Integration of PLSA into Probabilistic CLIR Model
— Yokohama National University at NTCIR4 CLIR —

Tetsu MURAMATSU and Tatsunori MORI
Graduate School of Environment and Information Sciences, Yokohama National University
79-7 Tokiwadai, Hodogaya, Yokohama 240-8501, Japan
{tetsu0730,mori}@forest.eis.ynu.ac.jp

Abstract

In this paper, we propose a method of Cross-Language Information Retrieval based on an integration of a probabilistic CLIR model and Probabilistic Latent Semantic Analysis (PLSA). PLSA is adopted to extract the information of translation probability from a parallel corpus. The information is utilized in a probabilistic CLIR model. Although the probabilistic CLIR model with PLSA is quite effective, it takes very long time in the processing. We therefore introduce an approximation method based on a two-phased retrieval model in order to reduce the computational cost.

Using the model, we submitted runs for Japanese-to-English bilingual retrieval in CLIR task of NTCIR4.
Keywords: *PLSA, Probabilistic CLIR model, Two-phased retrieval.*

1 Introduction

In CLIR, the language of query is different from the language of documents. To cope with the issue, *query translation* is usually performed to reduce CLIR to a mono-lingual IR.

Here, there are two main problems in the CLIR based on query translation. First one is how to select a suitable translation. Although the translation candidate of a certain word can be obtained from a translation dictionary, we need additional information to select a suitable translation from the candidates. In order to disambiguate translation, many previous works introduced techniques using information of translation probability[5]. Takano et al.[6] proposed a method to obtain translation probability accurately by introducing unobserved variables into probabilistic model using Probabilistic Latent Semantic Analysis (PLSA)[2, 3].

The second problem is how to use information of translation probability in retrieval models. Although we can pick up one translation from candidate lists according to translation probability, the probabilistic information of other candidates is wastefully discarded after the selection. To solve the problem, there are several proposals of probabilistic retrieval model that utilize all of information of translation probability in retrieval process[8, 1].

Based on the above discussion, we propose a method of Cross-Language Information Retrieval based on an integration of a probabilistic CLIR model and PLSA. In this model, PLSA is firstly applied to a parallel corpus in order to extract the information of translation probability. Then, the information is utilized in a probabilistic CLIR model.

The probabilistic CLIR model with PLSA is quite effective, but we found in NTCIR4 CLIR formal run that it takes very long time in the processing. We therefore propose an approximation method based on a two-phased retrieval model in order to reduce the computational cost. In the two-phased retrieval, the system firstly retrieves a certain amount of documents using simple non-probabilistic translation of query and the vector space model. Then it re-ranks the set of documents according to the proposed method.

2 System Description

2.1 Retrieval Model

We adopt the probabilistic retrieval model proposed by Xu et al.[8] in order to used information of translation probability effectively.

In the model, the similarity between a query and a document is regarded as a probability that a set of query words are observed in the translation of document as follows:

$$P(Q|d) \quad = \quad \prod_{t \in Q} \left[\alpha P(t|Mt) + (1-\alpha) \sum_{t \in d} P(t|d) \right] \tag{1}$$

where d is a document, Q of a set of query words, t is one of query words, $P(t|d)$ is the probability of

observing the word t in the translation of d, Mt is the model of the language of query, and $P(t|Mt)$ is the probability that the word t is observed in the language of query. Note that the first term in Formula (1) is introduced in order to cope with the zero frequency problem.

The term $p(t|d)$ in Formula (1) can be rewrite by introducing translation probability as follows.

$$P(t|d) \quad = \quad P(t|e)P(e|d) \tag{2}$$

where e is a word in the document d and $P(t|e)$ is the probability that the word e is translated into a word t. We finally obtain the following formula:

$$P(Q|d) = \prod_{t \in Q} \left[\alpha P(t) + (1-\alpha) \sum_{e \in d} P(t|e)P(e|d) \right] \tag{3}$$

In document retrieval, the set of documents are ranked in descending order of the probability of Formula (3) with respect to a set of query words.

2.2 Translation Probability

In this paper, the translation probability $P(t|e)$ is calculated from co-occurrence probability $P(t,e)$ that is estimated from a parallel corpus by using PLSA. If we have the information of $P(t,e)$, then the probability $P(t|e)$ can be derived as follows:

$$P(t|e) \quad = \quad \beta \frac{P(t,e)}{\sum_{t' \in T(e)} P(t',e)} + (1-\beta) \frac{P(t|M_t)}{\sum_{t' \in T(e)} P(t'|M_t)} \tag{4}$$

where $T(e)$ is a list of translation candidates of the word e.

While the first term corresponds to a translation probability obtained from a parallel corpus, the second term represents a default probability of translation based on the language model of the word t. The second term is necessary in order to estimate the translation probability of words that do not appear in the corpus.

2.2.1 Calculation of Co-occurrence Probability using PLSA

In our previous work[6], we proposed a method to extract translation probability from a parallel corpus by using PLSA. In the method, the probability $P(t,e)$ of co-occurrence of two words is indirectly estimated from the probability $P(w,d)$ of occurrence of a word, which can be derived from a document-word matrix by PLSA. In this paper, we try to calculate $P(t,e)$ directly from a word-word co-occurrence matrix using PLSA.

Firstly, a word-word co-occurrence matrix is obtained from a parallel corpus that consists of pairs of aligned sentences. Each of columns and rows of the matrix corresponds to one of word, and the element is the frequency of co-occurrence of two words. Since an aligned pair of sentences is regarded as a window of co-occurrence, words from different languages may co-occur in the matrix. We use 180,000 pairs of aligned sentences of Japanese articles from 'Yomiuri Shimbun Newspaper' and English articles from 'The Daily Yomiuri'[7].

Secondly, we calculate a co-occurrence probability $P(w_e, w_j)$ of two words w_e and w_j using PLSA from the co-occurrence matrix. According to PLSA, $P(w_e, w_j)$ is represented by introducing a set of unobserved variables as follows.

$$P(w_j, w_e) = \sum_{z \in Z} P(w_j|z)P(w_e|z)P(z) \tag{5}$$

The probabilities $P(w_j|z)$, $P(w_e|z)$ and $P(z)$ can be calculated from the co-occurrence matrix by using an EM algorithm in which the E-step:

$$P(z|w_j, w_e) = \frac{P(z)P(w_j|z)P(w_e|z)}{\sum_{z'} P(z')P(w_j|z')P(w_e|z')} \tag{6}$$

and the M-step:

$$P(w_j|z) \quad = \quad \frac{\sum_{w_e} n(w_j, w_e)P(z|w_j, w_e)}{\sum_{w'_j, w_e} n(w'_j, w_e)P(z|w'_j, w_e)} \tag{7}$$

$$P(w_e|z) \quad = \quad \frac{\sum_{w_j} n(w_j, w_e)P(z|w_j, w_e)}{\sum_{w_j, w'_e} n(w_j, w'_e)P(z|w_j, w'_e)} \tag{8}$$

$$P(z) \quad = \quad \frac{\sum_{w_j, w_e} n(w_j, w_e)P(z|w_j, w_e)}{\sum_{w_j, w_e} n(w_j, w_e)} \tag{9}$$

are iteratively calculated, where $n(w_j, w_e)$ is the number of co-occurrence of w_j and w_i in the co-occurrence matrix.

2.2.2 Calculation of Language model $P(t|M_t)$

In order to calculate $P(t|M_t)$, we use the documents of NTCIR3 corpus, Mainichi Shimbun Newspaper Articles, and the parallel corpus.

2.3 Two-phased Retrieval Model

In the formal runs, we tried to apply our model to the whole of document database. It however took very long time in processing, because the system based on the model has to calculate the probability $P(Q|d)$ for all documents. Moreover, the calculation of $P(Q|d)$ itself is also expensive, because the co-occurrence probability for all of combination of query words and words in a documents has to be computed.

In order to reduce the cost, we introduce a two-phased retrieval method after the formal run. In the two-phased retrieval, the system firstly retrieves

a certain amount of documents using simple non-probabilistic translation of query and the vector space model. Then it re-ranks the set of documents according to the proposed method. The process is detailed as follows.

1. Obtain a list of translation candidates for each query word by looking up a translation dictionary.

2. Concatenate those lists to make a list of translated words.

3. Submit the list of translated words to a monolingual IR system to retrieve n-best documents.

4. Calculate the probability $P(Q|d)$ for each of retrieved document.

5. Re-rank the n documents according to the probability.

2.4 Indexing and Dictionary

As a translation dictionary, we adopt EDR bilingual dictionary[4]. In the formal run, namely, under the condition of Japanese-to-English bilingual retrieval with the probabilistic model, English-to-Japanese translation was only necessary to perform retrieval, because words in English documents are translated into Japanese words. Thus, we only used words in the EDR English-Japanese dictionary for indexing. More precisely, as for English words to be indexed, we select the English nouns that appear in the EDR English-Japanese dictionary as entry words. As for Japanese words to be indexed, translation candidates in the EDR English-Japanese dictionary is used.

On the other hand, in the two-phased model, we have to use not only English-Japanese dictionary, but also Japanese-English dictionary as follows, because the system needs to translate Japanese query words into English in the first phase.

- Japanese words to be indexed
  - Nouns of entry words in the EDR Japanese-English dictionary
  - Translation candidates of English nouns in the EDR English-Japanese dictionary
- English words to be indexed
  - Nouns of entry words in the EDR English-Japanese dictionary
  - Translation candidates of Japanese nouns in the EDR Japanese-English dictionary

3 Experiment

We participated in T-runs and D-runs of Japanese-to-English Bilingual Language Information Retrieval (BLIR) in the CLIR task of NTCIR4. In the formal run, we applied our probabilistic retrieval model to whole of document database, and did not use the two-phased retrieval model.

In addition to that, we conducted some additional experiments with the two-phased retrieval model. The condition of experiments is detailed as follows.

- The experiment with the original probabilistic retrieval model (Formal runs)
  - Number of unobserved variables: $N_z = 500$
  - The EDR English-Japanese dictionary is only used as a dictionary.
- The experiment with the two-phased retrieval model (Additional runs)
  - Number of unobserved variables: $N_z = 500, 900$
  - Number of documents to be retrieved in the first phase: $N_d = 1000, 2000, 3000, 4000$.
  - The EDR English-Japanese and the Japanese-English dictionaries are used.

4 Results and Discussion

The results of experiments are shown in Figures 1, 2 and Table 1. Table 2 shows the results of the formal runs and the results with the two-phased retrieval model. In those figures and tables, the label 'Formal-run' expresses the results of two runs, FORES-J-E-T-03 and FORES-J-E-D-01. The label 'merged' means that the merged dictionary is used, and the label 'no merged' means that each of EDR dictionaries used as it is.

As shown in Figures 1 and 2, the two-phased retrieval model with the merged dictionary is the most accurate. While the original probabilistic retrieval model takes more than one hour per query, the two-phased model can reduce the processing time to about 30 seconds(in the case of $N_d = 1000$). As for the number of unobserved variables, N_z , the effectiveness does not degrade so much even if we decrease N_z. It means that $N_z = 500$ is enough to obtain translation probability, at least for the parallel corpus.

However, the precisions of the proposed system is still lower than the average of all participants. As shown in Table 2, the values of precisions are widely diverse query by query. While the system achieved more than 0.3 Average Precision for seven queries, for five queries it could not retrieve any correct document

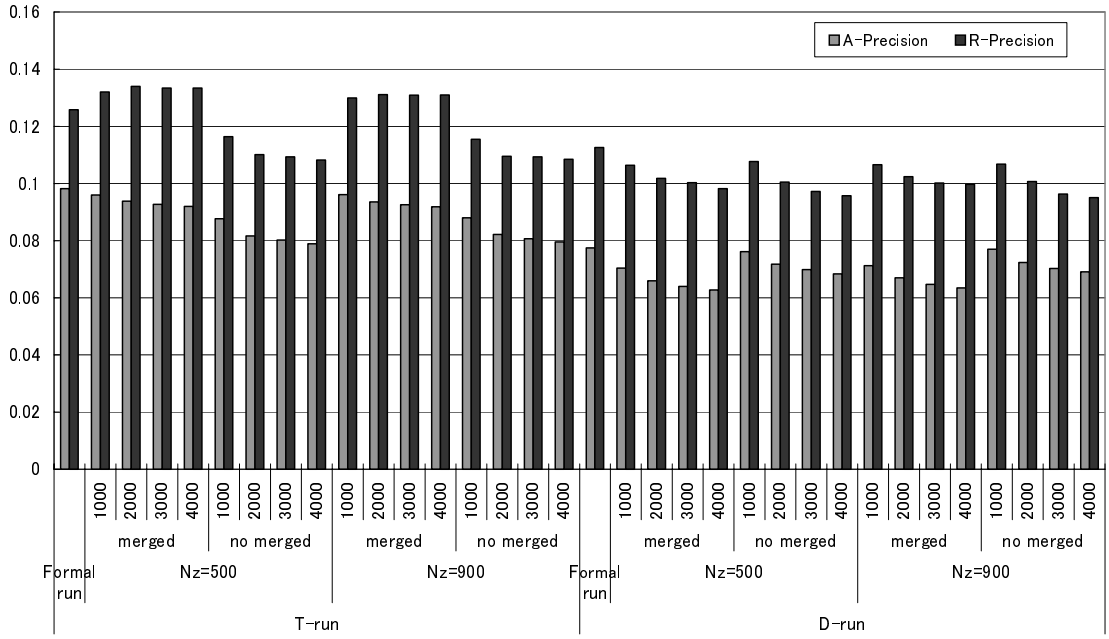

Figure 1. The Result of Rigid

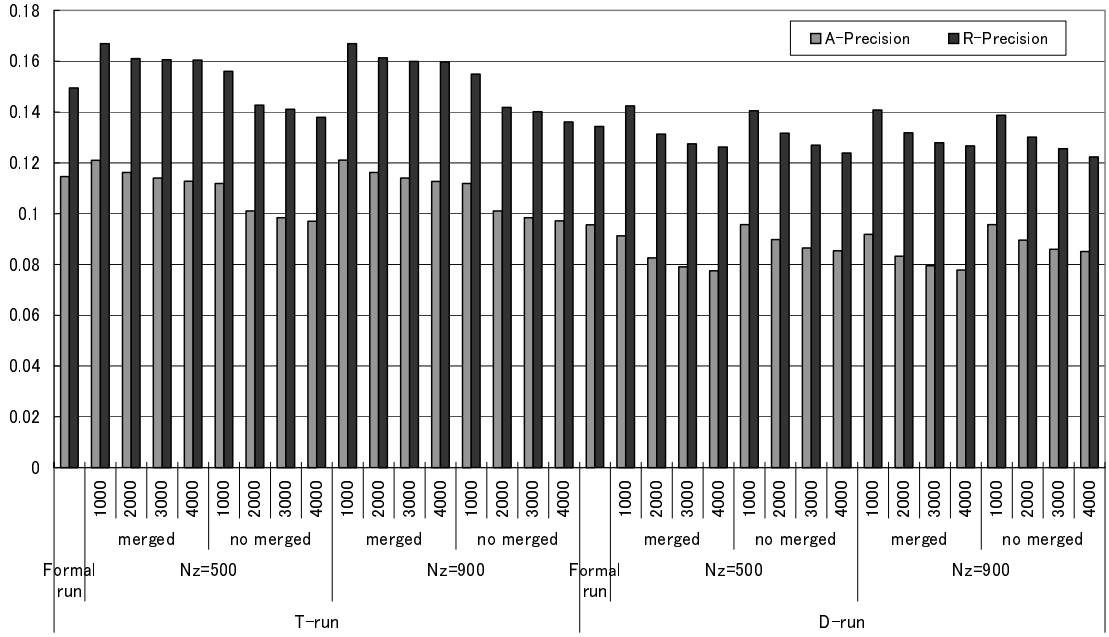

Figure 2. The Result of Relax

at all. In our experiment, we use only general translation dictionary. We have to use some specific dictionaries that have proper noun. Some mechanism for transliteration is also necessary to improve the effectiveness.

5 Conclusion

In this paper, we proposed a method of Cross-Language Information Retrieval based on an integration of a probabilistic CLIR model and PLSA. We also introduced an approximation method based on a two-phased retrieval model in order to reduce the computational cost. In the experiment of Japanese-to-English bilingual retrieval in CLIR task of NTCIR4, we showed the effectiveness of the two-phased retrieval model.

On the other hand, it is a problem that retrieval accuracy is still low. In our future work, we would like to consider the improvement of word translation using some proper noun dictionaries.

References

[1] M. Franz and J. S. McCarley. Arabic information retrieval at IBM. In *Proceedings of the eleventh Text Retrieval Conferene (TREC 2002)*, 2002.

[2] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of SIGIR '99: 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, 1999.

[3] T. Hofmann. Probablistic latent semantic analysis. In *Proceedings of the fifteenth Conference on Uncertainty in Artificial Intelligence (UAI '99)*, pages 289–269, 1999.

[4] Japan Electrronic Dictionary Research Institute. *EDR Electronic Dictionary Specification*, 1995.

[5] G. Kikui. Retrieving Documents Across Language-Barriers. *Journal of Japanese Society for Artificial Intelligence*, 15(4):550–558, July 2000. (in Japanese).

[6] Y. Takano, T. Muramatsu, and T. Mori. Cross-language information retrieval using segmented plsi. In *Proceedings of the 9th annual meeting of the association for Natural Language Processing, Japan*, pages 389–392, Mar. 2003.

[7] M. Utiyama and H. Isahara. Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, pages 72–79, 2003.

[8] J. Xu, R. Weischedel, and C. Nguyen. Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of SIGIR '01: 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 105–110, 2001.

Table 1. The Result with Two-phased Retrieval Model ($N_d = 1000$)

			Rigid		Relax		time
			A-Precision	R-Precision	A-Precision	R-Precision	(sec./query)
T-run	FORES-J-E-T-03		0.0982	0.1258	0.1146	0.1495	>1 hour
	Z=500	merged	0.096	0.132	0.121	0.1669	34.3
		no merged	0.0877	0.1164	0.1119	0.156	20.2
	Z=900	merged	0.0961	0.1299	0.1211	0.1669	29.3
		no merged	0.088	0.1155	0.1119	0.1549	23.2
D-run	FORES-J-E-D-01		0.0775	0.1126	0.0956	0.1343	>1 hour
	Z=500	merged	0.0704	0.1064	0.0913	0.1424	39.2
		no merged	0.0762	0.1077	0.0957	0.1405	28.0
	Z=900	merged	0.0713	0.1066	0.0919	0.1408	42.3
		no merged	0.077	0.1068	0.0957	0.1387	30.4

Table 2. Result of the run FORES-J-E-T-03

query	Relax		Rigid		query	Relax		Rigid	
number	A-Pre	R-Pre	A-Pre	R-Pre	number	A-Pre	R-Pre	A-Pre	R-Pre
2	0.0133	0	0.0104	0	32	0.0004	0	0.0006	0
3	0.0463	0.0833	0.0504	0.0714	33	0.0134	0.0483	0.0006	0
5	0.1021	0.2	0.0894	0.2143	34	0.0204	0.0192	0.0132	0.0345
6	0.0488	0	0.0347	0	35	0.0303	0.092	0.0164	0.0571
7	0.1113	0.1905	0.0994	0.1176	36	0.7885	0.7487	0.7265	0.7163
9	0.3198	0.3571	0.1291	0	37	0.0387	0.0864	0.0254	0.0377
10	0.0119	0	0.0101	0	39	0.2227	0.3265	0.1313	0.1889
11	0.1875	0.2258	0.1691	0.2143	40	0.0039	0	0.0029	0
12	0.0007	0	0.0008	0	41	0.0483	0.0989	0.0405	0.0789
13	0.0218	0.087	0.026	0.1176	42	0	0	0	0
14	0.2151	0.2955	0.2668	0.3636	43	0.0049	0.0286	0.0056	0.0323
15	0.2337	0.3625	0.1937	0.2865	45	0.0083	0.0535	0.0026	0.0127
16	0.053	0.1049	0.0245	0.0636	46	0.3255	0.375	0.2408	0.3056
17	0.2357	0.2682	0.0986	0.1053	47	0.1315	0.2612	0.1113	0.2227
18	0.0002	0.0114	0	0	48	0.7506	0.6744	0.6448	0.6182
19	0.0022	0.0295	0.0012	0.0198	49	0.0462	0.1484	0.0387	0.1304
20	0	0	0	0	50	0.3248	0.3797	0.3849	0.3911
21	0.1878	0.2574	0.2865	0.383	51	0.0237	0.0709	0.0218	0.0672
22	0.0017	0	0.0017	0	52	0.3045	0.575	0.129	0.1739
23	0.0434	0.0833	0.0258	0.0909	53	0.0652	0.1414	0.2445	0.381
24	0.1732	0.1839	0.1274	0.2647	54	0.0002	0.0082	0	0
25	0.0003	0	0.0002	0	55	0	0	0	0
26	0.0239	0.0519	0.0374	0.0294	56	0.2968	0.3608	0.1975	0.23
27	0	0	0	0	57	0.0092	0.0506	0.007	0.0578
28	0.0151	0.0526	0.0082	0.0714	58	0	0	0	0
29	0.01	0.0182	0.0079	0	59	0.2616	0.2237	0.1631	0.1728
30	0.5226	0.5667	0.5532	0.5893	60	0	0.0035	0.0001	0
31	0.0009	0.0192	0.0001	0.0051	total	0.1146	0.1495	0.0982	0.1258