Automatic Acquisition of Phrasal Knowledge for English-Chinese Bilingual Information Retrieval

Ming-Jer Lee
Institute of Information Science
Academia Sinica
mjlee@iis.sinica.edu.tw

Lee-Feng Chien
Institute of Information Science,
Academia Sinica
lfchien@iis.sinica.edu.tw

Abstract Extraction of phrasal knowledge, such as proper names, domain-specific keyphrases and lexical templates from a domain-specific text collection are significant for developing effective information retrieval systems for the Internet. In this paper, we are going to introduce our ongoing research on automatic phrasal knowledge acquisition for English-Chinese bilingual texts. The underlying techniques consist of adaptive keyphrase extraction, lexical template extraction, phrase translation extraction and high-order Markov language model construction. In addition to the increase of retrieval effectiveness, IR systems based on these techniques are expected able to perform much better in many aspects, such as automatic term suggestion, information filtering, text classification and cross-language information retrieval, etc.

1 Introduction

Capability of phrasal information extraction is crucial in an IR system for the increase of retrieval effectiveness and development of advanced searching techniques, such as automatic term suggestion and cross-language information retrieval [1,2]. Considering the inherent linguistic differences between English and Chinese languages[3] and the demand of high-perfomance phrasal knowledge acquisition, we are developing an approach in acquisition of English-Chinese bilingual information automatically from Internet resources. The proposed approach is formed as an abstract diagram shown in Fig. 1, in which an IR system is composed of a knowledge acquisition subsystem for extracting phrasal information on-line. The knowledge acquisition subsystem, which consists of keyphrase extraction module, lexical template extraction module, phrase translation extraction and language model construction module, as will be briefly described below, is served to increase the capability of phrasal information extraction, with the input of domain-specific texts in the IR system. At present, although there were some techniques have been successfully developed [4,5,6]at the initial stage, the whole research is still under exploration.

2 Overview of the Developing Techniques

PAT-tree-based High-order Markov Language Model Construction

Statistical N-gram language models are often used in many NLP systems to estimate the probability values or word associations of any word pairs or sequences. For reducing the complexity of model representation, bigram or trigram models are frequently used as an approximation. Of course, this will decrease the power of language modeling. According to our experiments, techniques such as PAT-tree indexing used for recording full-text documents in IR systems can be more efficient in representing high-order N-gram language models, especially for when the training corpus is large and dynamic[5]. The PAT tree actually provides indices to all possible streams of characters or words with an arbitrary length N, where N can be significantly large than 2 or 3, together with the frequency counts for these streams in the network resource databases. Since the content of the searching database can be taken as the corpus to train a domain-specific language model for the database. All of the required statistical parameters for a N-gram model can be extracted directly from the PAT tree. The language model can be easily adapted with the update of the database content and the corresponding PAT tree. The PAT-tree-based language model is efficient in modeling of phrasal information such as lexical patterns and proper names with the change of database contents. The techniques which will be described below actually rely on the basis.

Adaptive Keyphrase Extraction

The proposed method for adaptive keyphrase extraction is based on the PAT-tree language model, which can reduce the reliance of rigid lexicon and sophisticated word segmentation, and extract keyphrases has no limitation in length [6]. Since words in the Chinese language are formed by a sequence of characters, and words in a sentence are not clearly bounded by spaces as in English [3], all of the character strings in text must be taken as candidates of keywords or keyphrases. The proposed technique is a two-step automatic process. At first, in order to be able to extract character strings with correct and complete word boundaries, an effective context-dependency-based estimation method and filtering algorithm were developed based on the PAT trees. By examining the context dependency among the left and right adjacent patterns for each possible lexical patterns (highly-associated character strings) in the PAT-tree indices, most of the lexical patterns which are incomplete in lexical boundaries and semantics can be filtered out. Then, the remaining lexical patterns which are assumed to be complete in semantics will be further examined using a common-word lexicon, a set of lexical rules, a general-domain PAT tree and a keyphrase determination strategy in the second step. Only lexical patterns which are really specific and significant will be extracted as the keyphrases. Based on the proposed approach, exciting results have been achieved in different applications, such as book indexing, document classification and automatic term suggestion in information retrieval. At present, we are trying to extend the proposed technique for extracting keyphrases in English texts..

Lexical Template Extraction

In addition to the extraction of keyphrases for domain-specific term lexicon construction as in Fig.1, the proposed approach is expected able to identify lexical templates as linguistic cues for extracting significant keyphrases. A lexical template is defined to be a text pattern consisting of separate strings, which frequently occur together and have replaceable keyphrases in the middle. For example, "*approach to .... in...*" in the strings "*approach to* Internet searching *in* Chinese regions" and "*approach to* computer networking *in* application to" is an example of lexical templates. If such a template can be identified, many keyphrases, which are difficult to extract for their low occurrences or with frequently-used composed words, maybe can be treated in certain degree. The technique for such a purpose is still under developing. An initial idea is that, for each data stream with possibility of a keyphrase composed, it will construct a lattice of data streams from a set of top n similar sentences or sentence fragments, which will be retrieved from the searching database. A searching algorithm which can judge whether there are lexical templates constituted in the lattice are developing.

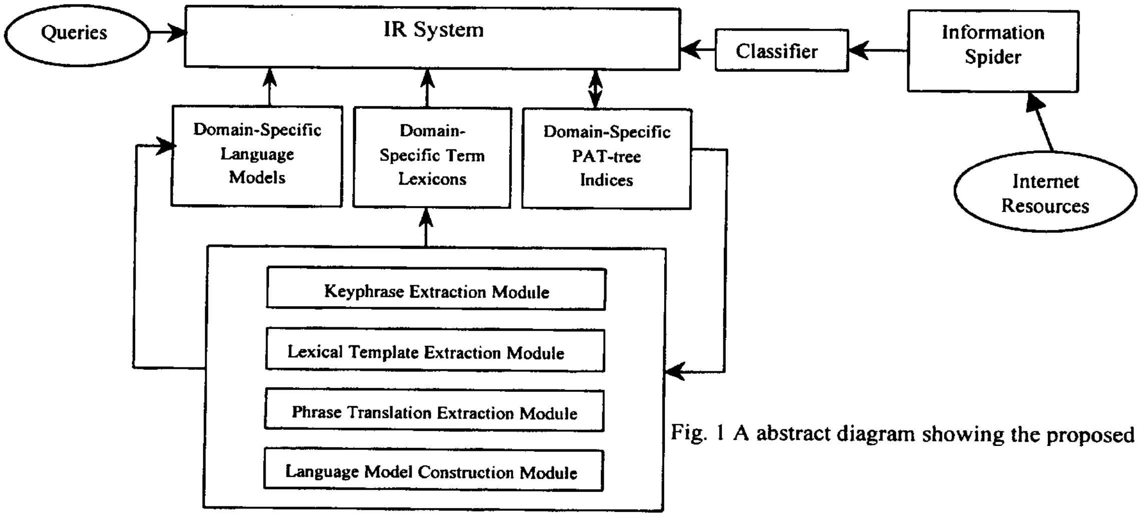
Phrase Translation Extraction

It is very important for cross-language information retrieval, if the developed techniques are capable of extracting representative phrase translations automatically from comparable bilingual text collection. A preliminary study for such a purpose is undertaking in the presented research. The developing technique is basically a three-step process for English-Chinese comparable texts: similar documents clustering, keyphrase extraction and phrase translation extraction. In the step of document clustering, each English document in the examined comparable text will be first translated into corresponding Chinese text and a feature vector regarding to the transcribed text will be assigned, in order to find similar Chinese documents with similarity estimation in Chinese. At the same time, each Chinese document in the text collection will be also assigned a feature vector with the same method. The clustering processing will be performed for each English document to find all of its similar Chinese documents as the output of the step.

After the similar Chinese documents have been found for each English document, the extraction of keyphrases will be performed. For each English document, we extract keywords or keyphrases from its original English text, and extract a set of keyphrases (possible translation equivalents) from its Chinese documents. The used English keyphrase extraction method here is typical, which is mainly based on analysis of POS tagging and compound nouns. But, considering the difficulty of Chinese text segmentation, the adopted Chinese keyphrase extraction is an effective and specially-designed approach [6]. Finally, for each of the extracted English keywords or keyphrases, we will determine translation equivalents in Chinese from the keyphrases extracted from its similar Chinese documents. The method used for determination is based on both linguistic analysis and statistical measurement.

References

1. Lisa Ballesteros and W. Bruce Croft, "Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval," ACM SIGIR'97, 84-91.
2. Douglas W. Oard, "Cross-language Text Retrieval Research in the USA," April 1997. (http://www.area.pi.cnr.it/ErcimDL/third-DELOS-workshop/Oard/oard-delos/paper.html)
3. K. L. Kwok, "Evaluation of an English-Chinese Cross-Lingual Retrieval Experiment". Working Notes on AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford University, 1997
4. Lee-Feng Chien, 'Fast and Quasi-Natural Language Search for Gigabytes of Chinese Texts", ACM SIGIR'95.
5. Lee-Feng Chien ,Sung-Chien Lin, et al., "Internet Chinese Information Retrieval Using Unconstrained Mandarin Speech Queries Based on A Client-Server Architecture and A PAT-tree-based Language Model", Proceedings of the 1997 IEEE International Conference on Acoustics, Speech and Signal Processing, German, pp. 1155-1158 (ICASSP'97).
6. Lee-Feng Chien. "PAT-tree-based Keyword Extraction for Chinese Information Retrieval," ACM SIGIR'97, 50-59.



Queries
IR System
Classifier
Information Spider
Domain-Specific Language Models
Domain-Specific Term Lexicons
Domain-Specific PAT-tree Indices
Internet Resources
Keyphrase Extraction Module
Lexical Template Extraction Module
Phrase Translation Extraction Module
Language Model Construction Module

Fig. 1 A abstract diagram showing the proposed approach.