Evaluating a Conceptual Indexing Method by Utilizing WordNet

Mustapha Baziz, Mohand Boughanem, Nathalie Aussenac-Gilles

IRIT/SIG
Campus Univ. Toulouse III
118 Route de Narbonne
F-31062 Toulouse Cedex 4
Email *{ baziz, boughane, aussenac }@irit.fr*

ABSTRACT. This paper describes our participation to the English Girt Task of CLEF 2005 Campaign. A method for conceptual indexing based on WordNet is used. Both documents and queries are mapped onto WordNet. Identified concepts belonging to WordNet synsets are extracted from documents and queries and those having a single sense are expanded. All runs are carried out using a conceptual indexing approach. Results prove a primacy of using queries from the title field of the topics and a slight gain of using stemming compared to the non stemming cases.

ACM Categories and Subject Descriptors

H3.3 **[Information Storage And Retrieval]:** Information Search and Retrieval; H.3.1 **[Content Analysis and Indexing]** – *Search process, Retrieval models.*

General Terms

Algorithms, Experimentation.

Keywords

Conceptual Indexing, WordNet, Documents and Query Expansion.

1. Introduction

The objective of our participation to the English GIRT task in 2005, was to evaluate the use of a conceptual indexing method based on the WordNet [3] lexical database. The technique consists in detecting mono and multiword WordNet concepts from both documents and queries and then in using them as a conceptual indexing space. Terms not recognized in WordNet (less than 8%) are also added to complete the representation. Even though they are not useful at the expansion stage, they are used to compare documents and queries at the searching stage.

This paper is organized as follows. In section 2, we describe the synoptic scheme of our system which includes the Mercure search engine . In section 3, the tests required for conceptual indexing are formally described: the concept detection and weighting methods in 3.1, and the disambiguation-expansion method in 3.2. Section 4 reports the official evaluation results compared with the median average obtained by all participating systems. Finally, section 5 gives some conclusions and prospects.

2. Overview of the Approach

In this section, we describe the conceptual indexing method based on WordNet. The principle involves, being given a document (resp. a query), mapping it onto WordNet and then to extract the concepts (mono and multi terms) that belong to WordNet and appear in the text of the document (resp. the query) [1]. The extracted concepts are then weighed and marked using part of speech information (POS) to facilitate their expansion. The *expansion* which we call *Short Expansion* (or SE) amounts to expanding from the document[1] mono sense WordNet terms (having only one sense) by using all of their synonyms extracted from the synset[2] they belong to, and only one of their hypernym concepts (belonging to their hypernym synset). The indexing method may or may not use expansion and stemming [5] (according to the run). It includes classical keywords indexing by adding the terms that do not belong to WordNet dictionary.

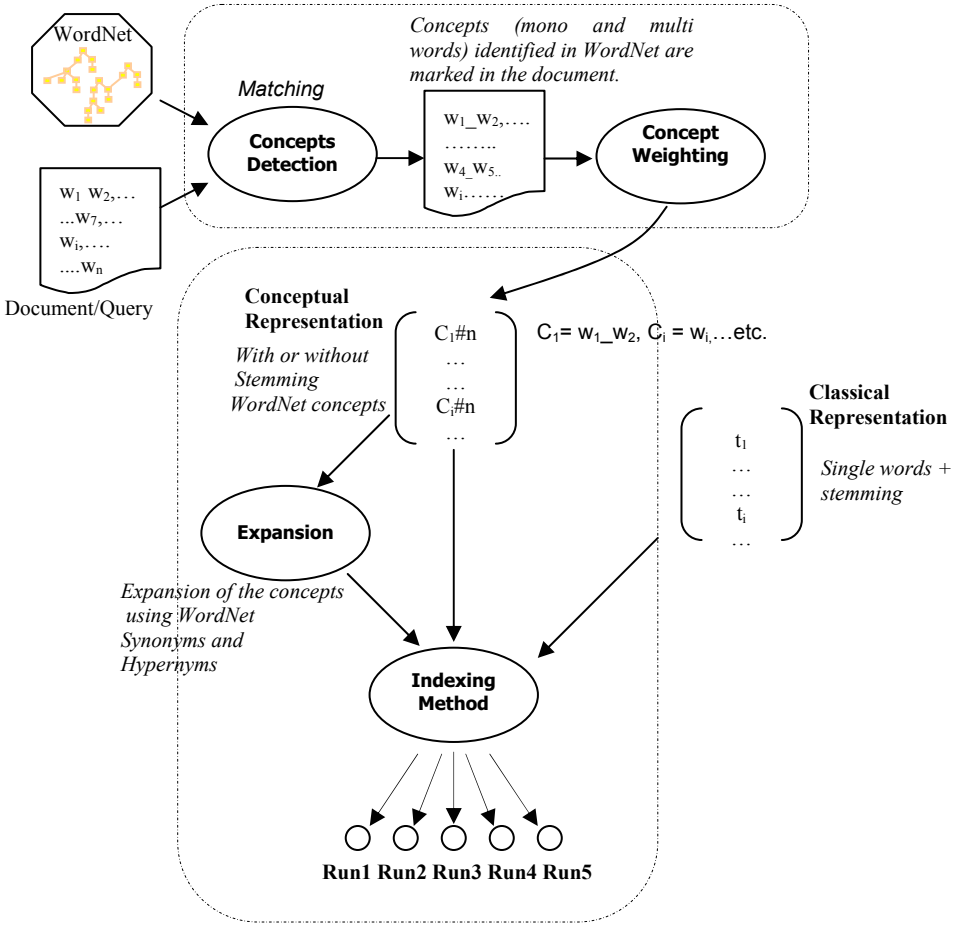

Figure1- Description of the indexing method used to generate the different runs.

[1] In the following, the word "document" will refer to both queries and documents in the collection.

[2] WordNet is organised around the notion of Synset (Synonym set). Each Synset contains terms that are synonyms in a given context. Synsets are interrelated by different relations like Hypernymy (Is-a).

A total of all, five runs were carried out. They are described in Table2 of section 3.
In the next section we will explain the main steps of our system: the concept detection and weighting methods used to carry out our experiments.

I. Detail of the approach

1.1.1 Concepts Detection

Concept detection consists of extracting mono and multiword concepts from documents and queries that correspond to nodes (synsets) in WordNet. Formally, let consider:

$$D= \{w_1, w_2, \ldots, w_n\} \quad (1)$$

the initial document composed of n single words. The result of the concept detection process will be a document D_c. It corresponds to:

$$D_c= \{c_1, c_2, \ldots, c_m, w'_1, w'_{2,\ldots,}w'_m\} \quad (2)$$

where $c_1, c_2, , c_m$ are concepts recognized as WordNet entries. These concepts could be mono or multiword. It may also happen that single words $w'_1, w'_{2,\ldots,}w'_{m,}$ of the initial document (query) do not belong to the WordNet vocabulary. They will not be used for expanding the document (the query). However, they will be added to the

group_president_and_chief_operating_officer_mike_cramer_called…
group_president_and_chief_operating_officer_mike_cramer_called
group_president_and_chief_operating_officer_mike_cramer
group_president_and_chief_operating_officer_mike
group_president_and_chief_operating_officer
group_president_and_chief_operating
group_president_and_chief
group_president_and
group_president
….
chief_operating_officer_mike_cramer_called
chief_operating_officer_mike_cramer_called
chief_operating_officer_mike_cramer
chief_operating_officer_mike
chief_operating_officer

Concept: **"chief_operating_officer#n**" detected

mike_cramer_called
mike_cramer_called
…

Figure2. Concept detection method by combining adjacent words.

final expanded document in order to be used at the search stage.
To detect concepts in the query, we use an ad hoc technique that relies solely on concatenation of adjacent words to identify compound (multiword) concepts in WordNet. In this technique, two alternative ways may be carried on. The first one would be projecting WordNet on the document : all WordNet multiword concepts are mapped onto the document and those occurring in it. This method has the advantage of creating a reusable resource (a document representation made out of WordNet concepts). Its drawback is the possibility to omit concepts which appear in the document and in WordNet under different forms. For example, if WordNet containts the multiword

concept *"solar battery"*, a simple comparison with document would miss the same concept appearing in its plural form *"solar batteries"*. The second way, which we adopt in our experiments, follows an the opposite path, projecting the document onto WordNet: for each multiword candidate concept derived by combining adjacent words in the document, we first question WordNet using these words just as they are, and then we use their base forms if necessary.

Word are combined, as shown in Figure1, according to the longest succession of words for which a concept is detected. In the example of Figure1, the longest concept *"chief_operating_officer#n"* (#n is used for the POS name) is selected although *"chief"* and *"officer"* could also be identified as single word concepts. This concept is defined by WordNet as follow:

> chief executive officer, CEO, chief operating officer -- (the corporate executive responsible for the operations of the firm; reports to a board of directors; may appoint other managers (including a president))

Example of a document after its projection onto WordNet:

In Figure 3 below, we can see a document example from the collection (named GIRT-EN19950120120), after its projection onto WordNet conceptual network. For example health_care_delivery#n is a concept that belongs to a WordNet synset identified in the document. Words that are not tagged (like "ddr" in this example) do not belong to WordNet terminology.

```
<DOC>
<DOCNO> GIRT-EN19950120120 </DOCNO>
<TITLE-EN>
   establishment#n and development#n of the health_care_delivery#n system#n
   in#n syria#n with regard_to#n morbidity#n especially#r infectious_disease#n
</TITLE-EN>
   ddr
   syria#n
   asia#n
   health_care_delivery#n system#n
   arab#n country#n
   historical#a development#n
   near_east#n
   contagious_disease#n
   developing#n country#n
   epidemiology#n
   morbidity#n
   health#n policy#n
   descriptive#a study#n
   medical#n sociology#n
   health#n policy#n
   sociology#n of developing#n country#n developmental#a sociology#n
</DOC>
```

Figure3. An example document from the collection after its projection onto WordNet.

The notations "#n", "#a", "#v", "#r" are used to indicate the part of speech (POS) of the terms belonging to WordNet. They refer respectively to names, adjectives, verbs and adverbs. For the moment, the POS is not used in the index. We need it only to expand the identified mono-sense WordNet terms.

1.1.2 WordNet Covering rate for Documents and Queries

As seen in the previous example, a large majority of the vocabulary used in the collection documents is covered by WordNet. Table1 summarizes the cover rate concerning both queries and documents. More than 92.87% of

the vocabulary used in the documents is covered by WordNet and 99.39% (so almost totality!) of the vocabulary used in the queries is covered.

Concerning compound concepts (or multiterms), they represent about 9% for the document and only 7.83% (0.52 compound term in average) for the queries. Multiterms have often only one sense. It is important to use them in our case, as only mono sense terms from the documents and the queries are expanded in our approach.

Table1. Statistics on using WordNet to index the English Girt Collection.

Total no of docs: 151319 Total no of queries: 25	Total number of TERMS (CLASSICAL)		WORDNET TERMS ONLY			
			All WN terms		WN compounds terms only	
	Documents	Queries(1)	Documents	Queries(1)	Documents	Queries(1)
Total no of terms	**5 118 187**	**166**	**4 753 566**	**165**	**456 715**	**13**
Average no of terms	**33.82**	**6.64**	**31.41**	**6.6**	**3.01**	**0.52**
% (Wn terms compared to the classical)	**-**	**-**	**92.87%**	**99.39%**	**8.92%**	**7.83%**

(1) Only Queries using both Title and Description fields (without expansion) are considered in the table.

1.1.3 Concepts Weighting

The extracted concepts (single or multiwords) are then weighted as in the classical keywords case according to a kind of TF.IDF which is also a variant of the OKAPI system [4].

Thus, a weight $Weight(t_i, d_j)$ of a term t_i in a document d_j is given by the following formula [2]:

$$Weight(t_i, d_j) = \frac{tf_{ij} * (h_1 + h_2 * \log(\frac{N}{n_i}))}{h3 + h4 * \frac{dl_j}{\Delta d} + h_5 * tf_{ij}} \qquad (3)$$

Where:

tf_{ij} : The frequency of the term t_i in the document d_j,

h_1, h_2, h_3, h_4, h_5 : Constants,

n_i : The number of documents containing the term t_i,

N : The total number of documents,

Δ_d : Average document length,

dl_j : Length of document d_j.

The objective of this measure is to attenuate the impact of terms having too much high frequency values.

3. Evaluation

We submitted five official runs to the monolingual English GIRT task ("GIRT_EN"): CWN_T, C_T, CWN_TD, CWNSE_T and CWNSE_TD. The runs are carried out by using title and/or description fields, using or not the term stemming and by performing or not expansion. They are summarized in Table2.

Table2. *Description of the official runs.*

Run	Description
CWN_T	Title field of the topics is used. No stemming is used for WordNet terms. No expansion is used.
C_T	Title field of the topics is used. Stemming for all terms. No expansion is used.
CWN_TD	Title and Description fields of the topics are used. No stemming is used for WordNet terms. No expansion is used.
CWNSE_T	Short Expansion (SE) is used in Queries. Title field of the topics is used. No stemming is used for WordNet terms.
CWNSE_TD	Short Expansion (SE) is used in Queries. Title and Description fields of the topics are used. No stemming is used for WordNet terms.

The results obtained by the different runs are summarized in Table3. It should be noticed that an error slipped into the program in the name of query 132 (named by error 232). Consequently, the query 132 is not evaluated at all. The first column of Table 3 gives the median average precision (MAP) obtained by our five official runs on all the queries. We give in the second column the same runs when using the query relevance file obtained after submission and with the query 132 corrected.

Table 3. *Official Results obtained for the five submitted runs compared to the median average.*

	Median Average Precision (MAP)	Non official Results	Increment (%)
CWN_T	0.3411	0.3762	+10,29%
C_T	0.3411	0.3765	+10,38%
CWN_TD	0.3223	0.3574	+10,89%
CWNSE_T	0.3251	0.3579	+10,09%
CWNSE_TD	0.3235	0.3563	+10,14%

Concerning the official results, as it can be shown in Figure4, the best results are obtained when using only the title field of the topics and stemming the extracted terms (run C_T). Followed by the run CWN_T where WordNet terms are not stemmed, and then the run CWNSE_T where a short expansion (by synonyms and one

hypernym) is applied to non polysemic terms. The two last runs (CWN_TD and CWNSE_TD) are obtained when both title and description fields are used to build the queries respectively with and without expansion.

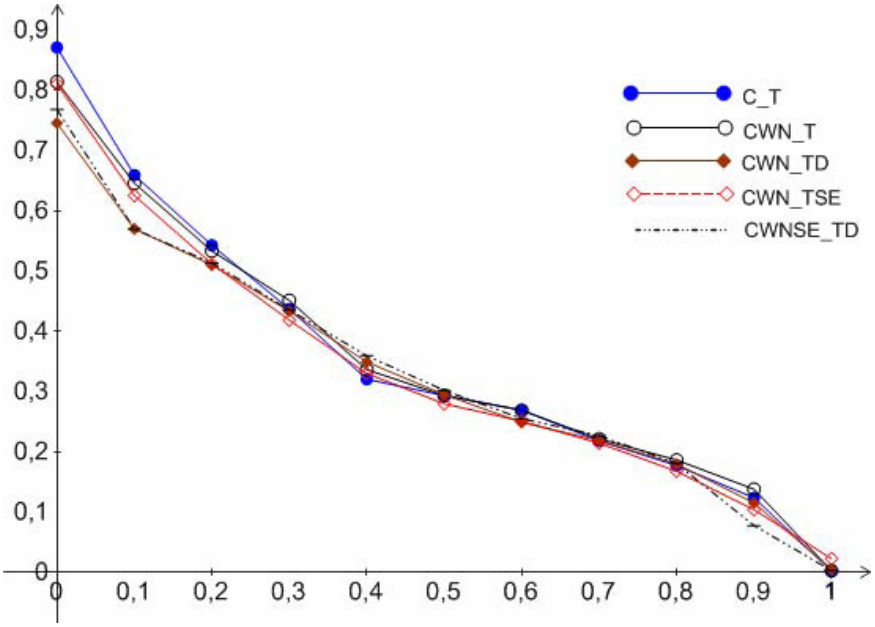

Figure4. Recall-Precision curves for the fives submitted runs.

Concerning the non official runs, the results follow the same logic while being better than the official ones. The fourth column of Table 3 gives the difference of the global results, for the five runs, between the submitted results and the results obtained after the error has been fixed. Roughly the official results could be enhanced by 10,36% in average for each run by using the query 132 and with changing nothing to the system.

The reason is that the omitted query (132) brings very good results, which also increases the global result. The detailed results of query 132 are given in Table 4 for the five runs.

Table4. Non official results for the omitted query 132

Run	Num	P5	P10	P15	P20	P30	P100	P1000	MAP
C_T	132	1.0000	1.0000	1.0000	1.0000	1.0000	0.9200	0.1440	0.8854
CWN_T	132	1.0000	1.0000	1.0000	1.0000	1.0000	0.8900	0.1470	0.8791
CWN_TD	132	1.0000	1.0000	1.0000	1.0000	1.0000	0.8900	0.1470	0.8773
CWNSE_T	132	1.0000	1.0000	1.0000	0.9500	0.9000	0.8500	0.1470	0.8196
CWNSE_TD	132	1.0000	0.9000	0.9333	0.9500	0.9333	0.3600	0.0540	0.8192

4. Conclusion

We have evaluated the performances of our conceptual indexing method which consists of matching documents and queries with WordNet. In this method, documents and queries are represented by WordNet nodes. The first remark, when comparing our submitted runs, is that using only title field (runs C_T and CWN_T) from the topics seems to bring the better results than using the title and description fields together. The second remark

concerns the use of term stemming. Results showed that stemming indexing terms (run C_T) is slightly better than not stemming them (run CWN_T) when we consider only the first retrieved documents. However, by using a more global judgment (MAP), both cases are close. Another remark concerns the Expansion method used in our experiments. Even though it is made so as to avoid the disambiguation problem (only mono sense terms are expanded), expansion does not seem to bring the best results. The best run is obtained without expansion and by using only the title field of the topics. However, the results obtained by the expansion method, when expanding titles, are better than those obtained when the description fields are used in addition to titles in the queries and without expansion. So we still believe that a more sophisticated expansion method could bring better results [1]. The specificity of the GIRT collection documents could also require some adaptation (to evaluate the usefulness of using hypernymy relation for example).

Another conclusion concerns the suitability of using WordNet for the domain specific collection. It appears that WordNet largely covers the vocabulary of the English GIRT collection (more than 90% for the documents and practically the entire vocabulary of the 25 used queries) and is suitable to be used for this particular collection.

References

1. Baziz M., Boughanem M. and Aussenac-Gilles Nathalie "The Use of Ontology for Semantic Representation of documents". In Proceeding of Semantic Web and Information Retrieval Workshop (SWIR) held in conjunction with the 27th ACM SIGIR Conference'04, July 25–29, 2004, Sheffield, United Kingdom.
2. Boughanem M., Julien C., Mothe J., Soulé-Dupuy C. "Mercure at TREC-8" Adhoc, Web, CLIR and Filtering tasks. Proceeding of Trec-8, (1999).
3. Miller G., Wordnet: A lexical database. Communication of the ACM, 38(11):39-41, (1995).
4. Okapi at TREC-6, Proceeding of the 6th International Conference on Text Retrieval TREC, Harman D.K. (Ed.), NIST SP 500-236, pages: 125-136, (1997).
5. Porter, M. An algorithm for suffix stripping. Program, 14(3):130-137, July, 1980.