Translating with Scarce Resources in Cross-Language Information Retrieval: A case Study on Japanese-English

FATIHA SADAT, [†1] MASATOSHI YOSHIKAWA [†2] and SHUNSUKE UEMURA [†1]

This paper seeks to present an approach to learning bilingual terminology from scarce resources in order to translate source query terms and retrieve target documents across languages. An extracted bilingual lexicon from comparable corpora will provide a valuable resource to enrich existing bilingual dictionaries. A linear combination involving the extracted bilingual terminology from comparable corpora, readily available bilingual dictionaries and transliteration is proposed. An application on Japanese-English language pair shows that this combination yields better translations and an effectiveness of information retrieval could be achieved across languages.

Keywords: Comparable corpora, Bilingual terminology extraction, Bilingual dictionaries enrichment, Translation, Cross-Language Information Retrieval.

1 Introduction

Large text corpora represent a crucial resource for the acquisition of bilingual terminology and the enrichment of multilingual lexical resources. Moreover, in recent years non-aligned comparable corpora have been an object of studies and research related to natural language processing and information retrieval [1, 2, 3, 6, 9, 11, 12, 13, 17, 18], because of their availability and easy accessibility through the World Wide Web.

In the present paper, our goal is to learn translation lexicons using scarce resources, i.e. readily available resources and possibly through the Internet. We are concerned by exploiting news articles as comparable corpora in order to translate terms in a source language to any specified target language. Our preliminary study is conducted on Japanese-English language pair using general-domain comparable corpora and could be extended to other languages and domains. Evaluations were conducted on Cross-Language Information Retrieval (CLIR) using large-scale test collection for Japanese and English.

The remainder of the present paper is organized as follows: Section 2 presents an overview of the proposed approach for bilingual terminology acquisition from comparable corpora. Linear combination to dictionary-based translation and transliteration is presented in Section 3. Experiments and evaluations in CLIR are discussed in Sections 4. Section 5 concludes the present paper.

2 An Overview of the Proposed Approach on Comparable Corpora

Unlike parallel texts, which are clearly defined as translated texts, there is a wide variation of non-parallel-ness in monolingual data. It can be manifested in the topic, the domain, the authors, the time period, etc. Comparable corpora are collections of texts from pairs or multiples of languages, which can be contrasted because of their common features. We rely on such comparable corpora for the extraction of bilingual terminology, in the form of translations and/or similar terms.

We follow the model proposed by [2, 6, 13]. First, word frequencies, context word frequencies

†1 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology (NAIST)
†2 名古屋大学情報連携基盤センター
Information Technology Center, Nagoya University

in surrounding positions (here three-words window) are computed following statistics-based metrics. Context vectors for each term in the source language and the target language are constructed. We use the *log-likelihood ratio* [4], which is expressed in equation (1) as follows:

$$LLR(w_i, w_j) = K_{11} log \frac{K_{11}N}{C_1R_1} + K_{12} log \frac{K_{12}N}{C_1R_2} + K_{21} log \frac{K_{21}N}{C_2R_1} + K_{22} log \frac{K_{22}N}{C_2R_2} \quad (1)$$

where,

$C_1 = K_{11} + K_{12}$, $C_2 = K_{21} + K_{22}$,

$R_1 = K_{11} + K_{21}$, $R_2 = K_{12} + K_{22}$,

$N = K_{11} + K_{12} + K_{21} + K_{22}$,

K_{11} = frequency of common occurrences of word w_i and word w_j,

K_{12} = corpus frequency of word w_i - K_{11},

K_{21} = corpus frequency of word w_j - K_{11},

$K_{22} = N - K_{12} - K_{22}$.

Next, context vectors of the target words are translated using a preliminary seed lexicon.

We consider all translation candidates, keeping the same context frequency value as the source term. This step requires a seed lexicon that will be enriched using the proposed bootstrapping approach of this paper.

Similarity vectors are constructed for each pair of source term and target term using the *cosine metrics* [15], which is expressed in equation (2).

$$Similarity\ (w_i, w_j) = \frac{\sum_k v_{ik} v_{jk}}{\sqrt{\sum_k v_{ik}^2 \sum_k v_{jk}^2}} \quad (2)$$

where,

v_{ik} represents co-occurrence frequencies in context vectors of the source term w_i with term w_k..

v_{jk} represents co-occurrence frequencies in context vectors of the target term w_j with term w_k..

Thus, similarity vectors are constructed to yield a probabilistic translation model $P_{comp}(t|s)$.

3 Different Translation Models and their Linear Combination

Combining different models has showed success in previous research [2].

We propose a combined model involving comparable corpora, readily available bilingual dictionaries as well as transliteration for the special phonetic or spelling representation of Japanese language, represented by the *Katakana* alphabet.

3.1 Bilingual Dictionary-based Translation

General-purpose dictionaries are basic source for translations and could be exploited for bilingual terminology extraction. The proposed dictionary-based translation model is derived directly from readily available bilingual dictionaries, by considering all translation candidates and their associated phrases, for each source entry.

Therefore, if a term *s* is represented by *N* sets of translation candidates in the bilingual dictionary and each set S_i $(i=1...n)$ contains a number M_i of translation candidates; then a translation candidate *t*, which appears in a number *K* of translation sets, will be represented by a probability $P_{dict}(t|s)$ as follows:

$$P_{dict}(t \mid s) = \sum_{i=1..k} \left[\frac{1}{N \times M_i} \right] \quad (3)$$

Note that stop words are discarded. For instance the Japanese term '取る' *(toru)* appears in the bilingual Japanese-English dictionary with 3 translation sets as follows:

取る: {take a rest} / {take pictures and move} / {take}

Here N =3, $M_1 = 2$, $M_2 = 3$ and $M_3 = 1$. Thus,

$$P_{dict}(take| 取る) = \frac{1}{N \times M_1} + \frac{1}{N \times M_2} + \frac{1}{N \times M_3} = 0.61$$

$$P_{dict}(rest| 取る) = \frac{1}{N \times M_1} = 0.16$$

$$P_{dict}(picture| 取る) = P_{dict}(move| 取る) = \frac{1}{N \times M_2} = 0.11$$

In this case, bilingual translation alternatives of

Japanese term '取る' *(toru)* are ranked and selected according to their probability values as follows: (*take*, 0.61), (*rest*, 0.16), (*picture*, 0.11), (*move*, 0.11), etc.

3.2 Transliteration

Transliteration is the phonetic or spelling representation of one language using the alphabet of another language. The special phonetic alphabet (here Japanese katakana) to foreign words and loanwords requires *romanization* or transliteration [8]. Japanese vocabulary is frequently imported from other languages, primarily (but not exclusively) from English. *Katakana*, the special phonetic alphabet is used to write down foreign words and loanwords, example names of persons and other terms. The English word '*computer*' is transliterated in Japanese katakana as 'コンピューター', as well '*engineer*' is transliterated as 'エンジニアー', and '*space shuttle*' is transliterated as 'スペースシャトル'. Named entities such as proper names of foreign (else than Japanese) persons, locations and organizations, are transliterated in Japanese. An example is '*Bill Clinton*' as named entities and transliterated in Japanese as 'ビルクリントン'.

Assume a source term *s* (written in katakana) is represented by *N* transliteration alternatives. Each transliteration *t* will be represented by a probability $P_{translit}(t|s)$ as follows:

$$P_{translit}(t \mid s) = \frac{1}{N} \qquad (4)$$

In the present paper, a transliteration system is used to convert terms written in katakana to their romaji forms, i.e., the alphabetical description of Japanese pronunciation and thus complete a transliteration. Note that if the transliteration system presents a unique transliteration for each Japanese term written in katakana, then $P_{translit}(t|s)$ will be equal to 1.

3.3 Linear Combination

A linear combination is completed using the three probabilistic translation models derived from the comparable corpora $P_{comp}(t|s)$, readily available bilingual dictionaries $P_{dict}(t|s)$ and the transliteration model $P_{translit}(t|s)$, as expressed in equation (5).

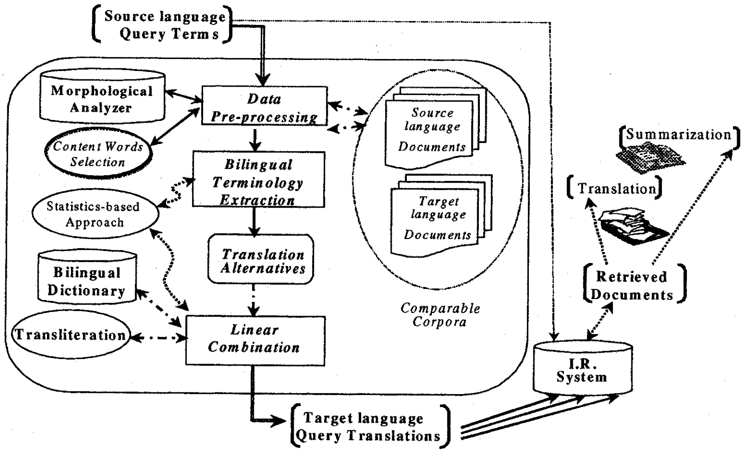

Fig. 1 Overview of the proposed combination of different translation models in Cross-Language Information Retrieval

$$P(t|s)=\alpha_1 P_{comp}(t|s)+\alpha_2 P_{dict}(t|s)+\alpha_3 P_{translit}(t|s) \quad (5)$$

where, $P_{comp}(t|s)$, $P_{dict}(t|s)$ and $P_{translit}(t|s)$ represent distribution probabilities derived from the comparable corpora, bilingual dictionaries and the transliteration model, respectively. Parameters α_1 to α_3 are models dependant and represent the importance of each translation strategy, with $\sum_{i=1,..3} \alpha_i = 1$.

Translation alternatives are ranked according to the combined probability. A fixed number of top-ranked translation candidates are selected and misleading candidates are discarded.

4 Experiments and Evaluations

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

4.1 Linguistic Resources

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

Morphological analyzers, *ChaSen* [1] version 2.2.9 [10] for texts in Japanese and *OAK*[2] [16] for English texts were used in linguistic pre-processing.

EDR [5] and *EDICT* [3] bilingual Japanese-English dictionaries were used in translation.

KAKASI[4], a language processing inverter and free software, available on the Internet was used in the transliteration process of Japanese terms written in katakana to English. Corrections on the transliteration were completed manually by a native Japanese language speaker.

NTCIR-2 [7], a large-scale test collection was used to evaluate the proposed strategies in CLIR.

SMART information retrieval system [14], which is based on vector model, was used to retrieve English documents.

4.2 Results and Evaluations

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 Japanese and English terms. Similarity vectors were constructed for Japanese-English pairs of terms.

We conducted experiments and evaluations on the monolingual and bilingual tasks of NTCIR test collection.

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.

Results and performances of different translation models and their combination are described in Table 1. Evaluations were based on the average precision, differences in term of average precision of the monolingual counterpart and the improvement over the monolingual counterpart. Fig. 2 represents the recall/precision curves for the proposed translation models and their linear combination using SMART retrieval system.

[1] http://chasen.aist-nara.ac.jp/

[2] http://nlp.cs.nyu.edu/oak/

[3] http://www.csse.monash.edu.au/~jwb/wwwjdic.htm

[4] http://kakasi.namazu.org/

Table. 1 Results and Evaluations on different tramslation models and their combination using NTCIR test collection

Translation Model		Avg. Precision	% Monolingual	% Difference Improvement		
ME -	*Monolingual English*	<u>**0.2683**</u>	<u>100</u>	-	-	-
DT -	*Dictionary and Transliteration*	<u>0.2279</u>	<u>84.94</u>	<u>- 15.05</u>	-	-
SCC -	*Comparable Corpora*	<u>0.1417</u>	<u>52.81</u>	<u>- 47.18</u>	<u>-37.82</u>	-
DT&SCC -	*Linear Combination*	<u>**0.2366**</u>	<u>**88.18**</u>	<u>**-11.81**</u>	<u>**+3.82**</u>	<u>**+66.97**</u>

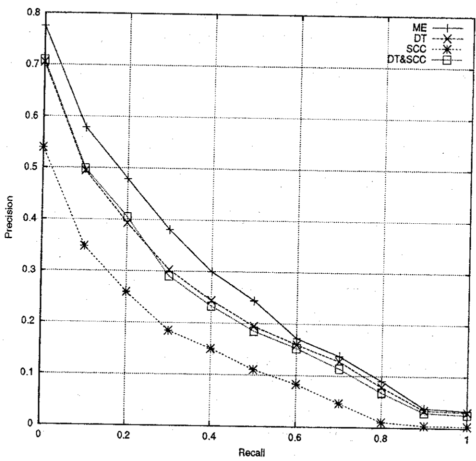

Fig. 2 Recall/Precision curves for the proposed translation model and their linear combination

The combined dictionary-based and transliteration model *'DT'* showed 84.94% improvement of the monolingual retrieval, while the comparable corpora-based model *'SCC'* showed a lower improvement in average precision compared to the monolingual retrieval and the combined dictionary-based and transliteration model *'DT'* with 52.81% of the monolingual retrieval. The proposed combination of comparable corpora, bilingual dictionaries and transliteration *'DT&SCC'* showed the best performance in terms of average precision with 88.18% of the monolingual counterpart, +3.82% compared to the dictionary-based method and +66.97 compared to the comparable corpora model taken alone.

5 Conclusion

We investigated an approach of extracting bilingual terminology from comparable corpora with an application on Japanese-English language pair. A combined model involving comparable corpora, readily available bilingual dictionaries and transliteration was found very efficient and could be used to enrich bilingual lexicons and thesauri. Most of the selected terms were considered as translation candidates or expansion terms in CLIR. Exploiting different translation models revealed to be effective.

Ongoing research is focused on transliteration of the special phonetic alphabet, *katakana* in the case of Japanese language. Techniques on phrasal translation will be investigated in order to select best phrasal translation alternatives in CLIR.

Acknowledgments

The present research is supported in part by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] Dagan, I., Itai, I.: Word Sense Disambiguation using a Second Language Monolingual Corpus. *Computational Linguistics*, Vol. 20, No. 4, pp. 563-596 (1994).

[2] Dejean, H., Gaussier, E., Sadat, F.: An Approach based on Multilingual Thesauri and

Model Combination for Bilingual Lexicon Extraction. *Proc.COLING'02*, pp. 218-224 (2002).

[3] Diab, M., Finch, S.: A Statistical Word-Level Translation Model for Comparable Corpora. *Proc. Conference on Content-based Multimedia Information Access RIAO* (2000).

[4] Dunning, T.: Accurate Methods for the Statistics of Surprise and Coincidence. *Computational linguistics*, Vol. 19, No.1, pp. 61-74 (1993).

[5] EDR.: Japan Electronic Dictionary Research Institute, Ltd. EDR electronic dictionary version 1.5 technical guide. *Technical report TR2-007, Japan Electronic Dictionary research Institute, Ltd* (1996).

[6] Fung, P.: A Statistical View of Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora. *In Jean Véronis, Ed. Parallel Text Processing* (2000).

[7] Kando, N.: Overview of the Second NTCIR Workshop. Proc. Second NTCIR Workshop on Research in Chinese and Japanese Text Retrieval and text Summarization (2001)

[8] Knight, K., Graehl, J.: Machine Transliteration. *Computational Linguistics*, Vol. 24, No. 4 (1998).

[9] Koehn, P., Knight, K.: Learning a Translation Lexicon from Monolingual Corpora. *Proc. ACL-02 Workshop on Unsupervised Lexical Acquisition* (2002).

[10] Matsumoto, Y., Kitauchi, A., Yamashita, T., Imaichi, O., and Imamura, T: Japanese morphological analysis system ChaSen manual. *Technical report NAIST-IS-TR97007, NAIST* (1997).

[11] Nakagawa, H.: Disambiguation of Lexical Translations Based on Bilingual Comparable Corpora. *Proc. LREC2000, Workshop of Terminology Resources and Computation WTRC2000*, pp.33-38 (2000).

[12] Peters, C., Picchi, E.: Capturing the Comparable: A System for Querying Comparable Text Corpora. *Proc. 3rd International Conference on Statistical Analysis of Textual Data*, pp. 255-262 (1995).

[13] Rapp, R.: Automatic Identification of Word Translations from Unrelated English and German Corpora. *Proc. EACL'99* (1999).

[14] Salton, G.: The SMART Retrieval System, Experiments in Automatic Documents Processing. *Prentice-Hall, Inc., Englewood Cliffs, NJ* (1971).

[15] Salton, G., McGill, J. Introduction to Modern Information Retrieval. *New York, Mc Graw-Hill* (1983).

[16] Sekine, S.: OAK System– Manual. *New York University,* (2001).

[17] Shahzad, I., Ohtake, K., Masuyama, S., Yamamoto, K.: Identifying Translations of Compound Using Non-aligned Corpora. *Proc. Workshop MAL*, pp. 108-113 (1999).

[18] Tanaka, K., Iwasaki, H.: Extraction of Lexical Translations from Non-Aligned Corpora. Proc. COLING'96 (1996).