

Using Multilingual Ontologies for Adaptive Web-Based Language Exploration

Ernesto W. De Luca, Stefan Hauke, Andreas Nürnberger and Stefan Schlechtweg
Otto-von-Guericke University of Magdeburg
Universitätsplatz 2, 39106 Magdeburg, Germany
Phone: +49-391-67-18290, Fax: +49-391-67-12018
deluca@iws.cs.uni-magdeburg.de

Abstract. In this paper we discuss approaches to use multilingual ontologies for combining language exploration and web searches. The proposed methods can be used in computer-assisted language learning (CALL) applications for teaching foreign languages as well as in multilingual information retrieval systems. One main idea is to support users in navigating information using semantic connections between words senses provided by multilingual ontologies. This can help a user to better understand the different meanings of a word in his native language, and even more important, to explore its meanings in a foreign language. Combined web searches can help in understanding meanings, since the search results provide examples for word and phrase usage. Furthermore, hit statistics of word co-occurrences in web pages provide hints about correct translations or word usage.

1 Introduction

Through the increasing globalization, people are forced nowadays to obtain and to process information not only in their native language, but also information provided in foreign languages. Especially, if people want to access and search in multilingual document collections, they have to have good language skills to discover the correct meaning of the concepts in the target language. Unfortunately, people have frequently a good passive understanding of a foreign language, but are very often not able to find the correct word sense translation. Thus, tools that are able to translate words and implicitly support language acquisition would be beneficial.

In order to support this need, we present in the following an approach that combines the use of multilingual ontologies with document or web searches. Thus, we consider the Web as a learning repository, where learners can find examples of word usage. The web documents are representative example of combination between words for finding the correct translation and the word-related relevant documents. This combination can be used in tools for language acquisition in computer-assisted language learning (CALL) environments [2, 20] or for cross-language retrieval, while solving some of the still existing problems of multilingual retrieval systems and at the same time implicitly supporting the user in language acquisition. Approaches like CALL applications are used for language teaching and learning in order to support language learners with computer technology. Usually these tools help the learner for

evaluating, reinforcing and presenting the learned topics essentially with interactive elements.

1.1 Navigating Multilingual Information

Currently, the possibility to navigate multilingual information is not yet so sophisticated that users can access information in the seamless and transparent way as they do in their mother tongue. There are different problems to be solved in order to enable users to access multilingual information. These are:

1. disambiguating the words,
2. translating the words, and
3. presenting the multilingual results appropriately.

The first point refers to the problem of polysemy (from Greek: “many meanings”) of a word. One example is given by the word *bank*. It has different meanings (bank of a river, bank to draw money, etc) that can be recognized only by the context. Humans often use polysemous words for searching for documents; a distinction of the related word senses is difficult [9]. If we want to work in a multilingual context, we have to disambiguate words both in the source (query) and in target languages (language used in documents or web pages) (see Section 2).

Second, if we want to retrieve documents in other languages, we have to translate the concepts of the intended word meanings. Machine translation should help in processing and delivering this information. However, as discussed in [13], due to its poor performance this approach cannot be seen as a realistic answer to the problem of query translations right now.

The problem of automatically finding multilingual documents dealing with the same topic over languages is not yet solved [13] and – due to the required manual efforts to correct wrong or incomplete translations – represents a time expensive solution. In our work, we use the word senses retrieved from the linguistic ontologies and their translation in order to provide an alternative solution to this problem (Section 3).

The last point is related to information presentation. The visualization of the information related to the words and their linguistic relations plays an important role. It should enable on the one hand a simple navigation through the huge number of relations and documents providing examples for word usage and, if possible, should restrict information only to the relevant word sense-related results. On the other hand it should encourage users to explore and, thus, learn languages using the words of interest and their network of linguistic relations.

In this paper we consider the problem of language exploration, trying to offer an intuitive adaptive visualization of word combinations and the related web documents. It means that we retrieve the word senses of a word (including their semantic word relations as well as translations) from the linguistic ontologies and the related matching documents from the Web that provide examples for word usage. Since we cannot expect a perfect word sense disambiguation, as described in more details in the fol-

lowing, a visualization that is adapted to context, languages and retrieved documents and that allows making use of possibly imperfect sense categorization is especially important. We discuss this issue in more detail in Sect. 3.

2 Word Sense Disambiguation (WSD) and Sense Translation

The automatic disambiguation of word senses is still a very interesting and challenging research tasks. Since the 1950's different researchers work on disambiguating words, sentences or documents for different purposes as there are machine translation, information retrieval and hypertext navigation, content and thematic analysis, grammatical analysis (Part-Of-Speech Tagging), speech or text processing [8]. However, a satisfying method with acceptable performance has not yet been found. A word sense disambiguation (WSD) process can be described as a two step process:

1. All the senses of the word relevant (at least) to the considered/current text or discourse are extracted/found (through lists, categories, ontologies, dictionaries, etc.).
2. The appropriate word sense (considering the context and the external knowledge resources) is assigned to the word.

A variety of association methods (knowledge-driven, data-driven or corpus-based WSD) can be used in order to assign a sense to each word occurrence. In the following, we briefly discuss some fundamental approaches.

2.1 Disambiguating the Meaning of a Word

Humans are able to disambiguate polysemous words using their world knowledge (e.g. situational or experience-dependent) about the related context, but in most cases they can do this using their linguistic context knowledge related strictly to the language [10]. In order to identify the meaning of a polysemous word in an automatic WSD task, we need to recognize and model this context. Basically, this can be done in two ways [8]:

1. As a bag of words (as in some window surrounding the searched word, as in a bag).
2. As relational information (including information about distance from searched word, syntactic relations, semantic categories, etc.).

The linguistic knowledge can be accessed using the knowledge-driven WSD approach [8]. Here, lexical resources [14] (like linguistic ontologies, machine readable dictionaries, thesauri or computational lexicons) provide linguistic information. In order to obtain a linguistic context description of different word senses we have to explore, e.g., linguistic ontologies using the word we are looking for, selecting the concepts based on the linguistic relations that define the different word senses and their linguistic context.

2.2 The Use of EuroWordNet

Some of the linguistic information required to disambiguate word senses, as we have discussed above, are provided in linguistic ontologies like EuroWordNet [22]. Besides, this resource can be used for text analysis, computational linguistics and many related areas [12]. EuroWordNet provides a list of word senses for each word, organized into synonym sets (SynSets), each representing one constitutional lexicalized concept. The EuroWordNet structure is the same as the Princeton WordNet [11] in terms of SynSets with different semantic relations between them. Each individual WordNet represents a unique language-internal system of lexicalizations. An Inter-Lingual-index (ILI) was introduced in order to interlink the WordNets. Thus, it is possible to access the concepts (SynSets) of a word sense in different languages, and, hence, to retrieve one and the same concept in different languages with its related translations and linguistic relations. The multilingual retrieval of a word sense (SynSet) is done using the ILI entries. When a synset, e.g. “bank” with the meaning “financial institution”, is retrieved in the English WordNet, its SynSet-ID can be used to retrieve the same concept in all other language-dependent WordNets (Italian, Spanish, French, etc.) that describe the same concept with the same ID, but naturally contain the word description in its specific language.

However, we encounter different problems if we use EuroWordNet for language acquisition or retrieval of documents, as we have discussed in more detail in [3]. WordNet and EuroWordNet provide a differentiation of word senses that is very often too fine grained [4]. For example, if we retrieve the term *rule*, we get 12 different meanings that would hardly be distinguished by someone who is merely interested in more general information about the usage and possible senses of the word ‘rule’. We discussed these problems in more detail in [3] and came to the conclusion that methods that allow a revision of the WordNet structure are required. One way to obtain a higher granularity is to merge SynSets if they describe a very similar meaning of the same word [5]. Such methods could be used for creating an adapted or reduced structure of the ontology hierarchy, having fewer word senses that are carrier of a more distinctive meaning. This reduced number of meanings can then be used in order to categorize the documents retrieved and a user can more easily select the sense he is looking for [3]. A first approach we presented to solve this problem is described in [5].

In the following, we focus on the use of EuroWordNet for retrieving documents that include examples of word usage in a multilingual framework. The methods for document search in the approaches we have developed so far are integrated in our search engine framework CARSA [1]. The word senses and their translations are obtained from the EuroWordNet Ontology.

2.3 Combining and Translating Combinations of Words

As we mentioned above, the word senses in EuroWordNet are also connected to their appropriated translation in different languages. However, if users want to restrict the the word usage of interest to a certain context, they usually use more than one term. The use of sets of words results in the problem that we now have to deal with all

possible combinations of all meanings of all terms. The number of combinations grows exponentially with the number of query words [10]. This makes it even more difficult to disambiguate all possible word combinations. However, in the following we show how we can use a visualization of search hits and combinations of the meanings of translated query terms to help the user in finding possibly correct translations.

A very specific example for word combinations are transparent compound words (e.g. “Holztür”, door made from wood) in German. Compound words are often not contained in linguistic ontologies such as EuroWordNet. However, the meaning of such a compound word can in many cases be obtained from the combination of the meanings of the word parts. If people, for example, do not know the meaning of a compound word they try to decompose it in order to extract the word sense. In order to understand the word sense, people frequently try to translate the individual word parts in their own language and then try to understand the linguistic context.

Another usage would be if people do not know how to translate a word/concept in another language with the intended meaning. Short dictionary explanations frequently do not help. In this case they might find it useful to navigate through all possible senses and automatically get examples for word usage from web documents. Besides, hit statistics about the co-occurrence of words in documents provide information about possibly correct translations. These and other examples are the typical user scenarios we cover with our tool.

3 Tools for Visualization and Navigation of Linguistic Ontologies

If we like to support a user in language acquisition and cross-language text retrieval the interface has to adapt to the languages involved, to the word senses and to the context during the disambiguation process. Given that we deal with (linguistic) relations and sets (or groups) of similar documents, we first briefly discuss related works dealing with the visualization and management of such relational features.

3.1 Related Work

In recent years, ontologies have been used as a way to share, reuse and process domain knowledge. Applications such as information management systems, semantic web services and electronic commerce adopt them for this purpose. But in order to manage and present information in the best way, we need tools that support the creation, visualization, and management of ontologies.

Several ontology editors have been already implemented for this task. OilEd [2] is an example of freeware ontology editor for building ontologies. Protégé [6] is another example of a free, open-source platform to construct knowledge-based applications with ontologies. Protégé can be extended by plug-ins that can be easily embedded. However, the main focus for our task at hand is not the management of an ontology, but to use the ontology as a resource to enable user friendly navigation through linguistic information. Therefore, the aspect of information presentation is much more important.

Several approaches to visualize sets and relations have already been proposed. In [2] different representations of sets with the help of number values, brightness value, different bars and circles are shown and their effectiveness is compared. The authors represent with the size of the circles the number of elements in the set. Also, [19] describes this as a good way to represent sets, but using this circle representation in combination with the Steven's psychophysical power function.

For representing linguistic relations much work has already been done. For example, the Visual Thesaurus is a tool that uses an exploration-oriented approach in order to allow a user to navigate through linguistic resources. When the user clicks on any related word of a concept that is currently in focus, this word moves to the center of the search work pane (see <http://www.visualthesaurus.com/>). Other work like in [17] is based on the same design principle for working with relations using circles and centering the search objects. To solve the problem of overlapping, transparency effects are used.

The authors of [6] developed VisDic for browsing and editing multilingual information taken from EuroWordNet. Here users can browse static information on text blocks. Another web interface for multilingual information browsing is presented in [18]. Here a parallel corpus annotated with MultiWordNet [16] can be browsed as well as the words with their related annotated word senses, but the corpus is very restricted. All accessible information is static. This interface is used only for a bilingual search in a closed domain. Other work dealing with the multilinguality and lexicography has shown that researchers in this area mostly deal with multilingual lexical resources or corpora only, without automatically retrieving web documents related to the query. None of the above mentioned tools seamlessly integrates navigation of linguistic ontologies with web searches.

3.2 MultiLexExplorer

MultiLexExplorer is a tool that combines the knowledge-driven WSD (see Section 2) with the knowledge-based text retrieval [13] approach in an interactive language learning framework. This tool is not intended as an environment for development of ontologies. It is designed to help the user in the language learning process. We use the linguistic ontologies in order to disambiguate documents (retrieved from the web or a local document collection) given the different meanings (retrieved from linguistic ontologies, in our case EuroWordNet [22]) of a term having unambiguous description in different languages. We focus especially on the integration of methods that support the adaptation of the system interface and the output to the current search context [1]. Using EuroWordNet for language exploration, we support the user in:

- exploring the linguistic context of a word in the general hierarchy,
- searching in different languages, e.g., by translating word senses using the interlingual index of EuroWordNet,
- disambiguating the word senses of different word combinations,
- interacting with the system changing the search context "center" of the original query and, thus, also the search words and the number of retrieved results,

- expanding the original query to restrict the number of retrieved documents, and in
- categorizing the retrieved web documents automatically using different categorization methods (e.g., as described in [3] and [4]).

In the following, we describe these aspects in more detail and present how MultiLexExplorer implements these functionalities.

3.2.1 Enabling Multilingual Ontology-based Exploration

MultiLexExplorer can help users in discovering languages, disambiguating meanings and combining words in order to learn about their correct translation. Figure 1 gives an overview of the tool. The different parts of the user interface are labelled and described in the following.

The interaction with the system starts by entering words the user is interested in (label a1) and configuring the services he wants to work with (label b). Users can explore the linguistic context of the word using the linguistic ontologies in their native language or choose another language (label a1). This linguistic context can interactively be modified using the interface by clicking on the linguistic relation boxes to include or exclude such relations (label e).

Users can decide to start a multilingual exploration search. One application could be given by experimenting with the disambiguation of translated german multi-word queries. Here, we first give the query words in one (source) language and choose the target language (label d). The interface automatically provides translations of all possible source language senses in the target language based on the ILI entries of EuroWordNet (cf. Section 2.2). In this way we can recognize the word senses of the different word combinations and disambiguate them with the help of the lexical resource. At this time all the query words are shown with selected linguistic relations (“Haus” in Figure 1 on the left hand side, and “Tür” on the right hand side) and their translation (in Figure 1 to the right of “Haus” or to the left of “Tür”) retrieved. The linguistic relations and the languages can be selected on the right hand side of the interface.

3.2.2 Using Multilingual Linguistic Information for Language Acquisition

The user can simultaneously explore word senses contained in the multilingual lexical resources and related web documents. Here, the tool combines the translated words (label c3) to search the Web for documents containing this combination of words. The number of hits is visualized by the size of the bigger circles (label c2). The tool automatically searches for all combinations between all senses and belonging synonyms. If a user selects one of these circles, search results are shown then in the style of any search engine (label f). The interface in this case shows the document hits distribution of the translations. Thus, if the user does not know the meaning of a composite word in German, he can try to understand the contextual meaning of the word without really knowing the word he is looking for. This can be very helpful

when looking for meanings that we don't know a priori but also for language acquisition. This help is given, for example, by browsing a second time the results retrieved from the web search.

Another possibility given to the user is to change the search context "center" during the search process (label c1). If users are looking for a certain word sense and during their search they realize that the context they are looking for is another as their primary search, they can switch it with a right click on every word (given from the linguistic relations) and combine this word with the word positioned on the opposite side of the interface. In our example if the user clicks on the word "Gebäude" and choose it as the new center, this will be combined with the word "Tür". The tool automatically adapts the visualization to this new selection and all word combinations are rebuilt for this new case and results are presented.

The word can also be removed from the work pane (label c1) or can be added to expand the query for retrieving a more restricted number of web documents (label a2) containing the query words and this added word. For example, if we choose the hyponym "Eingangstür" in Figure 1, we can add it to the query "Haus AND Tür" and related documents are retrieved.

MultiLexExplorer can also be used in combination with document classification methods described in [3]. This gives the possibility to users to better navigate the huge amount of documents contained in the Web, since the documents are now classified with the word senses of the lexical resource or with another categorization processes that can be selected as plug-ins (label g). The integrated plug-ins are connected to the CARSA [1] architecture that handles a pool of plug-ins, each containing

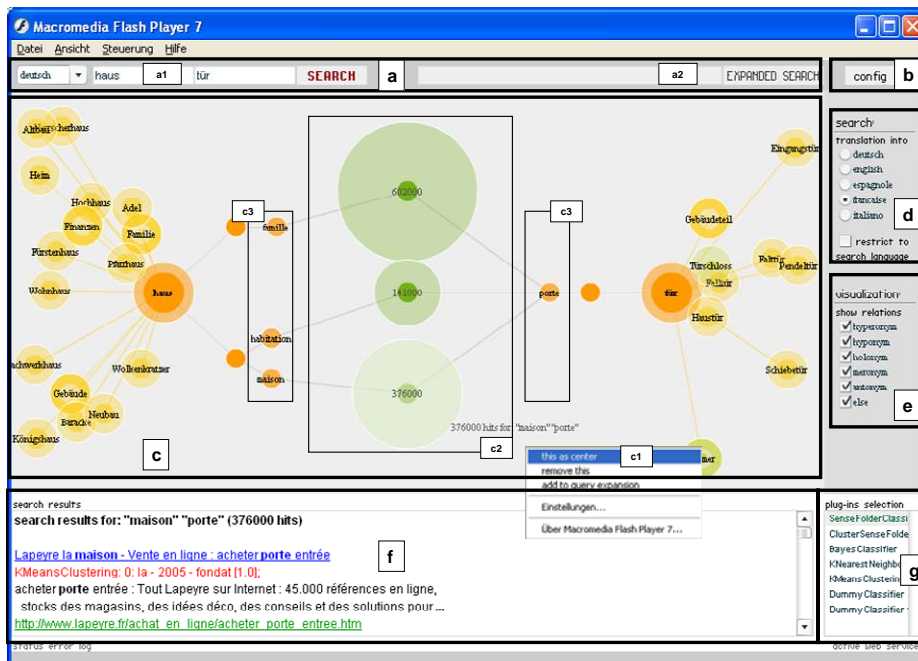


Figure 1. Screenshot of the MultiLexExplorer Tool.

one specific method for a specific part of the search process. This comprises plug-ins for accessing document collections for indexing and searching, plug-ins that provide methods to structure or to mine result sets as well as methods that pre-process the data such that it can be easily visualized by a client tool connected to the meta-searcher. Furthermore, plug-ins that provide access to ontologies (e.g., WordNet) that are required by other plug-ins (e.g., for multilingual support or semantic categorization) are integrated. The categorization plug-ins results are used then as additional information given in the second line of every document (label f). By using these classification methods users could select only the meaning they are interested in and then navigate only relevant documents. In this way we can categorize documents not only depending on the language, but also on the word sense.

4 Conclusions

In this paper, we have discussed different problems related to word sense disambiguation, language acquisition and cross-language text retrieval. Based on these discussions we proposed methods that can support a user in language acquisition by providing support for navigation in linguistic ontologies and by supporting him with interactive visualization techniques in finding the correct word sense translation. At the same time the proposed methods facilitate a user to efficiently retrieve examples for word usage in foreign language documents.

In future work we will study methods to further improve and personalize the applied word sense disambiguation methods and the integration in the user interface of the developed prototype. For example, information from a learner profile could be used to automatically modify the granularity of the senses that are distinguished by the system (see the discussion in Sect. 2.2). An advanced learner might be interested in very fine grained sense distinctions, while a beginner is usually more interested in learning quickly rough language concepts. Currently, it is only possible to manually adapt the granularity of word sense distinction.

The use of different ontologies is already possible through the CARSA web-service architecture [1]. Given the accessibility of our implementation via web-services, a Protégé plug-in could be made available for building knowledge-based tools and applications. In the near future, we are planning a user study in order to evaluate the performance of our tool.

References

1. K. Bade, E. W. De Luca, A. Nürnberger and S. Stober. CARSA - An Architecture for the Development of Context Adaptive Retrieval Systems, In *Proc. of the 3rd Int. Work. on Adaptive Multimedia Retrieval (AMR 2005)*, Springer-Verlag, 2006.
2. S. Bechhofer, I. Horrocks, C. Goble, and R. Stevens. OilEd: A reason-able ontology editor for the semantic web. In *KI-2001: Advances in Artificial Intelligence, LNAI 2174*, pages 396-408. Springer, 2001.

3. E. W. De Luca and A. Nürnberger. Improving Ontology-Based Sense Folder Classification of Document Collections with Clustering Methods, In: *Proc. of the 2nd Int. Workshop on Adaptive Multimedia Retrieval (AMR 2004)*, pp 72-86, Valencia 2004.
4. E. W. De Luca and A. Nürnberger., Supporting Mobile Web Search by Ontology-based Categorization. In: *Proc. of GLDV 2005*, 28-41, 2005.
5. E. W. De Luca and A. Nürnberger, The Use of Lexical Resources for Sense Folder Disambiguation. In *Proc. of the Work. Lex. Sem. Res. (DGfS-06)*, Bielefeld, Germany, 2006.
6. J. Gennari, M. A. Musen, R. W. Ferguson, W. E. Grosso, M. Crubezy, H. Eriksson, N. F. Noy, S. W. Tu. The Evolution of Protégé: An Environment for Knowledge-Based Systems Development. 2002.
7. A. Horák and P. Smrž. VisDic - Wordnet Browsing and Editing Tool. In: *Proceedings of the Second International WordNet Conference, GWC 2004*, 2004.
8. N. Ide and J. Véronis. Word Sense Disambiguation: The State of the Art. In: *Computational Linguistics*, Volume 14, Part 1, 1998.
9. R. Mihalcea and D. Moldovan. Automatic Generation of a Coarse Grained WordNet, in *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, pp.35-41, Pittsburgh, PA, June 2001.
10. G. A. Miller. Ambiguous Words. In: *Impacts Magazine*. Published on KurzweilAI.net, 2001.
11. G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. Miller. Five papers on WordNet. ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps. 1993.
12. J. Morato, M. Marzal, J. Lloréns and J. Moreiro. WordNet Applications. In: *Proc. of the 2nd Int. Conf. Global WordNet*, Brno, Czech Rep. 2004.
13. C. Peters and P. Sheridan. Multilingual Information Access. In: *Lectures on Information Retrieval, Third European Summer-School, ESSIR 2000*, Varenna, Italy, 2000.
14. W. Peters. Lexical Resources, In: *NLP group Department of Computer Science*, University of Sheffield, http://phobos.cs.unibuc.ro/roric/lex_introduction.html, 2001.
15. C. Peters, P. Clough, J. Gonzalo, G.J.F. Jones, M. Kluck, B. Magnini (Eds.). Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum (CLEF 2004), LNCS Vol. 3491, Springer-Verlag, 2004.
16. E. Pianta, L. Bentivogli and C. Girardi. MultiWordNet: Developing an aligned multilingual database. In *Proc. of the 1st Int. Global WordNet Conf.*, pp. 293-302, India, 2002.
17. S. Pierre. Revealicious revealing, the way you use [del.icio.us.](http://www.ivy.fr/revealicious/), <http://www.ivy.fr/revealicious/>, 2005.
18. M. Ranieri, E. Pianta and L. Bentivogli. Browsing Multilingual Information with the MultiSemCor Web Interface, In: *Proceedings of the LREC 2004 Satellite Workshop on The Amazing Utility of Parallel and Comparable Corpora*, Portugal, 2004, pp. 38-41.
19. H. Schumann and W. Müller. Visualisierung, Grundlagen und allgemeine Methoden. *Berlin Heidelberg : Springer*, 2000.
20. J.-B. Son (ed.) (2004) Computer-assisted language learning: concepts, contexts and practices. Lincoln, NE: iUniverse.
21. Thinkmap. Visual Thesaurus, <http://www.visualthesaurus.com>, 2005.
22. P. Vossen. EuroWordNet: a multilingual database for information retrieval. In: *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, Zurich, 1997.
23. M. Warschauer (1996). Computer Assisted Language Learning: an Introduction. In Fotos S. (ed.), *Multimedia Language Teaching* (pp. 3-20). Tokyo: Logos International.