Effects of Comparable Corpora on Cross-language Information Retrieval

Fatiha Sadat

University of Quebec in Montreal, Computer Science Department
201 President Kennedy avenue, Montreal, QC, Canada

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

1 Introduction

Large text corpora represent a crucial resource for bilingual terminology acquisition and multilingual lexical resources enrichment. Moreover, in recent years non-aligned comparable corpora have been an object of studies and research related to natural language processing and information retrieval (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. 2001; Tanaka and Iwasaki 1996), because of their availability and easy accessibility through the World Wide Web.

In the present paper, our goal consists on learning 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 a large-scale test collection NTCIR[1] for (Japanese, English) language pair. CLIR consists of retrieving documents written in one language using query terms in another language.

[1] http://research.nii.ac.jp/ntcir/

sadat F.
Effects of Comparable Corpora on Cross-Language Information Retrieval.
DOI: 10.5220/0003029200530059
In *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science (ICEIS 2010)*, page
ISBN: 978-989-8425-13-3

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, 5 and 6. Section 7 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 order to enrich existing bilingual dictionaries, thesauri and retrieve documents across different languages.

In the present study, we follow the proposed model by (Dejean et al. 2002; Fung 2000; Rapp 19992). First, word frequencies, context word frequencies in surrounding positions (here three-words window) are estimated 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* (Dunning 1993) as expressed in equation (1):

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

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* (Salton and McGill, 1983), as expressed in equation (2):

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

where,
v_{ik} represents co-occurrence frequencies in context vectors of the source term w_i with term w_k. and v_{jk} represents co-occurrence frequencies in context vectors of the target term w_j with term w_k..
Therefore, similarity vectors are constructed to yield a probabilistic translation model $P_{comp}(t|s)$ for bilingual terminology extraction from comparable corpora.

3 Linear Combination of Different Translation Models

Combining different models has showed success in previous research (Dejean et al. 2002). We propose a combined probabilistic translation 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.
Fig. 1 presents an overview of the proposed approach in CLIR combining different translation models such as the comparable corpora, bilingual dictionaries and transliteration.
General-purpose dictionaries are basic source of 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.
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 (Knight and Graehl 1998). 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.
Finally, translation alternatives are ranked according to the combined probability. A fixed number of top-ranked translation candidates are selected for each source term and misleading candidates are discarded.
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'.
Therefore, the combined probabilistic model will involve distribution probabilities 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 (3):

$$P(t|s) = \alpha_1 P_{comp}(t|s) + \alpha_2 P_{dict}(t|s) + \alpha_3 P_{translit}(t|s) \quad (3)$$

Parameters α_1 to α_3 are models dependant and represent the importance of each translation strategy, with $\sum_{i=1,..3} \alpha_i = 1$.

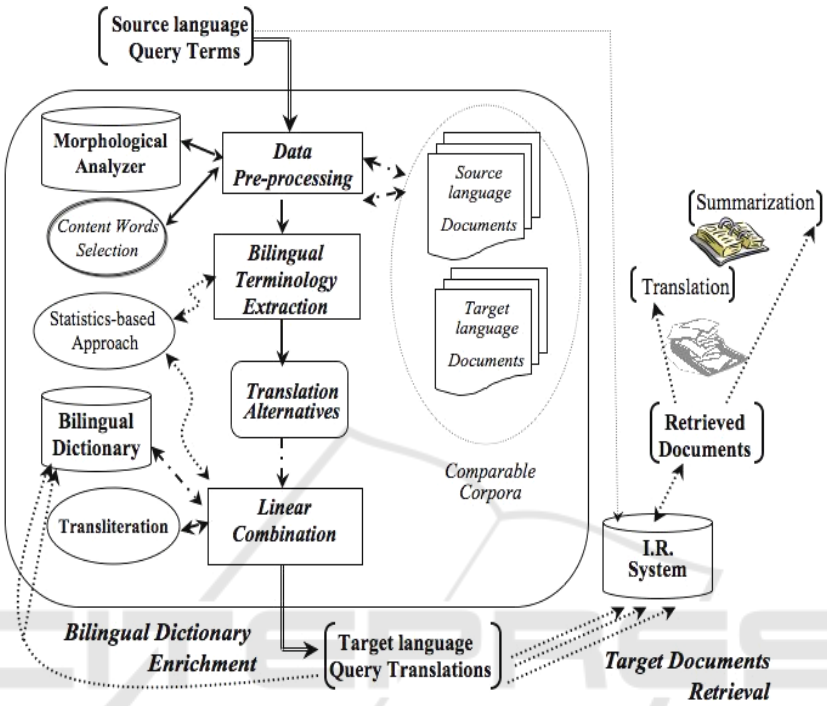

Fig. 1. Overview of the proposed approach combining different translation models in CLIR.

4 Experiments and Evaluation in CLIR

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. Moreover, documents of *NTCIR-2* test collection were considered as comparable corpora in order to cope with special features of the test collection during evaluations.

- Morphological analyzers, *ChaSen*[2] version 2.2.9 (Matsumoto et al. 1997) for texts in Japanese and *OAK*[3] (Sekine 2001) for English texts were used in linguistic pre-processing.
- *EDR* (EDR 1996) and *EDICT*[4] bilingual Japanese-English dictionaries were used in translation.
- *KAKASI*[5], 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 transliteration were completed manually by a native Japanese language speaker.
- *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 model, was used to retrieve English documents.

4.2 Results and Discussion

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.

Table 1. Results and Evaluations on different translation models and their combination.

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

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

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

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

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

Topics 0101 to 0149 were considered and key terms contained in 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.
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 Conclusions

We investigated the 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 and phrasal translation in CLIR.

References

1. Dagan, I., Itai, I. Word Sense Disambiguation using a Second Language Monolingual Corpus. *Computational Linguistics* 20(4): 563-596. (1994).
2. Dejean, H., Gaussier, E., Sadat, F. An Approach based on Multilingual Thesauri and Model Combination for Bilingual Lexicon Extraction. In *Proceedings of COLING'02*, Taiwan, pp 218-224. (2002)
3. Diab, M., Finch, S. A Statistical Word-Level Translation Model for Comparable Corpora. In *Proceedings of the Conference on Content-based Multimedia Information Access RIAO*. (2000
4. Dunning, T. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational linguistics* 19(1): 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. In Proceedings of the Second NTCIR Workshop on Research in Chinese and Japanese Text Retrieval and text Summarization, Tokyo. (2001)
8. Knight, K., Graehl, J. Machine Transliteration. Computational Linguistics 24 (4). (1998)
9. Koehn, P., Knight, K. Learning a Translation Lexicon from Monolingual Corpora. In *Proceedings of 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. In *Proceedings of 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. In *Proceedings of the Third 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. In *Proceedings of 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. (1999) Identifying Translations of Compound Using Non-aligned Corpora. In *Proceedings of Workshop MAL*, pp 108-113.
18. Tanaka, K., Iwasaki, H. Extraction of Lexical Translations from Non-Aligned Corpora. In *Proceedings of COLING'96*. (1996)