Bilingual Terminology Acquisition from Comparable Corpora and Phrasal Translation to Cross-Language Information Retrieval

Fatiha Sadat Nara Institute of Science and Technology Nara, 630-0101, Japan

Masatoshi Yoshikawa Nagoya University Nagoya, 464-8601, Japan

Shunsuke Uemura Nara Institute of Science and Technology Nara, 630-0101, Japan

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

Abstract

The present paper will seek to present an approach to bilingual lexicon extraction from non-aligned comparable corpora, phrasal translation as well as evaluations on Cross-Language Information Retrieval. A two-stages translation model is proposed for the acquisition of bilingual terminology from comparable corpora, disambiguation and selection of best translation alternatives according to their linguistics-based knowledge. Different re-scoring techniques are proposed and evaluated in order to select best phrasal translation alternatives. Results demonstrate that the proposed translation model yields better translations and retrieval effectiveness could be achieved across Japanese-English language pair.

1 Introduction

Although, corpora have been an object of study of some decades, recent years saw an increased interest in their use and construction. With this increased interest and awareness has come an expansion in the application to knowledge acquisition, such as bilingual terminology. In addition, non-aligned comparable corpora have been given a special interest in bilingual terminology acquisition and lexical resources enrichment (Dejean et al., 2002; Fung, 2000; Koehn and Knight, 2002; Rapp, 1999).

This paper presents a novel approach to bilingual terminology acquisition and disambiguation from scarce resources such as comparable corpora, phrasal translation through re-scoring techniques as well as 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 completed on a large-scale test collection, NTCIR for Japanese-English language pair.

2 The Proposed Translation Model in CLIR

Figure 1 shows the overall design of the proposed translation model in CLIR consisting of three main parts as follows:

- ***Bilingual terminology acquisition from bi-directional comparable corpora***, completed through a two-stages term-by-term translation model.

- ***Linguistic-based pruning***, which is applied on the extracted translation alternatives in order to filter and detect terms and their translations that are morphologically close enough, i.e., with close or similar part-of-speech tags.

- ***Phrasal translation***, completed on the source query after re-scoring the translation alternatives related to each source query term. The proposed re-scoring techniques are based on the World Wide Web (WWW), a large-scale test collection such as NTCIR, the comparable corpora or a possible interaction with the user, among others.

Finally, a *linear combination* to bilingual dictionaries, bilingual thesauri and transliteration for the special phonetic alphabet of foreign words and loanwords, would be possible depending on the cost and availability of linguistic resources.

2.1 Two-stages Comparable Corpora-based Approach

The proposed two-stages approach on bilingual terminology acquisition and disambiguation from comparable corpora (Sadat et al., 2003) is described as follows:

- Bilingual terminology acquisition from source language to target language to yield a first translation model, represented by similarity vectors $SIM_{S \rightarrow T}$.

- Bilingual terminology acquisition from target language to source language to yield a second translation model, represented by similarity vectors $SIM_{T \rightarrow S}$.

- Merge the first and second models to yield a two-stages translation model, based on bi-directional comparable corpora and represented by similarity vectors $SIM_{(S \leftrightarrow T}$.

We follow strategies of previous researches (Dejean et al., 2002; Fung, 2000; Rapp, 1999) for the first and second models and propose a merging and disambiguation process for the two-stages translation model. Therefore, context vectors of each term in source and target languages are constructed following a statistics-based metric. Next, context vectors related to source words are translated using a preliminary bilingual seed lexicon. Similarity vectors $SIM_{S \rightarrow T}$ and $SIM_{T \rightarrow S}$ related to the first and second models respectively, are constructed for each pair of source term and target translation using the cosine metric.

The merging process will keep common pairs of source term and target translation (s,t) which appear in $SIM_{S \rightarrow T}$ as (s,t) but also in $SIM_{T \rightarrow S}$ as (t,s), to result in combined similarity vectors $SIM_{S \leftrightarrow T}$ for each pair (s,t).The product of similarity values in vectors $SIM_{S \rightarrow T}$ and $SIM(_{T \rightarrow S}$ will yield similarity values in $SIM_{S \leftrightarrow T}$ for each pair (s,t) of source term and target translation.

2.2 Linguistics-based Pruning

Morphological knowledge such as Part-of-Speech (POS), context of terms extracted from thesauri could be valuable to filter and prune the extracted translation candidates. POS tags are assigned to each source term (Japanese) via morphological analysis.

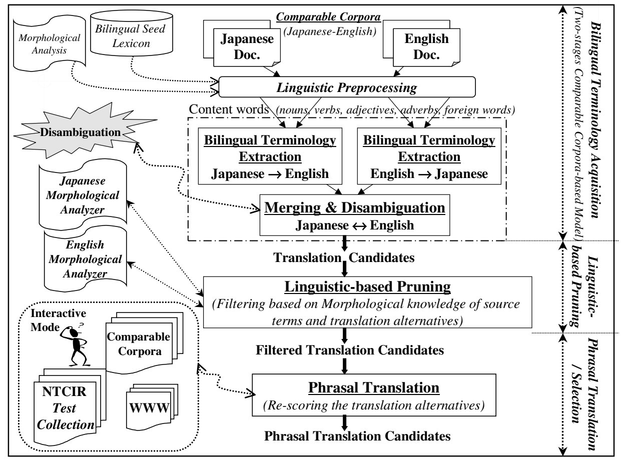

Figure 1: The Overall Design of the Proposed Model for Bilingual Terminology Acquisition and Phrasal Translation in CLIR

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 pair of languages, Japanese nouns and verbs are compared to English nouns and verbs, respectively. Japanese adverbs and adjectives are compared to English adverbs and adjectives, because of the close relationship between adverbs and adjectives in Japanese (Sadat et al., 2003).

Finally, the generated translation alternatives are sorted in decreasing order by similarity values and rank counts are assigned in increasing order. A fixed number of top-ranked translation alternatives are selected and misleading candidates are discarded.

2.3 Phrasal Translation

Query translation ambiguity can be drastically mitigated by considering the query as a phrase and restricting the single term translation to those candidates that were selected by the proposed combined statistics-based and linguistics-based approach (Sadat et al., 2003). Therefore, after generating a ranked list of translation candidates for each source term, re-scoring techniques are proposed to estimate the coherence of the translated query and decide the best phrasal translation.

Assume a source query Q having n terms $\{s_1 \ldots s_n\}$. Phrasal translation of the source query Q

is completed according to the selected top-ranked translation alternatives for each source term s_i and a re-scoring factor RF_k, as follows:

$$Q_{phras} = \sum_{k=1..thres} [Q_k(s_1..s_n) \times RF_k(t_1..t_n; s_1..s_n)]$$

Where, $Q_k(s_1..s_n)$ represents the phrasal translation candidate associated to rank k. The re-scoring factor $RF_k(t_1..t_n; s_1..s_n)$ is estimated using one of the re-scoring techniques, described below.

Re-scoring through the WWW

The WWW can be considered as an exemplar linguistic resource for decision-making (Grefenstette, 1999). In the present study, the WWW is exploited in order to re-score the set of translation candidates related to the source terms.

Sequences of all possible combinations are constructed between elements of sets of highly ranked translation alternatives. Each sequence is sent to a popular Web portal (here, *Google*) to discover how often the combination of translation alternatives appears. Number of retrieved WWW pages in which the translated sequence occurred is used to represent the re-scoring factor RF for each sequence of translation candidates. Phrasal translation candidates are sorted in decreasing order by re-scoring factors RF. Finally, a number (*thres*) of highly ranked phrasal translation sequences is selected and collated into the final phrasal translation.

Re-scoring through a Test Collection

Large-scale test collections could be used to re-score the translation alternatives and complete a phrasal translation. We follow the same steps as the WWW-based technique, replacing the WWW by a test collection and a retrieval system to index documents of the test collection.

NTCIR test collection (Kando, 2001) could be a a good alternative for Japanese-English language pair, especially if involving the comparable corpora.

Re-scoring through the Comparable Corpora

Comparable corpora could be considered for the disambiguation of translation alternatives and thus selection of best phrasal translations (Sadat et al., 2002). Our proposed algorithm to estimate the re-scoring factor RF, relies on the source and target language parts of the comparable corpora using statistics-based measures. Co-occurrence tendencies are estimated for each pair of source terms using the source language text and each pair of translation alternatives using the target language text.

Re-scoring through an Interactive Mode

An interactive mode (Ogden and Davis, 2000) could help solve the problem of phrasal translation. The interactive environment setting should optimize the phrasal translation, select best phrasal translation alternatives and facilitate the information access across languages. For instance, the user can access a list of all possible phrases ranked in a form of hierarchy on the basis of word ranks associated to each translation alternative. Selection of a phrase will modify the ranked list of phrases and will provide an access to documents related to the phrase.

3 Experiments and Evaluations 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. Collections of news articles from Mainichi Newspapers (1998-1999) for Japanese and Mainichi Daily News (1998-1999) for English were considered as comparable corpora. 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. NTCIR-2 (Kando, 2001) 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.

Thus, Content words (nouns, verbs, adjectives, adverbs) were extracted from English and Japanese texts. Morphological analyzers, ChaSen version 2.2.9 (Matsumoto and al., 1997) for texts in Japanese and OAK2 (Sekine, 2001) for texts in English were used in linguistic pre-processing. EDR (EDR, 1996) was used to translate context vectors of source and target languages.

First experiments were conducted on the several combinations of weighting parameters and schemes of SMART retrieval system for documents terms and query terms. The best performance was realized by ATN.NTC combined weighting scheme.

The proposed two-stages model using comparable corpora showed a better improvement in terms of average precision compared to the simple model (onestage comparable corpora-based translation) with +27.1% and a difference of -32.87% in terms of average precision of the monolingual retrieval. Combination to linguistics-based pruning showed a better performance in terms of average precision with +41.7% and +11.5% compared to the simple comparable corpora-based model and the two-stages comparable corpora-based model, respectively.

Applying re-scoring techniques to phrasal translation yields significantly better results with 10.35%, 8.27% and 3.08% for the WWW-based, the NTCIR-based and comparable corpora-based techniques, respectively compared to the hybrid two-stages comparable corpora and linguistics-based pruning.

The proposed approach based on bi-directional comparable corpora largely affected the translation because related words could be added as translation alternatives or expansion terms. Effects of extracting bilingual terminology from bi-directional comparable corpora, pruning using linguistics-based knowledge and re-scoring using different phrasal translation techniques were positive on query translation/expansion and thus document retrieval.

4 Conclusion

We investigated the approach of extracting bilingual terminology from comparable corpora in order to enrich existing bilingual lexicons and enhance CLIR. We proposed a two-stages translation model involving extraction and disambiguation of the translation alternatives. Linguistics-based pruning was highly effective in CLIR. Most of the selected terms can be considered as translation candidates or expansion terms. Exploiting different phrasal translation techniques revealed to be effective in CLIR. Although we conducted experiments and evaluations on Japanese-English language pair, the proposed translation model is common across different languages.

Ongoing research is focused on the integration of other linguistics-based techniques and combination to transliteration for *katakana*, the special phonetic alphabet to Japanese language.

References

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, Taipei, Taiwan.*

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

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

G. Grefenstette. 1999. The WWW as a Resource for Example-based MT Tasks. *In ASLIB'99 Translating and the Computer 21.*

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.*

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.*

W. C. Ogden and M. W. Davis. 2000. Improving Cross-Language Text Retrieval with Human Interactions. *In Proc. 33rd Hawaii International Conference on System Sciences.*

R. Rapp. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. *In Proc. European Association for Computational Linguistics EACL'99.*

F. Sadat, A. Maeda, M. Yoshikawa and S. Uemura. 2002. Exploiting and Combining Multiple Resources for Query Expansion in Cross-Language Information Retrieval. *IPSJ Transactions of Databases, TOD 15*, 43(SIG 9):39–54.

F. Sadat, M. Yoshikawa and S. Uemura. 2003. Learning Bilingual Translations from Comparable Corpora to Cross-Language Information Retrieval: Hybrid Statistics-based and Linguistics-based Approach. *In Proc. IRAL 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.*

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