corpus (Mitchell et al., 2003)[6]. Both the Newswire (NWIRE) and Broadcast News (BNEWS) sections where split into 60-20-20% document-based partitions for training, development, and blind testing, and later per-partition merged (MERGED) for a document source independent system evaluation. We computed relatedness scores for 282,658 word pairs in total. We report in Tables 2 and 3 the MUC score (Vilain et al., 1995) with performances above the baseline being highlighted in bold. This score is computed for those phrases which appear in both the key and the response. We discard therefore those responses not present in the key, as we are interested here in establishing the upper limit of the improvements given by our semantic features. In addition, we 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 given in the key.

The results show that WordNet and Wikipedia relatedness features tend to significantly increase performance on common nouns, that is, that both provide semantically relevant features for coreference resolution. As a consequence of having different knowledge sources accounting for the resolution of different RE types, the best results are obtained by (1) *combining features* generated *from different sources* and (2) *performing feature selection*. When combining different feature sources, we note an accuracy improvement on pronouns and common nouns, as well as an increase in F-measure due to a higher recall. The optimal system configurations always include features from both WordNet and Wikipedia. This supports the results of Table 1 where the best results were found by integrating relatedness scores from different sources, thus suggesting that WordNet and Wikipedia are *complementary knowledge sources*. More interestingly, it indicates that Wikipedia can indeed be used as a resource for large-scale NLP applications.

Conclusions

In this paper we investigated the use of Wikipedia for computing semantic relatedness measures and the application of these measures to a real-world NLP task such as coreference resolution. The results show that Wikipedia provides a suitable encyclopedic knowledge base for extracting semantic information. While using Wikipedia alone yields a slightly worse performance in our coreference resolution system as compared to WordNet, it showed nevertheless promising results. Also, by using Wikipedia we obtained the best semantic relatedness results on the 353-TC dataset. Even if the taxonomic categorization feature has been introduced into Wikipedia only two years ago, our results indicate that re-

	R	P	F_1	A_p	A_{cn}	A_{pn}
baseline	54.5	88.0	67.3	34.7	20.4	53.1
+WordNet	**56.7**	87.1	**68.6**	**35.6**	**28.5**	49.6
+Wikipedia	**55.8**	87.5	**68.1**	**34.8**	**26.0**	50.5
+SRL	**56.3**	**88.4**	**68.8**	**38.9**	**21.6**	51.7
all features	**61.0**	84.2	**70.7**	**38.9**	**29.9**	51.2

Table 3: Results ACE (merged BNEWS/NWIRE)

[6] We used the training data corpus only, as the availability of the test data was restricted to ACE participants. Therefore, the results we report cannot be compared directly with those using the official test data.

latedness computed using the Wikipedia taxonomy consistently correlates better with human judgements than a simple baseline based on Google counts, and better than WordNet for some datasets. In addition, just as WordNet, it can provide a useful knowledge source for adding semantic relatedness information to an NLP application such as a coreference resolution system.

What is most interesting about our results is that they indicate that a collaboratively created folksonomy can actually be used in AI and NLP applications with the same effect as hand-crafted taxonomies or ontologies. Even on a theoretical ground, it seems to be a wise choice to use knowledge generated collaboratively. This is because the Wikipedia folksonomy is created on a large scale by the very same people whose knowledge we try to model in our applications. So, it is no surprise that it also works.

Instead of letting a few ontology experts decide upon the structure of the world, its thorough description can be continuously approximated by a large number of people who collaborate on Wikipedia. Everyone contributes their expertise by describing different aspects of the world and categorizing them. More concretely, if a project is able to induce in just a couple of years a taxonomy able to compete with WordNet on linguistic processing tasks, and given its exponential growth rate, we can only expect a bright future for automatic knowledge mining techniques with Wikipedia[7].

Acknowledgments. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The second author has been supported by a KTF grant (09.003.2004). We thank our colleagues Katja Filippova and Christoph Müller and the three anonymous reviewers for their insightful comments.

References

Ahn, D., V. Jijkoun, G. Mishne, K. Müller, M. de Rijke & S. Schlobach (2004). Using Wikipedia at the TREC QA track. In *Proc. of TREC-13*.

Banerjee, S. & T. Pedersen (2003). Extended gloss overlap as a measure of semantic relatedness. In *Proc. of IJCAI-03*, pp. 805–810.

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

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

Bunescu, R. & M. Paşca (2006). Using encyclopedic knowledge for named entity disambiguation. In *Proc. of EACL-06*, pp. 9–16.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database.* Cambridge, Mass.: MIT Press.

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

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

[7]The WikiRelate! software described in this paper can be downloaded from `http://www.eml-research.de/nlp`.

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

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

Jarmasz, M. & S. Szpakowicz (2003). Roget's Thesaurus and semantic similarity. In *Proc. of RANLP-03*, pp. 212–219.

Kim, S. N. & T. Baldwin (2005). Automatic interpretation of noun compounds using WordNet similarity. In *Proc. of IJCNLP-05*, pp. 945–956.

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

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

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.

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

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

Patwardhan, S., S. Banerjee & T. Pedersen (2005). SenseRelate::TargetWord – A generalized framework for word sense disambiguation. In *Proc. of AAAI-05*.

Ponzetto, S. P. & M. Strube (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proc. of HLT-NAACL-06*.

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

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of IJCAI-95*, Vol. 1, pp. 448–453.

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

Seco, N., T. Veale & J. Hayes (2004). An intrinsic information content metric for semantic similarity in WordNet. In *Proc. of ECAI-04*, pp. 1089–1090.

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

Turney, P. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proc. ECML-01*, pp. 491–502.

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

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

Wu, Z. & M. Palmer (1994). Verb semantics and lexical selection. In *Proc. of ACL-94*, pp. 133–138.