Enhancing Cross-language Information Retrieval by an Automatic Acquisition of Bilingual Terminology from Comparable Corpora

Fatiha Sadat
Nara Institute of Science and Technology
8916-5 Takayama Ikoma, Nara
630-0101 Japan
fatia-s@is.aist-nara.ac.jp

Masatoshi Yoshikawa
Nagoya University
Furo-cho, Chikusa-ku
Nagoya, 464-8601 Japan
yosikawa@itc.nagoya-u.ac.jp

Shunsuke Uemura
Nara Institute of Science and Technology
8916-5 Takayama Ikoma, Nara
630-0101 Japan
uemura@is.aist-nara.ac.jp

ABSTRACT

This paper presents an approach to bilingual lexicon extraction from comparable corpora and evaluations on Cross-Language Information Retrieval. We explore a bi-directional extraction of bilingual terminology primarily from comparable corpora. A combined statistics-based and linguistics-based model to select best translation candidates to phrasal translation is proposed. Evaluations using a large test collection for Japanese-English revealed the proposed combination of bi-directional comparable corpora, bilingual dictionaries and transliteration, augmented with linguistics-based pruning to be highly effective in Cross-Language Information Retrieval.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval.

General Terms

Design, Experimentation.

Keywords

Comparable corpora, Cross-Language Information Retrieval, Bilingual lexicon extraction, Transliteration, Disambiguation, Part-of-Speech.

1. INTRODUCTION

Recently, 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 [1], [2], [4], [5].

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.
ACM 1-58113-646-3/03/0007.

In the present paper, we are concerned by exploiting scarce resources for bilingual terminology acquisition and disambiguation. Evaluations on Cross-Language Information Retrieval (CLIR), which consists of retrieving documents written in one language using queries written in another language, is another interest. We propose a novel approach to learning from comparable corpora and extracting a bilingual lexicon. A linear combination with dictionary-based translation and transliteration of foreign words and loanwords is provided. Linguistics-based pruning to select best translation alternatives is proposed as well. Finally, an application is conducted on NTCIR, a large test collection for Japanese, English language pair.

2. The PROPOSED APPROACH

We propose a two-staged approach for the acquisition and disambiguation of bilingual terminology from comparable corpora as follows:

- Bilingual terminology acquisition from source language to target language to yield a first probabilistic translation model $P_{s\rightarrow t}(t|s)$.
- Bilingual terminology acquisition from target language to source language to yield a second probabilistic translation model $P_{t\rightarrow s}(s|t)$.
- Merge the first and second models to yield a two-stages probabilistic translation model $P_{s\leftrightarrow t}(s|t)$, based on bi-directional comparable corpora.

We follow strategies of previous researches [1], [2], [5] for the first and second models. Therefore, context vectors for each source term and each target term are constructed, using statistics-based collocation criteria. Next, context vectors of 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 enrich using our proposed bootstrapping approach of the present study. Similarity vectors are constructed for each pair of source and target terms using the cosine metric.

Next, the constructed similarity vectors $SIM_{(S\rightarrow T)}$ and $SIM_{(T\rightarrow S)}$ for the first and second models respectively, are merged to result in bi-directional bilingual terminology acquisition from comparable corpora. 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 of both vectors $SIM_{(T\rightarrow S)}$ for a pair (s,t) and $SIM_{(S\rightarrow T)}$ for a

pair *(t,s)* will result in similarity values of vectors $SIM_{(S \leftrightarrow T)}$ for a pair *(s,t)*.

3. LINEAR COMBINATION

Combining different translation models has showed success in previous research [1]. We propose a combined translation model involving comparable corpora, readily available bilingual dictionaries and transliteration. The special phonetic alphabet (here Japanese *katakana*) to foreign words and loanwords requires *romanization* or transliteration [3]. Moreover, the proposed linear combination includes a linguistics-based pruning technique in order to filter the translation candidates on the basis of their morphological knowledge (POS, context), to select translation alternatives, which are morphologically close enough to the source term and to discard the misleading translation candidates. The combined probabilistic lexical model is represented by the following equation:

$$P(t \mid s) = \alpha_1 P_{comp}(t \mid s) + \alpha_2 P_{dict}(t \mid s) + \alpha_3 P_{translit}(t \mid s) + \alpha_4 P_{morph}(t \mid s)$$

where, $P_{comp}(t|s)$, $P_{dict}(t|s)$, $P_{translit}(t|s)$ and $P_{morph}(t|s)$ represent distribution probabilities derived from the comparable corpora, the bilingual dictionary, the transliteration model and the augmented model with linguistics-based pruning, respectively. Parameters α_1 to α_4 are models dependant and represent the importance of each translation strategy with $\sum_{i=1,..4} \alpha_i = 1$.

We propose two sorts of combinations as follows:

- *Pre-combination*, where the comparable corpora-based model is augmented with linguistics-based pruning, then combined with dictionary-based and transliteration models.
- *Post-combination*, where linguistics-based pruning is applied on the combined comparable corpora-based, dictionary-based and transliteration model.

Finally, 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 for phrasal translation and misleading translation candidates are discarded.

4. EVALUATION IN CLIR

A collection of news articles from *Mainichi Newspapers* (1998-1999) for Japanese and *Mainichi Daily News* (1998-199) for English was considered as comparable corpora. . We have also considered documents of *NTCIR* test collection as comparable corpora in order to cope with special features of the test collection during evaluations. Linguistic pre-processing was completed via morphological analyzers, *ChaSen* for Japanese texts and *OAK* for English texts. We focused on content words, i.e., nouns, adjectives, adverbs, verbs and foreign words. *EDR* and *EDICT* bilingual Japanese-English dictionaries were used in translation. *NTCIR* test collection and *SMART* retrieval system were used to evaluate the proposed strategies in CLIR. Results and performances of different models and combinations are described in Table 1.

Table 1. Results and evaluations on different translation models and combinations

	Method	Avg. Precision	% Mono.	% Diff-Improv
1	*Mono_Eng*	**0.2683**	**100**	**–**
2	*Dict+Translit*	0.2279	84.94	-15.05
3	*Uni-Corp*	0.1417	52.81	-47.18
4	*Bi-Corp*	0.1801	67.12	-32.87
5	*2+4*	**0.2721**	**101.41**	+ 1.41
6	*4+ Morph+2*	**0.2987**	**111.33**	**+11.33**
7	*4+2+ Morph*	**0.3030**	**112.93**	**+12.93**

The proposed two-stages model using comparable corpora '*4*' showed a better improvement in average precision compared to '*3*', the simple model (one stage) and approached the performance of the dictionary-based model *'2'* with 79.02%. Combined models '*2+4*' showed better performances, especially when using the linguistics-based pruning in pre-combination *'6'* and post-combination *'7'*. Our proposed combinations outperformed the monolingual retrieval '*1*' and confirm the intuition that monolingual performance is not necessarily the upper bound in CLIR.

5. CONCLUSION

We proposed and evaluated a novel approach to extracting bilingual terminology from comparable corpora in CLIR. Exploiting different translation models revealed to be highly effective. Ongoing research includes word sense disambiguation, phrasal translation and thesauri enrichment.

6. ACKNOWLEDGMENTS

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.

7. REFERENCES

[1] Dejean, H., Gaussier, E., and Sadat, F. An Approach based on Multilingual Thesauri and Model Combination for Bilingual Lexicon Extraction. In Proceedings of COLING 2002, pp. 218-224, 2002.

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

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

[4] Koehn, P., and Knight, K. Learning a Translation Lexicon from Monolingual Corpora. In Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition. 2002.

[5] Rapp, R. Automatic Identification of Word Translations from Unrelated English and German Corpora. In proceedings of EACL 1999.