A Conceptual Indexing Approach for the TREC Robust Task

Mustapha Baziz [1], Mohand Boughanem [1], Nathalie Aussenac-Gilles [2]

[1] IRIT-UPS
Campus Univ. Toulouse III
118 Route de Narbonne
F-31062 Toulouse Cedex 4

[2] IRIT-CNRS
Campus Univ. Toulouse III
118 Route de Narbonne
F-31062 Toulouse Cedex 4

ABSTRACT. This paper describes our participation to the TREC 2005 Robust Task. A method of conceptual indexing based on WordNet is used. Both documents and queries are mapped onto WordNet. Thus concepts belonging to WordNet synsets are identified and extracted whereas those having a single sense are expanded.

1. Introduction

The objective of our participation to the TREC 2005 Robust Task, was to evaluate the use of a conceptual indexing method based on WordNet [3] lexical database. The technique consists in detecting mono and multiword concepts corresponding to WordNet synsets from both documents and queries. Then these concepts are used as a conceptual indexing space. Terms not identified in WordNet are also added to complete the representation. Even though they are not useful during the expansion phase, they are used to compare documents and queries at the searching stage.

This paper is organized as follows. In section2, we describe the synoptic scheme of our system. In section3, the approach is detailed : our method for concept detection and weighting is described. Section4 presents the official evaluation results.

2. Overview of the Approach

In this section, we describe our conceptual indexing method based on WordNet. The principle consists in, being given a document (or a query), mapping it onto WordNet and then indetifying the WordNet concepts (mono and multi terms) from the words used in the text of the document (the query) [1]. The extracted concepts are then weighed and tagged using part of speech information (POS) in order to facilitate their expansion.

The *expansion* which we qualify *Short Expansion* (or SE) consists of expanding from the document and/or the queries the mono sense WordNet terms (having only one sense) by using all of their synonyms extracted from the synset[1] they belong to. The indexing method may use expansion and stemming or not [5], according to the used run. It combines classical keywords indexing to conceptual indexing by adding the terms not recognized by WordNet to the dictionary ? document representation ?.

A total of five runs were produced for this experiment. They are described in Table2 of section 3.

In the next section we will explain the main steps of our system which are the concept detection and weighting methods used to carry out our experiments.

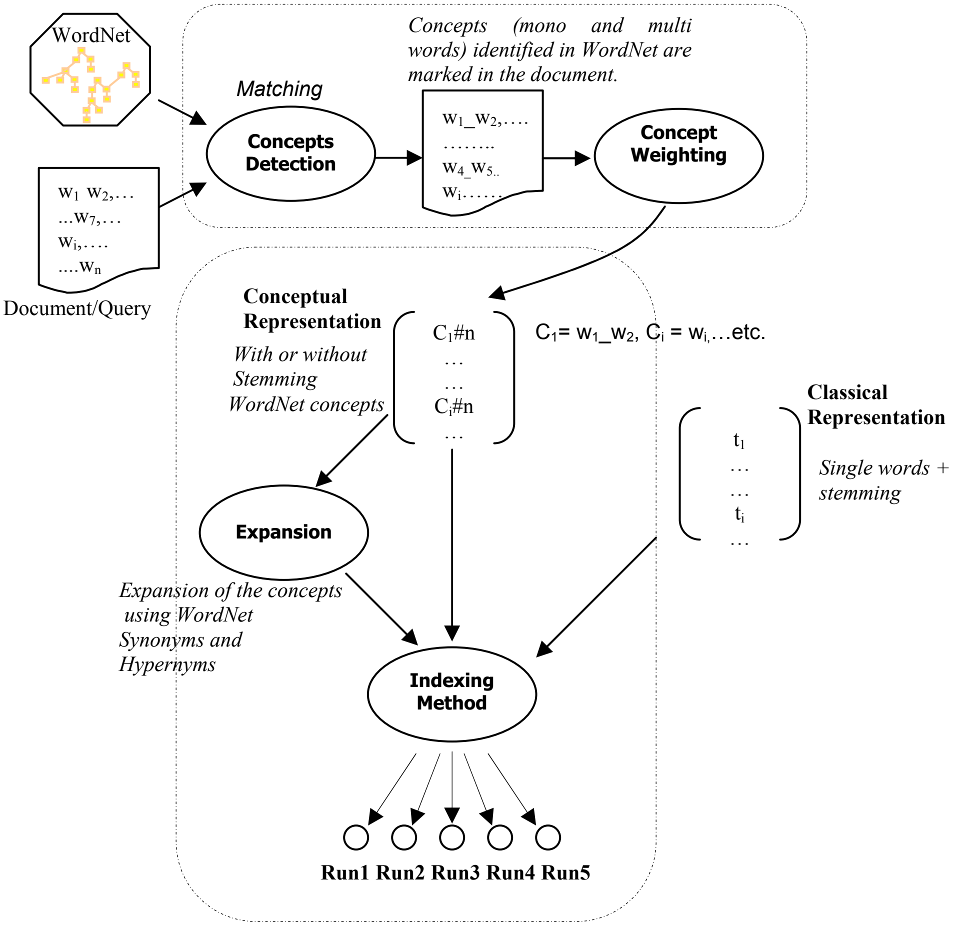

Figure1- Description of the indexing method used to generate the different runs.

3. Detail of the approach

3.1 Conceps Detection

[1] WordNet is organised around the notion of Synset (Synonym set). Each Synset contains concepts that are synonyms in a given context.

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

the

$$D= \{w_1, w_2, \dots, w_n\} \qquad (1)$$

initial document composed of n single words. The result of the concept detection process will be a document D_{c}.:

$$D_c= \{c_1, c_2, \dots, c_m, w'_1, w'_{2,\dots,}w'_{m'}\} \qquad (2)$$

where $c_1, c_2, , c_m$ are concepts identified as entries in WordNet. These concepts could be mono or multiword. It may also happen that single words $w'_1, w'_{2,\dots,}w'_{m'}$ of the initial document (query) do not belong to WordNet vocabulary. In that case, they will not be used for expanding the document (the query). However, they will be added to the final document (query) representation in order to be used at the search stage.

For detecting concepts in the document (query), we use an ad hoc technique that relies solely on the concatenation of adjacent words to identify compound (multiword) concepts of WordNet. In this technique, two alternative ways can be distinguished. The first one suggests to map WordNet on the document (the query) by extracting all multiword concepts from WordNet and then identifying those occurring in the document (query). This method has the advantage of creating a reusable resource. Its drawback is the possibility to omit concepts which appear in the query and in WordNet in different forms. For example, if WordNet contains the multiword concept *"solar battery"*, a simple comparison do not make it possible to identify the same concept appearing in its plural form *"solar batteries"* in the document or the query. The second way, which we adopt in our experiments, follows the opposite path, mapping the document (query) onto WordNet: for each multiword candidate concept derived by combining adjacent words in the document (query), we first question WordNet using these words just as they are, and then we use their base forms if necessary.

```
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.

Concerning word combination, as shown in Figure1, the longest successive word chain for which a concept is detected is selected. In the example of Figure1, the longest concept *"chief_operating_officer#n"* (#n is used for the

POS name) is considered 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))

3.2 Why using Multiword Concepts?

The extraction of multiword concepts in plain text is important to reduce ambiguity. These concepts generally are monosemous even though the words they contain can be individually ambiguous. For example, when taking each word separately in the concept *ear_nose_and_throat_doctor,* we have to disambiguate (according to WordNet) between 5 senses for the word *ear*, 13 senses (7 for the noun *nose* and 6 for the verb *nose*) for the word *nose*, 3 senses for *throat* (*and* is not used because it is a stop-word) and 7 senses (4 for the noun and 3 for the verb) for the word *doctor*. So, we would have a number of 5x13x3x7= 1365 possible combinations of candidate senses. But when considering all the words forming a single multiword concept (of course, the multiword concept must be recognized by WordNet), we will only have one sense.

In this example, the full concept (WordNet synset) and its definition (gloss in WordNet) are as follows:

The noun ear-nose-and-throat doctor has 1 sense
1. ENT man, ear-nose-and-throat doctor, otolaryngologist, otorhinolaryngologist, rhinolaryngologist -- (a specialist in the disorders of the ear or nose or throat.)

Statistics carried on WordNet2.0 presented in Table1, show that, from a total of 63,218 multiword concepts (composed of 2-9 words), 56,286 (89%) of them are monosemous. 6,238 have 2 senses (9.867%) and only 694 (0.506%) multiword concepts have more than 2 senses. Thus, the more there are multiword concepts in a document to be analyzed, the easier is their disambiguation.

Table1. *Polysemy repartition on multiword concepts in WordNet 2.0*

Number of senses	Number of multiword concepts (2-9 words)	%
1	56286	89,035%
2	6238	9,867%
3	375	0,593%
>=4	319	0,506%
Total	63218	100%

3.3 Example of Multiword Concepts Extracted from the Official Topics

All the concepts recognized in WordNet are identified. They should be mono or multiword. Below, examples of multiword concepts extracted from the official topics:

310	radio_wave	372	native_american	419	destructive_distillation
336	black_bear	374	nobel_prize	433	artistic_production
341	international_flight	374	prize_winner	435	population_growth
363	motor_vehicle	374	field_of_study	625	world_trade_center
367	old_fashioned,	383	mental_illness	639	growth_factor
367	body_of_water	393	mercy_killing	650	tax_evasion
367	fishing_vessel	393	life_support	689	family_planning

The extracted concepts are then weighed according to a kind of TF.IDF thatwe name CF.IDF. For a concept c_i composed of n words, its frequency in a query equals to the number of occurrences of a concept itself, and the one of all its sub-concepts. Formally:

$$cf(c_i) = count(c_i) + \sum_{sc\in sub(c_i)} \frac{length(sc)}{length(c_i)}\ count(sc) \qquad \textbf{(3)}$$

Where $length(c_i)$ represents the number of words that form c_i and $sub(c_i)$ is the set of all possible sub-concepts which can be derived from c_i: concepts of n-1 words from c_i, concepts of n-2, and all single words of c_i.

3.4 Example of concept weighting

Tf we consider the 3 word concept *"elastic potential energy"* in a given topic, its frequency is computed as follows:

> ***cf***(*"elastic potential energy") = count("elastic potential energy") + 2/3 count("potential energy")+1/3 count("elastic") + 1/3 count("potential") + 1/3 count("energy").*

Knowing that *potential energy* is itself also a WordNet multiword concept, here, it is a question of adding the number of occurrences of *potential energy* and not its frequency.

3.5 Example of a document after its projection onto WordNet:

In Figure3 below, we can see a document example from the collection (named XIE20000930.0016), after its projection onto WordNet. For example "costa_rica#n" is a concept that belongs to a WordNet synset identified in the document. Words that are not tagged (like "rojas" in this document example) do not belong to WordNet terminology. 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. At the moment, the POS tag is not used in the index. We need it only to expand the identified mono-sense WordNet terms.

```
(DOC
(DOCNO
 XIE20000930.0016
)DOCNO
(TITLE
 costa_rica#n n to host#n talks#n for colombian#n peace#n
)TITLE
(TEXT
 costa_rican#n n foreign_minister#n n roberto rojas was quote#v by the
 press#n as saying#n that the meeting#n will#n be hold#v between#r october#n 16
 and 18 in the central_american_country#n n but that the site#n had not
 be#v define#v yet
 rojas say#v costa_rican#n n president#n miguel angel#n rodriguez had
 say#v previously#r that costa_rica#n n would help#n colombia#n find#n the peace#n
 the costa_rican#n n government#n receive#v an official#n communication#n
 thursday#n from the colombian#n authorities#n authorize#v the meeting#n
 with representative#n of the revolutionary_armed_forces_of_colombia#n n
 farc#n and the national_liberation_army#n n eln rojas say#v
 the country#n will#n have#n an opportunity#n to promote#v its idea#n of
 agreement#n through#a dialogue#n costa_rica#n n has be#v the scenario#n and
 also#r play#v a role#n in the peace#n plan#n in central_america#n n and that
…
)TEXT
)DOC
```

4. Evaluation

We submitted five official runs to the Robust task: CKonT, CKonD, CKonTSE, CKSEonD et CKSEonTSE. The runs were carried out by using the title and/or description fields, by performing or not expansion. They are summarized in Table2. In the five (5) runs, WordNet terms are not stemmed, and those not belonging to WordNet are stemmed (classical indexing is used to no WordNet terms). An expansion method, named Short Expansion (SE), can be used to expand queries and/or documents. This method expands mono-sense WordNet terms with their synonyms, ie, with the terms belonging to their synset.

Table2. *Description of the official runs.*

Run	Description
CKonT	Title field of the topics is used. No expansion is used.
CKonD	Description field of the topics is used. No expansion is used.
CKonTSE	Title fields of the topics are used. Short expansion is used to queries.
CKSEonD	Description field of the topics is used. Short Expansion (SE) is applied to documents without queries.
CKSEonTSE	Short Expansion (SE) is used for both queries and documents. Title field of the topics is used.

The results obtained by the different runs are summarized in Table3. and the corresponding recall precision curves are given in Figure 3.

Table 3. *Official Results obtained for the five submitted runs.*

Precision at	**CKonT**	**CKonD**	**CKonTSE**	**CKSEonD**	**CKSEonTSE**
5 docs:	0.3920	0.3200	0.3480	0.3640	0.3720
10 docs:	0.3500	0.2880	0.3220	0.3080	0.3480
15 docs:	0.3347	0.2680	0.3040	0.2853	0.3333
20 docs:	0.3200	0.2620	0.3000	0.2770	0.3140
30 docs:	0.3007	0.2453	0.2807	0.2487	0.2807
100 docs:	0.2114	0.1748	0.2034	0.1870	0.2294
200 docs:	0.1616	0.1347	0.1568	0.1417	0.1763
500 docs:	0.1085	0.0894	0.1012	0.0950	0.1079
1000 docs:	0.0681	0.0569	0.0655	0.0603	0.0672
Average Pr.	**0.1511**	**0.1076**	**0.1415**	**0.1215**	**0.1572**

When one compares the overall results on the recall-precision curves, the best results are obtained with the **CKSEonTSE** run, which uses a Short Expansion (SE) method for both queries and documents. This run is followed by **CKonT** which, as the previous one, uses only the title field, but without any expansion. This second run (**CKonT)** brings better accuracy than the first one, **CKSEonTSE** , when one considers only precision at top 5, 10, 15, 20 and 30 documents.

It seems that expansion is worth only to find out more documents with a low similarity with the query. The other conclusion that can be drawn is that conceptual indexing brings better results with short and precise queries than with long one.

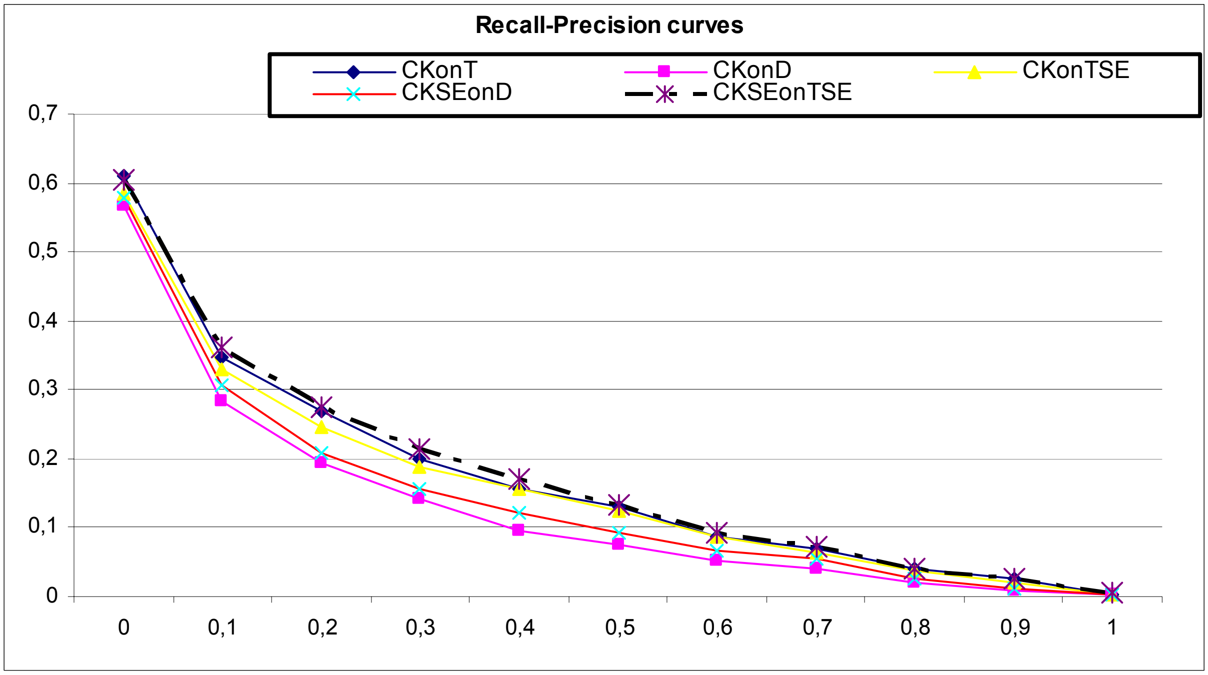

Figure 3. Recall-precision curves for the five submitted runs.

5. 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 a set of WordNet senses. The first remark, when comparing our submitted runs, is that a "careful" expansion method when applied for both queries and documents bring the best results. Indeed, this expansion method is made so as to avoid the disambiguation problem (only mono sense terms are expanded). The second remark concerns the use of the Description field. As shown in the results, using Description field in the queries brings better results when the documents of the collection are expanded.

References

1. Baziz M., Boughanem M. and Aussenac-Gilles N. “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).