IRIT at CLEF 2004:
The English GIRT task

Mustapha Baziz, Mohand Boughanem, Nathalie Aussenac-Gilles

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

ABSTRACT. This paper describes our participation to the monolingual English GIRT task. The main objectives of our experiments were to evaluate the use of Mercure IRS (designed at IRIT/SIG) on domain specific corpus. Two other techniques of automatic query reformulation using WordNet are evaluated.

1. Introduction

The objective of IRIT/SIG participation in 2004 was to evaluate the use of Mercure IRS on domain specific data. In addition to evaluate the Mercure system, two other techniques are experimented using WordNet. The first technique consists on detecting mono and multiword concepts from queries and then to weight them according to a proposed CF.IDF formula, a kind of TF.IDF. The second concerns disambiguation-expansion method consisting of selecting the closest synset (concept) to the initial query, from WordNet, to use for expanding the query.

This paper is organized as follows. In section2, the used Mercure IRS model is described. In section3, the additional tests are formally described: the concepts detection and weighting method from queries in 3.1, and the disambiguation-expansion method in 3.2. Section4 presents the official evaluation results compared with the median average obtained by all participating systems. Finally, section5 gives some conclusions and prospects.

2. Mercure Model

Mercure is an information retrieval system based on a connectionist approach and modelled by three-layered network (as shown in Figure1). The network is composed of a query layer (set of query terms), a term layer representing the indexing terms and a document layer [2].
Mercure includes the implementation of a retrieval process based on spreading activation forward and backward through weighted links. Queries and documents can be either inputs of the network. The links between two layers are symmetric and their weights are based on the TF.IDF measure inspired by the OKAPI [5] term weighting formula.

– The term-document link weights are expressed by:

$$d_{ij} = \frac{tf_{ij} * (h_1 + h_2 * \log(\frac{N}{n_i}))}{h_3 + h_4 * \frac{dl_j}{\Delta d} + h_5 * tf_{ij}} \qquad \textbf{(1)}$$

– The query-term links (at stage s) are weighted as follows:

$$q_{ui}^{(s)} = \begin{cases} \frac{nq_u * qtf_{ui}}{nq_u - qtf_{ui}} \; if\,(nq_u > qtf_{ui}) \\ qtf_{ui} \; otehrwise \end{cases} \qquad \textbf{(2)}$$

The query weights are based on spreading activation. Each neural node computes an input and spreads an output signal:

1. The query k is the input of the network. $Input_k$=1. Then, each neuron from the term layer computes an input value from this initial query:

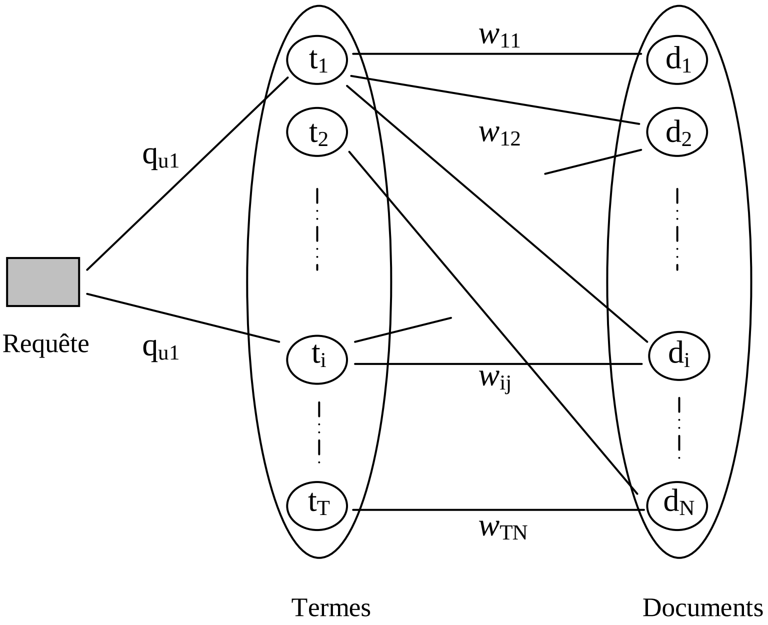


Figure1. *Mercure Model*

$$In(N_{ti}) = Input_k * q_{ki}^s \qquad (3)$$

The output value is computed as follows:

$$Out(N_{ti}) = g(In(N_{ti}) \qquad (4)$$

where g is the identity function.

2. These signals are propagated forward through the network from the term layer to the document layer. Each neuron computes an input and output value:

$$In(N_{dj}) = \sum_{i=1}^{T} Out(N_{ti}) * w_{ij} \qquad (5)$$

and,

$$Out(N_{dj}) = g(In(N_{dj})) \qquad (6)$$

The system output is:

$$Output_k (Out(N_{D1}), Out(N_{D2}),.., Out(N_{DN})$$

Notations:

T:	the total number of indexing terms,
N:	The total number of documents,
q_{ui}:	The weight of the term t_i in the query *u*,
t_i:	The term t_i,
d_j:	The document d_j
w_{ij}	The weight of the link between the term t_i and the document d_j,

dl_j Document length in words (without stop words),
Δd Average document length,
tf_{ij} The frequency of the term t_i in the document d_j,
n_i The number of documents containing term t_i,
nq_u The query length (number of unique terms)
qtf_{ui} Query term frequency

3. Overview of the additional tests

In this section, we describe two methods used for query processing based on WordNet. The first consists of concept detection and weighting from queries. The second method, disambiguation-expansion, tend to expand a query with its closest synset from WordNet [4].

3.1. Concepts detection and weighting

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

$$Q= \{w_1, w_2, \dots, w_n\} \qquad (7)$$

the initial query composed of n single words. The result of the concept detection process will be a query Q_c. It

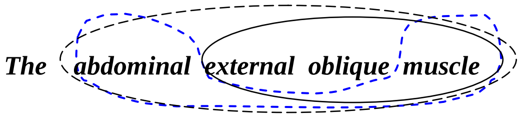


Figure2. *Example of text with different concepts*

corresponds to:

$$Q_c= \{c_1, c_2, \dots, c_m, w'_1, w'_{2,\dots,}w'_{m'}\} \qquad (8)$$

Where c_1, c_2, , c_m are concepts recognized as entries in WordNet. These concepts could be mono or multiword. It can also happen that single words w'_1, $w'_{2,\dots,}w'_{m'}$ of the initial query do not belong to ontology vocabulary. They will be used for disambiguating the query. They will then be added to the final expanded query.

For detecting concepts in the query, we use an ad hoc technique that relies solely on concatenation of adjacent words to identify compound (multiword) concepts of WordNet. In this technique, two alternative ways can be distinguished. The first one consists of projecting WordNet on the query by extracting all multiword concepts from WordNet and then identifying those occurring in the 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 with different forms. For example if WordNet recognizes a multiword concept *"solar battery"*, a simple comparison do not recognizes in the query the same concept appearing in its plural form *"solar batteries"*. The second way, which we adopt in this paper, consists in the opposite step, projecting the query on WordNet: for each multiword candidate concept derived by combining adjacent words in the query, we first question WordNet using these words just as they are, and then we use their base forms if necessary.

Concerning word combination, the principle consists in selecting the longest successive terms for which a concept is detected.

If we consider the example shown in Figure2, the sentence contains three (3) different concepts which are: *external oblique muscle*, *abdominal muscle* and *abdominal external oblique muscle*. The definition of the first concept according to WordNet is:

The noun abdominal muscle has 1 sense
1. abdominal, abdominal muscle, ab -- (the muscles of the abdomen);

This concept is not retained in our approach, because its words are not adjacent. The second *"external oblique muscle"* and the third *"abdominal external oblique muscle"* are synonyms, their definition is:

The noun external oblique muscle has 1 sense
1. external oblique muscle, musculus obliquus externus abdominis, abdominal external oblique muscle, oblique -- (a diagonally arranged abdominal muscle on either side of the torso)

The selected concept is associated to the longest multiword « *abdominal external oblique muscle* » which corresponds to the correct sense of the sentence. Remind that in words combination, the order must be respected (left to right) otherwise we could be confronted to the syntactic variation problem (*science library* is different from *library science*).

Example of multiword concepts extracted from the official topics:

103 live_in
105 on_the_job
106 multiple_sclerosis

109 animal_husbandry
114 federal_republic_of_germany
117 carbon_dioxide

124 telephone_interview
125 european_country
125 infant_mortality

The extracted concepts are then weighted according to a kind of TF.IDF, we 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:

$$cf(c_i) = count(c_i) + \sum_{sc \in sub(c_i)} \frac{length(sc)}{length(c_i)}\ count(sc) \qquad \textbf{(9)}$$

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.

Example:
if we consider a concept *"elastic potential energy"* in a given topic, composed of 3 words, 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 multiword concept and here, it is a question of adding the number of occurrences of *potential energy* and not its frequency.

3.2. Disambiguation-expansion using WordNet synset

Once mono and multiwords concepts of initial queries are extracted and weighted, an expansion process with WordNet synsets is carried out. As each recognized concept c_k (formula 8) could have several senses (a set R_{Syns} of synsets containing C_k):

$$R_{Syns}(C_k) = \left\{ C_k^1, C_k^2, ..., C_k^j, .., C_k^t \right\} \qquad \textbf{(10)}$$

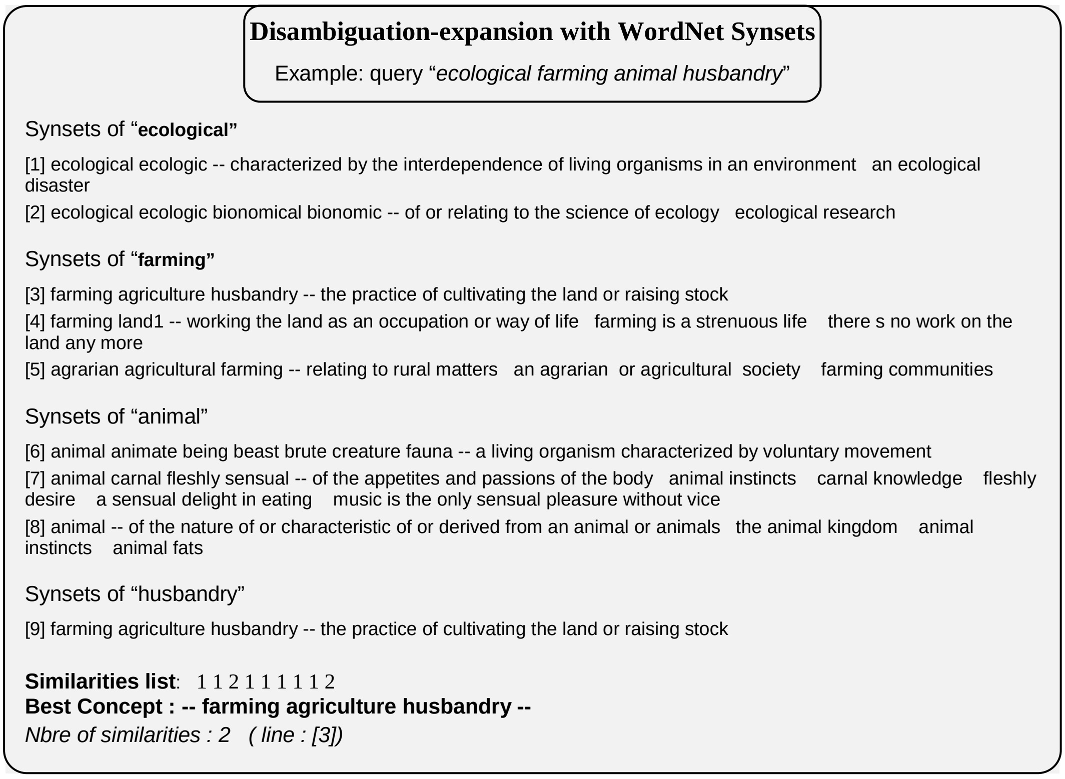
they are disambiguated using an adapted Lesk algorithm [3] which consists of overlapping each synset with the initial query. A concept-sense (synset) having the best overlapping (the greater number of common words) with the initial query is retained. Formally:

$$Best(R_{Syns}(C_k)) = \underset{k,j}{ArgMax} \left\| \left\{ C_k^1, C_k^2, ..., C_k^j, .., C_k^t \right\} \cap Q \right\| \qquad \textbf{(11)}$$

Example of Disambiguation

Let us consider a query:
Q=[* ecological farming animal husbandry*].
It contains 4 single-word concepts which are:
C_1= "ecological", C_2 = "farming", C_3 = "animal", C_4 = "husbandry".
The first concept "ecological" has two synsets ($R_{Syns}(C_1)$={[1], [2]}) which appear in lines noticed [1] and [2] of Figure3, the second "farming" has three synsets ($R_{Syns}(C_2)$ ={[3], [4], [5]}), the third "animal" has three ({[6], [7], [8]}) and the last concept "husbandry" has only one synset (at line [9]). As only one synset could be used for expanding the whole query in our "careful query expansion" approach, the best concept *Best(*$R_{syns}(C_k)$*)* which disambiguates the query Q is the synset of line [3] (or [9] which is identical to [3] in this example): farming

agriculture husbandry -- the practice of cultivating the land or raising stock . In our “careful expansion” method, synset without its glossary was used to expand the query, so farming agriculture husbandry. As the first and the last words already belong to the initial query, the final query will be expanded only with the word agriculture.



Disambiguation-expansion with WordNet Synsets

Example: query “*ecological farming animal husbandry*”

Synsets of “**ecological”**

[1] ecological ecologic -- characterized by the interdependence of living organisms in an environment   an ecological disaster

[2] ecological ecologic bionomical bionomic -- of or relating to the science of ecology   ecological research

Synsets of “**farming”**

[3] farming agriculture husbandry -- the practice of cultivating the land or raising stock

[4] farming land1 -- working the land as an occupation or way of life   farming is a strenuous life   there s no work on the land any more

[5] agrarian agricultural farming -- relating to rural matters   an agrarian or agricultural society   farming communities

Synsets of “animal”

[6] animal animate being beast brute creature fauna -- a living organism characterized by voluntary movement

[7] animal carnal fleshly sensual -- of the appetites and passions of the body   animal instincts   carnal knowledge   fleshly desire   a sensual delight in eating   music is the only sensual pleasure without vice

[8] animal -- of the nature of or characteristic of or derived from an animal or animals   the animal kingdom   animal instincts   animal fats

Synsets of “husbandry”

[9] farming agriculture husbandry -- the practice of cultivating the land or raising stock

Similarities list: 1 1 2 1 1 1 1 1 2
Best Concept : -- farming agriculture husbandry --
Nbre of similarities : 2   (line : [3])

Figure3. *Example of disambiguation-expansion using wordNet synsets.*

4. Evaluation

We submitted five official runs to the monolingual English GIRT task ("GIRT_EN"): Run1T, Run2TD, Run3TDfc, Run4TWN and Run4TDWN. They are described in Table1.

Table1. *Description of the official runs*

Run	Description
Run1T	Title part of the topics are used
Run2TD	Title and Description parts of the topics are used
Run3TDfc	Concept detection and weighting are used (Title and Description)
Run4TWN	Disambiguation-expansion method with WordNet Synsets is used (Title only)
Run4TDWN	Disambiguation-expansion method with WordNet Synsets is used (Title and Description)

The results obtained by the different runs are summarized in Table2. These results are compared in the third column (Increment) of Table2 with the median average precision (0.2990) obtained by all the systems that participated in the CLEF2004 GIRT task.

Table2. *Results obtained for the five official runs compared to the median average.*

	Average Precision	Increment (%)
Run1T	0.3740	+25.08%
Run2TD	0.3855	+28.92%
Run3TDfc	0.3764	+25.88%
Run4TDWN	0.3640	+21.73%
Run5TWN	0.3764	+25.88%

Roughly the obtained results are about +25% better than the median average obtained by all participating systems. These results show also that using WordNet in disambiguation-expansion and concepts frequencies do not enhance significantly the average precision even though the precision for the first retrieved documents (not reported here) are better in the case of Run5TWN. Detecting and weighting concepts method, to bring better results, should be enhanced and then applied to queries as well as to documents.

5. Conclusion and Future Work

We have evaluated the performances of our IRS (Mercure) in domain specific corpus, and a method for query reformulation based on concepts detection and weighting using WordNet synsets. In this method, multiword concepts are removed into single words in the final queries in order to be conforming to the used IRS indexing process. What is presented in this report is a part of a complete method achieved after our participation to 2004 CLEF campaign which is applied for queries and documents as well. This method is described in [1]. Next year, we intend to participate to CLEF with the new method.

6. 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., Dkaki T., Mothe J. and Soulé-Dupuy C. "Mercure at TREC-7" Proceeding of Trec-7, (1998).
3. Lesk, M. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone. In Proceedings of SIGDOC ’86, 1986.
4. Miller G., Wordnet: A lexical database. Communication of the ACM, 38(11):39-41, (1995).
5. 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).