The terminological tools and challenges of Asian languages term representation.

The Terminological Tools And Challenges Of Asian Languages Term Representation

Hui Li and Schubert Foo
Division of Information Studies
Nanyang Technological University, Singapore

Abstract

Asian languages, such as Chinese, Japanese and Korean (CJK) have one thing in common: they exhibit different linguistics characteristics from English and other western languages. This poses special difficulties since current term representation techniques used in English cannot be applied directly to these languages. Instead, segmentation is a first process in CJK information processing. Segmentation is a term representation technology to extract terminological information from such texts.

Using the case of the Chinese language, this paper reviews the development of terminological tools and the various approaches used in segmentation. Some online resources are introduced to highlight the state of development. Acknowledging that there is a direct relationship between segmentation effectiveness and terminological tools’ quality and complexity, the paper posits that the quality of terminological tools is far been from ideal and that existing resources are far been from adequate. A number of likely trends are identified, namely, the increasing emergence of new and various kinds of electronic terminological tools (CD-ROM or online resources), commercialization of electronic terminological tools, and improved quality through standardization, compatibility and integration of terminological tools, as well as the establishment of new cooperative efforts in creating and maintaining terminological tools.

Introduction

Many Asian languages, unlike its English counterpart, exhibits different linguistics characteristics that poses special difficulties so that current term representation techniques used in English cannot be applied directly to these languages. With the significant growth of Asian economic and social position as well as the increasing requirements for worldwide information exchange and knowledge transfer, research on Asian languages have attracted interest both within and outside Asia.
Asian languages, such as Chinese, Japanese or Korean (CJK for short) exhibits similar characteristics. Segmentation is a first process in CJK information processing. Segmentation is a term representation technology to extract terminological information from such texts. It is thought to play an important role in the fields of language understanding, human-machine interface, machine translation, information retrieval, language processing, artificial intelligence and cognitive psychology. Furthermore, segmentation techniques can also be used in western language phrase-based processing that shares similar problems as CJK. In previous work, little has been done on terminological tools although most segmentation approaches need to use various types of terminological tools. The objective of this paper is to bring attention to these terminological tools. It will use the case of the Chinese language to illustrate the developments of these tools, its problems and possible future directions and developments in terminological tools research and related activities.

**Basic Segmentation Approaches and Terminological Tools**

Text segmentation may be defined as a process of dividing original continuous sentences into meaningfully linguistic units, normally words (Nie et al. 1996; Wu/Tseng 1993, 1995). In the view of knowledge engineering, segmentation is the procedure based on the knowledge and knowledge understanding. The basic approaches of Chinese segmentation can be roughly divided into two groups: character-based approaches and word-based approaches as shown in Figure 1.

![Figure 1. Basic segmentation approaches and terminological tool](image)

Character-based approaches are purely mechanical processes that extract a certain number of characters from texts. It is, therefore, strictly not a segmentation technology. The majority of the segmentation approaches used in term representation are in the category of word-based approaches that can be further distinguished as statistics-based,
dictionary-based, hybrid, and heuristic approaches. In practical systems, statistics-based approaches are usually combined with dictionary-based approaches to become hybrid approaches, while heuristic approaches are dictionary-based but with the incorporation of some terminological knowledge and linguistic rules. Thus, dictionary-based approaches are the basic technology of all the segmentation approaches. At the same time, the dictionary is the basic terminological tool that is used in term representation. The heuristic approaches may involve more complex tools that include thesaurus and knowledge base that can help to recognize more terms according to the relationship between terms and linguistic knowledge respectively.

**Terminological Tools in Dictionary-Based Segmentation**

In using dictionary-based approaches, texts to be divided into segments are first matched against a dictionary. The dictionary is the most common terminological tool used in experimental and practical Chinese information processing systems. There are currently many alternative types of dictionaries used for different research groups although a few groups use the same dictionary in their systems.

According to the type of word collection in the dictionary, they can be identified as a complete dictionary that contains all possible words and phrases used in Chinese texts and component dictionary that stores word and phrase components such as morphemes and simple words. Complete dictionary aims to provide the general knowledge about Chinese words and to make the distinction of valid words. Theoretically, the dictionary should be as complete as possible. Nonetheless, it is infeasible and unnecessary to have a really complete dictionary containing all possible words used in Chinese texts as it is estimated that there are more then ten million words in Chinese language with new words continuing to be produced. In this situation, unknown words that are not stored in the dictionary always exist. Moreover, the more words collected in the dictionary, the higher the likelihood of resulting ambiguous words from segmentation. For instance, a Chinese character string, ABC, can be segmented differently to yield A/B/C, AB/C, A/BC or ABC. Such combinations are potential words in the complete dictionary! On the other hand, the component dictionary is developed for the purpose of alleviating the drawbacks of the complete dictionary by taking into account the characteristics of the Chinese language. Firstly, although the number of Chinese words is almost unlimited, the number of Chinese characters is finite. The 6763 characters in the GB2312-18 (Note: The GB code is the national standard adopted in China) can cover more than 99.99% of all the Chinese characters (Liu 1994). Secondly, Chinese words and phrases that exceed two characters can be formed by one or two character morphemes and words (Wang et al. 1990, Wu/Tseng 1995). However, the morphemes and simple words segmented need to be combined into complete word and phrases again with other non-linguistic or linguistic techniques. This approach is not a purely dictionary lookup and is less convenient to implement. It becomes more difficult when linguistic techniques such as semantic analysis, syntactic analysis and other linguistic knowledge are involved.

When conducting dictionary match, changing the direction of the text scan is proposed as a possible solution to improve results of term representation (Jie et al.1991, Wang et
The basic operations include forward scan (from the left to the right) and backward scan (from the right to the left). Normally, it is thought that leftward scan can produce more correct results than the rightward since compounds or phrases are normally right-headed, that is, important morphemes or words appear at the right of the compounds or phrases. According to the scan direction used in dictionary match, the dictionary can also be divided into common order dictionary and reverse dictionary (Liu 1994). However, these two scan manners cannot guarantee 100% accuracy of term representation. Sometimes both of them produce incorrect results. Furthermore, the reverse dictionary is more difficult to implement in practical systems. At the moment, the dictionaries used in practical systems are almost all common order dictionaries.

Terminological Tools in Heuristic Segmentation

Heuristic segmentation approaches usually apply linguistic rules and knowledge in an attempt to solve the ambiguity problems and detect unknown words that are not shown in the current dictionary. Unknown and ambiguous words constitute the most outstanding problem in current term representation techniques. According to the complexity of the knowledge, the terminological tools may be roughly divided to include thesaurus and knowledge base.

In Chinese text, a concept can be expressed using different words, while sometimes the same word may have different meanings in the context. A thesaurus that explores various relationships (equivalence, hierarchical and non-hierarchical) between terms can find more potential term knowledge in the document and improve the results of term representation. Using thesaurus in terminology and knowledge management for Chinese information has only a short history in mainland China. The first thesaurus was published in 1964 (Zhang et al. 1996). Since then, there are more than 100 thesauri compiled. However, the use of thesaurus in the term representation is still in the experimental stage. Only a couple of groups reported using thesaurus in their systems (Nie et al. 1996, Lim 1999). To some extent, their thesauri can be seen as an enhanced dictionary but with simple morphologic knowledge to describe terms’ correlation.

To further improve the effectiveness of term representation, many researchers introduce terminological knowledge base into their systems. Knowledge base is not the traditional terminological tools such as dictionary and thesaurus. It does not store terms directly but collect various linguistic phenomena to describe the rules of term-building, including the usage of every type of terms and the rules of words appearing in the text. The stored linguistic information and knowledge can help to recognize ambiguous terms and detect more potential terms missed by traditional terminological tools such as the dictionary. Knowledge base has become one of the most important terminological tools to push term representation performance close to perfect results. Based on the types and complexity of problems to be solved, the linguistic knowledge includes four groups of morphologic knowledge, sentence structure knowledge, semantic knowledge and pragmatic knowledge. Representatives that use knowledge base include ACTS (Wu/Tseng 1993, 1995), CLARIT (Tong/Zhai 1996), PIRCS (Kwok 1997a, 1997b, Kwok/Grunfeld 1996),

Similar to the dictionary, the linguistic units (terms) and the relationship between these units as well as other linguistic knowledge collected in the thesaurus and knowledge base should be as complete as possible. The current collection method of knowledge bases is normally based on the case study of the particular research group’s corpus. Although the use of a knowledge base is extremely useful for improving term representation performance, the accuracy can be impaired by the existence of too much knowledge that may interfere or conflict with each other. The processing time needed for segmentation is likely to increase as a result of extra processing and inference. Additionally, the success of utilising the knowledge will depend on the extent on how well the model has been designed, implemented and loaded with all the necessary knowledge (data).

**Terminological Tools on the Internet**

The terminological tools play a long-term and essential role in every application field of term representation. However, the creation and maintenance of these tools require huge investments of time and manpower. The accessibility of Internet, and the growing number of available online terminological tools have benefited researchers and organisations alike. Nowadays, many terminological tools for English and many other western languages can be found on the Web. In contrast, the number of such tools for Chinese (or CJK) is quite limited. The online Chinese terminological tools that can be found so far are all dictionaries. One of the most common one can be found at ftp://ftp.ifcss.org/pub/software/data/phrases.dat (GB code). Others include (Tsai 1998):

- Phonetic phrase database: [http://casper.beckman.uiuc.edu/~c-t sai4/chinese/dphphdb.zip](http://casper.beckman.uiuc.edu/~c-t sai4/chinese/dphphdb.zip) (Big-5 code).
- Tsai’s list of Chinese words: [http://casper.beckman.uiuc.edu/~c-t sai4/chinese/tsaiword.zip](http://casper.beckman.uiuc.edu/~c-t sai4/chinese/tsaiword.zip) (Big-5 code).

In addition to the above general dictionaries, a few subject-domain dictionaries exist:

• NetGlos - The multilingual glossary of Internet terminology:
• Network_term:

**Relationship Between Segmentation Effectiveness And Terminological Tools’ Quality**

The quality of terminological tools is affected by the three elements of complexity, data structure and system organisation. These elements will affect the accuracy and speed of text processing, which are thought as the most important criteria to evaluate the performance of term representation techniques. Sufficient complexity does not only mean that there are different types of terminological tools that exist, but also means that every type of terminological tool provides complete terminological information and knowledge for the term processing. Among the different types of terminological tools, dictionary is usually the first to be used, followed by the introduction of more complicated tools such as thesaurus and knowledge bases. As for the content, the dictionary should contain enough entries of words or word components; the thesaurus should provide sufficient relationship description among these words except the words themselves; and the knowledge base should present exhaustive terminological and linguistic characteristics that is found in the full-text documents. The reasonable arrangement of items and organisation of system will improve the matching process of these terminological tools and reduce the time of text processing.

Previous research reported five levels of the highest theoretical accuracy that can be obtained by using different terminological tools and linguistic knowledge (He et al. 1991). In this context, accuracy is actually expressed as the percentage of wrong segments so that a smaller ratio implies better performance and higher accuracy:

- 1st level: the highest accuracy is less than 1/150. It can be obtained by only using dictionary match. No linguistic knowledge is needed. A few ambiguity problems can be solved occasionally.
- 2nd level: the highest accuracy is about 1/689. It can be obtained by combining morphologic knowledge.
- 3rd level: the highest accuracy is about 1/2127. It can be obtained by introducing sentence structure knowledge. The three levels can solve more than 90% ambiguous words.
- 4th level: the highest accuracy is about 1/6250. It can be obtained by applying semantic knowledge.
- 5th level: even more accurate than 1/6250. It can be obtained by using pragmatic knowledge.

It can be seen that the complexity element does indeed affect the segmentation performance directly. The reports from practical systems confirm this relationship. Table 1 shows some representative results of using different terminological tools in Chinese
segmentation. The highest accuracy of using dictionary-only approach is reported as around 91% in Nie’s experiment (Nie et al. 1996). Few practical systems use a pure dictionary-only approach. Instead, a partial or complete knowledge base is commonly used. The Table also shows that the highest reported accuracy of applying knowledge base that includes complete terminological information is as high as 99.94% (Liu 1994).

Table 1. A range of terminological tools, their characteristics and performance

<table>
<thead>
<tr>
<th>System</th>
<th>Basic Terminological Tool</th>
<th>Complexity of Knowledge Base</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic Chinese Text Segmentation System (ACTS)</td>
<td>Component dictionary</td>
<td>Partial syntactic knowledge</td>
<td>90.00%</td>
</tr>
<tr>
<td>Nie’s dictionary approach</td>
<td>Complete dictionary</td>
<td>Morphological knowledge</td>
<td>91.00%</td>
</tr>
<tr>
<td>Association-Backtracting Word Segmentation (ABWS)</td>
<td>Complete dictionary</td>
<td>Morphological and syntactic knowledge</td>
<td>98.60%</td>
</tr>
<tr>
<td>Chinese Automatic Segmentation System (CASS)</td>
<td>Complete dictionary</td>
<td>Complete knowledge base</td>
<td>99.70%</td>
</tr>
<tr>
<td>The Modern Written Chinese Distinguishing Word System (CDWS)</td>
<td>Complete dictionary</td>
<td>Complete knowledge base</td>
<td>99.84%</td>
</tr>
<tr>
<td>Automatic expert system in written Chinese</td>
<td>Complete dictionary</td>
<td>Complete knowledge base</td>
<td>99.94%</td>
</tr>
</tbody>
</table>

Experimental research also reported exciting results using thesaurus. Furthermore, they expressed the belief that the thesaurus with higher quality will be extremely useful to improve the experiment results (Nie et al. 1996, Lim 1999).

Problems and Developments of Terminological Tools for Term Representation

Terminological tools are essential parts of term representation techniques in Chinese text processing. However, the currently usable tools that meet the requirements of term representation are very limited. The lack of usable tools has become one of the biggest obstacles of the research progress. Therefore, validation of terminological tools is considered as an urgent need that needs to be addressed in the terminology society. This section will explore the problems of current terminological tools and proposes a number of areas for development.

The electronic resources (including online resources) of terminological tools are lacking. Although many dictionaries and various thesauri in print exist, they cannot be used directly in current applications of automatic term representation technology. The electronic materials that meet the needs of term processing are quite limited. At the moment, the usable terminological tools that are distributed on the Internet are only
confined to dictionaries. Furthermore, the amount of dictionaries is limited. The majority of these dictionaries are in Big-5 code that is developed to process classical Chinese characters mainly used in Taiwan. Very few dictionaries in GB code exist despite it been commonly accepted as the standard code for modern Chinese characters in mainland China and many other overseas research groups, including the Chinese track of the established Text REtrieval Conference (TREC).

The quality of terminological tools needs to be improved. The traditional print-form terminological tools are governed by a strict set of standards and regulations in compilation and publication. The most important characteristics of terminological tools are standardization and authorization. However, few of the current terminological tools that are available online are from authorized sources, although they have a significant contribution to the term representation research. They are complied by individual research groups or researchers for their particular research purposes. Since the compilation of terminological tools is an extremely complicated systematic task, it is obvious that the product is far from ideal. Both the content and format of these available tools are in the experimental stage. This situation affects the quality of terminological tools to some extent. Different research purposes dictate different focus on terminological information and knowledge collection. This will in turn introduce the problem of incomplete knowledge and reduces reusability in other domains. The major problems in terminology tools that effect quality remain and include:

- Incorrect collection of illegal items such as non-word, incorrect relationship between words and incorrect description of linguistic rules and knowledge.
- Incomplete collection of terminological units (morphemes and words) and linguistic knowledge.
- Non-reasonable factors of terminological tools design (the size, items arrangement, and systematic organisation).

**Future Trends**

The development of terminological tools should keep the pace with the usage of the computer and Internet in terminological research and related activities. Such tools should be made available electronically. Computerization of the tools’ compilation, publication and utilisation is a major necessary development. With such computerization, a set of changes is likely to happen.

The accelerated use of computer paves the way for new and various kinds of electronic terminological tools (CD-ROM or online resources) for terminological research. The trend towards electronic commerce on the Internet is likely to invite commercialization of terminological tools. The high quality and sufficient complexity of terminological tools requires substantial investments of time, money and manpower. The creation of ideal terminological tools can only be stimulated by the corresponding return of commercial interest. Inevitably, commercialization of compilation and publication will become a reality. It will attract more participation of authorized compilers and publishers, more investments in the compilation, and ensure the quality and variety of available
terminological tools. Thus, researchers would have a larger pool of commercial and freeware products at their disposal.

Standardization, compatibility and integration of terminological tools will be improved in order to meet the needs of various terminological research and related activities. These three characteristics will be major considerations of newly printed terminological tools and emerging electronic terminological tools. These will contribute towards the improvement of the tools’ quality and reusability.

New co-operation in creating terminological tools will be established in order to offset the huge costs and investments. Due to the almost unlimited amount of words and linguistic characteristics, it takes significant effort to acquire the terminological information and knowledge as completely as possible. In addition, there is a continuing need for the regular update and production of new terms and terminological phenomena. Likewise, collection procedures and maintenance aspects must be continually supported. It is most unlikely that any individual groups or researchers can generate such ideal tools by their own efforts. Cooperation among vested parties is the possible solution to facilitate the creation and maintenance of terminological tools. This has been recognized very early on in the field of terminology. With the popular use of Internet, new and far-reaching co-operation will be established among the terminological societies all over the world. Finally, the collection, compilation and maintenance of terminological information will also benefit from such co-operative efforts.

Conclusion

Terminological tools are essential long-term basic tools for term representation research. Currently, many terminological tools have been applied in experimental and practical term representation systems. Among these tools, some online dictionaries are available on the Internet. However, the existing resources that can be used directly in term representation technology is limited and far from ideal.

The gaining prominence and popularity of Asian languages (with CJK in particular), coupled with computerization, commercial interest, and co-operation for creating new tools, will result in the creation of new types of high quality tools that meet the needs of term representation and other terminological related activities.

Bibliography


