QA with a Disambiguated Document Collection

Davide Buscaldi and Paolo Rosso

Natural Language Engineering Lab,

Dpto. de Sistemas Informáticos y Computación (DSIC),

Universidad Politécnica de Valencia, Spain

{dbuscaldi, prosso}@dsic.upv.es

Abstract

This report describes our approach to the Question Answering - Word Sense Disambiguation task. In our approach, disambiguated documents are used to improve the retrieval phase: this has been implemented by adding a WordNet expanded index to the document collection. This index contains synonyms, hypernyms and holonyms of the document words. Question words are searched for in both the expanded WordNet index and the default index. The obtained results do not show any improvement over the system that do not use the disambiguated collection. However, an analysis of the results shows that the average number of passages that contains the answer for each question is too small to detect any significative difference between the two systems.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; I.2 [**Artificial Intelligence**]: I.2.7 Natural Language Processing

General Terms

Measurement, Performance, Experimentation, Text Analysis

Keywords

Question Answering, Word Sense Disambiguation

1 Introduction

One of the most discussed issues in the field of Natural Language Processing is the usefulness of Word Sense Disambiguation (WSD), specifically if it can be used to improve the results in Information Retrieval or not. The QA-WSD task at CLEF[1] 2008 presented a great opportunity since it gave participants two disambiguated collections to perform tests on.

There were many options for the participation: for instance, use disambiguated questions or not, use the collection disambiguated using the method by the University of Basque Country (UBC) [1] or the method by the National University of Singapore (NUS) [4]. However, the rules allowed to submit only two runs: one without the use of semantic information and one with the use of semantic information. We decided to use standard (un-disambiguated) questions and the NUS disambiguated collection for the semantic-enabled run.

---

[1]http://www.clef-campaign.org

Our system is constituted by a modified version the QUASAR system described in [2], whose search engine (JIRS) has been replaced by Lucene[2]. In the configuration that uses semantics, the search index contains also terms that have been extracted from the hypernyms, holonyms, and synonyms of the document words by means of WordNet [3].

In the following section, we describe the system. In section 3 we describe the characteristics of our submissions and discuss the obtained results.

2 Our QA-WSD System

The system has some limitations due to the fact that it was developed for a past edition with different guidelines: it does not include an anaphora resolution system and it cannot answer list questions. We refer the reader to the description in [2] for a detailed description of the default system. In the following we will describe the WSD-based system only.

Previously to the indexing phase, all documents are split into sentences. These are used later to form the passages. In the indexing phase, we create two indices: the first one (*text*) contains all the terms of the sentence; the second one (expanded index, or *wn* index) contains all the synonyms of the disambiguated words; in the case of nouns and verbs, it contains also their hypernyms. For nouns, the holonyms (if available) are also added to the index. For instance, let us consider the following sentence from document `GH951115-000080-03`:

> Splitting the left from the Labour Party would weaken the battle for progressive policies inside the Labour Party.

The underlined words are those that have been disambiguated in the collection. For these words we can found their synonyms and related concepts in WordNet, as listed in Table 1.

Table 1: Expansion of terms of the example sentence. NA : not available (the relationship is not defined for the Part-Of-Speech of the related word).

lemma	ass. sense	synonyms	hypernyms	holonyms
split	4	separate part	move	NA
left	1	–	position place	–
Labour Party	2	labor party	political party party	–
weaken	1	–	change alter	NA
battle	1	conflict fight engagement	military action action	war warfare
progressive	2	reformist	NA	NA
policy	2	–	argumentation logical argument line of reasoning line	–

Therefore, the *wn* index will contain the following terms: *separate, part, move, position, place, labor party, political party, party, change, alter, conflict, fight, engagement, war, warfare, military action, action, reformist, argumentation, logical argument, line of reasoning, line.*

During the search phase, the *text* and *wn* indices are both searched for question terms. The top 20 sentences are returned for each question. The passages are built from these sentences,

---

by appending them the previous and next sentences in the collection. For instance, if the above example were a retrieved sentence, the resulting passage would be composed by the following sentences:

- `GH951115-000080-2` : "The real question is how these policies are best defeated and how the great mass of Labour voters can be won to see the need for a socialist alternative."

- `GH951115-000080-3` : "Splitting the left from the Labour Party would weaken the battle for progressive policies inside the Labour Party."

- `GH951115-000080-4` : "It would also make it easier for Tony Blair to cut the crucial links that remain with the trade-union movement."

Figure 1 shows the first 5 sentences returned for the question "What is the political party of Tony Blair?" using only the *text* index; in Figure 2 we show the first 5 sentences returned using the *wn* index.

The Labour Party , under Tony Blair , is poised to achieve political power for the first time in 16 years	GH950821-000120-5
No Headline Present Political peace : A truce is agreed on the Westminster front as party leaders John Major , Paddy Ashdown , and Tony Blair celebrate VJ-Day	GH950821-000164-0
Blair puts the family centre stage LABOUR leader Tony Blair , in a further move to occupy the political centre ground , yesterday staked his claim for Labour to be the ' ' party of the family	GH950330-000184-0
Blair should beware I AM mystified by your political editor ' s comment about Tony Blair ' ' dispatching the Bennite Marxist left ' ' and rescuing the Labour Party from the ' ' false historical perspective ' ' of a ' ' Marxist intellectual analysis ' ' (March 23	GH950327-000057-0
Blair wrestles with age-old links TRADE union leaders of all political persuasions have confronted Labour leader Tony Blair with a demand that the unions retain a 50 % voice in the party they created	GH951002-000214-1

Figure 1: Top 5 sentences retrieved with the standard Lucene search engine.

The Labour Party , under Tony Blair , is poised to achieve political power for the first time in 16 years	GH950821-000120-5
The Labour Party has been set a simple test by Tony Blair	GH950310-000026-16
No Headline Present Political peace : A truce is agreed on the Westminster front as party leaders John Major , Paddy Ashdown , and Tony Blair celebrate VJ-Day	GH950821-000164-0
' The investigation is understood to have the full support of Labour Party leader Tony Blair	GH950227-000161-5
On the eve of the Labour party conference , there were also sharp words for Tony Blair	GH951002-000228-9

Figure 2: Top 5 sentences retrieved with the WordNet extended index.

One noteworthy feature is that sentences retrieved with the expanded WordNet index are shorter, therefore allowing for more precise results.

3 Experiments

We submitted the two mandatory runs, one with the basic system (id: *nlel081enen*) that does not use semantic information, and one with the system described above (id: *nlel082enen*), using as collection the NUS-disambiguated collection. Of the 200 questions in the test set, only 49 had an answer in the disambiguated collection (the other questions had an answer in Wikipedia, which was not featured for the QA-WSD track), according to the organizers. However, we manually checked the data and found that it was possible to find an answer to 25 of the Wikipedia questions, bringing the number of questions with an answer in the collection to 74.

In Table 2 we show the results obtained by the two mandatory runs and another run that used the UBC-disambiguated collection (id: *nlel083enen*). The results of this last run are not official.

From the results we can say that the base system performed generally poor, although better than the system that included semantics. We calculated the number of question that could be

Table 2: Results obtained with the three runs over the 49 questions that had (officially) answer in the collection and all questions.

| run ID | 49 Questions | | | | All Questions | | | |
	R	X	U	Accuracy	R	X	U	Accuracy
nlel081enen	8	0	0	16.32	10	0	0	5.00
nlel082enen	7	0	0	14.29	8	0	1	4.00
nlel083enen	6	0	0	12.24	7	0	1	3.50

answered by our system, discarding the questions that do not have an answer in the collection and the questions whose answer type cannot be handled by our system, resulting in a maximum of 40 questions. Even with this reduced set of questions, the system did not perform well, obtaining a 25% accuracy over this set.

About the lower precision of the WSD based system, we carried out an analysis of the average number of passages that contained the answer for each of the questions. Of the 49 questions, only three answers were present in more than nine passages. The average number of passages containing the answer for each question in the remaining 46 questions is 2.04. This number justify both the bad overall performance of our system, which is based on redundancy, and the bad results obtained by the WSD based system: it could not find better passages simply because the relevant passages that could be retrieved were very few.

4 Conclusions

We presented a simple approach to take advantage of the disambiguated collection provided by the organizers of the QA-WSD track. It is based on an extended index that include synonyms, hypernyms and holonyms extracted by means of WordNet. However, the test set provided was not particularly fit to the task, with more than 75% of the questions not containing an answer in the collection. Moreover, the answers to questions that could be answered are contained in few passages, with the result that it cannot be demonstrated whether the use of semantic information proved useful or not.

Acknowledgements

References

[1] Eneko Agirre and Oier Lopez de Lacalle. UBC-ALM: Combnining k-NN with SVD for WSD. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007)*, pages 341–345, Prague, Czech republic.

[2] Davide Buscaldi, José Manuel Gómez, Paolo Rosso, and Emilio Sanchis. N-gram vs. keyword-based passage retrieval for question answering. In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval*, volume 4730 of *Lecture Notes in Computer Science*, pages 377–384. Springer, 2007.

[3] George. A. Miller. Wordnet: A lexical database for english. In *Communications of the ACM*, volume 38, pages 39–41, 1995.

[4] Yee Sen Chan, Hwee Tou Ng, and Zhi Zhong. US-PT: 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.