Prediction of Performance on Cross-lingual Information Retrieval by Regression Models

Kazuaki Kishida*, Kazuko Kuriyama**, Noriko Kando[+], Koji Eguchi[+]

*Faculty of Cultural Information Resources, Surugadai University / National Institute of Informatics (NII)
Hanno, Saitama 357-8555, Japan
kishida@surugadai.ac.jp

**Shirayuri College
Chofu 182-8525, Japan
kuriyama@shirayuri.ac.jp

[+]National Institute of Informatics (NII)
Tokyo 101-8430, Japan
{kando, eguchi} @nii.ac.jp

Abstract

The purpose of this paper is to examine empirically factors having effects on performance of cross-lingual information retrieval. In order to obtain experimental data, at the NTCIR-4 CLIR task, we submitted search results of Japanese monolingual run and three bilingual runs retrieving the Japanese document collection (i.e., Chinese-Japanese, Korean-Japanese and English-Japanese runs). It turns out that a regression model of which independent variables are "quality" of query translation and "difficulty" of the search in itself explains well variations of values of average precision by CLIR runs. The "quality" of translations was measured as a score assigned by a human assessor based on the degree to which each translation is coincident with the corresponding term in the Japanese topic that the task organizers provided, and the "difficulty" of the search was represented as a value of average precision by a run using the Japanese topic (i.e., monolingual run).

Keywords: *Cross-lingual information retrieval; Retrieval experiment, Performance, Regression analysis*

1. Introduction

Performance of cross-lingual information retrieval (CLIR) is largely dependent on quality or accuracy in the process of translating original queries. If perfect translations of query terms are obtained, the CLIR performance would approach to the level of monolingual retrieval. Thus much research effort has been dedicated to reduce automatically erroneous translations causing deterioration of CLIR performance.

The purpose of this paper is to explore regression models for predicting CLIR performance from information on quality of query translation. In order to obtain experimental data, we executed four types of search runs at the NTCIR-4 CLIR task as follows:

- Japanese to Japanese (J-J) runs
- Chinese to Japanese (C-J) runs
- Korean to Japanese (K-J) runs
- English to Japanese (E-J) runs

By assuming that results from the J-J monolingual searches provide an ideal performance, we can estimate the degree of dependency of CLIR (bilingual C-J, K-J and E-J runs) performance on the quality of automatic translation.

The rest of this paper is as follows. In section 2, regression models for predicting CLIR performance will be introduced. Section 3 will describe a retrieval system to be used for obtaining experimental data. Results from regression analysis will be presented in section 4. In section 5, we will discuss the results of analysis.

2. Regression models for predicating CLIR performance

2.1 Dependent variables

In the NTCIR-4 CLIR task, the <TITLE> field in each topic contains a few of words representing major concepts of the topic. For example, the topic 024

has three words in its <TITLE> field:

Illegal Tapping, Violation, Privacy.

It should be noted that "Illegal Tapping" is a compound word.

If automatic translations of these three English words into Japanese are perfectly identical with those in the <TITLE> filed of the corresponding Japanese topic prepared by the task organizers, performance of the E-J runs becomes inevitably equal with that of the J-J runs. We can assume that this is an ideal case. Meanwhile, if erroneous translations are generated in the automatic process, the E-J run would inevitably show lower performance than the J-J run.

Therefore, it is reasonable to use differences between two values representing performance of J-J and CLIR runs as a dependent variable in our regression model. That is, we suppose that

$$v - u = a + bx + e, \quad (1)$$

where u is a value of average precision for a topic by a CLIR run (C-J, K-J, or E-J), and v is the value by a J-J run. The Equation (1) includes an independent variable, x, which indicates the degree of "quality" in the process of translating the topic (see below), and a and b are regression coefficients (e is an error term).

Alternatively, by transposing v to the right side, we may be able to predicate directly a value of the variable u by a model such that

$$u = a + bx + cv + e, \quad (2)$$

where c is a regression coefficient. The model contains a value indicating performance by monolingual run, v, as an independent variable, which would explains the "difficulty" (or "easiness") of the topic. Therefore, we can interpret that Eq. (2) predicates performance of CLIR runs based on two factors: (1) "quality" of translation and (2) "difficulty" of the topic.

2.2 Independent variables

In order to measure the variable x, which represents the degree of "quality" in translation, human assessors have to score each translation according to a set of rules. In this paper, we use the following rules, in which judgments are made based on comparison between each translation and a corresponding Japanese word in the Japanese topic that the task organizers provided. For our convenience, we denote the corresponding Japanese word (single or compound) by "JT."

(A) In the case of a single word:
- If the word is identical with JT, score of the word is 1.0.
- If the word is synonymous with JT, score of the word is 0.8.
- If the word is almost same with JT but different *Kanji* characters are used, score of the word is 0.8.
- Otherwise, score of the word is 0.0.

(B) In the case of a compound word:
(B-1) The word is completely identical or different.
- If the compound word is completely identical with JT, score of the word is 1.0.
- If the compound word is completely different with JT, score of the word is 0.0.

(B-2) One or more components of the compound word are included in JT,
- If at least one component of the word is not included in JT, score of the word is 0.5.
- If at least one component of the word is synonymous with the corresponding component in JT, score of the word is 0.8.
- If at least one component of the word is almost same with the corresponding component in JT but different *Kanji* characters are used, score of the word is 0.8.

Rules for compound words are relatively more complicated. The rules in (B) are sequentially applied in the decreasing order of them. It should be also noted that different *Katakana* representations are considered as different words in this paper.

According to the rules (A) and (B), we can obtain "translation scores" of each word included in <TITLE> filed. Finally, the value of x is computed as a weighted average of the scores such that

$$x = \frac{w_1 s_1 + w_2 s_2 + ... + w_m s_m}{w_1 + w_2 + ... + w_m} \quad (3)$$

where

m is the number of distinct words included in the <TITLE> field,

s_j is a translation score of j-th word ($j = 1,..., m$),

w_j is a weight of j-th word ($j = 1,..., m$).

If we take that

$$w_1 = w_2 = ... = w_m = 1, \quad (4)$$

Eq. (3) reduces to a simple average. Otherwise, in order to consider "specificity" of each word, the idf method is applicable to the calculation of weights such that

$$w_j = \log \frac{N}{n_j} \quad (5)$$

where N is the total number of documents in the dataset and n_j is the number of documents in which the j-th word (JT) is appearing. As more specific words (e.g., proper nouns or technical terms) are not correctly translated, the value of x becomes smaller by using the idf factor.

2.3 Models

Since we have two regression models, (1) and (2), and two methods for weighting, (4) and (5), there are four models shown in Table 1.

Table 1 Regression models to be used

Weighting	Regression model	
	Eq.(1)	Eq.(2)
Eq.(4): simple	MODEL I(s)	MODEL II(s)
Eq.(5): idf	MODEL I(i)	MODEL II(i)

3. System and search runs

3.1 Retrieval system

3.1.1 Search engine. We used a search engine in ADOMAS (Advanced DOument MAnagement System) developed at Surugadai University in Japan.

3.1.2 Indexing system. In this study, only the Japanese document collection was targeted. Japanese words were extracted from texts of the documents and queries (topics) based on longest matching with entries in a machine-readable dictionary. We adopted strings matching with the entries as index terms. Unknown string between two known words was also used as an index term unless the unknown string consists of only *Hiragana* characters.

Furthermore, a heuristic rule was also applied, i.e., two adjacent words (known or unknown) are automatically combined into a compound word.

3.1.3 Retrieval model. In this study, a standard BM25 of Okapi weighting [1] was used with no special modification.

3.1.4 Pseudo-relevance feedback. In addition, a standard pseudo-relevance feedback (PRF) method was used in all search runs. We selected top 30 terms from a set of top-ranked 10 documents by an initial run based on the term weight,

$$r_t \times \log \frac{(r_t + 0.5)(N - R - n_t + r_t + 0.5)}{(N - n_t + 0.5)(R - r_t + 0.5)} \quad (6)$$

where

- r_t is the number of top-ranked documents including the term t,
- R is the number of top-ranked documents (i.e., $R = 30$ in this paper), and
- n_t is the number of documents including term t.

If the top-ranked term is already included in the set of search terms, term frequency in the query was increased 1.5 times. If not, the term frequency is set to 0.5.

3.1.5 Query translation. In the case of CLIR runs, all queries were translated into Japanese by the following commercial machine translation systems:

- For Chinese queries: Hourai for Windows
- For Korean queries: Kourai for Windows
- For English queries: PC-Transer

These MT systems are provided by Cross Language Inc. in Japan.

3.2 Search runs

We executed and submitted eight runs in total shown in Table 2.

Table 2 Submitted runs

Types	Topic field used	
	<TITLE>	<DESC>
J-J	NII-J-J-T-01	NII-J-J-D-02
C-J	NII-C-J-T-01	NII-C-J-D-02
K-J	NII-K-J-T-01	NII-K-J-D-02
E-J	NII-E-J-T-01	NII-E-J-D-02

4. Experimental results

4.1 Search performance

The Japanese collection includes 506,058 documents in total. The average document length is 155.99. Table 3 indicates values of mean average precision (MAP), and recall-precision curves of the eight runs are shown in Appendix 1.

Table 3 Search performance - MAP

Type	RunID	Rigid	Relaxed
J-J	NII-J-J-T-01	**0.2924**	0.4064
	NII-J-J-D-02	0.2740	0.3818
C-J	NII-C-J-T-01	**0.1746**	0.2294
	NII-C-J-D-02	0.1455	0.2036
K-J	NII-K-J-T-01	**0.2155**	0.2963
	NII-K-J-D-02	0.1894	0.2691
E-J	NII-E-J-T-01	**0.2266**	0.3143
	NII-E-J-D-02	0.2533	0.3403

If only MAP values of <TITLE>-runs for "Rigid" relevance are picked up, J-J is 0.2924, C-J is 0.1746, K-J is 0.2155 and E-J is 0.2266. Among CLIR runs, E-J is dominant, followed by K-J and C-J.

Appendix 2 shows values of $u - v$, i.e., the difference of values of these three CLIR runs from that of the monolingual J-J run. It turns out that CLIR runs outperform J-J run for a few topics.

4.2 Regression analysis

A human assessor assigned scores to each translation according to rules (A) and (B) described in section 2.2. For example, the results for topic 024 are shown in Table 4.

Since the NTCIR-4 CLIR test collection includes 55 topics for Japanese document sets and we executed three types of CLIR run (i.e., C-J, K-J and E-J), totally 165 (=55*3) observations are available for our regression analysis. Table 5 shows values of the squared correlation coefficients (R^2) and standard error of each model. As expected, the MODEL II explains more highly variations of CLIR performance

than the MODEL I. Meanwhile, the MODEL I(i) and II(i) incorporating the idf factor into weights show slightly better results than MODEL I(s) and II(s), respectively. We can not observe a significant effect of the idf factor. Actual regression models are as follows.

MODEL I(s):

$$v - u = -0.42339 + 0.420319x \quad (7)$$

MODEL I(i):

$$v - u = -0.41547 + 0.417349x \quad (8)$$

MODEL II(s):

$$u = -0.32081 + 0.402889x + 0.69687v \quad (9)$$

MODEL II(i):

$$u = -0.31297 + 0.397375x + 0.703277v \quad (10)$$

Further information on the regression models (7) to (10) is provided in Appendix 3.

Table 4 Example of translation scores: topic 024

words	DF*	Translation score		
		C-J	K-J	E-J
Illegal Tapping	6	1	0.5	0.5
Violation	2701	1	1	0
Privacy	2105	0	0.8	1
Value of x by Eq.(4)	-	0.67	0.77	0.50
Value of x by Eq.(5)	-	0.75	0.69	0.51

*Document frequency in the Japanese document collection.

Table 5 Summary of regression analysis

n=165

MODEL	R^2	Standard error
I(s): Eq.(1) and (4)	0.3256	0.1485
I(i): Eq.(1) and (5)	0.3388	0.1470
II(s): Eq.(2) and (4)	0.6373	0.1333
II(i): Eq.(2) and (5)	0.6422	0.1324

Table 6 Correlation matrix of variables in the MODEL II(i)

	x	v	u
x : translation score	1.0		
v : MAP by J-J runs	-0.08	1.0	
u : MAP by CLIR runs	0.40	0.66	1.0

Next, we shall examine the MODEL II(i) (Eq.(10)) in further detail. The model explains about 64% of the total sum of squares of MAP values by CLIR runs. It seems that the model can make a considerably accurate estimate of CLIR performance. Table 6 shows correlation coefficients between variables included in the MODEL II(i). There is almost no correlation between two independent variables, i.e., the quality of translation has no relationship with the difficulty of search. Therefore, the two variables contribute independently prediction of CLIR performance.

5. Discussion

As expected, it turned out that CLIR performance is able to be estimated from the degree of "quality" of translations and the degree of "difficulty" of the topic in itself. Actually, the MODEL II(i) explains about 64% of variations of MAP values by CLIR runs. Although this is not a novel discovery, the empirical findings would contribute to our understanding of nature of CLIR.

In this study, "erroneous" translations are operationally defined as those being different from Japanese words in the Japanese topic that the task organizers provided. However, such "erroneous" translations do not have always negative effects. As indicated in Appendix 2, CLIR runs in some topics show better performance than monolingual J-J runs. For example, in the topic 017, while average precision of the J-J run is only 0.122, that of the C-J run is 0.302 ("rigid," and <TITLE>-run). This is due to that the MT system generated two translations for a word, and that the one is identical with the Japanese word in the Japanese topic and another is a synonym to be useful for searching the topic. In CLIR, translation may have a kind of "side effect" causing higher performance.

6. Concluding remarks

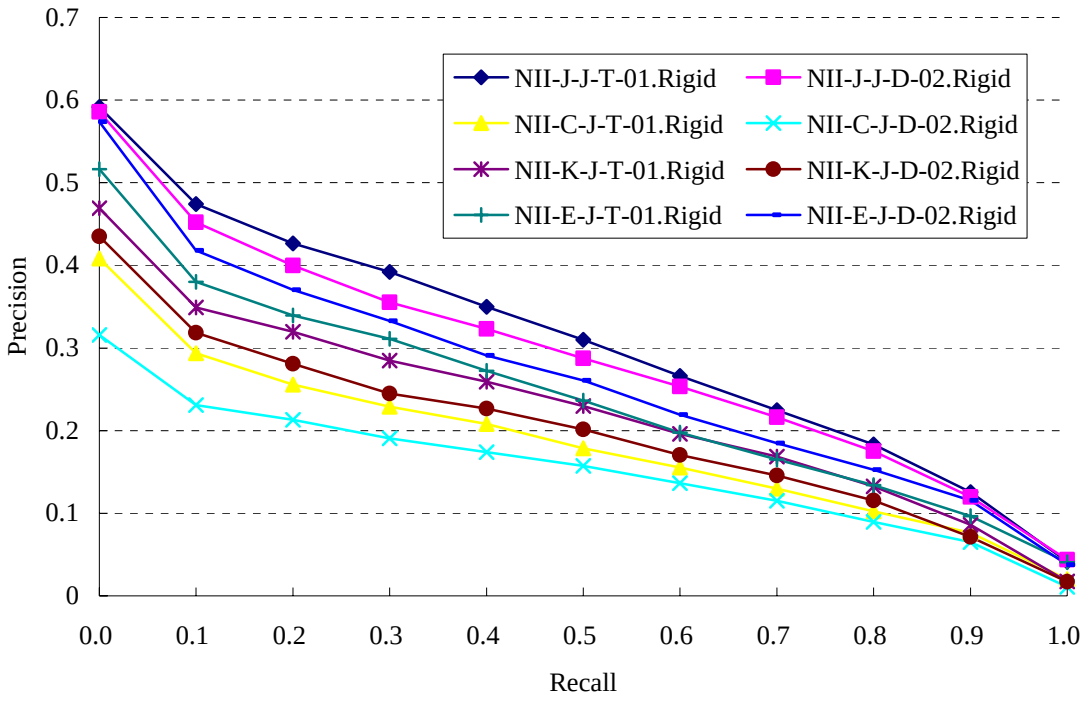
This paper attempts predicting CLIR performance from factors of "quality" of translations and "difficulty" of searching the topic. The regression model incorporating the two factors as independent variables explained about 64% of variations of CLIR performance. This means that the CLIR performance can be estimated considerably based on the two factors.

Of course, it should be noted that the regression models do not represent exactly a causal relationship between independent and dependent variable, and just only indicate a macro-level trend within the set of data we used. In this sense, further investigation using different data sets are needed.

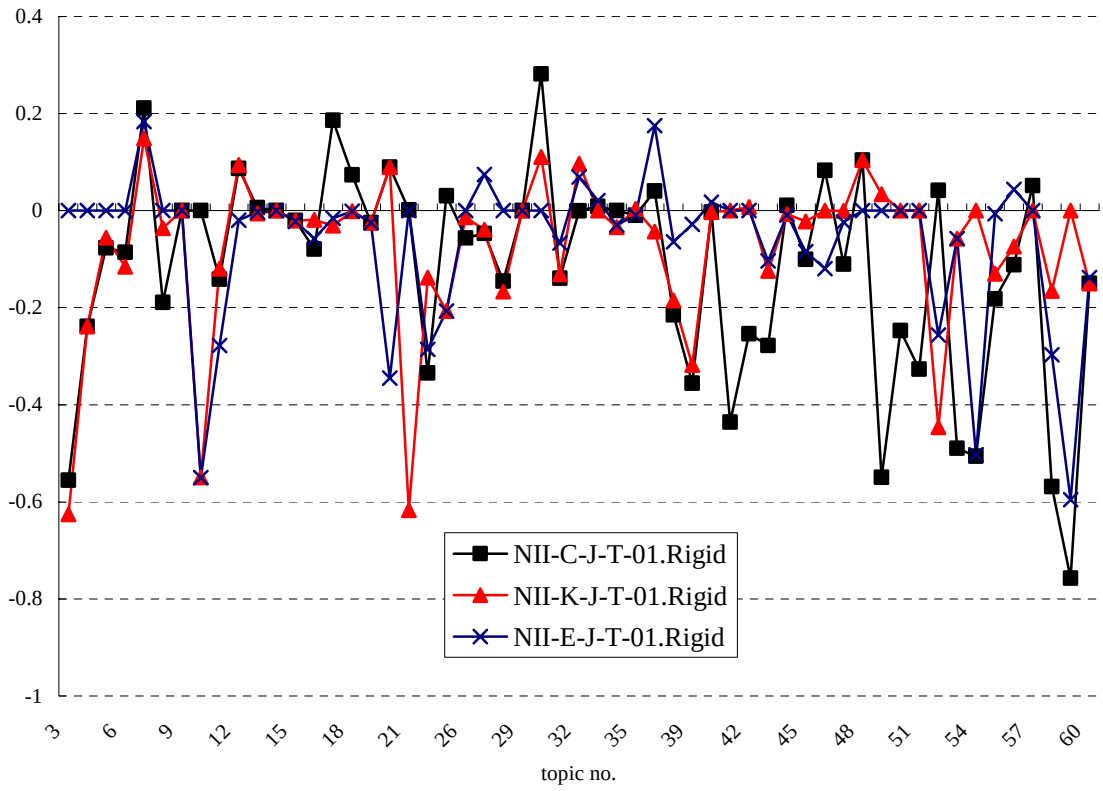
References

[1] Roberson, S. E., Walker, S., Jones, S., Han-cock-Beaulieu, M. M. & Gatford, M. (1995). Okapi at TREC-3. In Proceedings of TREC-3, Gaithersburg: MD, National Institute of Standards and Technology. http://trec.nist.gov/pubs/

Appendix 1: Recall-precision curves ("Rigid" only)



Appendix 2: Differences of MAP from monolingual runs



Appendix 3: Results of regression analysis

MODEL	Coefficients	Values	Standard error	t-value	probability
I(s)	*a*	-0.423390	0.039661	-10.6753	0.00000
	b (coefficient of *x*)	0.420319	0.047375	8.8721	0.00000
I(i)	*a*	-0.415470	0.037743	-11.0078	0.00000
	b (coefficient of *x*)	0.417349	0.045669	9.1386	0.00000
II(s)	*a*	-0.320810	0.039129	-8.19873	0.00000
	b (coefficient of *x*)	0.402889	0.042636	9.4496	0.00000
	c (coefficient of *v*)	0.696870	0.047869	14.5577	0.00000
II(i)	*a*	-0.312970	0.037765	-8.2875	0.00000
	b (coefficient of *x*)	0.397357	0.041263	9.6298	0.00000
	c (coefficient of *v*)	0.703277	0.047590	14.7778	0.00000