An API for Measuring the Relatedness of Words in Wikipedia

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

Abstract

We present an API for computing the semantic relatedness of words in Wikipedia.

1 Introduction

The last years have seen a large amount of work in Natural Language Processing (NLP) using measures of semantic similarity and relatedness. We believe that the extensive usage of such measures derives also from the availability of robust and freely available software that allows to compute them (Pedersen et al., 2004, WordNet::Similarity).

In Ponzetto & Strube (2006) and Strube & Ponzetto (2006) we proposed to take the Wikipedia categorization system as a semantic network which served as basis for computing the semantic relatedness of words. In the following we present the API we used in our previous work, hoping that it will encourage further research in NLP using Wikipedia[1].

2 Measures of Semantic Relatedness

Approaches to measuring semantic relatedness that use lexical resources transform these resources into a network or graph and compute relatedness using paths in it (see Budanitsky & Hirst (2006) for an extensive review). For instance, 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.

3 The Application Programming Interface

The API computes semantic relatedness by:

1. taking a pair of **words as input**;
2. **retrieving the Wikipedia articles** they refer to (via a disambiguation strategy based on the link structure of the articles);
3. **computing paths in the Wikipedia categorization graph** between the categories the articles are assigned to;
4. **returning as output the set of paths found, scored** according to some measure definition.

The implementation includes *path-length* (Rada et al., 1989; Wu & Palmer, 1994; Leacock & Chodorow, 1998), *information-content* (Resnik, 1995; Seco et al., 2004) and *text-overlap* (Lesk, 1986; Banerjee & Pedersen, 2003) measures, as described in Strube & Ponzetto (2006).

The API is built on top of several modules and can be used for tasks other than Wikipedia-based relatedness computation. On a basic usage level, it can be used to retrieve Wikipedia articles by name, optionally using disambiguation patterns, as well as to find a ranked set of articles satisfying a search query (via integration with the Lucene[2] text search engine). Additionally, it provides functionality for visualizing the computed paths along the Wikipedia categorization graph as either Java Swing components or applets (see Figure 1), based on the JGraph library[3], and methods for computing centrality scores of the Wikipedia categories using the PageRank algorithm (Brin & Page, 1998). Finally, it currently

[1] The software can be freely downloaded at http://www.eml-research.de/nlp/download/wikipediasimilarity.php.

[2] http://lucene.apache.org

[3] http://www.jgraph.com

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 49–52,
Prague, June 2007. ©2007 Association for Computational Linguistics

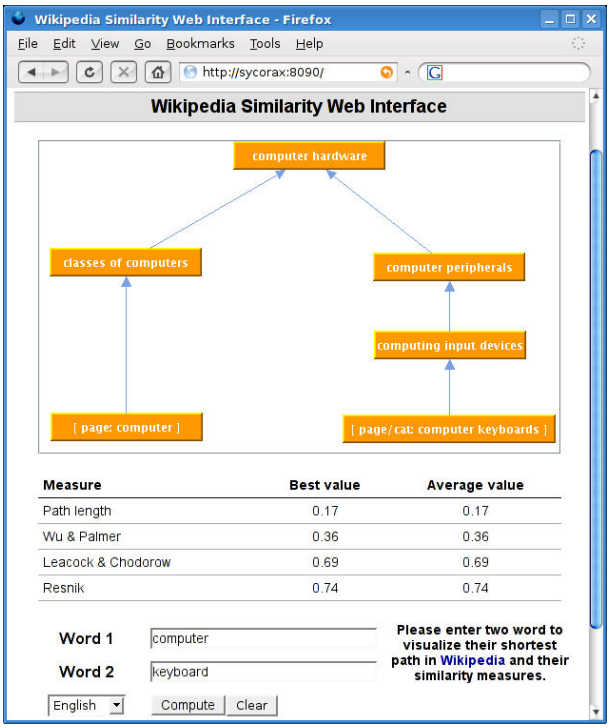

Figure 1: Shortest path between computer and keyboard in the English Wikipedia.

provides multilingual support for the English, German, French and Italian Wikipedias and can be easily extended to other languages[4].

4 Software Architecture

Wikipedia is freely available for download, and can be accessed using robust Open Source applications, e.g. the MediaWiki software[5], integrated within a Linux, Apache, MySQL and PHP (LAMP) software bundle. The architecture of the API consists of the following modules:

1. **RDBMS**: at the lowest level, the encyclopedia content is stored in 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**[6]: 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 interface 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 2. The higher input/output layer the user interacts with is provided by a Java API from which Wikipedia can be queried. The Java 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.

5 Using the API

The API provides factory classes for querying Wikipedia, in order to retrieve encyclopedia entries as well as relatedness scores for word pairs. In practice, the Java library provides a simple programmatic interface. Users can accordingly access the library using only a few methods given in the factory classes, e.g. `getPage(word)` for retrieving Wikipedia articles titled `word` or `getRelatedness(word1,word2)`, for computing the relatedness between `word1` and `word2`, and `display(path)` for displaying a path found between two Wikipedia articles in the categorization graph. Examples of programmatic usage of the API are presented in Figure 3. In addition, the software distribution includes UNIX shell scripts to access the API interactively from a terminal, i.e. it does not require any knowledge of Java.

6 Application scenarios

Semantic relatedness measures have proven useful in many NLP applications such as word sense disambiguation (Kohomban & Lee, 2005; Patwardhan et al., 2005), information retrieval (Finkelstein et al., 2002), information extraction pattern induction (Stevenson & Greenwood, 2005), interpretation of noun compounds (Kim & Baldwin, 2005), para-

[4] In contrast to WordNet::Similarity, which due to the structural variations between the respective wordnets was reimplemented for German by Gurevych & Niederlich (2005).

[5] `http://www.mediawiki.org`

[6] `http://search.cpan.org/dist/WWW-Wikipedia`

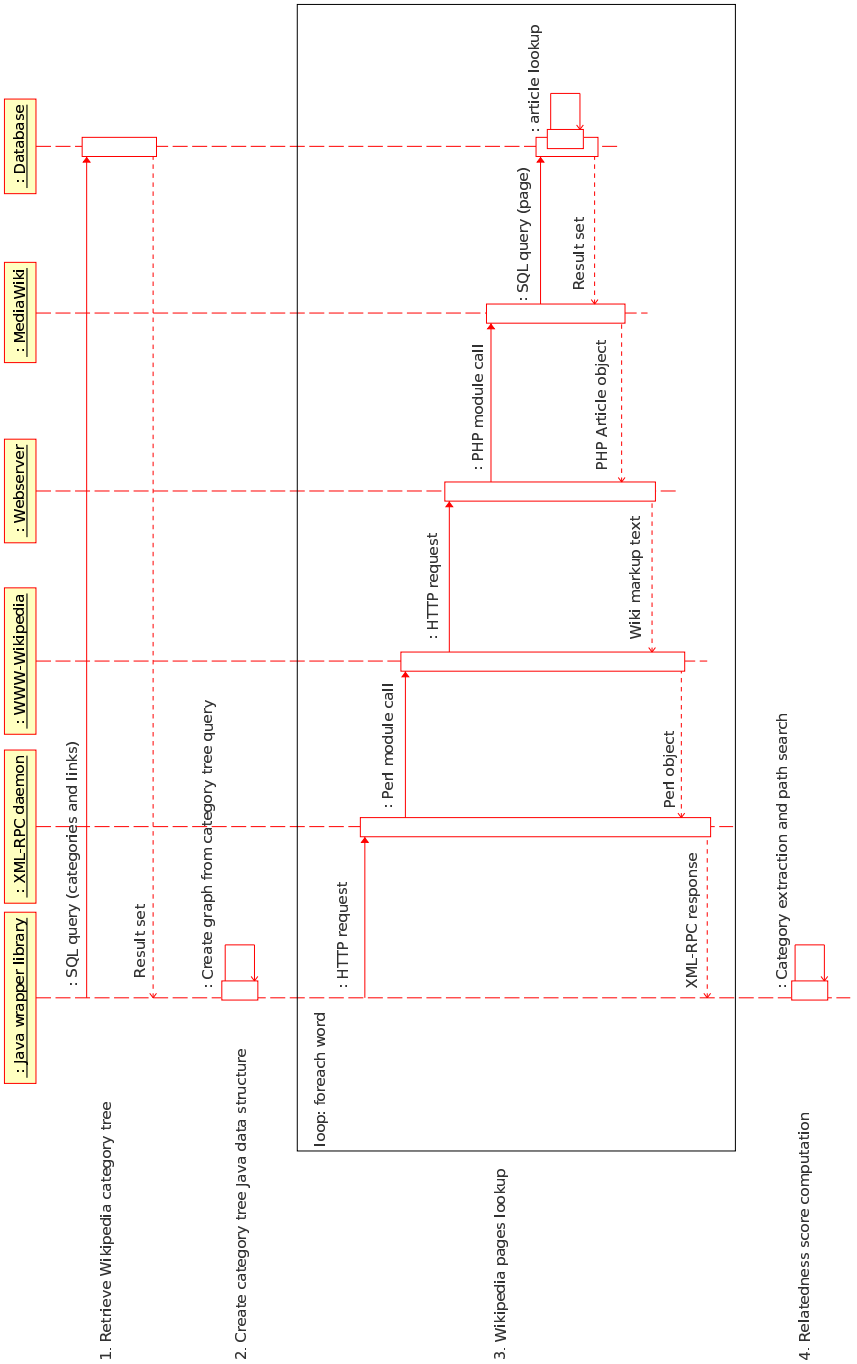

Figure 2: API processing sequence diagram. Wikipedia pages and relatedness measures are accessed through a Java API. The wrapper communicates with a Perl library designed for Wikipedia access and parsing through an XML-RPC server. WWW-Wikipedia in turn accesses the database where the encyclopedia is stored by means of appropriate queries to MediaWiki.

```
// 1. Get the English Wikipedia page titled "King" using "chess" as disambiguation
WikipediaPage page = WikipediaPageFactory.getInstance().getWikipediaPage("King","chess");

// 2. Get the German Wikipedia page titled "Ufer" using "Kueste" as disambiguation
WikipediaPage page = WikipediaPageFactory.getInstance().getWikipediaPage("Ufer","Kueste",Language.DE);

// 3a. Get the Wikipedia-based path-length relatedness measure between "computer" and "keyboard"
WikiRelatedness relatedness = WikiRelatednessFactory.getInstance().getWikiRelatedness("computer","keyboard");
double shortestPathMeasure = relatedness.getShortestPathMeasure();
// 3b. Display the shortest path
WikiPathDisplayer.getInstance().display(relatedness.getShortestPath());

// 4. Score the importance of the categories in the English Wikipedia using PageRank
WikiCategoryGraph<DefaultScorableGraph<DefaultEdge>> categoryTree =
        WikiCategoryGraphFactory.getCategoryGraphForLanguage(Language.EN);
categoryTree.getCategoryGraph().score(new PageRank());
```

Figure 3: Java API sample usage.

phrase detection (Mihalcea et al., 2006) and spelling correction (Budanitsky & Hirst, 2006). Our API provides a flexible tool to include such measures into existing NLP systems while using Wikipedia as a knowledge source. Programmatic access to the encyclopedia makes also available in a straightforward manner the large amount of structured text in Wikipedia (e.g. for building a language model), as well as its rich internal link structure (e.g. the links between articles provide phrase clusters to be used for query expansion scenarios).

Acknowledgements: 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 our colleagues Katja Filippova and Christoph Müller for helpful feedback.

References

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

Brin, S. & L. Page (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.

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

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.

Gurevych, I. & H. Niederlich (2005). Accessing GermaNet data and computing semantic relatedness. In *Comp. Vol. to Proc. of ACL-05*, pp. 5–8.

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.

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.

Kohomban, U. S. & W. S. Lee (2005). Learning semantic classes for word sense disambiguation. In *Proc. of ACL-05*, pp. 34–41.

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.

Mihalcea, R., C. Corley & C. Strapparava (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *Proc. of AAAI-06*, pp. 775–780.

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

Pedersen, T., S. Patwardhan & J. Michelizzi (2004). WordNet::Similarity – Measuring the relatedness of concepts. In *Comp. Vol. to Proc. of HLT-NAACL-04*, pp. 267–270.

Ponzetto, S. P. & M. Strube (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proc. of HLT-NAACL-06*, pp. 192–199.

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.

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.

Stevenson, M. & M. Greenwood (2005). A semantic approach to IE pattern induction. In *Proc. of ACL-05*, pp. 379–386.

Strube, M. & S. P. Ponzetto (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *Proc. of AAAI-06*, pp. 1419–1424.

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