A Corpus for Evaluating Semantic Multilingual Web Retrieval Systems: The Sense Folder Corpus

Ernesto William De Luca

Technical University of Berlin - DAI-Labor
Sekr. TEL 14 Ernst-Reuter-Platz 7, 10587 Berlin, Germany
Phone: +49 30 314 74074
Fax: +49 30 314 74003
ernesto.deluca@dai-lab.de

Abstract

In this paper, we present the multilingual *Sense Folder Corpus*. After the analysis of different corpora, we describe the requirements that have to be satisfied for evaluating semantic multilingual retrieval approaches. Justified by the unfulfilled requirements explained, we start creating a small bilingual hand-tagged corpus of 502 documents retrieved from Web searches. The documents contained in this collection have been created using Google queries. A single ambiguous word has been searched and related documents (approx. the first 60 documents for every keyword) have been retrieved. The document collection has been extended at the query word level, using single ambiguous words for English (argument, bank, chair, network and rule) and for Italian (argomento, lingua, regola, rete and stampa). The search and annotation process has been done both in a monolingual way for the English and the Italian language. 252 English and 250 Italian documents have been retrieved from Google and saved in their original rank. The performance of semantic multilingual retrieval systems has been evaluated using such a corpus with three baselines ("Random", "First Sense" and "Most Frequent Sense") that are formally presented and discussed. The fine-grained evaluation of the Sense Folder approach is discussed in details.

1. Introduction

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 used test collection 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. In this paper we first analyze different corpora including Senseval, Semcor, TREC data set and IR-Semcor and describe the requirements and limits of available corpora for evaluating semantic multilingual retrieval approaches.

2. Corpora Analysis

Before analyzing different kind of corpora used for Word Sense Disambiguation (WSD) or Information Retrieval (IR), some requirements should be discussed because of their importance for evaluating semantic multilingual retrieval approaches, in order to choose an appropriate corpus. In the following we summarize the requirements that have to be satisfied:

- *several languages* should be included to evaluate the multilingual property of our approach. Parallel Corpora could be helpful.

- *one dominant Domain annotation (Gliozzo and Strapparava, 2009)* describing the semantic domain of the document is required. A semantic domain is important in order to recognize the topic of a document.

- *one sense per Document annotation* describing the meaning of a given word sense of the document (Yarowsky, 1993). The annotation has to be document- and not only word-oriented. A corpus with annotation on the document level implicity supports the "one sense per document" assumption (Gale et al., 1992), where a given word in a document is supposed to have one prevalent meaning. we thus prefer such corpora.

- *Wordnet SynSet annotation* containing the SynsetID of a given word sense of the document. In order to use our approach based on lexical resources (at the moment on WordNet), the documents contained in a given corpus had to be annotated with WordNet SynSets. Thus, every document had to be at least annotated with one SynSet.

2.1. Word Sense Disambiguation Corpora

Senseval and Semcor are corpora that have been used for WSD evaluation and for learning classifiers (Strapparava et al., 2004). These two data sets are annotated at the sentence level with WordNet SynSets. But unfortunately, the word frequency of a term in a sentence is too low for having statistical significance. While Senseval is available for different languages, Semcor (Miller et al., 1993) covers only the English language. In order to see if this kind of data could be adapted for the evaluation purpose, we used these data, as an experimental setting for network induction (De Luca and Rügheimer, 2007). These data had to be pre-processed in order to extract the most relevant attributes. Different information could have been taken into account.

For these experiments the (Multi)SemCor data set (a bilingual version of Semcor (Bentivogli et al., 2005)) and the Brown2 subset of Senseval have been used. While Senseval has only as sentence level annotation, (Multi)SemCor is subdivided into paragraph, sentence and token level.

Whereas this structure can be used to determine collocations, annotations are only given on the token level so that all of the SemCor attributes refer to this last level.

We combined syntactic (POS) and semantic information (semantic domains extracted from WordNet) for model construction. In our experiments, we found indication for the necessity to employ more than one information source, because the two classes of features did not strongly interact. Only in few cases the POS attribute could be used for Word Sense Disambiguation though the discriminative power appeared to be high, if prediction via POS was applicable at all. However, the size of the considered data set and the number of features available at the given time suggest further experiments with extended data.

For the evaluation in this work, these data are not suitable because they are word- and not document- or domain-oriented.

2.1.1. Information Retrieval Corpora

Gonzalo et al. (Gonzalo et al., 1999) adapted Semcor 1.5 in order to build a test collection called *IR-Semcor* used for evaluating their approach. They split the documents in Semcor 1.5 to get coherent chunks of text for retrieval and added WordNet annotations. They obtained 171 fragments with an average length of 1131 words per fragment to which they also added a summary (human explanation) with lengths varying between 4 and 50 words and an average of 22 words per summary. Each of the summaries was hand-tagged with WordNet 1.5 senses, resulting in 254 documents. Both queries and documents of this corpus were hand-tagged with phrases, POS, and WordNet senses. The authors show that WSD is more beneficial than artificially ambiguous pseudo-words suggested by Sanderson (Sanderson, 2000). POS tagging (even if manually annotated) does not help improve retrieval, and phrases used as indexing terms are not useful enough, if partial credit is not given to the phrase components.

The IR-Semcor collection has been annotated manually in order to show that indexing with WordNet SynSets can give significant improvements to text retrieval even for large queries. This indexing approach works better than the synonymy expansion in Voorhees (Voorhees, 1994), probably because it identifies not only the synonym terms, but it also differentiates word senses.

At the moment the IR-Semcor corpus is not publicly available, so that the use of this resource for comparison was not possible for this work and the related properties could not be analyzed. Again, this corpus does not contain any document- or domain-annotation.

In contrast to the corpora described above, this kind of data have document annotation and are largely used in the information retrieval community. Because our purpose is not only evaluating an information retrieval system, but also a semantic multilingual information retrieval system, these data are 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 (Yarowsky, 1993) is not fulfilled, because more topics can be covered in one document.

Table 1: Sense Folder Corpus Overview

Language	Query Word	Documents	Word Sense
EN	argument	48	5
	bank	47	10
	chair	59	4
	network	57	3
	rule	41	12
IT	argomento	51	5
	lingua	50	4
	regola	47	3
	rete	47	9
	stampa	55	2

3. Building the Sense Folder Corpus

Justified by the unfulfilled requirements explained above, we decided to create our own multilingual *Sense Folder Corpus*. We started creating a small bilingual hand-tagged corpus of 502 documents retrieved from Web searches. The documents contained in this collection had been created using Google queries. A single ambiguous word has been searched and related documents (approx. the first 60 documents for every keyword) have been retrieved. The first search and annotation process has been done in a monolingual way for the English language. 252 English documents have been retrieved from Google and saved in their original rank (done by the Google PageRank algorithm (Page et al., 1999)). In a second step we analyzed the content of every single retrieved document, according to the query word and manually annotated it with the best matching word sense and domain contained in WordNet. Thus, every document contained in the collection has been annotated with only one WordNet domain label and one MultiWordNet query-dependent word sense label, respecting also the "one sense per document" assumption.

Then, the same search and annotation process has been applied for the Italian language. The English document collection has been extended at the query word level. Like its English counterpart, the Italian query words were ambiguous. The results were 250 Italian documents that have also been retrieved from Google and saved in their original rank. The inventory of the ambiguous words contained in the corpus is listed in Table 1, where every query word is represented by its name, followed by the number of retrieved documents and the number of possible word senses (meanings) that it can be assigned to. The decision to use documents retrieved automatically from Google is motivated by the need of using real data.

4. Baselines

In order to evaluate the performance of semantic-based methods, we decided to use three baselines for comparing the hand-tagged semantic annotations contained in the Sense Folder Corpus with the automatic annotation done with the Sense Folder approach (De Luca, 2008). This approach is a semantic multilingual approach, where the Sense Folder Corpus has been used for evaluation.

Additionally, two word sense granularity levels have been considered. In some cases word senses belong to the same meaning and could be merged in one sense (De Luca and Nürnberger, 2006a). In this case a coarse-grained representation of the sense inventory is useful for making results more meaningful for evaluation. On the other hand, the sense distinction might not be detailed enough, so that a fine-grained representation is needed in order to distinguish the word senses on a more detailed level.

According to these distinctions also used in Word Sense Disambiguation Evaluation tasks, two different evaluations have been included. The fine-grained evaluation considers the distinction of word senses on a more detailed level (all senses are included). The coarse-grained evaluation treats a more general distinction of word senses as semantically related word senses are merged. These two evaluations have been combined within three baselines described in the following:

- A *Random Baseline* assuming a uniform distribution of the word senses.
- A *First Sense Baseline*, 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).
- A *Most Frequent Sense Baseline* 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.

To clarify the notation in this Section we will use the following abbreviations:

W a selection of query words.

S_w a set $\{s_1, s_2 \ldots s_n\}$ of word senses. For a given word $w \in W$ the elements $\{s_1, s_2 \ldots s_{n-1}\}$ represent the word senses in WordNet. we assume that the implicit ordering provided by the indices corresponds to the WordNet ranking such that s_1 is the first sense. An additional sense S_n denotes an unclassified word sense and is assigned the lowest priority in the ranking. In the evaluation this word sense is labeled with a "-1" value.

$D_{w,s}$ The subset of documents in D, in which the word $w \in W$ occurs in the sense $s \in S_w$.

Random Baseline The *Random Baseline* provides a simple boundary for classification performance. In eq. 1 the estimated relative frequency $\widehat{r}$ is calculated, with the assumption that for any given occurrence of a word w, all senses $s \in S_w$ are equally likely to be the correct ones. In such a situation a simple guessing would on average predict the correct sense for a fraction of

$$\widehat{r}_{\text{eq}}(w) = \frac{1}{|S_w|} \tag{1}$$

of the words total occurrences.

For example, given that the number of word senses S_w of the word "argument" are 6 (considering also the unclassified word sense), the expected proportion of correct guesses is $\frac{1}{6}$. This computation is applied in the same way for all words.

First Sense Baseline The *First Sense Baseline* is based on the first word sense of WordNet contained in the Sense Folder Corpus.

The senses in WordNet are ordered according to the frequency data contained in the manually tagged resource SemCor (Miller et al., 1993). In SemCor word senses are listed according to the order of their first occurrence. Senses that did not occur in the corpus are appended to that list in arbitrary order (McCarthy et al., 2004).

The term "First Sense heuristic" is chosen unfavorably, in the sense that it suggests the existence of an objective static ranking. But in reality the ordering of the senses is estimated from corpora assumed to reflect the overall distribution of word senses in language usage. That assumption is hardly justified, because corpora are build from different sources that strongly depend on the writing style of the authors and the intented recipients.

Additionally, sampling variance undermines the validity of the above assumption due to the limited size of available corpora. SemCor, on which the WordNet ranking is based, has a too small sample size to acquire statistically significant estimates for a general overall word sense distribution. Even in the ideal case of a close correspondence of the word sense distributions in the corpus and the considered language in general, the probability that the first sense corresponds to the most frequent one may be quite low. This is because even the most frequent word sense often represents only a small part of the overall population.

Consequently, one may obtain different rankings depending on the available resources. But an objective "First Sense" should only depend on the language and not on a specific corpus.

For the reasons described above, in this work the first sense is treated differently as the most frequent given in WordNet. In this work, different basic population is taken under consideration, because of the differences between the context in the lexical resources and the Sense Folder corpus.

In this case, the First Sense baseline has been computed as explained in Eq. 3. s_f is the first word sense ordered in WordNet that is used as Sense Folder annotation for a document D given a word w.

But unlike from the SemCor/WordNet First Sense based on (Miller et al., 1993), the first sense in this work is only taken, if the word sense describing the first subset of documents is not empty (as shown in Eq. 2).

$$f = \min_{i=1}^{|S_w|}\{i|D_{w,s_i} \neq \emptyset\} \tag{2}$$

$$\widehat{r}_{s_f}(w) = \frac{|D_{w,s_f}|}{|D_w|} \tag{3}$$

Table 2 shows the mapping between the Sense Folders for the term "argument" and the correspondent annotated documents. The First Sense baseline is computed by $\frac{|D_{w,s_f}|}{|D_w|}$.

Table 2: Sense Folders for the term "argument"

Sense Folder	0	1	2	3	4	-1
Documents	3	4	18	9	8	6

Since there are documents for every sense, the first sense equals the one in the WordNet ordering ("0"). Thus we obtain a value of $\frac{3}{48}$ corresponding to 3 out of 48 documents, where "argument" is used in the sense "0". This computation is applied for evaluation in the same way for all words in the Sense Folder Corpus.

Most Frequent Sense Baseline The *Most Frequent Sense 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 word sense disambiguation approaches and very difficult to outperform.

The high performance of this baseline is explained by the use of a-posteriori knowledge about the distribution of word senses in the document collection.

This baseline can be computed if hand-tagged data annotated with word senses are available and a predominant sense can be recognized between them (McCarthy et al., 2004).

In Eq. 4 we compute the Most Frequent Sense baseline.

$D_{w,s}$ denotes the set of documents annotated with a word sense s in S_w for a word w.

For the example given in Table 2 the most frequent sense occurrs in 18 documents of the test collection and is the third sense in the WordNet ordering. Thus, the value of the Most Frequent Sense baseline is $\frac{18}{48}$. Similarly one it can be computed for the remaining words.

Note that the Most Frequent Sense baseline is a theoretical baseline, since the most frequent sense is not known a priori. The inclusion of a-posteriori information also explains the high performance obtained with this baseline.

$$\widehat{r}_{s_{mf}}(w) = \max_{s \in S_w} \frac{|D_{w,s}|}{|D_w|} \tag{4}$$

Besides the baselines described above, we also computed the overall word- and language-dependent baseline. Eq. 5 shows this computation. In the case of English W is the selection of the query words $argument, bank, chair, network, rule$ and for Italian $argomento, lingua, regola, rete, stampa$.

The word-dependent baseline is computed for every w in W, where we multiply every $N(w)$ being the total number of documents containing a query word w by $\widehat{r}(w)$ that stands for any of the baselines considered in Eq. 1, 3 and 4. At the end we divide this product by the total number of the disambiguation processes. Because we disambiguate one word per document, this number coincides with the total number of documents. The language-dependent baseline is computed in the same way but considering all documents of the collection. The baselines acquired are:

$$\widehat{r}_{overall} = \frac{\sum_{w \in W} N(w) \cdot \widehat{r}(w)}{\sum_{w \in W} N(w)} \tag{5}$$

5. Fine-grained Evaluation

In order to evaluate Semantic Multilingual Web Retrieval Systems, we have to analyze different parameters like linguistic relations, document encodings, classification and clustering methods. In this paper we consider the linguistic relations: Synonyms (Syn), Coordinate Terms (Coord), Hyperonyms (Hyper), Hyponyms (Hypo), Glosses (Gloss) and Semantic Domains (Dom), in see their influence for the semantic classification of documents. In addition, the Semantic Domain Hierarchy (DomH) of MultiWordNet (Pianta et al., 2002) is taken into account. We analyze also different document encodings. This investigation provides a basis for selecting an optimal representation. The tf and $tf \times idf$ encoding (Manning and Schütze, 1999), as well as the *stemming* (stem) vs. *not stemming* (noStem) term features, describe different vector spaces for the document classification. Moreover, we extend the evaluation with the comparison of the different classification and clustering methods for determining the best performing one: the "pure" Sense Folder classification (SF) based on cosine similarity (De Luca and Nürnberger, 2006b), the k-Means clustering algorithm (KM) (Macqueen, 1967), the Expectation Maximization (EM) (Nigam et al., 2000) and the Density-Based (DB) (Friedman and Meulman, 2004) algorithms. These clustering methods are evaluated in combination with the linguistic relations to discover their most suitable configuration.

Table 3 shows the results of the overall evaluation (including all search words) of the Sense Folder approaches in a fine-grained setting. Results are presented as a number interval $[0, 1]$ that quantifies the proportion of correct classification results. The best result for every single combination is highlighted in bold.

Analyzing the different linguistic parameters, we can see that the use of hyperonyms or hyponyms negatively influenced the performance of the system. Normally, a hyperonym should be the broader term of a given word, so that the searched word can be recognized in its more general linguistic context. But these terms included in WordNet are listed in a hierarchical structure, so that when the intersection between two word senses is found, the terms found 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; these terms included in WordNet 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 clustering methods are not needed or their use affects negatively the classification.

Table 3: Fine-grained (Overall) Accuracy Evaluation Results of different parameter settings with baselines.

	Syn, Hypo, Hyper, Coord, Dom, Gloss	Syn, Hypo, Coord, Dom, Gloss	Syn, Hypo, Hyper, Dom, Gloss	Syn, Hypo, Dom, DomH Gloss	Syn, Hypo, Dom, Gloss	Syn, Dom, Gloss
SF[tf][noStem]	0.33	0.29	0.32	0.32	0.32	0.37
KM[tf][noStem]	0.32	0.33	0.37	0.34	0.34	**0.47**
EM[tf][noStem]	0.28	0.22	0.20	0.31	0.31	0.32
DB[tf][noStem]	0.31	0.31	0.33	0.31	0.31	0.37
SF[tf][stem]	0.35	0.32	0.33	0.35	0.35	0.37
KM[tf][stem]	0.33	0.37	0.33	0.41	0.41	**0.48**
EM[tf][stem]	0.20	0.29	0.24	0.28	0.28	0.27
DB[tf][stem]	0.33	0.30	0.27	0.32	0.32	0.37
SF[tfxidf][noStem]	0.31	0.34	0.33	0.34	0.34	0.35
KM[tfxidf][noStem]	0.38	0.39	0.31	0.37	0.37	**0.42**
EM[tfxidf][noStem]	0.35	0.28	0.32	0.28	0.28	0.27
DB[tfxidf][noStem]	0.37	0.39	0.35	0.32	0.32	0.35
SF[tfxidf][stem]	0.23	0.27	0.29	0.33	0.33	0.33
KM[tfxidf][stem]	0.29	0.31	0.35	0.35	0.35	**0.44**
EM[tfxidf][stem]	0.25	0.26	0.27	0.29	0.29	0.26
DB[tfxidf][stem]	0.27	0.29	0.29	0.31	0.31	0.33

When the automatic classification is compared within the baselines presented in Section 4., we can see that all combinations outperform all "Random" (0.16) and "First Sense" (0.18) baselines. Last, we can notice that the use of selected linguistic relations combined with clustering methods enhances the classification performance. Best results have been achieved combining synonyms, semantic domains and glosses (Syn, Dom, Gloss) with the k-Means clustering algorithm (KM). This combination outperformed also the "Most Frequent Sense" baseline (0.35).

The evaluation based on the Sense Folder Corpus can also be done for every single query word in a coarse-grained setting, depending on the purpose of the systems to be evaluated (De Luca, 2008).

6. Conclusions

In this paper, we presented the Sense Folder Corpus. This is at the moment bilingual corpus that is being extended with other languages and will be made available for evaluating semantic multilingual retrieval systems. The classification accuracy performance of every semantic multilingual system can be compared within three baselines ("Random", "First Sense" and "Most Frequent Sense") in a fine- and coarse-grained setting, as well as within the Sense Folder approches presented in (De Luca and Nürnberger, 2006b; De Luca, 2008).

7. References

Luisa Bentivogli, Emanuele Pianta, and Marcello Ranieri. 2005. MultiSemCor: an English Italian aligned corpus with a shared inventory of senses. In *Proceedings of the Meaning Workshop 2005*, Trento, Italy.

Ernesto William De Luca and Andreas Nürnberger. 2006a. 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.

Ernesto William De Luca and Andreas Nürnberger. 2006b. Using Clustering Methods to Improve Ontology-Based Query Term Disambiguation. *International Journal of Intelligent Systems*, 21:693–709.

Ernesto William De Luca and Frank Rügheimer. 2007. Discovering Linguistic Dependencies with Graphical Models. In *LWA 2007 Workshop Proceedings*, pages 119–125, Germany, September. Martin-Luther-University Halle-Wittenberg.

Ernesto William De Luca. 2008. *Semantic Support in Multilingual Text Retrieval*. Shaker Verlag, Aachen, Germany.

Jerome H. Friedman and Jacqueline J. Meulman. 2004. Clustering objects on subsets of attributes (with discussion). *Journal Of The Royal Statistical Society Series B*, 66(4):815–849.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the 4th DARPA Speech and Natrual Language Workshop*, pages 233–237.

Alfio Massimiliano Gliozzo and Carlo Strapparava. 2009. *Semantic Domains in Computational Linguistics*. Springer, Berlin.

Julio Gonzalo, Anselmo Peñas, and Felisa Verdejo. 1999. Lexical ambiguity and information retrieval revisited. In *Proceedings of the joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99)*, Maryland.

J. B. Macqueen. 1967. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*.

MIT Press, Boston, USA.

D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. 2004. Finding predominant word senses in untagged text. In *42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain.

George Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. 1993. A semantic concordance. In *Proceedings of DARPA Speech and Natural Language Workshop*.

Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2/3):103–134.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *First International Conference on Global WordNet*, Mysore, India.

T.G. Rose, M. Stevenson, and M. Whitehead. 2002. 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.

Mark Sanderson. 2000. Retrieving with good sense. *Information Retrieval*, 2(1):49–69.

Carlo Strapparava, Alfio Gliozzo, and Claudio Giuliano. 2004. Pattern Abstraction and Term Similarity for Word Sense Disambiguation:IRST at Senseval-3. In *Proceedings of SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain.

E. Voorhees. 1994. Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 61–69.

David Yarowsky. 1993. One sense per collocation. In *Proceedings of ARPA Human Language Technology Workshop*.