TREC-7 CLIR using a Probabilistic Translation Model

Jian-Yun Nie
Laboratoire RALI,
Département d'Informatique et Recherche opérationnelle,
Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal, Québec, H3C 3J7 Canada
nie@iro.umontreal.ca

In this report, we describe the approach we used in TREC-7 Cross-Language IR (CLIR) track. The approach is based on a probabilistic translation model estimated from a parallel training corpus (Canadian HANSARD). The problem of translating a query from a language to another (between French and English) becomes the problem of determining the most probable words that may appear in the translation of the query. In this paper, we will describe the principle of building the probabilistic model, and the runs we submitted using the model as a translation tool.

1. Introduction

For Cross-Language IR (CLIR) the solution that immediately comes to one's mind is to translate the information query using a machine translation (MT) system, and to submit the resulting translation to a classical monolingual IR system. In [Nie98], we compared this approach with the two following ones:

- using a bilingual dictionary;
- using a probabilistic translation model.

Our results on TREC-6 data showed that using a bilingual dictionary alone lead to poor performances; but using a probabilistic translation model, we obtained a performance close to those with commercial MT systems (LOGOS and SYSTRAN).

In TREC7, we used the same strategy. A probabilistic translation model is used to translate queries from a language to another (between English and French). The translation result is a list of words, together with a probability value. It is then submitted to a modified SMART system for retrieval.

Let us first give a brief description on how the probabilistic model is built, then we will describe our tests in Trec7.

2. A Probabilistic Translation Model

By translation model, we mean a mechanism which associates to each source language sentence (or query) **e** a probability distribution $p(\mathbf{f}|\mathbf{e})$ on the sentences (or queries) **f** of the target language. A precise description of a family of such models can be found in Brown & al. [Brown93]. The model we will be using for the experiments reported here is basically their "Model 1". In this model, a source **e** and its translation **f** are connected through an alignment **a**, that is a mapping of the words of **e** onto those of **f**. If $\mathbf{e} = e_1, e_2, \ldots, e_l$ and $\mathbf{f} = f_1, f_2, \ldots, f_m$ then $\mathbf{a}_j$ will be used to refer to the particular position in **e** that is connected with position j in **f** (for example, $\mathbf{a}_2 = 4$ expresses the fact that f_2 is connected with e_4) and e_{aj} will be used to refer to the word in **e** at position $\mathbf{a}_j$.

The probability $p(\mathbf{f}|\mathbf{e})$ is decomposed as a sum over all possible alignments:

$$p(\mathbf{f}|\mathbf{e}) = \Sigma_{\mathbf{a}\in \mathbf{A}} p(\mathbf{f}, \mathbf{a}|\mathbf{e})$$

The conditional probability of **f** under alignment **a** given **e** can be analyzed as follows:

$$p(\mathbf{f},\mathbf{a}|\mathbf{e}) = p(\mathbf{f}|\mathbf{a},\mathbf{e})\ p(\mathbf{a}|\mathbf{e}) = K_{\mathbf{e},\mathbf{f}}\ p(\mathbf{f}|\mathbf{a},\mathbf{e})$$

The latter equality stems from the fact that in model 1, all alignments are considered equiprobable. Consequently $p(\mathbf{a}|\mathbf{e})$ is a constant $K_{\mathbf{e},\mathbf{f}}$ equal to 1 over the total number of alignments.

The core of the model is $t(f|e)$, the lexical probability that some word e is translated as word f. The value of $p(\mathbf{f}|\mathbf{a},\mathbf{e})$ depends mostly on the product of the lexical probabilities of each word pair connected by the alignment:

$$p(\mathbf{f}|\mathbf{a},\mathbf{e}) = C_{\mathrm{f,e}} \prod_{j=1,m} t(f_j|e_{aj})$$

where $C_{\mathbf{f},\mathbf{e}}$ is a constant that accounts for certain dependencies between the respective lengths of sentences **e** and **f** (mostly irrelevant here).

The probability of observing word f_j in **f** under a particular alignment **a** is:

$$p(f_{j,}|\mathbf{a},\mathbf{e}) = t(f_j|e_{aj})$$

And the probability of observing word f_j in **f** under any alignment is:

$$p(f_j|\mathbf{e}) = \Sigma_{i=1,l}\ t(f_j|e_i)$$

Since all alignments are considered equiprobable, we can simply sum up the values obtained by connecting f_j to each word e_1, e_2, …, e_l of **e**. In other words, the probability of observing a particular word in a given position in **f** is established as the total of the lexical contributions of each word of **e**.

The parameters of our translation model are estimated from a bilingual parallel corpus in which each sentence has been aligned with the corresponding sentence(s) of the other language. Such alignments can be produced using algorithms such as the one described in [Simard92]. Given such alignments we can estimate reasonable values for the parameters $t(f|e)$ using the Expectation Maximization algorithm, as described in [Brown93]. The model used in the experiments reported here has been trained using 8 years of the Canadian Hansard (parliamentary debates), that is, approximately 50 million words in English and in French.

We noticed in [Nie98] that the probabilistic model cannot distinguish true translation words from those only statistically associated, in particular, when the source words have low occurrence frequency in the training corpus. In order to solve this problem, we enforced, in a query translation, the “probability” of the words that are recognized as translations of some query words in a bilingual dictionary. This leads to a combined approach. Our experiments with TREC-6 data showed that this combination is very effective. In general, we obtained about 5% increase in average precision over the approach using the probabilistic model alone. On TREC-6 data, we used a small bilingual dictionary with less than 8000 words. It is showed that when the rate of enforcement was set at 0.02 we obtained the best performance. For TREC-7 experiments, we used a larger bilingual dictionary (a terminology database) with about 1.2 million entries (most of them are compound terms). The enforcement rate has been set at 0.01 because we have now much more translations added in.

3. Experiments

We used a modified version of SMART system [Buckley85] for monolingual document indexing and retrieval. The *ltc* weighting scheme is used for documents. For queries, we used the probabilities provided by the probabilistic model, multiplied by the *idf* factor. From the translation words obtained, we retained the 50 most probable words. This limit in number allows us to eliminate many noisy words in the translation that are simply statistically related to query words. The setting of 50 has been shown to be reasonable on Trec6.

Before indexing, a text is first stemmed as follows: According to a probabilistic tagging, each word is first associated with a (or several) grammatical category. It is then transformed into a canonical, citation form. For example, nouns and (French) adjectives are transformed into their masculine singular form, and verbs are transformed into their infinitive forms.

Our initial goal of participating in TREC-7 is to re-evaluate how effective the cross-language IR based on the probabilistic translation model is. So, we first submitted the following 4 runs:

- RaliAPf2e: Using French queries to retrieve AP English documents. This run only uses the probabilistic model;
- RaliDicAPf2e: The same as above, but it combines the probabilistic model and the bilingual dictionary.
- RaliSDAe2f: Using English queries to retrieve SDA French documents. This run only uses the probabilistic translation model.
- RaliDicSDAef: The same as above, but using the combined approach.

Later on, we also submitted two other runs in which SDA French documents and AP English documents are merged.

- RaliDicE2EF: Using English queries to retrieve English and French.
- RaliDicF2EF: Using French queries to retrieve English and French documents.

In these two runs both the probabilistic model and the bilingual dictionary are used.

<u>Simple CLIR</u>

The monolingual runs are performed using *ltc-ltc* weighting with SMART. The CLIR runs used *mtc-ltc* weighting. Table 1 shows the performances obtained in comparison with the monolingual runs on the same collection.

As we can see, the CLIR effectiveness is comparable to the monolingual runs. This is quite surprising because on TREC-6 data, the same approach led to performances of about 80% of the monolingual runs. What is even more surprising is the better performances obtained in French to English CLIR, than the English to English monolingual run. A possible explanation, in addition to some slight differences between the original English and French queries, is that the probabilistic translation allows us to include some very useful related words or synonyms. This phenomenon has been observed in a number of queries.

	Mono (F-F)	RaliDic	Rali
Rel.	991		
Rel.Ret.	806	792	784
0.0	0.6559	0.5983	0.5653
0.1	0.4829	0.4481	0.4598
0.2	0.4456	0.3973	0.3980
0.3	0.3808	0.3310	0.3118
0.4	0.3168	0.3003	0.2906
0.5	0.2623	0.2623	0.2563
0.6	0.1930	0.2010	0.1946
0.7	0.1544	0.1684	0.1597
0.8	0.1156	0.1390	0.1330
0.9	0.0726	0.0723	0.0763
1.0	0.0100	0.0159	0.0217
Avg.prec.	0.2658	0.2551	0.2491

SDA English to French retrieval (E-F)

	Mono (E-E)	RaliDic	Rali
Rel.	1689		
Rel.Ret.	1231	1381	1416
0.0	0.6128	0.6441	0.6543
0.1	0.4552	0.5042	0.5343
0.2	0.3987	0.4395	0.4549
0.3	0.3696	0.4155	0.4209
0.4	0.3418	0.3949	0.3943
0.5	0.3060	0.3362	0.3399
0.6	0.2487	0.2709	0.2796
0.7	0.2242	0.2365	0.2434
0.8	0.1938	0.1970	0.2046
0.9	0.1277	0.1495	0.1448
1.0	0.0699	0.0741	0.0727
Avg.prec.	0.2864	0.3186	0.3229

AP French to English retrieval (F-E)

Table 1. English to French and French to English runs.

Let us illustrate this by the following examples.

Query 30: Famine in Sudan
famine=0.154774
soudan=0.129183
étude=0.075273
étudier=0.023295
sévir=0.010796
pouvoir=0.010070
victime=0.007599
présenter=0.007366
port-soudan=0.006182
soudanais=0.006059
effectuer=0.005955
pressant=0.005752
trois=0.005726
secours=0.005652
seulement=0.004535
publier=0.004400
lutter=0.004091
signaler=0.003660

query 40: Concorde Supersonic Jet
français=0.042530
développement=0.037197
avion=0.033672
supersonique=0.030150
concorde=0.029113
réaction=0.027248
colombie-britannique=0.026678
pouvoir=0.014754
coopératif=0.013521
opération=0.012910
utiliser=0.010497
identifier=0.010412
activité=0.009597
question=0.009530
jet=0.009523
venu=0.009260
concorder=0.009003
britannique=0.008239

We observe that the top-ranked French words found for these queries are highly relevant to the original English queries. Some related words are also found. For example, “victime”, “port-soudan” and “soudanais” in query 40, and “avion”, “réaction” in addition of the true translation “supersonique” and “jet”.

However, we can notice several translation problems.

- Some non-significant common words such as “pouvoir” (can) have been included in a number of translations. These words, however, cannot be put in the stop list because they are meaningful in some cases (“pouvoir” may also mean “power”). This problem can be partly solved by including *idf* factor in the final weighting. In the final vector obtained with *mtc* weighting, the word “pouvoir” appears at 26th rank.
- Due to the particularities of the training parallel corpus, the word “British” in query 40 (in the description field) is translated first by “colombie-britannique” with a much higher probability than “britannique”. This phenomenon caused more problems than the previous one because *idf* cannot decrease their importance in the final vector. In the final vector, “colombie-britannique” is the 5th most important term.
- Many unrelated words appear in the translation because they occur often in a sentence that is aligned with one containing a word of the original query. For example, we can notice “effectuer” (carry out) in the translation of query 30, and “activité” (activity) in that of query 40.

In order to compare with an MT system, we translated the queries with the Systran system. The translated queries processed as in the monolingual runs. The following table shows the performances obtained.

	E-F	F-E
Trec7	0.2206	0.3185

Table 2. Average precision using MT

We can see that the probabilistic translation model performed slightly better than the Systran system under the same conditions. This confirms the same conclusion we drawn in [Nie98] using the Trec6 data.

Merging runs

Our emphasis in this Trec CLIR track has been put on simple CLIR without merging. The merging run has been submitted at the last minute. We did not spend much time to define a reasonable merging strategy. We used a very simple approach: The original queries (English or French) are used to retrieve documents in the same language from one of the two collections (AP or SDA), and the translated queries are used to retrieve documents in the other collection. Retrieved documents from the two collections are re-ranked according to their similarities to the queries.

The problem we were facing with is that the similarities obtained in monolingual IR and CLIR are not comparable. Words in vectors are weighted in very different ways. In monolingual runs, the SMART’s *ltc* scheme is used, whereas in the CLIR runs, the weight is a combination of translation probability and *idf*. The direct merging of the two document lists resulted in a very unbalanced ranking of AP and SDA documents: Either we have many AP documents at the top level, or the SDA documents at the top level.

In order to solve partly the incompatibility of similarities, we chose to use *mtc* for queries and *ltc* for documents in both cases. The documents from the two runs seem to be more balanced in the merged result, although not completely. Typically, we still observed that the similarities in the monolingual answer set are more distanced between the top and bottom than in the CLIR answer set.

Table 2 shows a comparison of these two merging runs with other runs in this category.

		Rel. Ret. @ 1000					Avg. Prec.				
Topic	Rel.	Best	Med.	Worst	E-EF	F-EF	Best	Median	Worst	E-EF	F-EF
26	12	11	9	0	9	7	0.1200	0.0342	0.0000	0.1019 (>)	0.0947
27	84	80	42	10	49	51	0.3200	0.0911	0.0133	0.1506 (>)	0.1428
28	157	147	118	18	118	126	0.6815	0.3748	0.0148	0.3748 (=)	0.3863
29	19	19	12	2	12	11	0.9060	0.5824	0.0005	0.5335 (<)	0.4357
30	133	133	130	12	133	133	0.5784	0.4053	0.0111	0.5784 (B)	0.6521
31	227	202	165	62	194	190	0.4660	0.3548	0.1029	0.3526 (<)	0.3543
32	57	57	56	12	56	56	0.8428	0.7565	0.0117	0.7469 (<)	0.6779
33	87	83	66	10	40	48	0.6516	0.2657	0.0158	0.0501 (<)	0.0918
34	11	11	11	2	11	6	0.1218	0.0341	0.0013	0.0341 (=)	0.0157
35	74	60	46	13	46	32	0.1520	0.0975	0.0078	0.0763 (<)	0.0337
36	114	108	84	15	84	79	0.6559	0.3349	0.0091	0.2560 (<)	0.1497
37	44	37	14	0	14	19	0.3675	0.0247	0.0000	0.0115 (<)	0.0179
38	147	144	135	16	135	138	0.6794	0.4239	0.0027	0.5330 (>)	0.5413
39	35	32	16	2	32	30	0.1223	0.0609	0.0012	0.0500 (<)	0.1432
40	43	43	38	1	42	41	0.7890	0.5626	0.0000	0.7890 (B)	0.5999
41	290	277	239	4	239	222	0.7337	0.4104	0.0001	0.4104 (=)	0.4165
42	55	50	31	6	41	39	0.3246	0.0622	0.0288	0.0604 (<)	0.1078
43	242	142	96	6	7	137	0.2112	0.0549	0.0005	0.0005 (W)	0.1976
44	6	6	4	1	4	4	0.4095	0.2565	0.0003	0.2711 (>)	0.1826
45	47	47	45	6	23	22	0.7028	0.3158	0.0010	0.2114 (<)	0.2653
46	2	2	1	0	2	1	0.0083	0.0026	0.0000	0.0026 (=)	0.0006
47	140	139	136	25	136	137	0.6186	0.3568	0.0331	0.3568 (=)	0.5304
48	101	93	48	15	69	77	0.6608	0.1277	0.0232	0.1277 (=)	0.3266
49	130	109	100	8	102	94	0.4673	0.1905	0.0004	0.1905 (=)	0.1506
50	216	158	116	10	119	135	0.3943	0.1529	0.0042	0.1529 (=)	0.1736
51	43	42	39	14	41	42	0.7903	0.5803	0.0366	0.6307 (>)	0.5110
52	51	49	33	6	33	33	0.5317	0.1429	0.0012	0.0952 (<)	0.0838
53	113	36	17	2	7	50	0.0575	0.0060	0.0001	0.0007 (<)	0.0569
Avg.	92.41	79.9	63.69	9.59	62	67.59	0.4609	0.2435	0.0111	0.2465	0.2622

Runs : E-EF = RaliDicE2EF,
F-EF = RaliDicF2EF

Table 3. Merging runs with English and French documents

For the E2EF run, the comparison with other runs is shown in the following table. The average precision for all the queries is about the same as the median.

Best	> median	= median	< median	Worst
2	5	8	12	1

Table 4. Comparison with other participants

Although the merge run on English and French documents using French queries is not an official category, we also provide this run in the above table in order to compare with the CLIR run using English queries. In the French to English/French run, we obtained slightly better average precision on all the queries.

The difference between the two runs is the sharpest for query 43. After analyzing the query, we found that the poor performance in E-EF run was due to a mistake in manipulating the original query. The original query has been wrongly altered, so that the monolingual retrieval did not find any relevant document for this query. After correcting the situation, we obtained an average precision of 0.1636 for this query in monolingual run, and 0.1472 in the merge run. This is above the median level.

The medium performance of the merge run is not surprising to us. In choosing *mtc* weighting for monolingual run, we knew that the effectiveness will drop (this has been tested on Trec6 data). This, in addition to the still unbalanced ranking of SDA and AP documents in the final list, greatly affected the merge run.

4. Final remarks

Our participation to the TREC-7 CLIR track is to verify the effectiveness of our approach using a probabilistic translation model. Our previous experiments with TREC-6 data [Nie98] showed that CLIR using this approach may match and even surpass that using commercial MT systems. The tests in Trec7 confirmed this once more. However, there are several problems in the translation model used. We will try to improve the model and its application to CLIR in the future.

In comparison with the best performances of CLIR, our results are still low. The main reason lies in the global setting of the system. The weighting schemes we used are not the most effective. In the future, we will try to use better weighting scheme such as *ltu* or *Okapi* formula. Despite this, our comparison with the monolingual runs and the runs using Systran still hold. They are carried out under the same condition. So we expect to have the same comparison with new weighting schemes or other system setting.

References

[Brown93] P. F. Brown, S. A. D. Pietra, V. D. J. Pietra, and R. L. Mercer, The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, vol. 19, pp. 263-312 (1993).

[Buckley85] C. Buckley, *Implementation of the SMART information retrieval system*. Cornell University, Technical report 85-686, (1985).

[Nie98] J.Y. Nie, P. Isabelle, P. Plamondon, G. Foster, Using a probabilistic translation model for cross-language information retrieval, *Sixth workshop on Very Large Corpora*, Montreal, pp. 18-27, (1998)

[Simard92] M. Simard, G. Foster, P. Isabelle,Using Cognates to Align Sentences in Parallel Corpora, *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal (1992).