IRn in the CLEF Robust WSD Task 2008

Sergio Navarro, Fernando Llopis, Rafael Muñoz
Natural Language Processing and Information Systems Group
University of Alicante, Spain
snavarro,llopis,rafael@dlsi.ua.es

Abstract

This paper describes our participation in the Robust WSD Task within the CLEF 2008. The aim of this pilot task is exploring methods which can take profit of WSD information in order to improve the IR systems. In our approach we have used a passage based system jointly with a WordNet based expansion method for the collection documents and the queries using the two WSD systems runs provided by the organization. Furthermore we have experimented with two well known relevance feedback methods - LCA and PRF -, in order to figure out which is more suitable to take profit of the WSD query expansion based on Wordnet. Our best run has obtained a 4th place in the competition with a value of 0.4008 MAP. We conclude that LCA fits better than PRF to this task. And that our WSD expansion is useful for some query subsets. In future works we will study the features of the query subsets for which the performance of our system decreases.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.2 Information Storage H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Information Retrieval, PRF, LCA, WordNet, Automatic Query Expansion, Relevance Feedback, WSD

1 Introduction

The aim of the CLEF Robust WSD Task task is exploring the contribution of Word Sense Disambiguation (WSD) to monolingual and multilingual Information Retrieval, in order to find successful methods to take profit of WSD information which helps the systems to increase their levels of robustness.

We are researching in the IR area, and it is so common to find - specially in collections of image annotations - documents which use narrow texts. It has a direct impact over the textual retrieval. Indeed, the problem of mismatch between a concept in a query and in a document, when it is expressed with different terms than found in the collection, is aggravated in this type of collections with small sized documents. Despite the fact that relevance feedback is a good tool for improving the results, it often shows unpredictable behaviour. Which makes us look for

alternative and complementary methods. We believe that the solution pass through the use of external resources. Since that these narrow collections usually do not reflect as many relations between different related terms as in a standard collection - where usually there are more terms related that have at least a document where they are coocurring -.

Bearing in mind the efficiency of the system we have worked in a method which do not have a great cost in the retrieval phase for our system. Thus, we have used a simple strategy of term expansion for the collection documents and the queries, which is based on the WSD systems offered by the organization.

This paper is structured as follows: Firstly, it presents the main characteristics of the IR-n system focusing on the documents and query expansion strategy, and the relevance feedback strategies, then it moves on to explain the experiments we have made to evaluate the system, and finally it describes the results and conclusions.

2 The IR-n System

In our approach, we used IR-n - an information retrieval system based on passages -. Passage-based IR systems treats each document as a set of passages, with each passage defining a portion of text or contiguous block of text. Unlike document-based systems, these systems can consider the proximity of words with each other, that appear in a document in order to evaluate their relevance [4].

The IR-n passage-based system differs from other systems of the same category with regard to the method proposed for defining the passage - that is - using sentences as unit. Thus, passages are defined by a number of consecutive sentences in a document [4].

IR-n uses stemmer and stopword lists to determine which information in a document will be used for retrieval. For a list of stemmers and stopwords used by IR-n, see www.unine.ch/infor/clef.

IR-n uses several weighting models. Weighting models allow the quantification of the similarity between a text - a complete document or a passage in a document - and a query. Values are based on the terms that are shared by the text and query and on the discriminatory importance of each term.

2.1 Expansion based on WordNet (WN) using WSD

This method, is an attempt to manage the information provided by a WSD system in order to overcome the problem of the mismatch between a concept in a query and in a document, and the problems derived from the natural language ambiguity.

The system expands terms within the queries and the collection documents. To carry out the expansion, it first selects the most likely WN synset returned by the WSD system - in the event of a tie it selects all the synsets with the maximum probability -. And afterwards, it generates the term expansion using all synonyms belonging to the selected synset/s .

In the phase of selecting the synset of a term, optionally IR-n can use two WSD systems in order to limit the synset selection only to those synsets which have been ranked as the most likely by one of the two WSD, and that at least has been ranked at second place by the other WSD system.

Finally, IR-n uses a parameter which allow, to configure the weight assigned for the terms added to the query.

2.2 Relevance Feedback

Most IR systems use relevance feedback techniques [3]. These systems usually employ local feedback. The local feedback assumes that top-ranked documents are relevant. The added terms are, therefore, common terms from the top-ranked documents. Local feedback has become a widely used relevance feedback technique. Although, it can deter retrieval, in case most of the

top-ranked documents are not relevant, results in TREC an CLEF conferences show that is an effective technique [7].

In past works [5] we noticed that in spite of the improvements in the general results brought by the relevance feedback - we used PRF [6] relevance feedback strategy -, this process also adds wrong terms for the expansion in some of the cases. Therefore we decided to focus part of our efforts on finding an alternative strategy for the relevance feedback, Thus, we are comparing in this CLEF edition PRF with Local Context Analisy (LCA) [7], as alternate strategy.

In the selection of terms, PRF gives more importance to those terms which have a higher frequency in the top relevant documents than in the whole collection. An alternative query expansion method relies on the Local Context Analysis (LCA), based on the hypothesis that a common term from the top-ranked relevant documents will tend to co-occur with all query terms within the top-ranked documents. That is an attempt to avoid including terms from top-ranked, non-relevant documents in the expansion. Furthermore, in the case of polysemous words, this method will help to retrieve documents more related to the sense of the query, since it is logical to think that the user will use words from the domain associated with this sense to complete the query. Indeed we think that in this year participation it could bebetter to use a method based on the terms of the query as LCA, since that the expanded terms based on WN used in the query and in all the documents, could boost performance of this relevance feedback strategy, improving its ability for skipping non relevant documents.

The IR-n architecture allows us to use query expansion based on either the most relevant passages or the most relevant documents.

3 Training

IR-n is a parameterizable system, which means that it can be adapted in line with the concrete characteristics of the task at hand. The parameters for this configuration are the number of sentences that form a passage, the weighting model to use, the type of expansion, the number of documents/passages on which the expansion is based, the weight used for the WN based expanded terms, the average number of words per document and the WSD system used.

This section describes the training process that was carried out in order to obtain the best possible features for improving the performance of the system.

The collections and resources are described first, and the next section describes specific experiments.

3.1 Data Collection

The organization has provided topics and document collections from previous CLEF campaigns - from year 2001 to year 2006 - which were annotated by two different systems for word sense disambiguation (WSD) developed by a group of the University of Barcelona (UBC) [1] and a group of the National University of Singapore (NUS) [2]. The documents are in English, and the topics in both English and Spanish.

The Table 1 and Table 2 show us the queries and collections used for training and test phases.

Table 1: Training Query Set and Data Collections

CLEF Year	**Topics No.**	**Collections**
2001	41-90	L.A. Times 94
2002	91-140	L.A. Times 94
2004	201-250	Glasgow Herald 95

Where the Topics No. column, is the range of queries used from the CLEF Ad-hoc task competition of the indicated competition year - CLEF Year -.

Table 2: Test Query Set and Data Collections

CLEF Year	Topics No.	Collections
2003	141-200	L.A. Times 94 and Glasgow Herald 95
2005	251-300	L.A. Times 94 and Glasgow Herald 95
2006	301-350	L.A. Times 94 and Glasgow Herald 95

3.2 Experiments

The experiment phase aims to establish the optimum values for the configuration of the system for the collection.

Below is a description of the input parameters of the system:

- **The Passage size (ps)**: Number of sentences in a passage.
- **Weight Model (wm)**: We used DFR weighting model.
- **Relevance Feedback (relFB)**: Indicating which relevance feedback uses the system - PRF or LCA.
- **Relevance Feedback parameters**: If *exp* has value 1, this denotes we use relevance feedback based on passages. But, if *exp* has value 2, the relevance feedback is based on documents. Moreover, *num* denotes the number of passages or documents that the relevance feedback will use from the textual ranking and finally, *term* indicates the k terms extracted from the best ranked passages or documents from the original query.
- **WSD system used for the expansion of the Collection (WSDCOL)**: Indicate which WSD system has been used or if none has been used for the documents expansion.
- **WSD system used for the expansion of the Query (WSDQuery)**: Indicate which WSD system has been used or if none has been used for the query expansion.
- **Weight for the WN based Expanded Terms (wWN)**: Is the weight used for the expanded terms using WN within the query.

Our participation is limited to the English monolingual task. Thus, for the experiments we have worked with DFR as the weighting schema. We have taken this decision based on the good results obtained in previous works for this language [5].

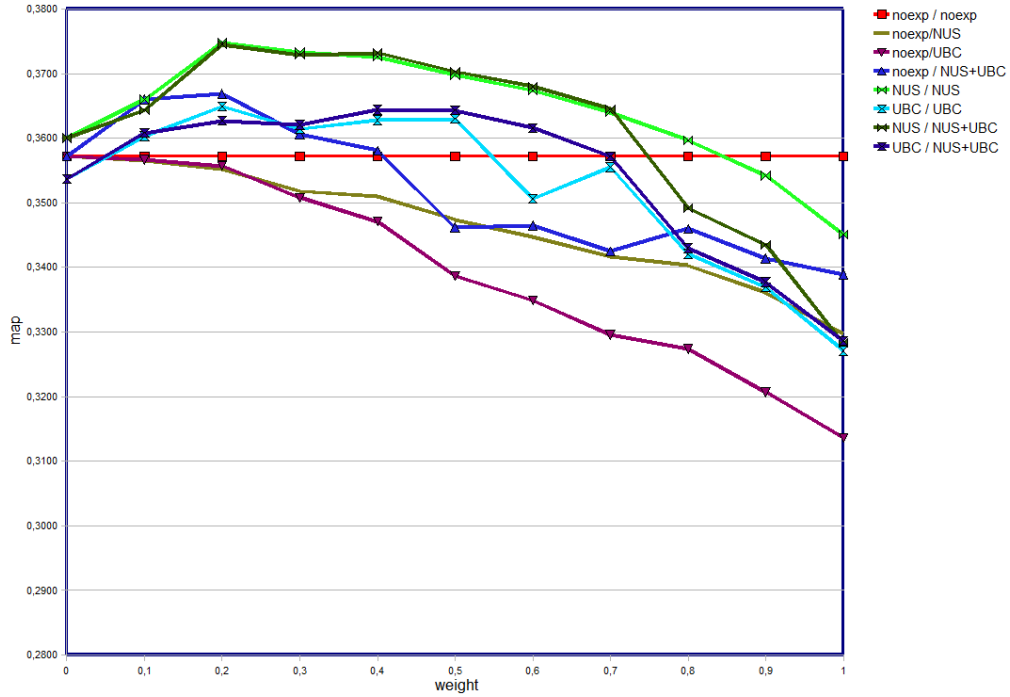
We started the experiments looking for which was the best configuration for the collection, in order to use it as baseline in our participation. Table 3 shows the best configurations obtained with our passages based system without using neither relevance feedback techniques nor query expansion based on WN.

Table 3: Best Baseline Runs

ps	c	avgld	map	recall
5	5.5	750	0.3556	0.8874
4	**5.5**	**750**	**0.3572**	**0.8865**
3	6	850	0.3498	0.8840
2	8.5	750	0.3332	0.8806
1	2	750	0.3133	0.8679

We have used the best run configuration - in MAP terms - , which uses a passage size of four sentences, as the base configuration for the training phase. The next experiments have added different combinations of values of relevance feedback and WN expansion parameters. In an attempt

Figure 1: MAP evolution with WN expansion



of having an overview of the effect of use the WN expansion we show in the Figure 1 a graphic with different combinations of WSDCOL and WSDQuery parameters - $WSDCOL/WSDQuery$ - with the best baseline run configuration, and the effect over the map measure of using a range of values between 0 and 1 for the weight of the WN expanded terms of the query.

We can see that the worst results - under the baseline results - are obtained for those runs which use only expansion for the query - not for the collection -. The best run is the one which uses NUS WSD system for expand the query and the collection. An finally we saw the method of mixing the two WSD does not improve the run which only uses NUS WSD system. Due to time restriction we do not have results mixing WSD systems for the expansion of the collection.

The next table - Table 4 - shows MAP and Recall values for the best runs for each combination of WSDCOL and WSDQuery parameters.

Table 4: Best Runs without Relevance Feedback

WSD **COL**	**WSD Query**	**wWN**	**map**	**recall**
no	no	0	0.3572	88.65
no	NUS	0.1	0.3565	89.67
no	UBC	0.1	0.3567	89.52
no	NUS+UBC	0.2	0.3669	92.02
NUS	**NUS**	**0.2**	**0.3748**	**92.38**
UBC	UBC	0.2	0.3649	91.31
NUS	NUS+UBC	0.2	0.3745	92.26
UBC	NUS+UBC	0.4	0.3644	90.60

In the Table 5, we show the best results obtained using LCA and PRF with each one of the

best runs of the Table 4.

Table 5: Best Relevance Feedback Runs

relFb	WSD COL	WSD Query	wWN	exp	num	k	map	recall
no	no	no	0	0	0	0	0.3572	0.8865
prf	no	no	0	2	5	15	0.3719	0.9040
lca	no	no	0	1	5	15	0.3833	0.9201
prf	NUS	no	0	2	5	5	0.3626	0.9381
lca	NUS	no	0	2	15	10	0.3845	0.9393
prf	NUS	NUS	0.2	2	10	10	0.3756	0.9417
lca	**NUS**	**NUS**	**0.2**	**2**	**15**	**10**	**0.3949**	**0.9417**
prf	UBC	no	0	2	10	10	0.3651	0.9250
lca	UBC	no	0	2	15	10	0.3668	0.9321
prf	UBC	UBC	0.2	2	10	10	0.3761	0.9262
lca	UBC	UBC	0.2	2	20	10	0.3803	0.9286

As we forecasted we can observe that the best MAP results are obtained using LCA. Indeed, the major improvement respect PRF occurs with the run witch use NUS WSD system.

4 Results in 2008 Robust WSD Task

The organization of the task only have allowed to send 4 submission using WSD runs and 4 without using WSD. Thus, we have sent to the task two runs without WSD: the baseline, and the best run which used only LCA. And for the 4 WSD runs, we have sent the best run without relevance feedback and the three best runs using relevance feedback.

Due to problems related with using an incomplete test query set - we submitted our runs out of time -. Thus, it has made that our results does not appear between the official task results. Table 6 and Table 7 show the results obtained in the tasks Monolingual without WSD and in the task with WSD respectively. The results are ordered by MAP. Also, we can see in this table our ranking position in MAP terms within the competition results.

Table 6: Results in 2008 Robust WSD Task - Without WSD runs -

runName	relFb	WSD WSD COL	WSD WSD Query	rk CLEF map	map	gmap	recall
TestIRnSinColLCA	lca	no	no	3	0.4008	0.1514	0.8851
TestIRnSinCol	no	no	no	10	0.36610	0.1473	0.8851

The best run submitted by the participants without using WSD in the competiton has obtained a value of 0.4515 of MAP.

The best run submitted by the participants using WSD has obtained a value of 0.4499 of MAP.

On the one hand, opposite to what happens in training phase, all the runs which have used WSD have obtained results for all the measures under the results of the run which have used LCA without WSD. On the other hand these results show us that LCA as in the training phase always improves the results respect the same configuration without its use.

Table 7: Results in 2008 Robust WSD Task - WSD runs -

runName	relFb	WSD WSD COL	WSD WSD Query	rk CLEF map	map	gmap	recall
TestIRnUBC_0.2_LCA	lca	UBC	UBC	13	0.3748	0.1361	0.8768
TestIRnNUSSoloCol_LCA	lca	NUS	no	14	0.3726	0.1384	0.8722
TestIRnNUS_0.2_LCA	lca	NUS	NUS	15	0.3720	0.1389	0.8761
TestIRnNUS_0.2	no	NUS	NUS	16	0.3664	0.1471	0.8669

5 Conclusion and Future Work

We conclude from those results that in spite our WSD approach has showed good results with the training set, we have doubts about its suitability in general for all kind of queries. Since that we have obtained contradictory results in the competition. In future works we will try to research the causes of its behaviour with the competition query set, analysing the possible error sources - the method itself, the wordnet organization or errors in disambiguation by the WSD systems - and its relation with the features of the queries.

6 Acknowledgement

This research has been partially funded by the Spanish Government within the framework of the TEXT-MESS (TIN-2006-15265-C06-01) project and by European Union (EU) within the framework of the QALL-ME project (FP6-IST-033860).

References

[1] Eneko Agirre and Oier Lopez de Lacalle. Combining k-nn with svd for wsd. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007)*, pages 341–345, Prague, Czech Republic, 2007.

[2] Yee Seng Chan, Ng Hwee Tou, and Zhong Zhi. Exploiting parallel texts for word sense disambiguation in the english all-words tasks. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007)*, pages 253–256, Prague, Czech Republic, 2007.

[3] Aitao Chen and Fredric C. Gey. Combining Query Translation and Document Translation in Cross-Language Retrieval. In Carol Peters, Julio Gonzalo, Martin Braschler, and et al., editors, *4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Lecture notes in Computer Science*, Lecture notes in Computer Science, Trondheim, Norway, 2003. Springer-Verlag.

[4] Fernando Llopis. *IR-n: Un Sistema de Recuperacin de Informacin Basado en Pasajes*. PhD thesis, University of Alicante, 2003.

[5] Sergio Navarro, Fernando Llopis, Rafael Muñoz, and Elisa Noguera. Information Retrieval of Visual Descriptions with IR-n System based on Passages. In *In on-line Working Notes, CLEF 2007*, 2007.

[6] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[7] Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, 2000.