UFRGS@CLEF2009: Retrieval by Numbers

Thyago Bohrer Borges, Viviane P. Moreira
Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil
[tbborges, viviane]@inf.ufrgs.br

Abstract

For UFRGS's participation on CLEF's Robust task, our aim was to compare retrieval of plain documents to retrieval using information on word senses. The experimental run which used word-sense disambiguation (WSD) consisted in indexing the synset codes of the senses which had scores higher than a predefined threshold. The documents in both baseline and WSD runs were indexed by Zettair. The metric for comparing queries and documents was OkapiBM25. The results of the experiments show that only 47 topics were helped by the strategy, while 103 had their performances worsened. A statistical t-test has shown that the experimental run which did not use WSD information significantly outperformed the one which did. A deeper analysis of our results and a set of further experiments are now under preparation.

Categories and Subject Descriptors

H.3.1 [Content Analysis and Indexing]: Linguistic processing. H.3.4 [Systems and Software]: Performance evaluation

Free Keywords

Experimentation, performance measurement

1 Introduction

This paper reports on experiments submitted to CLEF 2009 Robust track. The aim of the task is to assess the validity of using word-sense disambiguated data for Information Retrieval.

The goal of our experiment is to perform query expansion using WordNet senses that were assigned the highest scores for each word form in the texts.

The remainder of this paper is organised as follows: Section two describes our experimental runs and presents the results. Section 3 discusses future experiments which we plan to carry out. Section 4 presents the conclusions.

2 Experiments

2.1 Description of Runs and Resources

We worked on the English news collections composed by LA Times 94 and Glasgow Herald 95. There are 169,477 documents in total. Three versions of the collection were available: a "plain" version, and two versions with word-sense disambiguation (WSD) data.

Using the WSD documents (UBC version), we created a document collection composed by the synset codes of all WordNet senses which exceeded an arbitrary threshold (set to 0.30). WordNet is an lexical base, in which nouns, verbs, adjectives and adverbs are grouped in sets called "synsets". Figure 1 shows an example of an input word found in a document and the result of the processing that extracts the synset codes. If a term did not have a synset code, or a sense scoring higher than the threshold, we kept the original word form (i.e. the contents of the <WF> tag.

Input	Output
<TERM ID="C041-27" LEMA="report" POS="VBP"> <WF>report</WF> <SYNSET SCORE="**0.393362015980332**" CODE="**00655029**-v"/> <SYNSET SCORE="0" CODE="00653609-v"/> <SYNSET SCORE="0" CODE="00653917-v"/> <SYNSET SCORE="0" CODE="00655324-v"/> <SYNSET SCORE="**0.606637984019668**" CODE="**00653371**-v"/> <SYNSET SCORE="0" CODE="00653772-v"/> </TERM>	00655029 00653371

Figure 1 – Original term with WSD information and the output of pre-processing

The same approach was used in the documents was applied for building the queries from the topics. The queries we built automatically, using the title and description fields.

```
<top>
<num>10.2452/141-AH</num>
<EN-title>
04968965 02310834 02311368 Kiesbauer </EN-title>
<EN-desc>
01456625 00483900 01538749 00488684 01124979 04480483 00242644 05448780  05115901 04968965
02310834 02311368 03433996 03482557 04745188 02486167 04733874 PRO7 07222682 Arabella
Kiesbauer. </EN-desc>
</top>
```

Figure 2 – Example of query topic

The IR system we used was Zettair (Zettair), which is a compact and fast search engine developed by RMIT University (Australia) distributed under a BSD-style license. Zettair implements a series of IR metrics for comparing queries and documents. We used Okapi BM25 as some preliminary tests we performed on other data collections showed it achieved the best results.

We have submitted one baseline runs indexing the plain collection and one run using the WSD-annotated documents. There was a bug in the code that generated our WSD run, so we also report on a third (unofficial) run (WSD2) which has the correct data. The details of the runs are shown in Table 1.

Table 1 - Details of the test collections for the monolingual runs

RunID	Total number of terms	Number of distinct terms	Average number of terms per document
baseline	595,025	88,797,697	523
WSD1	518,993	91,642,665	553
WSD2	497,659	91,719,598	553

The table shows that the number of the total terms in the WSD run was smaller than in the baseline run. However, the opposite has happened with the number of distinct terms. The average number of terms per documents was higher on the WSD run as in many cases, more than one sense was kept for a term.

2.2 Results

Our results are summarised in Table 2 and Figure 3. The baseline run clearly outperformed the WSD run. A t-test using the average precision of the 160 queries has yielded a *p*-value of 0.0045, showing that the baseline was significantly better than the WSD run. The Recall-Precision curves on Figure 3 also show that the baseline was better in all recall levels. The superiority of the baseline is also reflected on the number of relevant documents retrieved and on precision at different cut-off points.

Table 2 –Summary of the Results

RunID	MAP	Relevant Retrieved	Precision at 10
baseline	0.3160	3290	0.3582
WSD2	0.2547	2870	0.2902

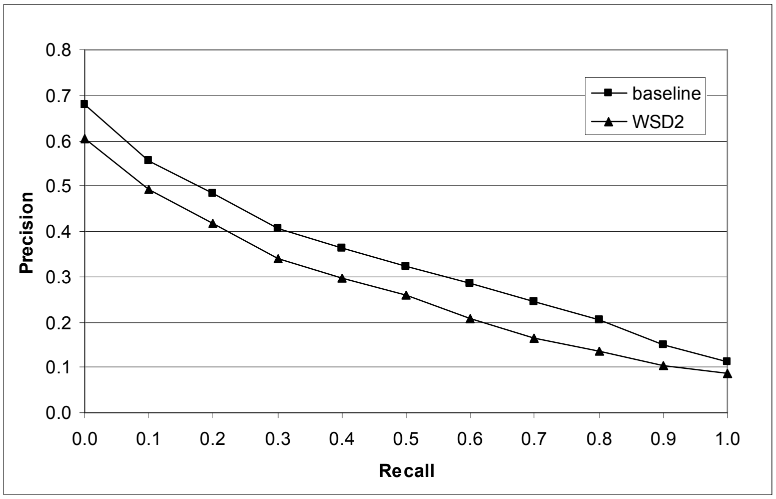

Figure 3 – Recall-Precision curves for the baseline and WSD run

A topic-by-topic analysis has shown that ten queries had the same average precision in both runs, 47 improved with WSD information, and 103 were better in the baseline run. Table 3 shows the top ten topics which were helped by the addition of WSD information and Table 4 shows the ten topics that were most harmed. A more detailed topic-by-topic analysis will be performed so that we can identify common patterns in the topics which had their performances improved and the ones which had their results worsened by the addition of WSD information.

Table 3 – Ten topics with the biggest increase in MAP with the addition of WSD information

Topics	Baseline	WSD2	Diff
10.2452/171-AH	0.0677	1.0000	0.9323
10.2452/177-AH	0.1112	0.9118	0.8006
10.2452/198-AH	0.2500	1.0000	0.7500
10.2452/190-AH	0.3101	0.9821	0.6720
10.2452/182-AH	0.0447	0.5913	0.5466
10.2452/306-AH	0.5000	1.0000	0.5000
10.2452/265-AH	0.0954	0.5797	0.4843
10.2452/153-AH	0.0000	0.4494	0.4494
10.2452/164-AH	0.0406	0.4221	0.3815
10.2452/183-AH	0.0406	0.3970	0.3564

Table 4 – Ten topics with the biggest decrease in MAP with the addition of WSD information

Topics	Baseline	WSD_WF	Diff
10.2452/162-AH	1.0000	0.0333	0.9667
10.2452/173-AH	1.0000	0.0714	0.9286
10.2452/180-AH	0.9240	0.0013	0.9227
10.2452/170-AH	1.0000	0.1687	0.8333
10.2452/181-AH	0.7607	0.0948	0.6659
10.2452/294-AH	0.5715	0.0560	0.5155
10.2452/340-AH	0.6393	0.1345	0.5048
10.2452/184-AH	0.5052	0.0410	0.4642
10.2452/143-AH	0.6160	0.1921	0.4239
10.2452/180-AH	0.6791	0.2572	0.4219

3 Further Experiments

The experiments reported here were a starting point and we plan to investigate some aspects further. First, we only worked with the UBC data. It would be interesting also to do experiments with the NUS collection to enable some comparisons.

We arbitrarily chose a threshold of 0.30 for the synset codes to be maintained. The idea is to try different thresholds and assess how they impact the results.

We also plan to investigate different strategies for query expansion using synonyms and related terms extracted from WordNet.

4 Conclusions

This paper described the experiments performed by our group for CLEF 2009 Ad hoc Robust task. We compared an experimental run in which we indexed the plain documents with an experimental run in which we took WSD information into consideration. The results have shown that the baseline (plain) run has outperformed the WSD run.

We plan to do further experiments as there are many issues which are worthy of a more detailed investigation.

Acknowledgements

This work was partially supported by CNPq.

References

WordNet. Retrieved 01/03/09, 2009, from http://wordnet.princeton.edu/

Zettair. Retrieved 11/06/07, 2007, from http://www.seg.rmit.edu.au/zettair/