

The Sense Folder Approach for Generic and Domain-Specific Retrieval Tasks

Ernesto William De Luca

Otto-von-Guericke University of Magdeburg
39106 Magdeburg
ernesto.deluca@ovgu.de

Frank Rügheimer

Biologie Systématique, Institut Pasteur
75015 Paris, France
frueghei@pasteur.fr

Abstract

In this work, we present and evaluate a new approach to semantic search. This approach is distinguished by pointing users to semantic concepts that offer possible refinements of their query. In parallel a combination of information retrieval and machine learning strategies is applied to annotate and filter documents with respect to those semantic categories. In an outlook on recent work, we describe how the approach can be applied and extended as an aid to find information relevant to specific biological questions in publication databases.

1 Introduction

Search engines, such as Google and Yahoo have become an essential tool for the majority of Web users for finding information in the huge amount of documents contained in the Web. Even though, for most ad-hoc search tasks [Baeza-Yates and Ribeiro-Neto, 1999], they already provide a satisfying performance, certain fundamental properties still leave room for improvement. For example, users get lost in navigating the huge amount of documents available on the Web and are obliged to scan the list of all retrieved documents, in order to find the relevant ones. This can partially be attributed to the possibly misleading statistics that leads to semantically inhomogeneous result sets. Basically, results are computed from word frequencies and link structures, but other factors, such as sponsored links and ranking algorithms, are also taken into account.

More recent approaches try to categorize documents automatically with clustering methods. For instance, Vivísimo [Koshman *et al.*, 2006] organizes search results into categories (hierarchical clusters), basing on textual similarity. These methods only consider the word distribution in documents without taking into account linguistic criteria derived from the underlying query, such as different meanings of a term. Therefore, the assigned categories usually do not represent the categories a user is expecting for the query at hand.

In general, the search process starts when a user provides a list of keywords and the system returns a list of documents ordered by the degree of similarity to the applied query. This means that if the keywords are well chosen, an appropriate list of results is frequently provided.

However, if the result list covers different meanings (if the search terms are ambiguous) or topics (if the search terms are used in different domains), then documents related to the corresponding categories appear rather unsorted in the result list.

Linguistic information (e.g. semantics) can provide valuable support for the user's search process. For instance, retrieved documents could be grouped by the meanings of the query. The user could choose one of these meanings and navigate only the documents related to it.

In addition, users search the Web and formulate their queries in their own language. But when they are unsuccessful, i.e. when their query does not match any results, they may also search and read documents in a foreign language [Peters and Sheridan, 2000].

In the light of the current lack of readily available tools that actively support researchers in finding desired information, given e.g. a gene name, we argue that semantic support would be a valuable addition to biological research tools. We therefore transfer the Sense Folder approach to biological data and set out to develop a search engine for biological publication databases.

1.1 Preliminaries

In order to better understand the semantic-based *Sense Folder approach* (see Section 2) presented in this paper, we first introduce some preliminary definitions (see Section 1.1) and related work (see Section 1.2). Then, the system architecture (see Section 2.1), semantic support methods (see Section 2.2) and the related user interface (see Section 2.5) are discussed. *Document classification* (see Section 2.3) and *clustering* (see Section 2.4) techniques used for filtering documents semantically are explained. Finally, the approach is compared and evaluated with respect to various baselines (see Section 3). The paper finishes with the discussion of ongoing research, where this semantic-based approach is applied for biological domain-specific tasks (see Section 4) and some concluding remarks (see Section 5).

Lexical resources, containing the different meanings of the words and the related linguistic relations, provide the semantic information needed for categorizing documents. For successful information retrieval it is crucial to represent documents in an adequate form with suitable attributes. In order to semantically compare documents, similarity measures have to be applied. The approach should be evaluated

using appropriate performance measures.

Lexical Resources

Lexical resources are a special type of language resources [Cole *et al.*, 1997] that provide linguistic information about words. Lexical resources are used in this work for supporting the user with semantic information during the search process, as discussed in the following.

WordNet [Miller *et al.*, 1990; Fellbaum, 1998] is one of the most important English lexical resources available and can be used for text analysis and many related areas [Morato *et al.*, 2004], like word sense identification, disambiguation, and information retrieval [Vintar *et al.*, 2003]. WordNet provides a list of word senses for each word, organized into synonym sets (SynSets), each carrying exactly one meaning. Different relations link the SynSets to two types of linguistic relations, the first type is represented by lexical relations (e.g. synonymy, antonymy and polysemy), and the second by semantic relations (e.g. hyponymy and meronymy). Glosses (human descriptions) are often (about 70% of the time) associated with a SynSet [Ciravegna *et al.*, 1994].

We decided to use Wordnet for retrieving the meanings related to the queries and the related linguistic relations.

Vector Space Model

The vector space model [Salton, 1971; Salton and Lesk, 1971] is the most frequently used statistical model for ad-hoc retrieval and represents a user query q and a document d_i as vectors in a multi-dimensional linear space. Each dimension corresponds to characteristics of a word in a document (word occurrence, word frequency, etc.). For instance, if word occurrence is used, each dimension takes boolean values, while the use of (weighted) relative term or word frequency, such as tf or $tf \times idf$, leads to a real-valued vector space $\mathcal{V} = [0, 1]^n$.

Thus, the document vectors can be represented with attributes of terms in a document, such as term frequency (tf) or inverse document frequency (idf) [Salton and Buckley, 1988].

Cosine Similarity

Similarity between documents is assumed to coincide with the similarity in the vector space measured for instance using *cosine similarity* [Manning and Schütze, 1999]. This is a measure of similarity between two vectors of n dimensions computed by finding the angle between them. The approach relies on the assumption that a relevant document d_i and the corresponding query q are linked by common terms. These common word occurrences are expected to be reflected by a document vector \vec{d}_i that is close to the query vector \vec{q} . That way, the task of finding relevant documents, given a query, is reduced to the identification of the document vectors that form the smallest angles with q . Therefore, the cosine similarity is defined as follows:

$$sim(d_i, q) = \frac{d_i \cdot q}{|\vec{d}_i| \times |\vec{q}|}. \quad (1)$$

Performance Measures

Generally, queries are usually less than perfect for two reasons: first of all, they retrieve some irrelevant documents and secondly, they do not retrieve all the relevant documents. In order to evaluate the effectiveness of a retrieval system, different measures can be used [Baeza-Yates and Ribeiro-Neto, 1999]. The measure chosen for the evaluation of the classification performance in this work is the

accuracy that is the proportion of the total number of correct predictions.

1.2 Related Work

In this section related work that evaluate and compare different parameter settings is presented.

Agirre and Rigau [Agirre and Rigau, 1996] analyze WordNet relations for WSD and evaluate different combinations for disambiguating words using a conceptual density algorithm. They show that some relations such as meronymy (has-part relation) do not improve the performance as expected. They also point out that in WordNet not all semantic relations are available for all words, which might result in significant classification problems, since one disambiguating class might be described more specific than another class.

Larsen and Aone [Larsen and Aone, 1999] propose the application of clustering methods with a vector space model and tf or $tf \times idf$ document representations to overcome the information overload problem. They conclude that weighting terms by $tf \times idf$ works better than weighting by tf , except for corpora with a very small number of documents. But the influence of linguistic relations of words for classification is not taken into account.

In recent work Patwardhan and Pedersen [Patwardhan and Pedersen, 2006] employ so-called *context vectors* for Information Retrieval, including only WordNet glosses, because these descriptions are considered to contain content rich terms that better allow to distinguish concepts than a generic corpus.

2 Semantic-based Search and Disambiguation: The Sense Folder Approach

The idea of this approach is to use lexical resources in order to disambiguate/filter documents retrieved from the Web, given the different meanings (e.g. retrieved from lexical resources) of a search term, and the languages the users are able to speak. For achieving this goal different approaches are combined. *Word Sense Disambiguation* approaches are used for recognizing the contexts of the words contained in the query. *Document Categorization* techniques are used for collecting similar documents in semantic groups. *Semantic* and *Multilingual Text Retrieval* methods are developed because of the need of supporting humans to filter and retrieve relevant documents from the huge amount of data available using semantics and multilingual knowledge [De Luca and Nürnberger, 2006c; 2006b]. This section summarizes the core of this paper, the *Sense Folder Approach*. After defining a *Sense Folder*, the system architecture and the related user interface are described. Then, the classification and clustering methods used are discussed in more detail.

Sense Folder Definition Given a query term q , a *Sense Folder* is a container (prototype vector) that includes all selected linguistic information (linguistic context) of one sense of the query term retrieved from lexical resources.

2.1 System Architecture

Figure 1 gives an overview of the Sense Folder system architecture (and the related disambiguation process). The process starts after the user submits a query through the user interface (see [De Luca and Nürnberger, 2006a] and

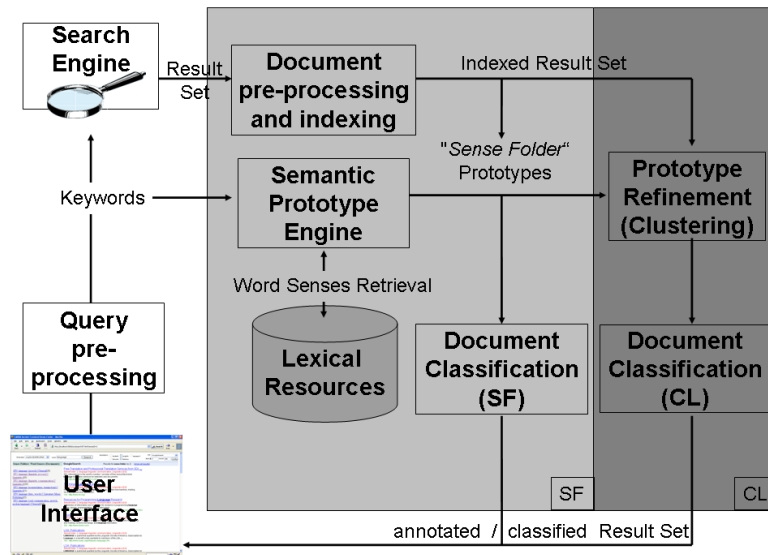


Figure 1: Overview of the Sense Folder Approach.

Section 2.5). For every word contained in the query a pre-processing step is applied (see Section 2.2).

After the query has been processed, the user keywords are simultaneously sent to the search engine and to the *Semantic Prototype Engine*. While documents are retrieved, pre-processed and indexed, for every search term the different meanings of a term and the related linguistic relations are retrieved from the lexical resource. Using these linguistic relations a query term can be expanded with words defining the context for each of its meanings, thus forming *Sense Folders*. Based on this information, semantic prototype vectors describing each semantic class are constructed.

Then, based on this information for each document (retrieved from the search engine) the similarity to the Sense Folder prototypes is computed and the semantic class with the highest similarity to the considered document is assigned. This first categorization method is called “*pure*” *Sense Folder* (SF) classification approach (see Section 2.3).

Afterwards, clustering algorithms (see Section 2.4) are applied in order to fine tune the initial prototype vectors of each Sense Folder using the distribution of documents around the initial prototype vectors, i.e., we expect that in a Web search usually a subset of documents for each possible meaning of a search term is retrieved. Thus, each subset forms a cluster in document space describing one semantic meaning of this term. This additional clustering step (CL) has been introduced in order to enhance the semantic-based classification (only based on lexical resources) by considering also similarities in-between documents.

In contrast to the approach presented in [Agirre and Rigau, 1996] that only used a small window of words around the considered term in order to disambiguate its meaning, the assumption of this paper is that the meaning of a search term used in a Web page can be defined based on the whole document, since Web pages are usually very short and usually cover only one semantic topic. The assumption that words have only one sense per document in a given collocation is proven by experiments presented in [Gale *et al.*, 1992; Yarowsky, 1993].

2.2 Semantic-based Support: the Query Pre-Processing

In the following section three approaches that can be applied to queries and documents are described, in order to support users in the semantic-based searching process. Specifically, the goal is to improve the semantic search process; therefore several problems have to be addressed, before the semantic classification of documents is started. When users mistype in writing the query, the system has to be able to give correction alternatives, recognizing the etymology of the query words (e.g. using stemming methods) or recognizing named-entities to continue the semantic-based search. The semantic-based search differs from the “normal” search, because users are “redirected” to semantic concepts that could describe their query. This “redirection” is provided on the left side of the user interface (see Figure 2), where suggestions are generated by the system as described in the following.

Spelling Correction An important task for retrieving the relevant documents related to the query is to identify the misspelled words and correct them for a correct interpretation. In this work, the use of a spell-checker (implemented in a joint work [Ahmed *et al.*, 2007]) supports the user during the search process, not only because it performs an efficient correction, but also because it can “redirect” the user to a semantic search. Thus, if the user types a word that is not contained in the lexical resource used, the system can suggest other “similar” words (concepts), according to the words found by the spell checker. Then, a semantic classification is started using the words selected by the user [De Luca and Nürnbergger, 2006c].

Stemming Because stemming methods are supposed to be suitable for reducing words to their base form, the Snowball stemmer [Porter, 2001] has been integrated. It includes a range of stemmers for different languages (e.g. the Porter stemmer for English, but also stemmers for French, German, Italian and Spanish). The aim of this approach is to improve performance merging similar variants of a word (sharing the same meaning) in one meaning. Stemming

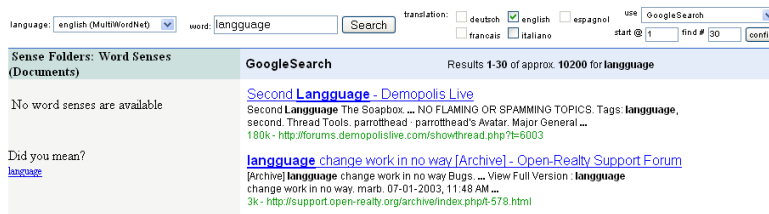


Figure 2: Semantic-based Support: the Query Pre-Processing.

methods do not analyze texts morphologically, but they try to reduce words in an etymological way to their base form; the stems are the result of such process. These stemmers can be used in order to help users in finding the base form of their keywords and “redirect” their search to the concept expressed in its base form. These base forms can be used by the users for the semantic-based search.

Named-Entity Recognition Because query words used for searching documents are not only common words, but represent also locations, organization, time expressions, and proper nouns, a named-entity recognizer (NER) has been added to the system, in order to support the user, if the search engine cannot disambiguate this kind of information. NER information is, in many cases, not ambiguous (e.g. a proper noun) and cannot be used for a semantic-based document search. The Stanford NER [Finkel *et al.*, 2005] can be used as a support for recognizing named-entities, directing the user to the semantic search. If the user types, for example, only the name “Java,” the NER should recognize the meaning of this instance and suggest more than one disambiguation possibilities (e.g. when “Java.” is related to the concept “island” or to the concept “programming language”).

2.3 Sense Folder Classification (SF)

As discussed above, the user query is simultaneously submitted to the search engine, which is providing a set of search results and to the *Semantic Prototype Engine* that retrieves the linguistic information used for creating the vectors describing the disambiguating semantic classes. Each element in the vector corresponds to a term i in the document, while the size of the vector is defined by the number of words n occurring in the considered document collection (dictionary). The weights of the elements depend on the tf or $tf \times idf$ [Salton and Buckley, 1988] and can be combined with stemming methods. Once the vector space description for each document is computed, the documents are classified by computing the similarity to each prototype vector describing the disambiguating semantic classes and assigning the class with the highest similarity to the considered document. The cosine similarity measure (see Section 1.1) is used in this case for the “pure” (semantic) Sense Folder classification.

2.4 Sense Folder Clustering Methods (CL)

After the document vectors are assigned to their respective WordNet class, clustering methods are used to tune/refine the classification results. Clustering methods in this work use a small number of labeled documents (Sense Folders) with a large pool of unlabeled documents. Three clustering algorithms have been implemented and evaluated. The first is an unsupervised method (k-Means clustering

| #Word Sense (Synonyms) [Domain] |
|--|
| #0 chair (professorship) [Pedagogy] |
| #1 chair [Furniture] |
| #2 chair (electric chair, death chair, hot seat) [Law] |
| #3 chair (president, chairman, chairwoman, chairperson) [Person] |

Table 1: WordNet noun collocation of the term “chair”

[Manning and Schütze, 1999]), while the last two are semi-supervised (Expectation-Maximization Clustering [Dempster *et al.*, 1977] and Density-Based Clustering [Friedman and Meulman, 2004]).

The k-Means clustering algorithm uses the number of classes obtained from WordNet for the number of clusters k as cluster centers. These classes are the so-called Sense Folders.

The Sense Folders are also used as “labeled data” for training the Expectation-Maximization clustering algorithm, i.e. in every cycle loop of the clustering process (in contrast to the k-Means clustering, where the Sense Folders are used only for the initialization) and the parameter λ is used as weight for the unlabeled/unclassified documents. If the parameter $\lambda=0$ the structure of unlabeled data is neglected and only the labeled data are considered for the estimation, while if the parameter $\lambda=1$ the information about the labeled class is neglected and only the structure is considered.

The initialization of the Density-based algorithm is the same of the k-Means algorithm. But the difference is due to the use of k neighbors and a parameter λ that is added. The use of such an algorithms is due to the assumption that data points that lie nearly together possess similar characteristics leads to an algorithm, where every data point is influenced by data points in its local neighborhood.

For a more detailed discussion about the clustering algorithms, the reader should refer to [De Luca, 2008].

2.5 Sense Folder User Interface

The semantic-based approach presented in this paper should simplify the search process by providing users with explicit information about ambiguities and this enables them to easily retrieve the subset of documents they are looking for. Labels defining the disambiguating classes are added to each document of the result set. This semantic information is assigned by the Sense Folder classification and clustering methods and appended as semantic annotation to the document as shown in Figure 3.

Thus, the visualization of such additional information gives the possibility to the user to filter the relevant query-related results by semantic class. For instance, if a user types, for example, the word “chair”, he/she has the possibility to obtain four different semantic classes based on the noun collocations of this word (included in WordNet) as shown in Table 1. These classes represent the different meanings of the search terms given by the user and are pro-

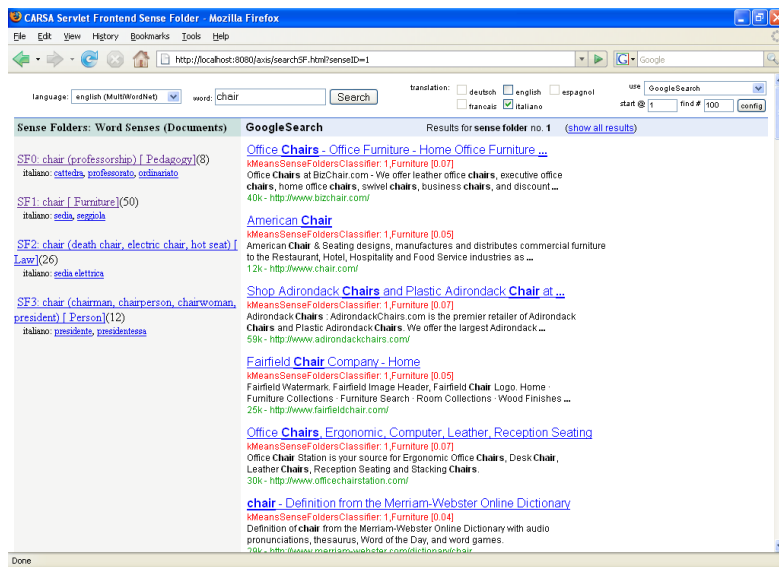


Figure 3: Semantic-based Search Engine.

vided on the left side of the interface (see Figure 3). If a user selects one of these meanings (e.g. the Sense Folder meaning SF1), the documents related to this concept are shown/filtered. Thus, users do not need to scan all documents, but they can browse only the documents related to the SF1 meaning of their query.

3 Evaluation

This section summarizes the results of a detailed fine-grained evaluation of different parameter settings. Various linguistic relations are combined with the different document attributes explained in Section 1.1. These are measured and compared to each other in order to recognize their best combination. Sense Folder classification and clustering methods are added and evaluated against three different baselines (*Random*, *First Sense* and *Most Frequent Sense*).

3.1 Baselines

According to WSD evaluation tasks, the approach has been evaluated in a fine-grained framework that considers the distinction of word senses on a more detailed level (all senses are included). This evaluation has been combined within three baselines described in the following:

- A *Random Baseline* (Random) assuming a uniform distribution of the word senses. This baseline provides a simple boundary for classification performance.
- A *First Sense Baseline* (FS), i.e. the score achieved by always predicting the first word sense, according to a given ranking, of a given word in a document collection. The First Sense baseline is often used for supervised WSD systems [McCarthy *et al.*, 2004]. This baseline is based, for this work, on the first word sense of WordNet contained in the Sense Folder Corpus [De Luca, 2008].
- A *Most Frequent Sense Baseline* (MFS) based on the highest a-posteriori word sense frequency, given a word in a document collection, i.e. the score of a *theoretically* best result, when consistently predicting the same word sense for a given word. This baseline is

based on the highest frequency of a word sense contained in a document collection. It is the best possible result score when consistently predicting one sense. It is often used as a baseline for evaluating WSD approaches and very difficult to outperform [McCarthy *et al.*, 2004].

3.2 Corpora Analysis

In order to evaluate this approach we analyzed different corpora to check if they were appropriate for the evaluation problem at hand. Many collections are already available in order to measure the effectiveness of information retrieval systems. Examples are given by the *Reuters Corpus* (RCV1), containing a large collection of high-quality news stories [Rose *et al.*, 2002], or the Reuters-21578 and Reuters-22173 data being the most widely test collection used for text categorization. Another collection is the *Text REtrieval Conference* (TREC) data collection, having the purpose to support information retrieval research by providing an infrastructure for large-scale evaluation of text retrieval methodologies. Because our purpose is not only evaluating an information retrieval system, but also a semantic information retrieval system, these data sets are unfortunately not appropriate for this task. They do not provide any semantic information based on a given query word, resulting that they are a document- and not query-oriented collection. No WordNet annotations are included and the “one sense per document” assumption is not fulfilled, because more topics can be covered in one document.

Since none of the available benchmark collections was appropriate for our task, we decided to use the multilingual *Sense Folder Corpus* created for evaluating semantic-based retrieval systems [De Luca, 2008]. This is a small bilingual (english and italian) hand-tagged corpus of 502 documents retrieved from Web searches created using Google queries. A single ambiguous word (“argument, bank, chair, network, rule” in english and “argomento, lingua, regola, rete, stampa” in italian) has been searched and related documents (approx. the first 60 documents for every keyword) have been retrieved. Every document contained in the collection has been annotated with only one WordNet domain

| Linguistic Relations | Document Encoding | Clustering |
|-------------------------------------|-------------------------------|-------------------------------|
| Synonyms (<i>Syn</i>) | | No Clustering (SF) |
| Domain (<i>Dom</i>) | Stemming (<i>Stem</i>) | K-Means Clustering (KM) |
| Hyponyms (<i>Hypo</i>) | No Stemming (<i>NoStem</i>) | Modified EM Clustering (EM) |
| Hyperonyms (<i>Hyper</i>) | | Density-Based Clustering (DB) |
| Coordinate Terms (<i>Coord</i>) | | |
| Domain Hierarchy (<i>DomH</i>) | | |
| Human descriptions (<i>Gloss</i>) | | |

Table 2: Components of Fine-Grained Evaluation Procedures

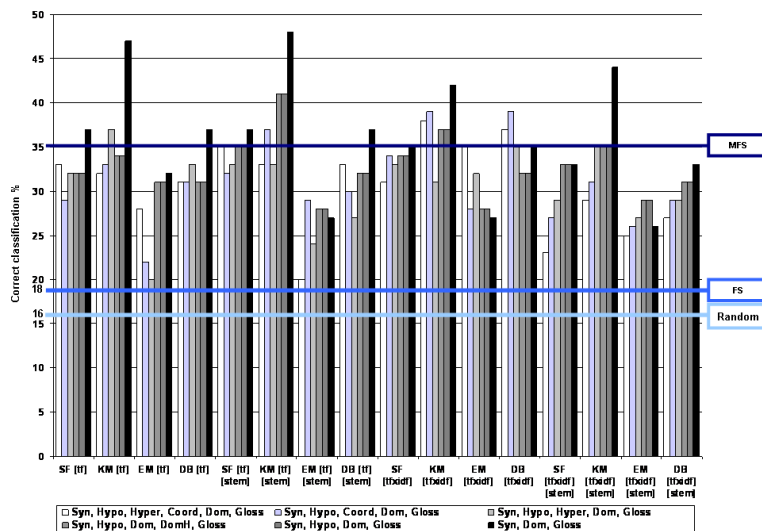


Figure 4: Sense Folder Fine-grained Accuracy Evaluation Results of different parameter settings with baselines.

label and one MultiWordNet query-dependent word sense label, respecting also the “one sense per document” assumption [Gale *et al.*, 1992; Yarowsky, 1993].

3.3 Parameter Settings

Table 2 outlines all components used for the fine-grained evaluation presented in this paper. Different combinations of linguistic relations (synonyms, coordinate terms, hyperonyms, hyponyms, glosses and semantic domains, and the semantic domain hierarchy) are taken into account, in order to assign documents to the semantic class they belong to.

Different encodings are considered in order to find an optimal representation. The *tf* and *tf × idf* encoding, as well as the *stemming* vs. *not stemming* term features, describe different vector spaces for the document classification. The resulting parameter settings are: *tf*-based (*Tf*) and respective stemmed one (*Tf+Stem*), *tf × idf*-based (*Tf×Idf*) and respective stemmed one (*Tf×Idf+Stem*).

Three Sense Folder clustering methods have been implemented. The k-Means clustering algorithm (KM) does not require any parameters, because the number *k* of cluster centers, corresponds to the number of word senses available in WordNet used as initial prototypes for the clustering process.

The DB Clustering method (DB) uses Sense Folders as initial prototypes. The parameter λ has been set to 0.9 and the *n* parameter that represents the number of neighbors to be considered, is set to 2.

The Expectation-Maximization(EM)- λ algorithm adopts the Sense Folders in the role of “labeled data,” whereas the vectors representing the documents supply the “unlabeled data”. The weight parameter λ for the unclassified data

points is set to 0.9.

The parameter settings for the last two clustering algorithms (EM- λ and DB) have been set according to the evaluation results presented in [Honza, 2005].

3.4 Fine-grained Accuracy Evaluation

Analyzing the results presented in Figure 4, we can notice that a slight overall improvement is shown when stemming methods are applied in conjunction with the *tf* measure. We can see that the “pure” Sense Folder classification in some cases is already sufficient for classifying documents in the correct meaning. But in most cases clustering methods improve classification considerably.

When the automatic classification is compared within the baselines, we can see that all combinations outperform all “Random” and “First Sense” baselines.

Analyzing the linguistic relations in more detail, we can notice that the use of hyperonyms or hyponyms negatively influence the classification performance. Normally, a hyperonym should be the broader term of a given word that generalize the related linguistic context. But these terms included in WordNet are at the end too general and make the disambiguation of word senses more difficult. As a rule, a hyponym should narrow down the distinct word senses describing the search word more specifically; but these WordNet terms are not significant enough to split them.

When such linguistic information is combined with clustering methods, in some cases, the classification performance is strongly enhanced, because similar documents are recognized. Sometimes this semantic information already contained in lexical resources is sufficient to recognize the linguistic context of a document given a query, so that clus-

| Notion | General IR | Biological Domain |
|----------------------------|------------|---|
| ambiguous description | word | gene name |
| referenced entity | word sense | gene/protein/protein acting in particular biological role |
| relations between entities | WordNet | Gene Ontology |

Table 3: Corresponding Notions for General Information Retrieval and Gene Names in Biological Domain

tering methods are not needed or their use negatively affects the classification.

The recognition of the correct word sense given a query for retrieving only relevant documents is fundamental. This is needed to better support users in the search process, showing only the filtered relevant documents.

Summarizing, the fine-grained classification works better with the tf -based document representation and stemming methods than with the $tf \times idf$ -based document representation. This is most likely to be attributed to the idf measure that cannot be estimated very good and be meaningful adopted, because the document collection is still relative small.

4 Semantic-based Support in Biology: The use of the GENE Ontology

Many of the difficulties encountered in web information retrieval are paralleled when querying biological databases. With the recent development in experimental techniques, finding the most relevant pieces of information in the growing pool of published biological knowledge becomes increasingly difficult. From this situation arises the need for suitable search engines that support biologists in submitting queries that are adapted to very particular information needs.

The products of many genes, for instance, are used in two or more often otherwise unrelated roles in the organism. Such multiple roles arise, for instance, from alternative splicing (single gene gives rise to several proteins) or due to the coded proteins possessing several active sites. In biological databases this is reflected by pairs of biological process annotations for the same element with neither term being registered as a specialization of the other in the Gene Ontology [Ashburner *et al.*, 2000]. A researcher, however, would often only be interested in material on one of these roles, so the possibility to additionally restrict queries is desirable.

A second problem arises from the fact that many genes were discovered and named independently before being identified as referring to identical or analogue entities. This leads to synonymous gene names. In practice such synonyms are treated by mapping known alternative gene names to a standard descriptor. In rare cases two or more different genes share at least one name though (gene homonyms). This situation results in an ambiguity comparable to the different biological functions discussed above, though this time the ambiguity also extends to the genetic level as well, as the genes in questions correspond to different locations in the genome.

In all those cases the Sense Folder approach allows to guide and refine searches allowing to focus on results from the desired context only (compare Section 2.5). Moreover, with the Gene Ontology, a structure that allows to expand the list of search terms is already available.

Finally we need to address the problem of associating potential (query results) with individual Sense Folders. Fortunately this task is completely analogous to the general

case so the approach described in Section 2.3 can be applied. The observed analogies are summarized in Table 3.

5 Concluding Remarks

In this paper, an approach for semantic-based search and support has been presented. It combines different techniques to provide the extended functionality of “directing” users to semantic concepts that help increase the specificity of their queries. On the application level, semantic support is provided via additional components that have been integrated into the user interface of search engines. We evaluated the performance of the presented semantic-based approach, and conclude that the accurate selection of linguistic relations, clustering methods and document encodings strongly influences the fine-grained classification results. All baselines are outperformed and best results are achieved when combining the linguistic relations (synonyms, domains and glosses) with the k-Means Clustering algorithm, representing the document vectors as tf -based vectors and applying stemming methods. An outlook about the employment of such an approach for biological tasks is discussed and the tasks in this domain-specific context found to be comparable to those of the more general information retrieval problem.

References

- [Agirre and Rigau, 1996] Eneko Agirre and German Rigau. Word sense disambiguation using conceptual density. In *Proceedings of COLING’96*, pages 16–22, 1996.
- [Ahmed *et al.*, 2007] Farag Ahmed, Ernesto William De Luca, and Andreas Nürnberger. MultiSpell: an N-Gram Based Language-Independent Spell Checker. In *Proceedings of Eighth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2007)*, Mexico City, Mexico, 2007.
- [Ashburner *et al.*, 2000] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000.
- [Baeza-Yates and Ribeiro-Neto, 1999] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, Addison-Wesley, New York, 1999.
- [Ciravegna *et al.*, 1994] Fabio Ciravegna, Bernardo Magnini, Emanuele Pianta, and Carlo Strapparava. A project for the construction of an italian lexical knowledge base in the framework of wordnet. Technical Report IRST 9406-15, IRST-ITC, 1994.
- [Cole *et al.*, 1997] R. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, and V. Zue. *Survey of the State of the Art in*

- Human Language Technology*. Center for Spoken Language Understanding CSLU, Carnegie Mellon University, Pittsburgh, PA, 1997.
- [De Luca and Nürnberger, 2006a] Ernesto De Luca and William De Luca and Andreas Nürnberger. A Word Sense-Oriented User Interface for Interactive Multilingual Text Retrieval. In *Proceedings of the Workshop Information Retrieval In conjunction with the LWA 2006, GI joint workshop event 'Learning, Knowledge and Adaptivity'*, Hildesheim, Germany, 2006.
- [De Luca and Nürnberger, 2006b] Ernesto De Luca and William De Luca and Andreas 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, 2006c] Ernesto De Luca and William De Luca and Andreas Nürnberger. The Use of Lexical Resources for Sense Folder Disambiguation. In *Workshop Lexical Semantic Resources (DGfS-06)*, Bielefeld, Germany, 2006.
- [De Luca, 2008] Ernesto William De Luca. *Semantic Support in Multilingual Text Retrieval*. Shaker Verlag, Aachen, Germany, 2008.
- [Dempster et al., 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*(39(1)):1–38, 1977.
- [Fellbaum, 1998] Christiane Fellbaum. *WordNet, an electronic lexical database*. MIT Press, 1998.
- [Finkel et al., 2005] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, Morristown, NJ, USA, 2005. Association for Computational Linguistics.
- [Friedman and Meulman, 2004] Jerome H. Friedman and Jacqueline J. Meulman. Clustering objects on subsets of attributes (with discussion). *Journal Of The Royal Statistical Society Series B*, 66(4):815–849, 2004.
- [Gale et al., 1992] William A. Gale, Kenneth W. Church, and David Yarowsky. One sense per discourse. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pages 233–237, 1992.
- [Honza, 2005] Frank Honza. Clustering mit a-priori anahmen über clusterspezifische attributwichtigkeiten. Master's thesis, University of Magdeburg, Germany, 2005.
- [Koshman et al., 2006] Sherry Koshman, Amanda Spink, and Bernard J. Jansen. Web searching on the vivisimo search engine. *J. Am. Soc. Inf. Sci. Technol.*, 57(14):1875–1887, 2006.
- [Larsen and Aone, 1999] Bjornar Larsen and Chinatsu Aone. Fast and effective text mining using linear-time document clustering. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 16–22, New York, NY, USA, 1999. ACM Press.
- [Manning and Schütze, 1999] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Boston, USA, 1999.
- [McCarthy et al., 2004] D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. Finding predominant word senses in untagged text. In *42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, 2004.
- [Miller et al., 1990] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Five papers on WordNet. *International Journal of Lexicology*, 3(4), 1990.
- [Morato et al., 2004] Jorge Morato, Miguel Angel Marzal, Juan Llorens, and José Moreiro. Wordnet applications. In Masaryk University, editor, *Proceedings of the 2nd Global Wordnet Conference 2004*, 2004.
- [Patwardhan and Pedersen, 2006] Siddharth Patwardhan and Ted Pedersen. Using WordNet Based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, Trento, Italy, 2006.
- [Peters and Sheridan, 2000] Carol Peters and Páraic Sheridan. Multilingual information access. In *Lectures on Information Retrieval, Third European Summer-School, ESSIR 2000, Varenna, Italy, 2000*.
- [Porter, 2001] M.F. Porter. Snowball: A language for stemming algorithms. Technical report, Open Source Initiative OSI, 2001.
- [Rose et al., 2002] T.G. Rose, M. Stevenson, and M. Whitehead. The reuters corpus volume 1 - from yesterday's news to tomorrow's language resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, 2002.
- [Salton and Buckley, 1988] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.
- [Salton and Lesk, 1971] G. Salton and M. E. Lesk. *Computer evaluation of indexing and text processing*, page 143180. Prentice-Hall, Inc. Englewood Cliffs, 1971.
- [Salton, 1971] G. Salton. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.
- [Vintar et al., 2003] S. Vintar, P. Buitelaar, and M. Volk. Semantic relations in concept-based cross-language medical information retrieval. In *Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining*, Croatia, 2003.
- [Yarowsky, 1993] David Yarowsky. One sense per collocation. In *Proceedings of ARPA Human Language Technology Workshop*, 1993.