Mercure at CLEF-1

M. Boughanem, N.Nassr

IRIT/SIG
Campus Univ. Toulouse III
118, Route de Narbonne
F-31062 Toulouse Cedex 4
Email : trec@irit.fr

1 Summary

This paper describes the tests performed by our team in CLEF programme. These tests were done using Mercure system and concern : Multilingual, Bilingual and Monolingual tasks. The section 2 presents the Mercure system. The section 3 describes our general approach to CLIR. The section 4 gives the details of the experiments and the results.

2 Mercure model

Mercure is an information retrieval system based on a connectionist approach and modelled by a multi-layered network. The network is composed of a query layer (set of query terms), a term layer representing the indexing terms and a document layer [3],[2].

Mercure includes the implementation of a retrieval process based on spreading activation forward and backward through the weighted links. Queries and documents can be either inputs or outputs of the network. The links between two layers are symmetric and their weights are based on the $tf * idf$ measure inspired from the OKAPI [4] term weighting formula.

- the term-document link weights are expressed by:

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

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

$$q_{ui}^{(s)} = \begin{cases} \frac{nq * qtf}{nq - qtf} \; si \; (nq > qtf) \\ qtf \; otherwise \end{cases} \quad (2)$$

The query evaluation is based on spreading activation. Each node computes an input and spreads an output signal [2].

2.1 Query evaluation

A query is evaluated using the spreading activation process described as follows :

1. The query Q_u is the input of the network. Each node from the term layer computes an input value from this initial query: $In(t_i) = q^s_{ui}$ and then an activation value : $Out(t_i) = g(In(t_i))$ where g is the identity function.

2. These signals are propagated forwards through the network from the term layer to the document layer. Each document node computes an input : $In(d_j) = \sum_{i=1}^{T} Out(t_i) * w_{ij}$ and then an activation , $Out(d_j) = RSV(Q_u, d_j) = g(In(d_j))$.

   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 term frequency of t_i in the document D_j,
   n_i: the number of documents containing term t_i,
   nq: the query length, (number of unique terms)
   qtf: query term frequency.

3 General Clir Methodology

Our CLIR approach is based on query translation. It is illustrated by figure 1.

Indexing : a separate index is built for the documents in each language. English words are stemmed using Porter algorithm, French words are stemmed using a truncature (7 first characters), no stemming for the German and Italian words. The German and Italian stoplists were downloaded from Internet.

Translation : is based on "dictionaries". For the CLEF1 experiments, three bilingual dictionaries were used all of which were actually simply a list of terms in language $l1$ that were paired with some equivalent terms in language $l2$. Table 1, shows the source and the number of entries in each dictionary.

Desambiguisation when multiple translations exist for a given term they are generally relevant only in a specific context. The disambiguisation consists of selecting the terms that are in the context of the query. We consider that a context of a given query can be represented by the list of its terms. The desambiguisation process consists of building a context of the target query and using this context to desambiguate the list of substitutions resulting from the query source translation.

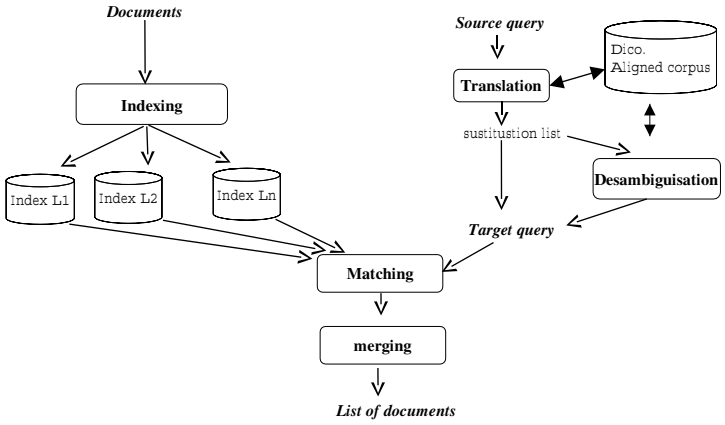

Figure 1: General CLIR approach

Type	Source	nb. entries
E2F	http://www.freedict.com	42443
E2G	http://www.freedict.com	87951
E2I	http://www.freedict.com	13478

Table 1: Dictionaries characteristics

A context of the target query is built using an aligned corpus. It consists of selecting the best terms appearing in the top (X=12) documents in target language aligned to the top (X=12) retrieved by the query source. The terms are sorted according the following formula :

$$score(t_i) = \sum_{d_k \in D_x} d_{ik}$$

D_x : set of aligned documents to those retrieved by the source query,
d_{ik} : weight of term t_i in document d_k.

The desambiguisation of the translated query consists of retaining only terms that appear in the list of terms of the target context. However, if a specific term has an unique substitution this term is retained even though it not exists in the context of the target query. Note that in this process all the terms appearing in the target context are retained we do not select only the best translation as it is done in some other works [1].

4 Experiment and Results

4.1 Multilingual experiment

Two runs using English topics and retrieving documents from the pool of documents in all four languages (German, French, Italian and English), were submitted. The queries were

translated using the downloaded dictionaries. No desambiguisation, all the translated words were retained in the target queries. The runs were performed by doing individual runs for pair languages and merging the results to form the final ranked list. Two merging strategies were tested :

- naive strategy : all the documents resulting from the pair searches join a final list. These documents are then sorted according to their RSV. The top 1000 were submitted.
- normalised strategy : each list of retrieved documents resulting from the pair search was normalised. The normalisation consists simply of dividing the RSV of each document by the maximum of RSVs in that list. The documents of the different lists are then merged and sorted according to their normalised RSV. The final list corresponds to the top 1000 documents.

Two runs were submitted : irit1men2a based on normalised merging and irit2men2a based on naive merging.

	irit1men2a	irit2men2a
better than median Avg. Prec. :	15 (best 0)	16 (best 0)
worse than median at Avg. Prec. :	25 (worst 2)	24 (worst 1)

Table 2: Comparison with Median at average precision

Table 2 compares our runs against the published median runs. We notice that for both runs the number of topics better and less than median are slightly the same.

Run-Id	P5	P10	P15	P30	Exact	Avg. Prec.
irit1men2a	0.3750	0.3250	0.2900	0.2433	0.1996	0.1519
irit2men2a	0.3950	0.3400	0.3017	0.2500	0.2284	0.1545

Table 3: Comparisons between the merging strategies

Table 3 compares the merging strategies. It can be seen that the naive strategy is slightly better than the normalised strategy in the top document, and at Exact precision but no difference at average precision. Nothing was gained from the normalised strategy.

The impact of the merging strategy.

Pair language	P5	P10	P15	P30	Exact	Avg. Prec.
E2F (34 queries)	0.2941	0.2118	0.1824	0.1353	0.2185	0.2046
E2G (37 queries)	0.2378	0.2189	0.1910	0.1396	0.1683	0.1489
E2I (34 queries)	0.1882	0.1647	0.1333	0.0843	0.1877	0.1891
E2E (33 queries)	0.5091	0.4212	0.3677	0.2798	0.4490	0.4611

Table 4: Results of pair search

Table 4 shows the results of pair language (example, E2F means English queries translated to French and compared to French documents, etc.). We can easily notice that the monolingual (E2E) search performs much more better than all the pair (E2F, E2G, E2I) searches. Moreover, all the pair searches (except E2G) have their average precision better than the best multilingual search. The merging strategy caused the loss of relevant documents, Table 5 shows the total number of relevant in the pair list and the numuber of document which was kept in the final list lost when merging. Relevant documents were lost from all the pair lists.

	E2E	E2F	E2I	E2G
Rel. Ret. by pair list	554	389	228	467
Rel. kept in the final list	500	281	152	296
Rel. lost.	54	107	76	171

Table 5: Comparison between the number of relevant in Pair and Multilingual lists

4.2 Bilingual experiment

The bilingual experiment was carried on using F2E free dictionary + desambiguisation. The desambiguisation was performed using WAC (Word-wide-web Aligned Corpus) parallel corpus built by RALI Lab (http://www-rali.iro.umontreal.ca/wac/).

irit1bfr2en	
better than median Avg. Prec. :	22 (best 3)
worse than median at Avg. Prec. :	11 (worst 2)

Table 6: Comparative bilingual F2E results at average precision

Table 6 compares our run against the published median runs. Most queries give results better than the median and 3 were the best.

Run-id (33 queries)	P5	P10	P15	P30	Exact	Avg. Prec.
Dico+Des.	0.3152	0.2636	0.2182	0.1636	0.2841	0.2906
Dico	0.2788	0.2515	0.2000	0.1566	0.2685	0.2741
Impr (%)	13	4.8	9	4.5	5.8	6

Table 7: Impact of the desambiguisation

Table 7 compares the results between the runs Dico+desambiguisation and Dico only. The desambiguisation is effective the average precision improves of 6%.

4.3 Monolingual experiments

Three runs were submitted in monolingual tasks : iritmonofr, iritmonoit, iritmonoge

First of all, we notice clearly that the monolingual search is much better than both the multilingual and the bilingual searches. Secondly, French monolingual results seem to be better than both Italian and the German. Italian results are better than German. These

Run-id	P5	P10	P15	P30	Exact	Avg. Prec.
iritmonofr FR (34 queries)	0.4765	0.4000	0.3510	0.2637	0.4422	0.4523
iritmonoit IT (34 queries)	0.4412	0.3324	0.2490	0.1637	0.4182	0.4198
iritmonoge GE (37 queries)	0.4108	0.3892	0.3550	0.2766	0.3197	0.3281

Table 8: Comparison between monolingual search

runs were done using exactly the same procedures the only difference concerns the stemming which was used only for French.

5 Acknowledgements

This work was in part supported by the EC through the 5th framework, Information Societies Technology programme (IRAIA Project, IST-1999-10602, http://iraia.diw.de).

References

[1] L. Ballesteros, W. Croft. *Resolving Ambiguity for Cross-Language Retrieval* in Proceedings of the 21st ACM SIGIR'98, pages, 64-71.

[2] M. Boughanem, C. Chrisment & C. Soule-Dupuy, *Query modification based on relevance backpropagation in Adhoc environment*, Information Processing and Managment. April 1999.

[3] M. Boughanem, T. Dkaki, J. Mothe & C. Soule-Dupuy, *Mercure at trec7*, Proceedings of the 7th International Conference on Text REtrieval TREC7, E. M. Voorhees and Harman D.K. (Ed.), NIST SP 500-236, Nov. 1997.

[4] S. Robertson and al *Okapi at TREC-6*, Proceedings of the 6th International Conference on Text REtrieval TREC6, Harman D.K. (Ed.), NIST SP 500-236, Nov. 1997.