Exploring and Visualizing Wordnet Data with GermaNet Rover

Marie Hinrichs **Richard Lawrence** **Erhard Hinrichs**

University of Tübingen, Germany
{marie.hinrichs, richard.lawrence, erhard.hinrichs}@uni-tuebingen.de

Abstract

This paper introduces GermaNet Rover, a new web application for exploring and visualizing GermaNet, the German wordnet. Rover provides semantic relatedness calculations between concept pairs using six algorithms, and allows regular expression or edit distance lookup in combination with other search constraint options. Visualizations include a concept's position in the hypernym graph and the shortest path between concepts. Rover provides easy access to these features and is available as a CLARIN resource.

1 Introduction

GermaNet[1] is a lexical-semantic network (a *wordnet*) for the German language. Wordnets have been developed for many languages, often modeled on the Princeton WordNet® for English (Fellbaum, 1998). A number of wordnets are available as CLARIN[2] resources – for example plWordNet (Maziarz et al., 2016), Estonian Wordnet (Kahusk and Vider, 2005; Pedersen et al., 2013), FinnWordNet (Lindén and Carlson., 2010), Open Dutch Wordnet (Postma et al., 2016), and BulNet (Rizov and Dimitrova, 2016).

There are many web applications for browsing wordnets, several of which include visualizations. But current tools generally lack several features which would be useful for researchers. For example, many tools only include simple searching capabilities, and lack a means of specifying non-exact, "fuzzy" search terms. Many tools also lack a means of comparing the concepts in the wordnet, although wordnet-based measures of semantic similarity have been available for some time. To the best of our knowledge, plWordNet offers the only web application which presents semantically similar words for a given concept, and only one similarity score is presented.

GermaNet Rover[3] is a new web application for exploring and visualizing GermaNet. Rover provides advanced search features, including searching by regular expression or edit distance, as well as a variety of other search constraints. It allows comparing concepts in the wordnet via six different measures of semantic relatedness, which can be viewed in the web interface for individual pairs of concepts, or processed in batch via file upload. Rover also offers visualizations of the network structure on which these measures are based, including of a concept's position in the wordnet and of the shortest paths between two concepts. Rover provides user-friendly access to these features and is available as a CLARIN resource.

2 Background

A wordnet is a data set representing a network of semantic relationships in a language. In a wordnet, words or *lexical units* are grouped into sets of synonyms called *synsets*. Each synset represents a single concept of the language, which may be variously expressed by the lexical units it contains. Relations

[1] https://uni-tuebingen.de/en/142806
[2] https://www.clarin.eu/
[3] https://weblicht.sfs.uni-tuebingen.de/rover/

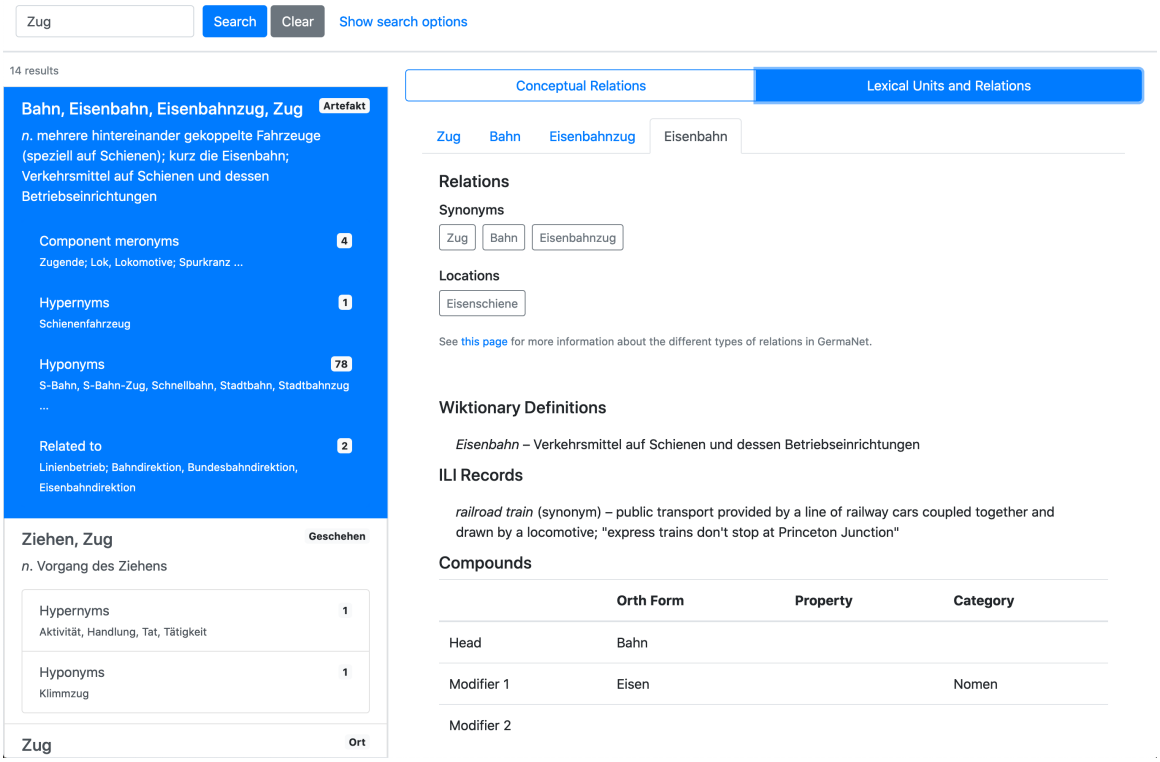

Figure 1: The Synset Search interface, with results of a search for *Zug* "train". Summaries of synsets in the search results appear on the left. Details about the selected synset appear on the right.

between synsets, called *conceptual relations*, are a wordnet's primary representation of semantic relationships. GermaNet also contains *lexical* relations, which hold between individual lexical units, rather than synsets.

There are several types of conceptual relations in GermaNet. The primary relation is *hypernymy*, an asymmetric predication relation ('X is Y but Y is not necessarily X'). For example, the synset containing *Haustier* "pet" has as hypernym the synset containing *Tier* "animal". Other conceptual relations in GermaNet include *causation*, *entailment*, and *meronomy*, the latter of which represents whole-part relationships (e.g. *Pedal* "pedal" is part of a *Fahrrad* "bike").

The hypernymy relation is uniquely important because every synset is guaranteed to have at least one hypernym, except for the artificial root synset. The synsets thus form a connected directed graph with edges given by the hypernymy relation. This graph has a mostly-hierarchical structure, in which more general concepts may be thought of as being higher in the hierarchy, with more specific concepts falling under them. This graph structure is the basis for several measures in the literature of semantic relatedness between concepts.

3 GermaNet Rover

GermaNet Rover is a web application that enables researchers to explore the semantic relationships and other data in GermaNet. Rover currently has two main features: Synset Search and Semantic Relatedness. Both features also provide relevant visualizations of the local structure of GermaNet. We describe these features in more detail below.

3.1 Synset Search

The Synset Search feature (see Figure 1) allows searching for synsets in GermaNet, browsing the relationships between synsets, and viewing detailed data about individual synsets, including a visualization of its place in the hypernym relation.

Searches match a search term against all the synsets in GermaNet. The search space can be further

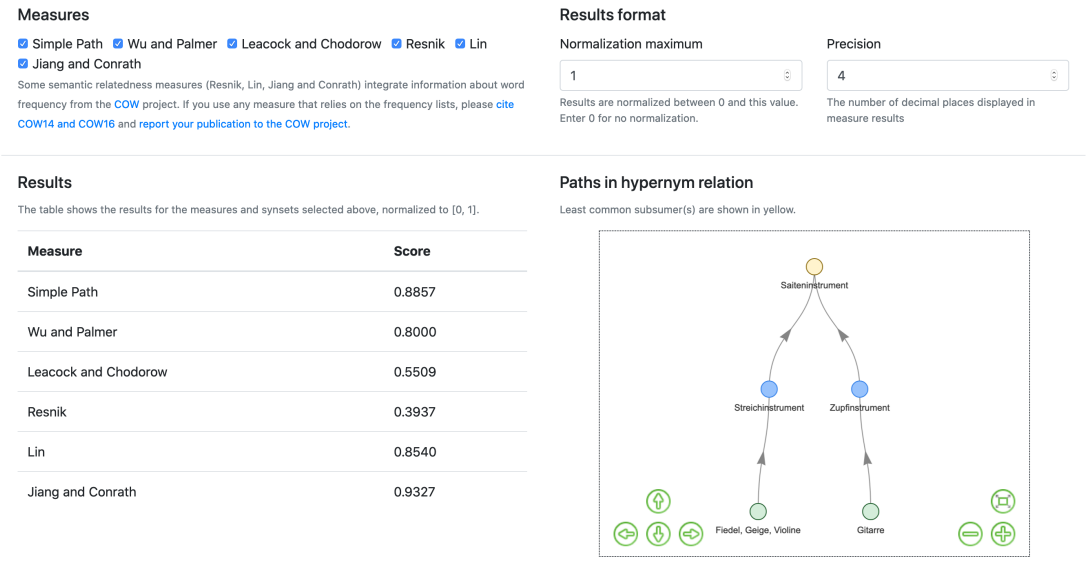

Figure 2: The Visualize Relatedness interface in the Semantic Relatedness feature, comparing *Gitarre* "guitar" and *Fiedel, Geige, Violine* "violin". The results table displays relatedness scores for each selected measure. The network diagram displays the shortest path between the two synsets via their least common subsumer.

constrained to particular grammatical categories (e.g. adjectives), semantic classes (e.g. *Bewegung* "motion"), or orthographic variants (including forms used prior to the orthography reform of 1996, and variant forms of both the current and old forms). Fuzzy matching of the search term is supported through three different methods: by ignoring capitalization in the search term, by specifying a numerical edit distance from the search term, or by treating the search term as a regular expression. The search parameters are matched against the individual lexical units in a synset; any synset containing at least one lexical unit matching the search parameters will be returned as a result. For each synset in the result set of a search, a brief summary appears with the synset's word category and class, any definitions or paraphrases associated with the synset, and the type and number of the synset's conceptual relations.

Selecting one of these summaries displays more detailed information about the synset, separated into two tabs. The Conceptual Relations tab offers a detailed presentation of all of the synset's conceptual relations, organized by type. For each related synset, the lexical units of that synset are displayed as a series of buttons which navigate to the related synset, which allows progressively exploring the network via its relations. The Conceptual Relations tab also displays a network diagram which visualizes the local structure of the hypernymy relation between the selected synset and the artificial root synset.

The Lexical Units and Relations tab is further broken down into one detail tab for each lexical unit in the selected synset. These detail tabs display any information present in GermaNet about these individual lexical units, including their relations to other lexical units, example sentences with associated frame types, and associated records in other data sets.

3.2 Semantic Relatedness

The Semantic Relatedness feature allows calculating the relatedness between pairs of synsets using several different algorithms. The measures currently implemented are derived from Wu and Palmer (1994), Jiang and Conrath (1997), Leacock and Chodorow (1998), Lin (1998), and Resnik (1999). An additional *Simple Path* measure calculates the length of the path between the two synsets via the hypernymy relation, relative to the length of the longest such path. Because the different measures have different minimum and maximum values, the calculations are normalized to a common interval, so that the results can be more easily compared.

The Semantic Relatedness feature provides two different interfaces to work with these measures: Visualize Relatedness and Batch Processing. On the Visualize Relatedness tab (see Figure 2), two specific

synsets are selected using two independent search inputs, similar to those in the Synset Search feature. The semantic relatedness of the two selected synsets is then calculated according to the implemented measures, and the results are displayed in a table. The display of the results table can be controlled by selecting different measures and adjusting the normalization interval.

An important notion underlying each of the implemented measures is the *least common subsumers* of the selected synsets. These are the two synsets' nearest common ancestors in the hypernymy relation. The Visualize Relatedness tab also displays a network diagram which visualizes the shortest paths between the two selected synsets via their least common subsumers. Both the results table and the network diagram are updated automatically whenever different synsets or measures are selected, allowing quick exploration of how the different measures behave in relation to the structure of the hypernym graph.

The Batch Processing tab can be used to calculate the relatedness of many synsets at once by uploading a file containing word pairs. The batch processor searches for synsets containing the words in each pair, and then compares each pair of synsets containing the two words. (Thus, a single word pair can generate multiple synset comparisons.) A configuration section in the input file allows specifying the same options as are available on the Visualize Relatedness tab, including constraints for the searches, a selection of individual measures, and the normalization interval for the calculations. Results are returned in a tabular format that can easily be processed with a spreadsheet program or custom scripts.

3.3 Advances in Rover

Previous work had also enabled exploring and visualizing the data in GermaNet (Finthammer and Cramer, 2008; Henrich and Hinrichs, 2010). Rover builds both on that work and on several other internal projects, offering new advantages and opportunities for researchers working with wordnet data.

Unlike previous tools for GermaNet, Rover is structured as a web application, with a browser-based interface. This eliminates the need for users who just want to explore the data set to download or install special software, and allows the data and backend web server to be hosted on CLARIN infrastructure.

The user interface of Rover has been implemented using a new open source library called germanet-common[4], which provides abstractions for querying and displaying different types of wordnet data from a web server via a JSON API. Due to the design of this library, the Rover web interface would be relatively easy to port to other wordnets' infrastructure, and could provide a basis for exploring and visualizing data in other wordnets. The library also makes it easy to build other kinds of web interfaces for wordnets, beyond the features available in Rover.

The implementation of Rover has also brought new developments in the GermaNet Java API[5], the central library used by Rover's web server. This library implements the fuzzy matching capabilities in the Synset Search feature and the various measures in the Semantic Relatedness feature. Since it is released annually alongside the data set, these features are not limited to the Rover interface and will also be available to researchers working with the GermaNet data via this API.

4 An Application: Morphological Productivity

Beyond simple uses as a dictionary or thesaurus, Rover also supports more advanced linguistic research. This section presents an example of a research question that can quickly be studied with Rover.

Morphological productivity concerns the extent to which a morpheme can productively form compounds. The regular expression and grammatical category support in the Synset Search feature provides a way to quickly get a sense of the relative productivity of a given morpheme. For example, how productive is the adjective *reich* "rich" in comparison to its antonym *arm* "poor"?

To answer this question, one might start by searching with the regular expressions `.+reich` and `.+arm`. The leading `.+` in these search terms matches any non-empty string of characters, so they will match adjective compounds like *erfolgreich* "successful" and *abgasarm* "low-emission". The need to refine the searches will quickly become obvious from the results: one will want to exclude nouns like *Frankreich* "France" by limiting the results to adjectives, as well as adjectives that are compounds of

[4] `https://github.com/Germanet-sfs/germanet-common`
[5] `https://github.com/Germanet-sfs/GermaNetApi/`

warm rather than *arm* by adjusting the search term to `[^w]+arm`. With these refinements, one learns that *reich* is about three times as productive as *arm* in adjective compounds (138 vs. 46 results). Rover is designed to support this kind of interactive exploration and refinement, which can help researchers in their study of the German language.

5 Future work

Rover is still being actively developed, and several features are planned for future work. We are investigating the possibility of including additional measures, based on word embeddings, in the Semantic Relatedness feature. We also hope to further improve the network visualizations. Feedback from CLARIN researchers and beta-testers will inform this work.

Acknowledgements

The research described in this paper was carried out as part of the CLARIN-D project, which is funded by the German Federal Ministry of Education and Research (BMBF).

References

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Marc Finthammer and Irene M. Cramer. 2008. Exploring and Navigating: Tools for GermaNet. In *Proceedings of the 6th Language Resources and Evaluation Conference*.

Verena Henrich and Erhard Hinrichs. 2010. GernEdiT: A graphical tool for germanet development. In *Proceedings of the ACL 2010 System Demonstrations*, pages 19–24. Association for Computational Linguistics.

Jay J. Jiang and David W. Conrath. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the 10th Research on Computational Linguistics International Conference*, pages 19–33, Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing.

Neeme Kahusk and Kadri Vider. 2005. TEKsaurus—the Estonian WordNet online. In *The Second Baltic Conference on Human Language Technologies: Proceedings.*, pages 273–278.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In *WordNet: An Electronic Lexical Database*. MIT Press.

D. Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, San Francisco, CA, USA.

Krister Lindén and Lauri Carlson. 2010. Finnwordnet — wordnet påfinska via översättning. *LexicoNordica — Nordic Journal of Lexicography*, 17:119–140. In Swedish with an English abstract.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plWordNet 3.0 – a Comprehensive Lexical-Semantic Resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pages 2259–2268, Osaka, Japan.

Bolette S. Pedersen, Lars Borin, Markus Forsberg, Neeme Kahusk, Krister Lindén, Jyrki Niemi4 Niklas Nisbeth, Lars Nygaard, Heili Orav, Eirikur Rögnvaldsson, and Mitchel Seaton. 2013. Nordic and Baltic wordnets aligned and compared through "WordTies". In *Proceedings of the 19th Nordic Conference of Computational Linguistics*, pages 147–162.

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open Dutch Wordnet. In *Proceedings of the Eighth Global Wordnet Conference*, Bucharest, Romania.

P. Resnik. 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, July.

Borislav Rizov and Tsvetana Dimitrova. 2016. Hydra for Web: A Browser for Easy Access to Wordnets. In *Proceedings of the Eighth Global Wordnet Conference*, Bucharest, Romania.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Las Cruces, New Mexico, June. Association for Computational Linguistics.