Mercure at CLEF-2

N.NASSR, M.BOUGHANEM

IRIT/SIG
Campus Univ. Toulouse III
118, Route de Narbonne
F-31062 Toulouse Cedex 4
Email : {nassr, boughane}@irit.fr
Tel : 05-61-55-63-22 Fax: 05-61-55-63-23

1 Summary

This paper presents the experiments undertaken by our team (IRIT team) in multilingual, bilingual and monolingual tasks at CLEF programme. Our approach to CLIR is based on query translation. In bilingual experiment a dictionary is used to translate the queries from French to English and two techniques for desambiguiation were tested: aligned corpus and dictionary strategy. Desambiguiation technique is applied to select the best terms from the (translated) targed queries. All these experiments were done using Mercure system [2] which is presented in section 2 of this paper. The section 3 describes our general CLIR methodology, and finally, section 4 describes experiments and results performed at CLEF programme.

2 Mercure model

2.1 Model description

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 [1],[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[3] 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:

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

2.2 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:
   $In(t_i) = q_{ui}$ 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_u: the query length, (number of unique terms)
   qtf_{ui}: query term frequency.

3 General Clir Methodology

Our CLIR approach is based on query translation. It is illustrated by three main steps: Indexing, Translation and Dismabiguation described as follows:

- **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, Italian and Spanish words. The German, Italian and Spanish stoplists were downloaded from Internet.

- **Translation** : is based on "dictionaries". For the CLEF2 experiments, five 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.

Type	Source	nb. entries
E2F	http://www.freedict.com	42443
E2G	http://www.freedict.com	87951
E2I	http://www.freedict.com	13478
E2S	http://www.freedict.com	20700
F2E	http://www.freedict.com	35200

Table 1: Dictionaries characteristics

- **Desambiguiation**: when multiple translations exist for a given term, desambiguiation was performed by selecting the bests target query terms equivalent for each source query term.

Two strategies of desambiguiation were tested. The first one is based on aligned corpus and the second one is based on dictionary.

The first desambiguiation based on aligned corpus consist of:

1. Retrieving the top documents (X=20) for each source query term t_i in aligned corpus.
2. Retrieving the top documents (X'=20) for each translation t_{ij} for t_i in the same aligned corpus. t_{ij} is one translation for t_i among another.
3. Desambiguiation of the translated query consist of matching the retrieval documents (profiles) from the different translation against the source query profile. The best terms are the terms which have the best matchnig.

The second desambiguiation based on dictionary is described as follows :

1. Each source query term t_i is translated in target language using bilingual dictionary.
2. Each translation t_{ij} from t_i is transalted in source language using bilingual dictionary. t_{ij} is the one translation for t_1 among another.
3. The desambiguiation of the transalted query consist of retaining only target terms that return the source query term.

However if a specific term has an unique substitution this term is retained in all cases.

4 Experiment and Results

4.1 Multilingual experiment

One run iritmuEn2A using English topics and retrieving documents from the pool of documents in all four languages (German, French, Italian, Spanish and English), was submitted. The queries were translated using the downloaded dictionaries. No desambiguiation, all the translated words were retained in the target queries. The run was performed by doing individual runs for pair languages and merging the results to form the final ranked list.

Run-Id	P5	P10	P15	P30	Exact	Avg. Prec.
iritmuEn2A(50 queries)	0.4040	0.3520	0.3173	0.2760	0.1509	0.1039
Pair language	P5	P10	P15	P30	Exact	Avg. Prec.
E2F (49 queries)	0.2204	0.2102	0.1823	0.1415	0.2005	0.2044
E2S (49 queries)	0.3633	0.3265	0.3116	0.2537	0.2589	0.2281
E2I (47 queries)	0.1872	0.1596	0.1475	0.1255	0.1320	0.1321
E2E (47 queries)	0.5149	0.4085	0.3518	0.2716	0.4564	0.4863

Table 2: Comparison results of pair search and multilingual list

Table 2 shows the results of pair languages (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, E2S) searches. Moreover, all the pair searches have their average precision better than the multilingual search. The merging strategy caused the loss of relevant documents.

4.2 Bilingual experiment

Two runs using French topics and retrieving documents from the pool of document in English language, were submitted. The bilingual experiment was carried on using French to English free dictionary + desambiguiation. Two desambiguiation strategies were tested :

- Aligned corpus strategy : desambiguiation based on aligned corpus was performed using WAC (Word-wide-web Aligned Corpus) parallel corpus built by RALI Lab (http://www-rali.iro.umontreal.ca/wac/).
- (English-French) dictionary strategy: desmabiguiation based on dictionary was performed using free (English-French) dictionary. Table 1, shows the source and the number of entries in (English-French) dictionary.

Two runs were submitted: irit1bFr2En where desambiguiation based on dictionary and irit2bFr2En where desmabiguiation based on aligned corpus

Official results

Run-Id	P5	P10	P15	P30	Excat	Avg.Prec.
irit1bFr2En	0.3660	0.2979	0.2468	0.1844	0.3258	0.3294.
irit2bFr2En	0.3787	0.2957	0.2440	0.1794	0.3250	0.3398.

Table 3: Comparison between the desmabiguiation strategies

Table 3 compares the desambiguiation strategies. It can be seen that the desambiguiation based on aligned corpus is slightly better than the desambiguation based on dictionary at average precison but no difference at excat precision.

Non official results

Run-id (33 queries)	P5	P10	P15	P30	Exact	Avg.Prec
irit1bFr2En	0.2638	0.1915	0.1660	0.1312	0.2304	0.2375
Dico	0.3660	0.2936	0.2397	0.1809	0.3161	0.3305
Impr (%)	-27,92	-34.77	-30,74	-27.47	-27.11	-28.13
irit2bFr2En	0.3787	0.3043	0.2496	0.1851	0.3249	0.3436
Dico	0.3660	0.2936	0.2397	0.1809	0.3161	0.3305
Impr (%)	3.46	3.64	4.13	2.32	2.78	4

Table 4: Impact of the desambiguiation

Table 4 compares the results between the runs irit1bFr2En and irit2bFr2En (Dictionary+desambiguiation) and Dictionary only. It can be seen that the desambiguiation based on aligned corpus is better than the dictionary and the desambiguiation based on dictionary. The desambiguiation based on aligned corpus is effective the average precision improves of 4%.

4.3 Monolingual experiments

Four runs were submitted in monolingual tasks : iritmonoFR, iritmonoIT, iritmonoGE, iritmonoSP

Table 5 shows that French monolingual results seem to be better than both Italian, Spanish and the German. Italian results are better than Spanish and German. Spanish results are better than German. These runs were done using exactly the same procedures the only difference concerns the stemming which was used only for French. We notice clearly that the monolingual search is much better than both the multilingual and the bilingual searches.

Run-id (33 queries)	P5	P10	P15	P30	Exact	Avg. Prec.
iritmonoFR FR (49 queries)	0.4286	0.3898	0.3483	0.2830	0.3565	0.3700
iritmonoIT IT (47 queries)	0.4723	0.3894	0.3574	0.2730	0.3568	0.3491
iritmonoGE GE (49 queries)	0.4327	0.3816	0.3442	0.2884	0.2736	0.2632
iritmonoSP SP (49 queries)	0.4694	0.4347	0.4082	0.3626	0.3356	0.3459

Table 5: Comparison between monolingual search

5 Conclusion

In this paper we have presented, our experiments for CLIR at CLEF programme.
In multilingual IR, we showed that the merging strategy caused the loss of relevant documents, In bilingual IR, we showed that the desambiguiation technique based on aligned corpus for translated queries is effective. Results of experiments have also showed that using free dictionaries are fesaible, and desambiguiation based on aligned corpus give the good results even though the documents of aligned corpus are independent from those of database.

References

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

[2] 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.

[3] 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.