Learning Bilingual Translations from Comparable Corpora to Cross-Language Information Retrieval: Hybrid Statistics-based and Linguistics-based Approach

Fatiha Sadat
Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi
Nara, 630-0101, Japan

Masatoshi Yoshikawa
Nagoya University
Furo-cho, Chikusa-ku,
Nagoya, 464-8601, Japan

Shunsuke Uemura
Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi,
Nara, 630-0101, Japan

{fatia-s, uemura}@is.aist-nara.ac.jp, yosikawa@itc.nagoya-u.ac.jp

Abstract

Recent years saw an increased interest in the use and the construction of large corpora. With this increased interest and awareness has come an expansion in the application to knowledge acquisition and bilingual terminology extraction. The present paper will seek to present an approach to bilingual lexicon extraction from non-aligned comparable corpora, combination to linguistics-based pruning and evaluations on Cross-Language Information Retrieval. We propose and explore a two-stages translation model for the acquisition of bilingual terminology from comparable corpora, disambiguation and selection of best translation alternatives on the basis of their morphological knowledge. Evaluations using a large-scale test collection on Japanese-English and different weighting schemes of SMART retrieval system confirmed the effectiveness of the proposed combination of two-stages comparable corpora and linguistics-based pruning on Cross-Language Information Retrieval.

Keywords: *Cross-Language Information Retrieval, Comparable corpora, Translation, Disambiguation, Part-of-Speech.*

1 Introduction

Researches on corpus-based approaches to machine translation (MT) have been on the rise, particularly because of their promise to provide bilingual terminology and enrich lexical resources such as bilingual dictionaries and thesauri. These approaches generally rely on large text corpora, which play an important role in Natural Language Processing (NLP) and Information Retrieval (IR). Moreover, non-aligned comparable corpora have been given a special interest in bilingual terminology acquisition and lexical resources enrichment (Dagan and Itai, 1994; Dejean et al., 2002; Diab and Finch, 2000; Fung, 2000; Koehn and Knight, 2002; Nakagawa, 2000; Peters and Picchi, 1995; Rapp, 1999; Shahzad and al., 1999; Tanaka and Iwasaki, 1996).

Unlike parallel corpora, comparable corpora are collections of texts from pairs or multiples of languages, which can be contrasted because of their common features, in the topic, the domain, the authors or the time period. This property made comparable corpora more abundant, less expensive and more accessible through the World Wide Web.

In the present paper, we are concerned by exploiting scarce resources for bilingual terminology acquisition, then evaluations on Cross-Language Information Retrieval (CLIR). CLIR consists of retrieving documents written in one language using queries written in another language. An application is conducted on NTCIR, a large-scale data collection for (Japanese, English) language pair.

The remainder of the present paper is organized as follows: Section 2 presents the proposed two-stages approach for bilingual terminology acquisition from comparable corpora. Section 3 describes the integration of linguistic knowledge for pruning the translation candidates. Experiments and evaluations in CLIR are discussed in Sections 4. Section 5 concludes the present paper.

2 Two-stages Comparable Corpora-based Approach

Our proposed approach to bilingual terminology acquisition from comparable corpora (Sadat et al., 2003; Sadat et al., 2003) is based on the assumption of similar collocation, i.e., If two words are mutual translations, then their most frequent collocates are likely to be mutual translations as well. Moreover, we apply this assumption in both directions of the corpora, i.e., find translations of the source term in the target language corpus but also translations of the target terms in the source language corpus. The proposed two-stages approach for the acquisition, disambiguation and selection of bilingual terminology is described as follows:

- Bilingual terminology acquisition from source language to target language to yield a first translation model, represented by similarity $SIM_{S \to T}$.
- Bilingual terminology acquisition from target language to source language to yield a second translation model, represented by similarity $SIM_{T \to S}$.
- Merge the first and second models to yield a two-stages translation model, based on bi-directional comparable corpora and represented by similarity $SIM_{S \leftrightarrow T}$.

We follow strategies of previous researches (Dejean et al., 2002; Fung, 2000; Rapp, 1999) for the first and second translation models and propose a merging strategy for the two-stages translation model (Sadat et al., 2003).

First, word frequencies, context word frequencies in surrounding positions (here three-words window) are computed following a statistics-based metrics, the log-likelihood ratio (Dunning, 1993). Context vectors for each source term and each target term are constructed. Next, context vectors of the target words are translated using a preliminary bilingual dictionary. We consider all translation candidates, keeping the same context frequency value as the source term. This step requires a seed lexicon, to expand using the proposed bootstrapping approach of this paper. Similarity vectors are constructed for each pair of source term and target term using the cosine metric (Salton and McGill, 1983).

Therefore, similarity vectors $SIM_{S \to T}$ and $SIM_{T \to S}$ for the first and second models are constructed and merged for a bi-directional acquisition of bilingual terminology from source language to target language. The merging process will keep common pairs of source term and target translation *(s,t)* which appear in $SIM_{S \to T}$ as pairs of *(s,t)* but also in $SIM_{T \to S}$ as pairs of *(t,s)*, to result in combined similarity vectors $SIM_{S \leftrightarrow T}$ for each pair *(s,t)*.The product of similarity values of both similarity vectors $SIM_{S \to T}$ for pairs *(s,t)* and $SIM_{T \to S}$ for pairs *(t,s)* will result in similarity values in vectors $SIM_{S \leftrightarrow T}$.

Therefore, similarity vectors of the two-stages translation model are expressed as follows:

$$SIM_{S \leftrightarrow T} = \{(s, t, sim_{S \leftrightarrow T}(t|s)) \mid (s, t, sim_{S \to T}(t|s)) \\ \in SIM_{S \to T} \ \wedge \ (t, s, sim_{T \to S}(s|t)) \in SIM_{T \to S} \\ \wedge \ sim_{S \leftrightarrow T}(t|s) = sim_{S \to T}(t|s) \times sim_{T \to S}(s|t)\}$$

3 Linguistics-based Pruning

Combining linguistic and statistical methods is becoming increasingly common in computational linguistics, especially as more corpora become available (Klanvans and Tzoukermann, 1996; Sadat et al., 2003). We propose to integrate linguistic concepts into the corpora-based translation model. Morphological knowledge such as Part-of-Speech (POS) tags, context of terms, etc., could be valuable to filter and prune the extracted translation candidates. The objective of the linguistics-based pruning technique is the detection of terms and their translations that are morphologically close enough, i.e., close or similar POS tags. This proposed approach will select a fixed number of equivalents from the set of extracted target translation alternatives that match the Part-of-Speech of the source term.

Therefore, POS tags are assigned to each source term (Japanese) via morphological analysis. As well, a target language morphological analysis will assign POS tags to the translation candidates. We restricted the pruning technique to nouns, verbs, adjectives and adverbs, although other POS tags could be treated in a similar way. For Japanese-English[1] pair of languages, Japanese nouns (名詞) are compared to English nouns (NN) and Japanese verbs (動詞) to English verbs (VB). Japanese adverbs (副詞) are compared to English adverbs (RB) and adjectives (JJ); while, Japanese adjectives (形容詞) are compared to English adverbs (RB) and adjectives (JJ). This is because most adverbs in Japanese are formed from adjectives. Thus. We select pairs or source term and target translation *(s,t)* such as:

POS(s) = 'NN' and	POS(t) = ' 名詞'
POS(s) = 'VB' and	POS(t) = ' 動詞'
POS(s) = 'RB' and	[POS(t) = ' 副詞' or ' 形容詞']
POS(s) = 'JJ' and	[POS(t) = ' 形容詞' or ' 副詞']

Japanese foreign words (tagged FW) were considered as loanwords, i.e., technical terms and proper nouns imported from foreign languages; and therefore were not pruned with the proposed linguistics-based technique but could be treated via transliteration.

The generated translation alternatives are sorted in decreasing order by similarity values. Rank counts are assigned in increasing order, starting at 1 for the first sorted list item. A fixed number of top-ranked translation alternatives are selected and misleading candidates are discarded.

In order to demonstrate the procedure of our translation model, we give an example in Japanese and explain how the English translations are extracted, disambiguated and selected and how the phrasal translation is constructed.

Given a simple Japanese query ' アジア競技大会は、アジア最大のスポーツ競技会である' *(ajia kyougi taikai wa, ajia saidai no supoutsu kyougikai de aru).*

After segmentation, removing stop words and keeping only content words (nouns, verbs, adverbs, adjectives and foreign words), the associated list of Japanese terms becomes' アジア, 競技, 大会, アジア, 最大, スポーツ, 競技, 会' $(ajia, kyougi, taikai, ajia, saidai, supoutsu, kyougi, kai)$. The combined translation model is applied on each source term of the associated list and top-ranked word translation alternatives are selected according to their highest similarities as follows:

' アジア' $(ajia)$:{(asia, 1.035), (assembly, 0.0611), (city, 0.0589), (event, 0.0376), etc.}

' 競技 ' $(kyougi)$: {(competition, 0.057), (sport, 0.0561), (representative, 0.0337), (international, 0.0331), etc.}

' 大会 ' $(taikai)$: {(meeting, 0.176), (tournament, 0.0588), (assembly, 0.0582), (dialogue, 0.0437), etc.}

' 最大' $(saidai)$: {(general, 0.0459), (great, 0.0371), (famous, 0.0362), (global, 0.0329), (group, 0.032), (measure, 0.0271), (factor, 0.0268), etc.}

' スポーツ' $(supoutsu)$: {(sport, 1.098), (union, 0.0399), (day, 0.0392), (international, 0.0375), etc.}

' 会' (kai): {(taikai, 0.0489), (great, 0.0442), (meeting, 0.0365), (gather, 0.0348), (person, 0.0312), etc.}

The phrasal translation associated to the Japanese query is formed by selecting a number of top-ranked translation alternatives (here set to 3) and illustrated as follows: '*asia assembly city competition sport representative meeting tournament assembly general great famous sport union day taikai great meeting*'.

Linguistics-based pruning was applied on the Japanese terms and the extracted English translation alternatives. *Chasen* morphological analyzer (Matsumoto and al., 1997)for Japanese has associated POS tags as 名詞 (noun) to all Japanese terms:

アジア $(ajia)$	名詞-固有名
競技 $(kyougi)$	名詞-サ変接続
大会 $(taikai)$	名詞-一般
最大 $(saidai)$	名詞-一般
スポーツ $(supoutsu)$	名詞-一般
会 (kai)	名詞-一般

Therefore, English translation alternatives associated with POS tags as nouns (NN) via a morphological analyzer for English (Sekine, 2001)are selected and translation candidates having POS tags other than NN (noun) are discarded. Selected translation alternatives for the Japanese noun 最大 $(saidai)$ become '*group, measure, factor*'. As well, the

[1]English POS tags NN refers to noun, VB to verb, RB to adverb, JJ to adjective; while Japanese POS tags 名詞 refers to a noun, 動詞 to a verb, 副詞 to an adverb and 形容詞 to an adjective, with respect to their extensions.

Japanese term ' 会' (kai) is associated to the English translations: '*taikai, meeting, person*'.

The phrasal translation associated to the Japanese query after the linguistics-based pruning is illustrated as follows: '*asia assembly city competition sport representative meeting tournament assembly group measure factor sport union day taikai meeting person*'.

Possible re-scoring techniques could be applied on phrasal translation in order to select best translation alternatives among the extracted ones.

4 Experiments and Evaluations

Experiments have been carried out to measure the improvement of our proposal on bilingual terminology acquisition from comparable corpora on Japanese-English tasks in CLIR, i.e. Japanese queries to retrieve English documents.

4.1 Linguistic Resources

Collections of news articles from *Mainichi Newspapers* (1998-1999) for Japanese and *Mainichi Daily News* (1998-199) for English were considered as comparable corpora, because of the common feature in the time period and the generalized domain. We have also considered documents of *NTCIR-2* test collection as comparable corpora in order to cope with special features of the test collection during evaluations.

Morphological analyzers, *ChaSen version 2.2.9* (Matsumoto and al., 1997) for texts in Japanese and *OAK2* (Sekine, 2001) were used in the linguistic preprocessing.

EDR bilingual dictionary (EDR, 1996) was used to translate context vectors of source and target languages.

NTCIR-2 (Kando, 2001), a large-scale test collection was used to evaluate the proposed strategies in CLIR.

SMART information retrieval system (Salton, 1971), which is based on vector space model, was used to retrieve English documents.

4.2 Evaluations on the Proposed Translation Model

We considered the set of news articles as well as the abstracts of NTCIR-2 test collection as comparable corpora for Japanese-English language pairs. The abstracts of NTCIR-2 test collection are partially aligned (more than half are Japanese-English paired documents) but the alignment was not considered in the present research to treat the set of documents as comparable. Content words (nouns, verbs, adjectives, adverbs) were extracted from English and Japanese corpora. In addition, foreign words (mostly represented in katakana) were extracted from Japanese texts. Thus, context vectors were constructed for 13,552,481 Japanese terms and 1,517,281 English terms. Similarity vectors were constructed for 96,895,255 (Japanese, English) pairs of terms and 92,765,129 (English, Japanese) pairs of terms. Bi-directional similarity vectors (after merging and disambiguation) resulted in 58,254,841 (Japanese, English) pairs of terms.

Table 1 illustrates some situations with the extracted English translation alternatives for Japanese terms 映画 ($eiga$), using the two-stages comparable corpora approach and combination to linguistics-based pruning. Using the two-stages comparable corpora-based approach, correct translations of the Japanese term 映画 ($eiga$) were ranked in top 3 ($movie$) and top 5 ($film$). We notice that top ranked translations, which are considered as wrong translations, are related mostly to the context of the source Japanese term and could help the query expansion in CLIR. Combined two-stages comparable corpora with the linguistics-based pruning shows better results with ranks 2 ($movie$) and 4 ($film$).

Japanese vocabulary is frequently imported from other languages, primarily (but not exclusively) from English. The special phonetic alphabet (here Japanese $katakana$) is used to write down foreign words and loanwords, example names of persons and others. Katakana terms could be treated via *transliteration* or possible *romanization*, i.e., conversion of Japanese katakana to their English equivalence or the alphabetical description of their pronunciation. Transliteration is the phonetic or spelling representation of one language using the alphabet of another language (Knight and Graehl, 1998).

4.3 Evaluations on SMART Weighting Schemes

Conducted experiments and evaluations were completed on NTCIR test collection using the monolin-

Table 1: An example for the two-stages comparable corpora translation model and linguistics-based pruning

Japanese Term	Two-stages Comparable Corpora: English Translation	Similarity Value	Rank	Linguistics-based Pruning: English Translation	Similarity Value	Rank
映画 ($eiga$)	famous	0.449	1			
	picture	0.361	2	picture	0.361	1
	movie	***0.2163***	**3**	***movie***	***0.2163***	**2**
	oscar	0.1167	4	oscar	0.1167	3
	film	***0.1116***	**5**	***film***	***0.1116***	**4**

gual English runs, i.e., English queries to retrieve English documents and the bilingual Japanese-English runs, i.e., Japanese queries to retrieve English document. Topics 0101 to 0149 were considered and key terms contained in the fields, title $<TITLE>$, description $<DESCRIPTION>$ and concept $<CONCEPT>$ were used to generate 49 queries in Japanese and English.

There is a variety of techniques implemented in SMART to calculate weights for individual terms in both documents and queries. These weighting techniques are formulated by combining three parameters: *Term Frequency* component, *Inverted Document Frequency* component and *Vector Normalization* component.

The standard SMART notation to describe the combined schemes is "XXX.YYY". The three characters to the left (XXX) and right (YYY) of the period refer to the document and query vector components, respectively. For example, ATC.ATN applies augmented normalized term frequency, $tf \times idf$ document frequency *(term frequency times inverse document frequency components)* to weigh terms in the collection of documents. Similarly ATN refers to the weighting scheme applied to the query.

First experiments were conducted on several combinations of weighting parameters and schemes of SMART retrieval system for documents terms and query terms, such as ATN, ATC, LTN, LTC, NNN, NTC, etc. Best performances in terms of average precision were realized by the following combined weighting schemes: ATN.NTC, LTN.NTC, LTC.NTC, ATC.NTC and NTC.NTC, respectively.

The best weighting scheme for the monolingual runs turned out to be the ATN.NTC. This finding is somewhat different from previous results where ANN (Fox and Shaw, 1994), LTC (Fuhr and al., 1994) weighting schemes on query terms, LNC.LTC (Buckley and al., 1994) and LNC.LTN (Knaus and Shauble, 1993) combined weighting schemes on document terms and query terms showed the best results. On the other hand, our findings were quite similar to the result presented by Savoy (Savoy, 2003), where the ATN.NTC showed the best performance among the existing weighting schemes in SMART for English monolingual runs.

Table 2 shows some weighting schemes of SMART retrieval system, among others. To assign an indexing weight w_{ij} that reflects the importance of each single-term T_j in a document D_i, different factors should be considered (Salton and McGill, 1983), as follows:

- within-document term frequency tf_{ij}, which represents the first letter of the SMART label.
- collection-wide term frequency df_j, which represents the second letter of the SMART label. In Table 2, $idf_j = \log \frac{N}{F_j}$; where, N represents the number of documents and F_j represents the document frequency of term T_j.
- normalization scheme, which represents the third letter of the SMART label.

4.4 Evaluations on CLIR

Bilingual translations were extracted from comparable corpora using the proposed two-stages model. A fixed number (set to five) of top-ranked translation alternatives was retained for evaluations in CLIR.

Results and performances on the monolingual and bilingual runs for the proposed translation models and the combination to linguistics-based pruning are described in Table 3. Evaluations were based on the average precision, differences in term of average precision of the monolingual counterpart and the improvement over the monolingual counterpart. As

Table 2: Weighting Schemes on SMART Retrieval System

SMART Label	Weighting Scheme
NNN	$w_{ij} = tf_{ij}$
ATN	$w_{ij} = idf_j \times [0.5 + \frac{tf_{ij}}{2 \times max\_tf_i}]$
LTN	$w_{ij} = idf_j \times [ln(tf_{ij}) + 1.0]$
LTC	$w_{ij} = \frac{idf_j \times [ln(tf_{ij})+0.1]}{\sqrt{\sum_{k=1}^{n} [idf_k \times (ln(tf_{ik})+0.1)]^2}}$
ATC	$w_{ij} = \frac{idf_j \times (0.5+\frac{tf_{ij}}{2 \times max\_tf_i})}{\sqrt{\sum_{k=1}^{n} [idf_k \times (0.5+\frac{tf_{ik}}{2 \times max\_tf_i})]^2}}$
NTC	$w_{ij} = \frac{idf_j \times tf_{ij}}{\sqrt{\sum_{k=1}^{n} [idf_k \times tf_{ik}]^2}}$

well, evaluations using R-precision are illustrated in Table 3.

Figure 1 represents the recall/precision curves of the proposed two-stages comparable corpora-based translation model and combination to linguistics-based pruning, in the case of ATN.NTC weighting scheme.

The proposed two-stages model using comparable corpora '*BCC*' showed a better improvement in terms of average precision compared to the simple model '*SCC*' (one stage, i.e., simple comparable corpora-based translation) with +27.1%. Combination to Linguistics-based pruning showed the best performance in terms of average precision with +41.7% and +11.5% compared to the simple comparable corpora-based model '*SCC*' and the two-stages comparable corpora-based model '*BCC*', respectively, in the case of ATN.NTC weighting scheme.

Different weighting schemes of SMART retrieval system showed an improvement in term of average precision for the proposed translation models '*BCC*' and '*BCC+Morph*'.

The approach based on comparable corpora largely affected the translation because related words could be added as translation alternatives or expansion terms. The acquisition of bilingual terminology from bi-directional comparable corpora yields a significantly better result than using the simple model. Moreover, the linguistics-based pruning technique has allowed an improvement in the effectiveness of CLIR.

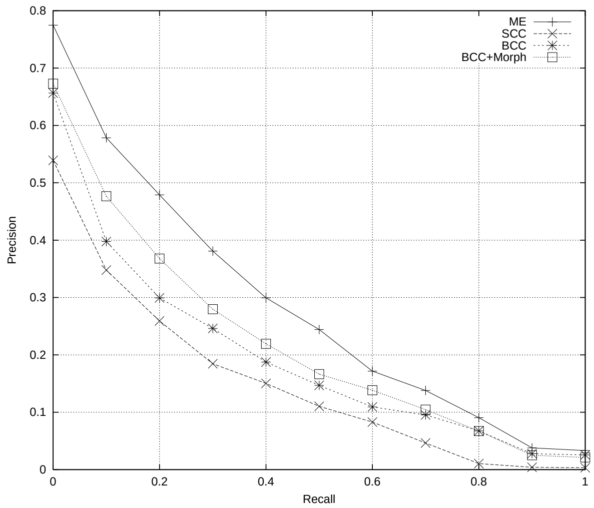

Figure 1: Recall/Precision curves for the proposed translation models and combination to linguistics-based pruning (*weighting scheme = ATN.NTC*)

Finally, statistical *t-test* (Hull, 1993) was carried out in order to measure significant differences between paired retrieval models. The improvement by using the proposed two-stages comparable corpora-based method '*BCC*' was statistically significant (p-value=0.0011). The combined statistics-based and linguistics-based pruning '*BCC+Morph*' was

Table 3: Best results on different weighting schemes for the proposed translation models and the linguistics-based pruning

Weighting Models	Average Precision, % Monolingual, and % Improvement				R-Precision, % Monolingual, and % Improvement			
	ME *(Monolingual English)*	**SCC** *(Simple Comp. Corpora)*	**BCC** *(Two-stages Comp. Corpora)*	**BCC+Morph** *(Linguistics-(based pruning)*	**ME** *(Monolingual English)*	**SCC** *(Simple Comp. Corpora)*	**BCC** *(Two-stages Comp. Corpora)*	**BCC+Morph** *(Linguistics-(based pruning)*
ATN.NTC	**0.2683** **(100%)**	0.1417 (52.81%) (-47.18%)	0.1801 (67.12%) (-32.87%)	**0.2008** **(74.84%)** **(-25.16%)**	**0.2982** (100%)	0.1652 (55.34%) (-44.6%)	0.2143 (71.86%) (-28.13%)	**0.2391** (80.18%) (-19.82%)
LTN.NTC	0.2236 (100%)	0.091 (40.69%) (-59.3%)	0.1544 (69.05%) (-30.94%)	0.1729 (77.32%) (-22.67%)	0.2508 (100%)	0.1339 (53.39%) (-46.61%)	0.1823 (72.69%) (-27.31%)	0.2066 (82.37%) (-17.62%)
LTC.NTC	0.1703 (100%)	0.0787 (46.21%) (-53.78%)	0.1138 (66.82%) (-33.17%)	0.1327 (77.92%) (-22.08%)	0.1943 (100%)	0.0966 (49.71%) (-50.28%)	0.1396 (71.85%) (-28.15%)	0.1663 (85.59%) (-14.41%)
ATC.NTC	0.1665 (100%)	0.0707 (42.46%) (-57.53%)	0.1091 (65.52%) (-34.47%)	0.1252 75.19% (-24.8%)	0.2004 (100%)	0.0923 (46.05%) (-53.94%)	0.1368 (68.26%) (-31.73%)	0.1481 (73.9%) (-26.1%)
NTC.NTC	0.1254 (100%)	0.0575 (45.85%) (-54.15%)	0.073 (58.21%) (-41.78%)	0.0915 (72.96%) (-27.03%)	0.154 (100%)	0.079 (51.3%) (-48.7%)	0.0989 (64.22%) (-35.78%)	0.1175 (76.3%) (-23.7%)

found statistically significant (p-value= 0.05) over the monolingual retrieval '*ME*'.

5 Conclusions and Future Work

Dictionary-based translation has been widely used in CLIR because of its simplicity and availability. However, failure to translate words and compounds as well as limitations of general-purpose dictionaries especially for specialized vocabulary are among the reasons of drop in retrieval performance especially when dealing with CLIR. Enriching bilingual dictionaries and thesauri is possible through bilingual terminology acquisition from large corpora. Parallel corpora are costly to acquire and their availability is extremely limited for any pair of languages or even not existing for some languages, which are characterized by few amounts of Web pages on the WWW. In contrast, comparable corpora are more abundant, more available in different domains, less expensive and more accessible through the WWW.

In the present paper, we investigated the approach of extracting bilingual terminology from comparable corpora in order to enrich existing bilingual lexicons and thus enhance Cross-Language Information Retrieval. We proposed a two-stages translation model consisting of bi-directional extraction, merging and disambiguation of the extracted bilingual terminology. A hybrid combination to linguistics-based pruning showed its efficiency across Japanese-English pair of languages. Most of the selected terms could be considered as translation candidates or expansion terms in CLIR.

Ongoing research is focused on the integration of transliteration for the special phonetic alphabet. Techniques on phrasal translation will be investigated in order to select best phrasal translation alternatives in CLIR. Evaluations using other combinations and more efficient weighting schemes that are not included in SMART retrieval system such as OKAPI, which showed great success in information retrieval, are among the future subjects of our research on CLIR.

Acknowledgements

The present research is supported in part by the Ministry of Education, Culture, Sports, Science and Technology of Japan. Our thanks go to all reviewers for their valuable comments on the earlier version of this paper.

References

C. Buckley, J. Allan and G. Salton. 1994. Automatic Routing and Ad-hoc Retrieval using Smart. *Proc. Second Text Retrieval Conference TREC-2*, pages 45–56,

I. Dagan and I. Itai. 1994. Word Sense Disambiguation using a Second Language Monolingual Corpus. *Computational Linguistics*, 20(4):563–596.

H. Dejean, E. Gaussier and F. Sadat. 2002. An Approach based on Multilingual Thesauri and Model Combination for Bilingual Lexicon Extraction. *In Proc. COLING 2002*.

M. Diab and S. Finch. 2000. A Statistical Word-level Translation Model for Comparable Corpora. *Proc. of the Conference on Content-based Multimedia Information Access RIAO*.

T. Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational linguistics* 19(1).

EDR. 1996. Japan Electronic Dictionary Research Institute, Ltd. EDR electronic dictionary version 1.5 EDR. *Technical guide. Technical report TR2-007*.

A. E. Fox and A. J. Shaw. 1994. Combination of Multiple Searches. *Proc. Second Text Retrieval Conference TREC-2*, pages 243–252.

N. Fuhr, U. Pfeifer, C. Bremkamp, M. Pollmann and C. Buckley. 1994. Probabilistic Learning Approaches for Indexing and Retrieval with the TREC-2 Collection. *Proc. Second Text Retrieval Conference TREC-2*, pages 67–74.

P. Fung. 2000. A Statistical View of Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora. *In Jean Veronis, Ed. Parallel Text Processing*.

D. Hull. 1993. Using Statistical Testing in the Evaluation of Retrieval Experiments. *Proc. ACM SIGIR'93*, pages 329–338.

N. Kando. 2001. Overview of the Second NTCIR Workshop. *In Proc. Second NTCIR Workshop on Research in Chinese and Japanese Text Retrieval and Text Summarization*.

J. Klavans and E. Tzoukermann. 1996. Combining Corpus and Machine-Readable Dictionary Data for Building Bilingual Lexicons. *Machine Translation*, 10(3-4):1–34.

D. Knaus and P. Shauble. 1993. Effective and Efficient Retrieval from Large and Dynamic Document Collections. *Proc. Second Text Retrieval Conference TREC-3*, pages 163–170.

K. Knight and J. Graehl. 1998. Machine Transliteration. *Computational Linguistics*, 24(4).

P. Koehn and K. Knight. 2002. Learning a Translation Lexicon from Monolingual Corpora. *In Proc. ACL-02 Workshop on Unsupervised Lexical Acquisition*.

Y. Matsumoto, A. Kitauchi, T. Yamashita, O. Imaichi and T. Imamura. 1997. Japanese Morphological Analysis System ChaSen Manual. *Technical Report NAIST-IS-TR97007*.

H. Nakagawa. 2000. Disambiguation of Lexical Translations based on Bilingual Comparable Corpora. *Proc. LREC2000, Workshop of Terminology Resources and Computation WTRC2000*, pages 33–38.

C. Peters and E. Picchi. 1995. Capturing the comparable: A System for Querying Comparable Text Corpora. *Proc. 3rd International Conference on Statistical Analysis of Textual Data*, pages 255–262.

R. Rapp. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. *In Proc. European Association for Computational Linguistics*.

F. Sadat, M. Yoshikawa and S. Uemura. 2003. Enhancing Cross-language Information Retrieval by an Automatic Acquisition of Bilingual Terminology from Comparable Corpora. *In Proc. ACM SIGIR 2003*, Toronto, Canada.

F. Sadat, M. Yoshikawa and S. Uemura. 2003. Bilingual Terminology Acquisition from Comparable Corpora and Phrasal Translation to Cross-Language Information Retrieval. *In Proc. ACL 2003*, Sapporo, Japan.

G. Salton. 1971. The SMART Retrieval System, Experiments in Automatic Documents Processing. *Prentice-Hall, Inc., Englewood Cliffs, NJ.*

G. Salton and J. McGill. 1983. Introduction to Modern Information Retrieval. *New York, Mc Graw-Hill.*

J. Savoy. 2003. Cross-Language Information Retrieval: Experiments based on CLEF 2000 Corpora. *Information Processing & Management* 39(1):75–115.

S. Sekine. 2001. OAK System-Manual. *New York University.*

I. Shahzad, K. Ohtake, S. Masuyama and K. Yamamoto. 1999. Identifying Translations of Compound using Non-aligned Corpora. *Proc. Workshop MAL*, pages 108–113.

K. Tanaka and H. Iwasaki. 1996 Extraction of Lexical Translations from Non-aligned Corpora. *Proc. COLING 96*.