A Word Sense-Oriented User Interface for Interactive Multilingual Text Retrieval

Ernesto William De Luca and Andreas Nürnberger
Faculty of Computer Science
University of Magdeburg
39106 Magdeburg, Germany
{deluca, nuernb}@iws.cs.uni-magdeburg.de

Abstract

In this paper we present an interface for supporting a user in an interactive cross-language search process using semantic classes. In order to enable users to access multilingual information, different problems have to be solved: disambiguating and translating the query words, as well as categorizing and presenting the results appropriately. Therefore, we first give a brief introduction to word sense disambiguation, cross-language text retrieval and document categorization and finally describe recent achievements of our research towards an interactive multilingual retrieval system. We focus especially on the problem of browsing and navigation of the different word senses in one source and possibly several target languages. In the last part of the paper, we discuss the developed user interface and its functionalities in more detail.

1 Introduction

The internet comprises of mainly English documents, but the amount of documents in other languages grows daily. Therefore, the internet is likely to change very quickly from an English language medium to a multilingual information and communication service. Most people have a good passive understanding of a foreign language, but are not usually in the situation to formulate search queries in this foreign language as good as in their mother tongue. Considering that people want to access multilingual information, the importance of their ability of language understanding increases rapidly. At the moment the support provided to navigate multilingual information is not yet so sophisticated that users can access documents over the internet in the seamless and transparent way as they do in their mother tongue.

In order to enable users to access multilingual information, different problems have to be solved: disambiguating the query words, translating the query words, as well as categorizing and presenting the results appropriately. In the following, we briefly discuss these aspects.

1.1 Disambiguating the Query Words

Humans often use polysemous words for searching for documents. Unfortunately, a distinction of the related word senses is difficult [Miller, 2001]. A word is polysemous if it has different meanings (polysemy from Greek poly = *many* and semy = *meanings*). When people search for documents related, e.g., to the word *bank*, they will find different documents related to different meanings of this word (bank as a financial institution, bank as a seat, etc.). Humans are able to disambiguate these polysemous words using their knowledge about the related context, but mostly they can do this using their linguistic context knowledge related strictly to the language [Miller, 2001]. Reading the documents retrieved, they can assign the word sense to its linguistic context. In order to identify the meaning of a polysemous word in an automatic word sense disambiguation task, this linguistic context has to be considered. Working in a multilingual context, words have to be disambiguated both in the native and in other languages (see Section 2.1).

1.2 Translating the Query Words

Retrieving documents in other languages, we have to translate the concepts of the search keywords. Machine translation should help in processing and delivering this information. But as discussed in, e.g., [Peters and Sheridan, 2000], this approach cannot be viewed as a realistic answer to the problem of query translations right now.

The problem of automatically matching documents and queries over languages is not properly solved yet, and therefore it has to be done manually to a great extent. In Section 2 the use of query-related word senses retrieved from the lexical resources and their translation as an alternative solution to this problem is discussed.

1.3 Categorizing and Visualizing the Results

User studies have shown that categorized information can improve the retrieval performance for a user. Thus, interfaces providing category information are more effective than pure list interfaces for presenting and browsing information as shown, for example, in [Dumais *et al.*, 2001], where the effectiveness of different interfaces for organizing search results was evaluated. Users were 50% faster in finding information organized into categories. Similar results based on categories used by Yahoo were presented in [Labrou and Finin 1999].

Motivated by these evaluations, we developed methods in order to provide additional disambiguating information to the documents of a result set retrieved from a search engine in order to enable categorization, restructuring or filtering of the retrieved document result set. Since we cannot expect a perfect word sense disambiguation or categorization of results, an adaptive and error tolerant visualization is required. Thus, the retrieval of information should be supported by an appropriate interactive visualization of results and categories.

2 Word Sense Disambiguation (WSD) and Translation (WST)

The automatic disambiguation of word senses is still a very interesting and challenging research task. Since the 1950's different researchers try to disambiguate words, sentences or documents for different purposes as machine translation, information retrieval and hypertext navigation, content and thematic analysis, grammatical analysis (Part-Of-Speech Tagging), speech or text processing [Ide and Véronis, 1998]. In general terms a word sense disambiguation (WSD) process can be described by two steps:

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

For disambiguating word senses a variety of association methods (knowledge-driven, data-driven or corpus-based WSD) can be used [Ide and Véronis, 1998]. So far, we only used the knowledge-driven WSD approach, i.e. we make use of linguistic information contained in lexical resources [Peters, 2001], like machine readable dictionaries, thesauri or computational lexicons, in order to obtain a linguistic context description of the word senses. Therefore, lexical resources have to be (automatically) explored using the query words, selecting the concepts based on the linguistic relations that define the different word senses and their linguistic context.

In order to identify the meaning of a polysemous word in a WSD task, we need to recognize also its linguistic context. For this purpose the linguistic context is used in two ways:

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

The linguistic context knowledge can be accessed from an information retrieval system using the knowledge-driven WSD approach mentioned above.

In order to use linguistic resources for a multilingual approach we have to retrieve not only the concept (word sense) with its linguistic relations, but also its related translations. Some of the linguistic information and the related translations required to disambiguate word senses, as we discussed above, are provided in lexical resources like EuroWordNet [Vossen, 1997]. Besides, this resource can be used for text analysis, computational linguistics and many related areas [Morato *et al.*, 2004]. In the following, we briefly describe the use of EuroWordNet for document retrieval in a multilingual Framework.

2.1 The use of EuroWordNet

Given that we want to retrieve from the web different documents in different languages, we have to analyze the different linguistic contexts of a word in these languages. Therefore, we decided to use the EuroWordNet multilingual lexical database. Its basic structure is the same as the Princeton WordNet [Miller *et al.*, 1993] in terms of SynSets with different semantic relations between them.

EuroWordNet consists of a set of language specific WordNets. Each individual WordNet represents a unique language-internal system of lexicalizations. The Inter-Lingual-Index (ILI) was introduced in order to connect the WordNets of the different languages. Thus, it is possible to access the concepts (SynSets) of a word sense in different languages. It means that we can retrieve one and the same concept in different languages with its related translations and linguistic relations.

In addition to the Inter-Lingual-index, there is also a Domain-Ontology and a Top-Concept-Ontology related to this lexical database. The shared Top-Ontology is a superordinate hierarchy of 63 semantic distinctions for the most important language independent concepts (e.g. Artifact, Natural, Cause, Building) and is interconnected with the ILI through the WordNet-Offsets. Hereby, a common semantic framework for all the languages is given, while language specific properties are maintained individually. The Domain-Ontology was created for use in information retrieval settings in order to obtain specific concepts (only implemented exemplary for the computer terminology). Figure 1 gives an overview over the architecture of the EuroWordNet whereby the single components and its relations are represented.

However, different problems related to the use of (Euro)WordNet for information retrieval have been encountered as discussed in more detail, e.g., in [Mihalcea and Moldovan 2001; Morato *et al.*, 2004; De Luca and Nürnberger, 2006c]. One main problem is that the differentiation of word senses is very often too fine grained for typical information retrieval tasks. One way to obtain a higher granularity is to merge SynSets if they describe a very similar meaning of the same word [De Luca and Nürnberger, 2006a]. For web search, such methods could be used for creating a reduced structure of the ontology hierarchy, having fewer word senses that are carrier of a more distinctive meaning, in order to categorize the documents retrieved [De Luca and Nürnberger, 2006c]. We described a first approach to solve this problem in [De Luca and Nürnberger, 2006a] in a monolingual task.

When we deal with EuroWordNet, these problems persist, and other problems come along. In general, the problem of automatically finding translation of word senses can be solved using such a resource. The use of the Inter-Lingual-Index helps for this purpose. However, the coverage of language-dependent word senses varies from language to language, i.e. from ~20.000 (german) to 150.000 (english) Synsets. Using this lexical resource, we have to take into account the missing (or incomplete) translations contained in the lexical resource, apart from the lexical gaps (word senses that exist in a language and not in another).

2.2 The use of the CARSA Search Engine Framework

The document search in our approach is done using our search engine framework CARSA [Bade *et al.*, 2005]. CARSA is a web services based architecture, which supports the development of context based information retrieval systems. The idea of these systems is to support a user in his search process by, e.g., adapting the search results as well as the interface itself to user specific needs and interests. We decided to divide query results set processing (the information to be presented) from the interface design (information presentation) in order to simplify the

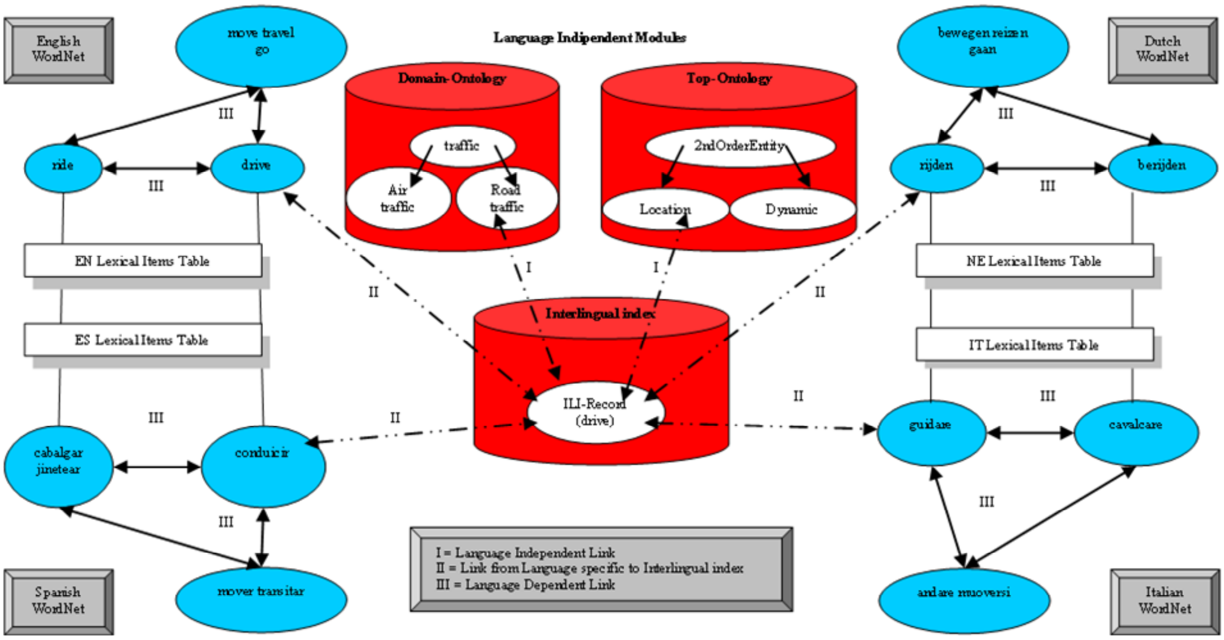


Figure 1. EuroWordNet Architecture (see [Vossen, 1997]).

development of retrieval systems for, e.g., different desktop as well as mobile devices. The central component of the retrieval system is a meta search engine providing methods to restructure and annotate result sets of user queries. Search engines (e.g. Google, local searchers) and the user interfaces are connected to the system by Web services. Using this modular implementation, it is easily possible to extend the system by additional search engines or to integrate different interfaces. An overview of the system architecture of CARSA is given in [Bade *et al.*, 2005].

3 Cross-Language Text Retrieval

After having described the problem of disambiguating word senses in general, we focus in the following on the disambiguation of the words in documents retrieved from an information retrieval system, and particularly in the multilingual framework we want to deal with.

In general an information retrieval system tries to find and retrieve relevant documents related to a user query, with documents and query being in the same language [Abdelali *et al.*, 2003]. Dealing with a multilingual document collection naturally brings up new questions. Being able to read a document in a foreign language does not always imply that a user can formulate appropriate queries in that language as well. Users find cross-language text retrieval particularly useful when they can express their information needs effectively in their mother tongue, while handling with languages they are less confident with [Oard, 1997].

In [Peters and Sheridan, 2000] different methods for multilingual information access are described, addressing the problems of accessing, querying and retrieving useful documents contained in different collections in several languages at any level of specificity, including different computational linguistic processing. The authors distinguish three main approaches for multilingual information access:

1. Machine translation techniques
2. Corpus-based techniques
3. Knowledge-based techniques

They argue that full machine translation (MT) can not be seen as a realistic answer to the problem of matching documents and queries over languages. One weakness of present fully automatic machine translation systems is the limitation of producing high quality translations only in specific domains. Such approaches could substitute every possible translation for a polysemous word, thus increasing recall at the expense of precision. In addition, it does not represent a cost-effective solution for query translation either [Oard and Dorr 1996].

The corpus-based approaches analyze automatically large collections of text with statistical methods. Here, the semantic is given only by translated sentences related across the languages and these approaches are applicable only in a restricted domain.

Since we want to avoid the use of large corpora and translation methods that are not yet providing sufficient quality, our focus is on the use of knowledge-based approaches to enable multilingual information access. These approaches use ontologies, dictionaries (bi- or multilingual) or thesauri in order to enable cross-language text retrieval. Thus, we first try to find all word senses, then retrieve the appropriate translation from the lexical resource and finally categorize documents using the (most likely) proper word sense. Finally, we have to visualize the results according to the user needs as described in more detail in Sect. 5.

For a more detailed description of the three fundamental approaches for multilingual information access, we refer to [Peters and Sheridan, 2000], where a detailed explanation and several references are given.

4 Combining Word Sense Disambiguation within Cross-Language Text Retrieval

In order to combine the word sense disambiguation process within a cross-language retrieval system, we have developed, so far, several tools, e.g., [De Luca and Nürnberger, 2005] and evaluated different disambiguation approaches [De Luca and Nürnberger, 2006a and 2006c]. In the following, we discuss some of the most important aspects. For more details see the referenced publications.

4.1 Tools

The first visualization interface for multilingual search, MultiLexExplorer [De Luca *et al.*, 2006], was developed with a focus on multilingual *explorative* search. MultiLexExplorer combines word sense disambiguation with a text retrieval approach in an interactive framework. It uses lexical resources to support a user in disambiguating documents (retrieved from the web or a local document collection) given the different meanings (retrieved from lexical resources, in our case EuroWordNet) of a search term having unambiguous description in different languages. By visualizing search results grouped by keyword combinations and word senses, the user can discover languages using lexical resources for disambiguating meanings, combining words and their translation. The translations of all possible source language senses are provided in the target language based on the ILI entries of EuroWordNet (see Section 2.1).

The LexiRes tool [De Luca and Nürnberger, 2006b] provides the possibility of restructuring the word senses provided by a lexical resource for information retrieval purposes. Users are usually interested in a small list of meanings with very distinctive features. Since many lexical resources, especially WordNet, provide frequently too fine grained word sense distinctions, we implemented the tool LexiRes that gives the possibility to navigate lexical information, helping authors of already available lexical resources in deleting or restructuring concepts using automatic or manual merging methods, e.g., as described in [De Luca and Nürnberger, 2006a].

These tools were first steps before implementing the interface presented in this paper. Both tools were used to deal with multilingual queries and documents and helped in finding an appropriate visualization of the results and word senses.

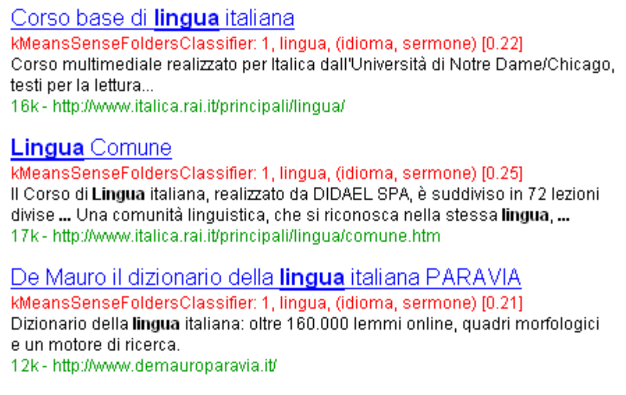
4.2 Document Disambiguation and Classification Using the Sense Folder Approach

In this section we briefly introduce the functionality of the Sense Folder Disambiguation. This approach is used to classify documents in *Sense Folders*, which are defined based on context descriptions obtained by merging information from word senses (retrieved from WordNet) with associated linguistic relations as proposed in [De Luca and Nürnberger, 2006c].

First of all we want to semantically disambiguate the query terms (used in the retrieved documents) using WordNet. Therefore, we categorize documents with respect to the meaning of a query term using different linguistic relations retrieved from EuroWordNet. These relations provides us with words defining the context of the query term in order to create *Sense Folders* for its different meanings. Thus, for each (EuroWordNet-) sense of a query term, a *Sense Folder* (prototypical word vector) is created containing:

- all synonyms (the SynSet)
- all hypernyms (the superordinate word), i.e. dividing senses/categories where hypernyms intersect,
- all hyponyms (the subordinate word),
- the belonging glosses (description of the SynSet elements by words that are frequently used in this specific semantic context),
- and the belonging word domain (word context).

These Sense Folders are compared within the words contained in the documents and are used in order to categorize and annotate retrieved documents with their best matching Sense Folder. Every document is first assigned to its most similar Sense Folder and afterwards this classification is revised by a clustering process in order to improve the disambiguation performance [De Luca and Nürnberger, 2006c]. Labels defining the disambiguating classes are then added to each document of the result set. The visualization of such additional information (Fig. 2) should enable a simple navigation through the huge number of documents and, if possible, should restrict information only to the relevant query-related results.



Corso base di **lingua** italiana
kMeansSenseFoldersClassifier: 1, lingua, (idioma, sermone) [0.22]
Corso multimediale realizzato per Italica dall'Università di Notre Dame/Chicago, testi per la lettura...
16k - http://www.italica.rai.it/principali/lingua/

Lingua Comune
kMeansSenseFoldersClassifier: 1, lingua, (idioma, sermone) [0.25]
Il Corso di **Lingua** italiana, realizzato da DIDAEL SPA, è suddiviso in 72 lezioni divise **...** Una comunità linguistica, che si riconosca nella stessa **lingua**, **...**
17k - http://www.italica.rai.it/principali/lingua/comune.htm

De Mauro il dizionario della **lingua** italiana PARAVIA
kMeansSenseFoldersClassifier: 1, lingua, (idioma, sermone) [0.21]
Dizionario della **lingua** italiana: oltre 160.000 lemmi online, quadri morfologici e un motore di ricerca.
12k - http://www.demauroparavia.it/

Figure 2 Annotation/classification example searching for the term ‘lingua’

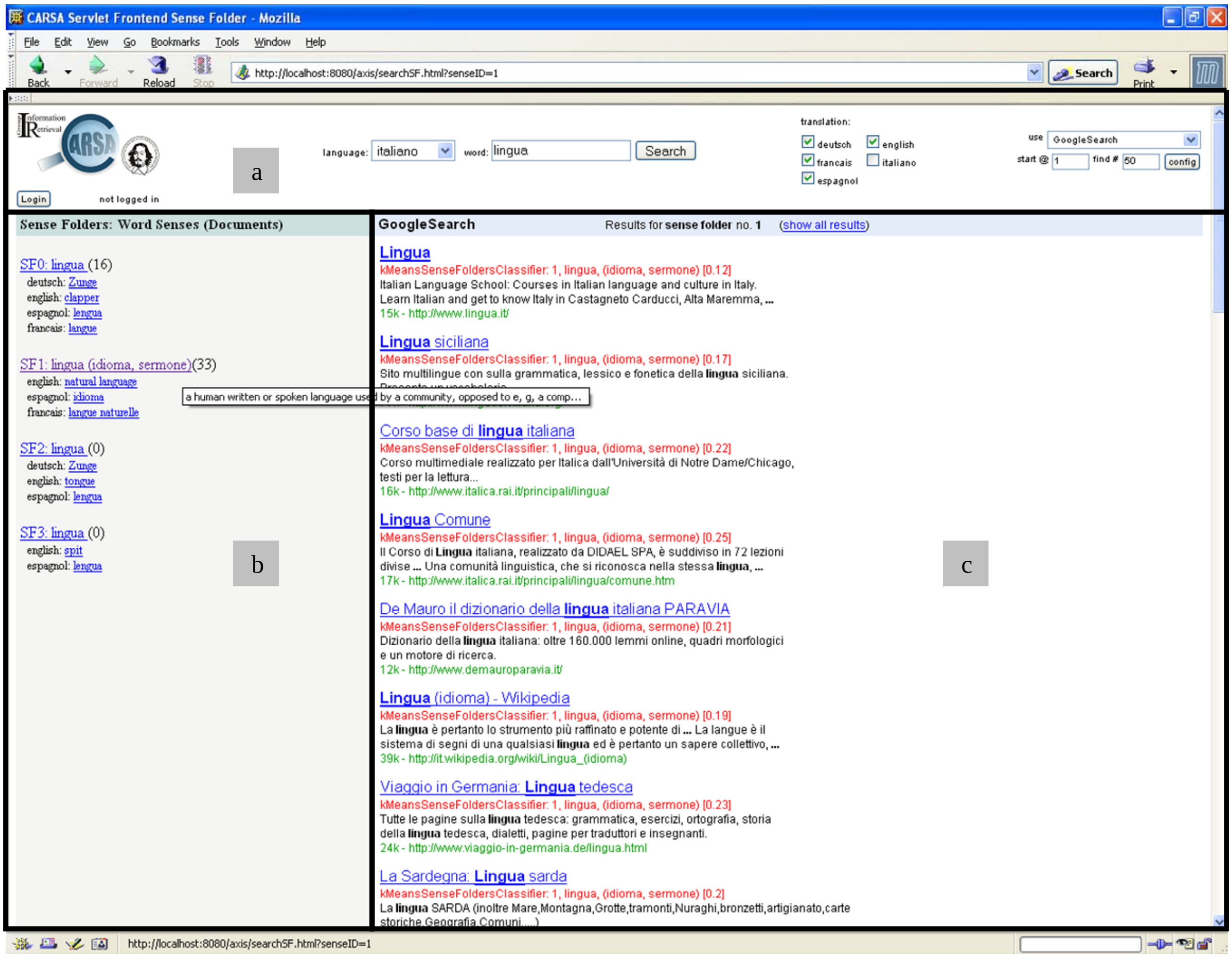
Figure 2 shows the implemented categorization techniques combining the knowledge-driven WSD with the knowledge-based text retrieval approach integrated in the developed user interface. The lexical resources are used in order to disambiguate documents (retrieved from the web) given the different meanings (retrieved from lexical resources, in this case EuroWordNet) of a search term. These techniques were combined with clustering processes that strongly improved the overall classification performance. While the pure Sense Folder based approach correctly classified 42% of the documents of a small benchmark dataset, the clustering process was able to assign approximately 70% of the documents to the correct class [De Luca and Nürnberger, 2006c]. More details about these approaches can also be found in [De Luca and Nürnberger, 2005 and 2006c].

5 The User Interface

In the following, we describe an approach for combining cross-language text retrieval, word sense disambiguation and document classification to provide a user-oriented presentation of the search results. We first briefly discuss related work, then we present the implemented user interface that gives the possibility of an interactive multilingual search, and finally we discuss a first evaluation of the automatic merging methods in this setting.

5.1 Related Work

Different work has already been done in dealing with word senses, clustering and multilingual queries. For example, in [Mihalcea and Moldovan, 2001 and Peters *et al.*, 1998] approaches for automatic sense clustering with EuroWordNet were presented. Methods for collapsing similar meanings for query expansion have been discussed in [Moldovan and Mihalcea, 2000].


CARSA Servlet Frontend Sense Folder - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back
Forward
Reload
Stop
Search
Print
Information Retrieval
CARSA
Login
a
b
c
Search
config

http://localhost:8080/axis/searchSF.html?senseID=1

not logged in

language: italiano word: lingua

translation:
☑ deutsch ☑ english
☑ francais ☐ italiano
☑ espagnol

use GoogleSearch
start @ 1 find # 50

Sense Folders: Word Senses (Documents)

SF0: lingua (16)
deutsch: Zunge
english: clapper
espagnol: lengua
francais: langue

SF1: lingua (idioma, sermone)(33)
english: natural language
espagnol: idioma
francais: langue naturelle

a human written or spoken language used by a community, opposed to e, g, a comp...

SF2: lingua (0)
deutsch: Zunge
english: tongue
espagnol: lengua

SF3: lingua (0)
english: spit
espagnol: lengua

GoogleSearch Results for **sense folder** no. **1** (show all results)

Lingua
kMeansSenseFoldersClassifier: 1, lingua, (idioma, sermone) [0.12]
Italian Language School: Courses in Italian language and culture in Italy.
Learn Italian and get to know Italy in Castagneto Carducci, Alta Maremma, ...
15k - http://www.lingua.it/

Lingua siciliana
kMeansSenseFoldersClassifier: 1, lingua, (idioma, sermone) [0.17]
Sito multilingue con sulla grammatica, lessico e fonetica della **lingua** siciliana.

Corso base di **lingua** italiana
kMeansSenseFoldersClassifier: 1, lingua, (idioma, sermone) [0.22]
Corso multimediale realizzato per Italica dall'Università di Notre Dame/Chicago,
testi per la lettura...
16k - http://www.italica.rai.it/principali/lingua/

Lingua Comune
kMeansSenseFoldersClassifier: 1, lingua, (idioma, sermone) [0.25]
Il Corso di **Lingua** italiana, realizzato da DIDAEL SPA, è suddiviso in 72 lezioni
divise ... Una comunità linguistica, che si riconosca nella stessa **lingua**, ...
17k - http://www.italica.rai.it/principali/lingua/comune.htm

De Mauro il dizionario della **lingua** italiana PARAVIA
kMeansSenseFoldersClassifier: 1, lingua, (idioma, sermone) [0.21]
Dizionario della **lingua** italiana: oltre 160.000 lemmi online, quadri morfologici
e un motore di ricerca.
12k - http://www.demauroparavia.it/

Lingua (idioma) - Wikipedia
kMeansSenseFoldersClassifier: 1, lingua, (idioma, sermone) [0.19]
La **lingua** è pertanto lo strumento più raffinato e potente di ... La langue è il
sistema di segni di una qualsiasi **lingua** ed è pertanto un sapere collettivo, ...
39k - http://it.wikipedia.org/wiki/Lingua_(idioma)

Viaggio in Germania: **Lingua** tedesca
kMeansSenseFoldersClassifier: 1, lingua, (idioma, sermone) [0.23]
Tutte le pagine sulla **lingua** tedesca: grammatica, esercizi, ortografia, storia
della **lingua** tedesca, dialetti, pagine per traduttori e insegnanti.
24k - http://www.viaggio-in-germania.de/lingua.html

La Sardegna: **Lingua** sarda
kMeansSenseFoldersClassifier: 1, lingua, (idioma, sermone) [0.2]
La **lingua** SARDA (inoltre Mare,Montagna,Grotte,tramonti,Nuraghi,bronzetti,artigianato,carte
storiche,Geografia,Comuni,...)

http://localhost:8080/axis/searchSF.html?senseID=1

Figure 3. Multilingual User Interface

Peters *et al.*, [1998] developed automatic methods for grouping senses into more coarse-grained sense groups. They started clustering, for example, word senses sharing the same hypernym (calling them sisters) that occurs between two or more senses hyperonyms.

In [Mihalcea and Moldovan, 2001] it was also motivated that there is no need of a fine grain distinction between concepts n a retrieval setting. In this work the authors propose an approach to collapse synsets having very similar meaning or deleting synsets that are rarely used. The similar meanings are collapsed here for query expansion. Also approaches for semantic indexing, e.g. [Gonzalo *et al.*, 1998], show that there is no need for such fine distinctions of word senses.

Different to the approaches mentioned above, we are not working on methods for query expansion, since we think that these approaches usually restrict the result set to much (and thus reduce the recall) and furthermore frequently change the original meaning of the query and thus do not reflect the users intention any longer [Gonzalo *et al.*, 1998]. Our goal is to support the user in the retrieval process by semantic annotation and structuring of result sets without modification of the initial query.

5.2 Interacting with the Multilingual User Interface

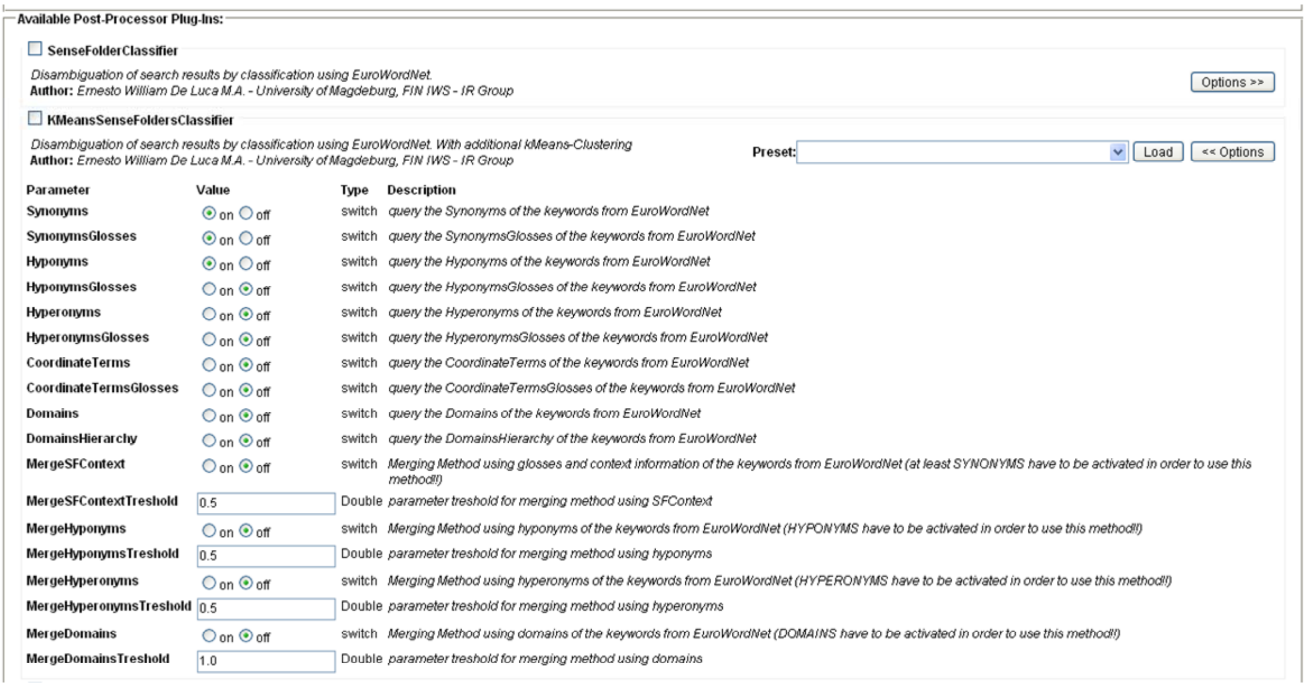
Figure 3 shows the user interface that is divided in three main parts:

a) The Search and Configuration Dialogue
b) The Word Sense Presentation
c) The Result List Presentation

A user that interacts with this user interface has to first configure the search before starting (label a). First of all the user chooses the language he wants to search with (as a “starting” source language). Afterwards, other languages (target languages that the user usually is able to speak and understand) can be selected for initializing the multilingual search. Furthermore, a user can also configure which search engines should be used. Of course, default settings are provided.

The configuration dialogue for the linguistic parameters for classification (see Section 5.3) can be started clicking on the button *config* (see Figure 4). However, this dialogue is recommended only for expert users. Here, the user can choose not only the linguistic parameters, but also the classification methods that should be used by the system. Presets for classification are implemented as a default.

After having configured the search parameters, the user can type in the query terms that he is interested in. These keywords are sent to the CARSA meta-search engine and to the Ontology Engine. The meta-search engine retrieves the documents. The documents are retrieved implicitly language-dependently. It means that when we start a search in Italian, we only retrieve Italian documents since the selected search terms are in italian. The Ontology En-



Available Post-Processor Plug-Ins:

□ SenseFolderClassifier

Disambiguation of search results by classification using EuroWordNet.
Author: *Ernesto William De Luca M.A. - University of Magdeburg, FIN IWS - IR Group*

□ KMeansSenseFoldersClassifier

Disambiguation of search results by classification using EuroWordNet. With additional kMeans-Clustering
Author: *Ernesto William De Luca M.A. - University of Magdeburg, FIN IWS - IR Group*

Parameter	Value	Type	Description
Synonyms	● on ○ off	switch	*query the Synonyms of the keywords from EuroWordNet*
SynonymsGlosses	● on ○ off	switch	*query the SynonymsGlosses of the keywords from EuroWordNet*
Hyponyms	● on ○ off	switch	*query the Hyponyms of the keywords from EuroWordNet*
HyponymsGlosses	○ on ● off	switch	*query the HyponymsGlosses of the keywords from EuroWordNet*
Hyperonyms	○ on ● off	switch	*query the Hyperonyms of the keywords from EuroWordNet*
HyperonymsGlosses	○ on ● off	switch	*query the HyperonymsGlosses of the keywords from EuroWordNet*
CoordinateTerms	○ on ● off	switch	*query the CoordinateTerms of the keywords from EuroWordNet*
CoordinateTermsGlosses	○ on ● off	switch	*query the CoordinateTermsGlosses of the keywords from EuroWordNet*
Domains	○ on ● off	switch	*query the Domains of the keywords from EuroWordNet*
DomainsHierarchy	○ on ● off	switch	*query the DomainsHierarchy of the keywords from EuroWordNet*
MergeSFContext	○ on ● off	switch	*Merging Method using glosses and context information of the keywords from EuroWordNet (at least SYNONYMS have to be activated in order to use this method!!)*
MergeSFContextTreshold	0.5	Double	*parameter treshold for merging method using SFContext*
MergeHyponyms	○ on ● off	switch	*Merging Method using hyponyms of the keywords from EuroWordNet (HYPONYMS have to be activated in order to use this method!!)*
MergeHyponymsTreshold	0.5	Double	*parameter treshold for merging method using hyponyms*
MergeHyperonyms	○ on ● off	switch	*Merging Method using hyperonyms of the keywords from EuroWordNet (HYPERONYMS have to be activated in order to use this method!!)*
MergeHyperonymsTreshold	0.5	Double	*parameter treshold for merging method using hyperonyms*
MergeDomains	○ on ● off	switch	*Merging Method using domains of the keywords from EuroWordNet (DOMAINS have to be activated in order to use this method!!)*
MergeDomainsTreshold	1.0	Double	*parameter treshold for merging method using domains*

Figure 4. Parameter Configuration

gine is concerned with the process of retrieving the word senses related to the query and the related translations used for the Sense Folder Disambiguation as described in Sect. 4.2. The retrieved word senses are used to filter the results and present them annotated by their meaning. Therefore, every document is labeled with the best matching Sense Folder (see Sect. 4.2). Every Sense Folder is used as class containing the related retrieved documents (label b). A glossary entry is shown in order to help a user in understanding what the word sense means. Such a glossary entry is activated on the mouse rollover event. It is always in English and retrieved from the EuroWordNet ontology.

If we click for example on the Sense Folder 1 (SF1) on the left side of the user interface, the system will show to the user only the documents that are classified by the system as belonging to that word sense, in this case 33 documents that are presented on the right side of the interface (label c), if the user clicks on the word sense. It means that if we are only interested in the documents related to the word sense "natural language" in Italian, we do not have to scan all results in order to retrieve the documents we are interested in. We can just browse the documents related to this word sense; in our case only 33 of 50. However, we like to emphasize that the word sense categorization is not perfect as already mentioned in Sect. 4.2. Therefore, we are still working on visualization methods that are better able to deal with this uncertain classification.

As we can see from Figure 3 not all senses are covered from the documents. It means that when we were looking for documents related to the word sense "lingua" (SF3) in the sense of "spit", we wouldn't find with the first search any related documents.

This interface gives the possibility of a multilingual search. As we can see, every Sense Folder has a translation related to the word senses retrieved for the languages chosen at the beginning of the search process. As we said before, the use of EuroWordNet implies missing (or incomplete) translations and lexical gaps. It means that not all word senses have a 1:1 translation in all foreign languages selected. However, considering our example above, where we are looking for the word sense "lingua" (SF3) in the sense of "spit", we can just click, for example, on the English translation of the word sense, to start automatically a new search with "spit" as a new search word in the English document web collection. The user interface presents then the new word senses related to the query word "spit" and filters the new retrieved documents to the correspondent Sense Folder. Obviously, here a new word sense disambiguation and retrieval process is started.

5.3 Search Configuration

Given that we want to use the word senses for filtering the documents with respect to their meaning, we have to configure the search with the document classification. The user can configure the Sense Folder Classification (or the classification supported by the clustering methods, as in the example with k-Means Clustering) choosing the parameters that characterize the word sense classes used as described in Sect. 4.2. Here the user can choose to activate any linguistic relation and merging method. Choosing the merging methods, thresholds can also be defined. Figure 4 shows the parameter configuration dialog that can be interactively be modified from the user. Depending on which linguistic parameters have been activated, the system classifies the documents. A first evaluation of the combination of the merging parameters has been described in [De Luca and Nürnberger, 2006a].

5.4 Evaluation of the Linguistic Parameters

In the following, we show a first evaluation of the combination of the linguistic parameters. Table 1 shows the results of this evaluation.

For our experimental studies we chose the pre-classified BankSearch web page collection [Sinka and Corne, 2002] consisting of 10,000 web documents classified into 10 equally-sized categories each containing 1,000 web

	SynHyperHypoGlo	SynHyperGlo	SynHypoGlo	SynGlo	SynHyperHypo
SF (Single SynSet „operation“)	**0.42**	0.38	0.40	0.32	0.29
CL (Single SynSet „operation“)	0.55	0.47	**0.54**	0.46	0.37
SF (merged SynSet „operation“)	**0.42**	0.39	0.40	0.30	0.22
CL (merged SynSet „operation“)	0.67	0.66	**0.82**	0.47	0.10
SF(Single SynSet „rule“)	0.36	0.28	**0.43**	0.33	0.26
CL (Single SynSet „rule“)	0.58	0.28	**0.68**	0.52	0.21
SF (merged SynSet „rule“)	0.40	0.31	**0.45**	0.36	0.27
CL (merged SynSet „rule“)	0.79	0.26	**0.87**	0.60	0.19

SF =Sense Folder Classification CL= k-Means Clustering with Sense Folders
Syn=Synonyms Hyper= Hyperonyms Hypo= Hyponyms Glo=Human descriptions

Table 1 Evaluation of linguistic parameters

documents. To each category one of four distinct themes, namely Banking and Finance, Programming Languages, Science, and Sport was assigned.

For the evaluation, we selected the subset of documents that contain the words “rule” or “operation”. The obtained documents were categorized using the pure Sense Folder classification (SF) approach and the clustering (CL) approach. We compared these two different automatic classification with the classification contained in the dataset (based on themes). Since these themes match nicely with the possible meanings of the term “rule” or “operation” described in WordNet (see Table 2 and Table 3), we first run the evaluation using all the SynSet available (related to the themes, but used as “Single SynSets”) and then merging them, mapping one or more SynSets to one Theme (“Exact Match, merged SynSets”). It means that we had first 6 SynSets (not merged) of the two word senses and 4 SynSets after merging semantically very similar word senses (For details on merging SynSets see [De Luca and Nürnberger, 2006a]). We consider in the following SynSet #0 as correctly classifying documents assigned to the banking and finance theme, SynSet #1 for the programming theme, SynSet #2 for the science and SynSet #3 for the sport theme. The SynSets that are considered not belonging to any of the themes have been removed. If the term “rule” or the term “operation” occurs in a document of this dataset it is usually used in the sense of the assigned theme. We can notice here that the best combination is almost always when we use only the combination of the linguistic relations (synonymy, hyponymy and gloss-description) with the merged form of the word senses.

6 Conclusions and Future Work

In this paper we presented a multilingual user interface that helps users in the search process considering the languages they can speak and the word senses they want to navigate in order to retrieve the documents they are looking for. Therefore, we integrated different word sense disambiguation methods in order to automatically categorize retrieved documents with respect to the sense in which a query term is used within the document. The results are presented in groups that can be accessed interactively. Even though, the performance of the word sense disambiguation methods is not yet sufficient and has to be improved, the interface already provides additional information that can help a user in browsing multilingual search results.

In future work, the usability of the current user interface has to be evaluated in order to better understand the needs of users working in a multilingual environment.

Furthermore, the use of EuroWordNet is very helpful, but we are thinking of implementing methods to extend this ontology, because only the English language has more or less acceptable coverage of the language.

The merging methods applied to the word senses can be helpful for a better document classification, but a deeper evaluation should be done and a more detailed analysis of the disambiguation performance is still necessary.

Wordnet SynSet	**Single SynSet**	**Exact Mapping (merged SynSet)**
#0 rule ruler (Metrology)	#1(2)	#1(2) Program
#1 rule formula (Sociology)	#0 (1)	#0(1) Banking
#2 rule regulation (behavior)	none	none
#3 rule formula (Mathematics)	#1 (3)	#1(2) Program
#4 principle rule (rule, law)	#2 (4)	#2 (3) Science
#5 principle rule (generalization)	#2 (5)	#2 (3) Science
#6 rule (religion)	none	none
#7 rule prescript (guide)	none	none
#8 rule (game, sport)	#3 (6)	#3(4) Sport
#9 rule linguistic rule (Linguistics)	none	none
#10 rule (legal authority)	none	none
#11 rule (History Time_Period)	none	none

Table 2. Comparison of WordNet SynSets and restructured SynSets for clustering for the word “rule”.

Wordnet SynSet	**Single SynSets**	**Exact Mapping (merged SynSet)**
#0. operation (being operative)	none	#3 (4) Sport
#1. operation (Commerce)	#0 (1)	#0 (1) Banking
#2. operation, functioning	none	#3 (4) Sport
#3. operation activity	#3 (5)	#3 (4) Sport
#4. operation (Computer Science)	#1 (2)	#1 (2) Program
#5. operation (Military)	none	none
#6. operation (Medicine)	#2 (3)	#2 (3) Science
#7. operation, procedure	#3 (6)	#3 (4) Sport
#8. process, operation, cognitive operation (Psychology)	#2 (4)	#2 (3) Science
#9. operation (Mathematics)	none	#1 (2) Program

Table 3. Comparison of WordNet SynSets and restructured SynSets for clustering for the word “operation”.

References

[Abdelali *et al.*, 2003] Abdelali, J. Cowie, D. Farwell, B. Ogden and S. Helmreich. Cross-Language Information Retrieval using Ontology In: *Proc. of the Conference TALN 2003,* France, 2003.

[Bade *et al.*, 2005] K. Bade, E. W. De Luca, A. Nürnberger and S. Stober. CARSA - An Architecture for the Development of Context Adaptive Retrieval Systems, In: *Proceedings of the 3rd International Workshop on Adaptive Multimedia Retrieval (AMR05).*

[De Luca and Nürnberger, 2005] E. W. De Luca and A. Nürnberger. A Meta Search Engine for User Adaptive Information Retrieval Interfaces for Desktop and Mobile Devices In: *Proc. of the Workshop on Personalized Information Access (PIA 2005),* In Conj. with the Int. Conference on User Modelling (UM'05), UK, 2005.

[De Luca and Nürnberger, 2006a] E. W. De Luca and A. Nürnberger. Rebuilding Lexical Resources for Information Retrieval using Sense Folder Detection and Merging Methods. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006).* Genova, Italy, 2006.

[De Luca and Nürnberger, 2006b] E. W. De Luca and A. Nürnberger. LexiRes: A Tool for Exploring and Restructuring EuroWordNet for Information Retrieval". In: *Proceedings of the Workshop on Text-based Information Retrieval (TIR-06).* In conjunction with the 17th European Conference on Artificial Intelligence (ECAI'06). Riva del Garda, Italy / Aug 29th, 2006.

[De Luca and Nürnberger, 2006c] E. W. De Luca and A. Nürnberger. Using Clustering Methods to Improve Ontology-Based Query Term Disambiguation. In: *International Journal of Intelligent Systems, Volume 21, 693-709, John Wiley & Sons*, 2006.

[De Luca *et al.*, 2006] E. W. De Luca, S. Hauke, A. Nurnberger and S. Schlechtweg. Using Multilingual Ontologies for Adaptive Web-based Language Exploration. In: *Proceedings of the International Workshop on Applications of Semantic Web Technologies for E-Learning (SW-EL06).* In Conjunction with the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2006). Dublin, Ireland, 2006.

[Dumais *et al.*, 2001] S. T. Dumais, E. Cutrell and H. Chen. Bringing order to the web: Optimizing search by showing results in context. In: *Proc. of the CHI'01*, 2001, 277-283.

[Gonzalo *et al.*, 1998] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran. Indexing with WordNet synsets can improve text retrieval. In: *Proceedings ACL/COLING Workshop on Usage of WordNet for Natural Language Processing*, 1998.

[Ide and Véronis, 1998] N. Ide and J. Véronis. Word Sense Disambiguation: The State of the Art. In: *Computational Linguistics*, Volume 14, Part 1, 1998.

[Morato *et al.*, 2004] J. Morato, M. Marzal, J. Lloréns and J. Moreiro. WordNet Applications. In: *Proc. of the 2nd Int. Conf. Global WordNet,* Brno, Czech Rep. 2004.

[Labrou and Finin 1999] Y. Labrou and T. Finin. Yahoo! as an ontology: using Yahoo! categories to describe documents. In: *Proc. of 8th Int. Conf. on Information and Knowledge Management,* 1999.

[Mihalcea and Moldovan, 2001] Rada Mihalcea and Dan Moldovan, Automatic Generation of a Coarse Grained WordNet. In: *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA, June 2001.

[Miller, 2001] G. A. Miller. Ambiguous Words. In: *Impacts Magazine*. Publ. on KurzweilAI.net, 2001.

[Miller *et al.*, 1993] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. Miller. Five papers on WordNet. ftp.cogsci.princeton.edu/pub/wordnet/5paper s.ps. 1993.

[Moldovan and Mihalcea, 2000] Dan Moldovan and Rada Mihalcea, Using WordNet and Lexical Operators to improve Internet Searches. In: *IEEE Internet Computing, vol. 4 no. 1*, January 2000.

[Oard, 1997] D. W. Oard. Alternative Approaches for Cross-Language Text Retrieval, College of Library and Information Services, University of Maryland, http://www.ee.umd.edu/medlab/filter/sss/papers/oard/paper.html, 1997.

[Oard, and Dorr 1996] D. W. Oard & B. J. Dorr, "A Survey of Multilingual Text Retrieval, UMIACS TR-96-19, University of Maryland, College Park, MD, 1996.

[Peters and Sheridan, 2000] C. Peters and P. Sheridan. Multilingual Information Access. In: *Lectures on Information Retrieval, Third European Summer-School,* ESSIR 2000, Varenna, Italy, 2000.

[Peters, 2001] W. Peters. Lexical Resources, In: *NLP group Department of Computer Science*, University of Sheffield, http://phobos.cs.unibuc.ro/roric/lexintroduction.html, 2001.

[Peters *et al.*, 1998] Peters, W., Peters, I., Vossen, P. Automatic sense clustering in EuroWordNet. In: *Proceedings of the 1st international conference on Language Resources and Evaluation*. Spain, 1998.

[Sinka and Corne, 2002] Sinka, M.P., Corne, D.W. A large benchmark dataset for web document clustering, in Abraham, A., Ruiz-del-Solar, J., Koeppen, M. (eds.), *Soft Computing Systems: Design, Management and Applications, Volume 87 of Frontiers in Artificial Intelligence and Applications* (2002) 881-890.

[Vossen, 1997] P. Vossen. EuroWordNet: a multilingual database for information retrieval. In: *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, Zurich, 1997.