Construction of Dictionary Network for Asian Languages

Thatsanee Charoenporn[1], Hitoshi Isahara[1], Virach Sornlertlamvanich[2]
[1]Thai Computational Linguistics Laboratory,
CRL Asia Research Center, Thailand
[2]Information Technology R&D Division,
National Electronics and Computer Technology Center, Thailand
E-mail: thatsanee@crl-asia.org, isahara@crl.or.jp, virach@nectec.or.th

1. Abstract

In this paper, we describe a concept of constructing a dictionary network for Asian Languages. We apply the corpus-based method to build, collect and term the lexicon for non-word-boundary languages. Based on some significant language features (such as mutual information, and entropy), C4.5 algorithm is used to extract word entries from a large amount of text collected from WWW. The text is semantically classified into classes. Considering the context, entries in each class are supposed to use for a specific sense. The extracted entries are consequently determined the senses according to the class that each belongs to. As a result, the entries determined the senses by their contexts are automatically developed from WWW text.

2. Introduction

A dictionary is an asset to natural language processing (NLP) systems. It could not deny that the quality of NLP is determined by either the completeness of data or the accuracy of the concept description. Even though constructing dictionary consumes a lot of budget and time, there is no any framework for developing a sharable dictionary resource. One of the major reasons is the un-resemblance of dictionary structure or the aspect of lexicon. Thai machine readable dictionary, therefore, occurred in many versions by many particular scope of each school of development. Besides, constructing method is still conducted in the classical way, that authors collect word entries by their own knowledge and codes the information of each word by referencing from others documentary sources. It would be very difficult to consolidate all up-to-date words and concepts by these means. To wrap up with these formal issues, we propose a novel model for constructing a corpus-based dictionary on the open format and open collaboration. The very huge corpus we have is the up-to-date online text, WWW. The procedure starts by collecting data automatically from WWW. Then word entries will be extracted from the non-boundary word data string by C4.5 learning algorithm. Simultaneously, each text is classified into group depending on its meaning or concept. For example, group of animal, group of astronomy, group of agriculture, group of equipment, and so on. In line with the hypothesis that each classified-text group has an identical relation, each group will represent one meaning or concept. For example if a word occurred in 3 classified-text group, this word will have at most 3 meaning. The scope of meaning of words is, therefore, determined by the sample sentences occurred in each group.

In this paper, we describe the procedure of the dictionary construction, starting with corpus collecting, word entry extraction, text classification, and finally meaning selection. The procedure is language independent. It can be applied to all non-word-boundary languages.

3. Dictionary Construction Procedure

To construct dictionary network, we start from Thai language first, and propose our language-independent method to Asian languages. The steps of dictionary construction are as these following steps; Document gathering, Term extraction, Text classification, Concept classification.

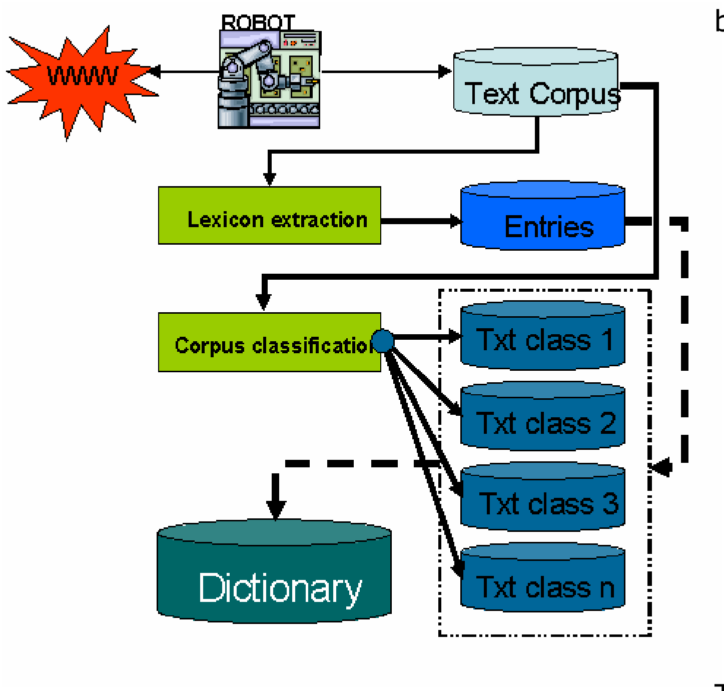


Figure 1 Dictionary Construction Procedure

3.1 Document gathering

Thai document on the defined WWW will be gathered automatically by our robotic system, the crawler. Only texts are extracted to make the collection.

3.2 Term extraction

The most difficult task for constructing a corpus-based dictionary for non-word-boundary languages such as Thai and almost every Asian language are the process of selecting "word" list or entries, since there is no clear criteria what words are.

Sornlertlamvanich et al. (2000) [5] provided an effective algorithm for corpus-based lexicon extraction for the non-word-boundary languages. The algorithm employed the C4.5 decision tree induction program as the learning algorithm for lexicon extraction.

The learned attributes, which are mutual information, entropy, word frequency, word length, functional words, first two and last two characters, are used in building decision tree by the following concepts;

- High mutual information implies that *xyz* co-occurs more than expected by chance. If *xyz* is a word, its Left mutual and Right mutual must be high.
- Entropy shows the variety of characters before and after a word. If *y* is a word, its left and right entropy must be high.
- Words tend to be used more often than non-word string sequences.
- Short strings are more likely to happen by chance. The long and short strings, therefore, should be treated differently.
- Functional words are filtered out for avoiding the misleading of the occurrences of string patterns.
- The considered string must be complied with the spelling rules of the language.

The approach yields about 85% and 56% in precision and recall measures respectively. Following this method, we can select new word entries for our dictionary easily, appropriately and completely without dictionary consulting.

3.3 Text classification

Text classification process will be applied to the gathered document to classify the documents. How to classify is on the process of learning.

By the literature review, we found that there are many approaches or methods on experiment for text classifying. For example; McCallum et al. (1999) classified text by Bootstrapping with Keywords, Expectation-

Maximization and Hierarchical Shrinkage. [3] It uses a small set of keywords per class. The keywords are used to assign approximates labels to the unlabeled documents by term matching, and the labels is the starting point for a bootstrapping process that learns a naïve Bayes classifier using EM and shrinkage.

Hotho et al. (2001) [1] proposed ontology-based text clustering method, by using a simple core ontology for restricting the set of relevant document features and for automatically proposing good aggregations. The aggregations are exploited by the standard clustering algorithm K-Means.

Letnattee et al. (2002) [2] proposed a method to classify Thai text by combining Homogeneous Classifiers for Centroid-based Text Classification by voting and cascading. And by their experiment, the method achieved good performance in either accuracy or time complexity.

Our text classification method will follow these formal systems. The output of this step is the groups of text classified by our method.

3.4 Concept classification

Now we come to the part of defining word's meaning by using concept classification. By the formal step, we got "word entries", and "classified-text groups" for meaning identification.

We identify the meaning of "words" by using the sample sentences available in each "classified-text groups". For example: we got a word "ม้า" [ma:]. It will be linked to the texts in 7 classified-text groups according to its sense as followings;

"land animal"
"marine animal"
"fish"
"star" or "astronomy"
"game"
and "equipment".

The lexicographer or only the native speaker will learn its meaning from the sentences in the text groups. And also human being is now needed to select the appropriate sentence for describing each meaning.

We provide a simple editor for meaning selecting as shown in Figure 2. The text extracted from each group will be loaded to the upper part of the editor. And lexicographer will select the sentences from each group to each sense. The output of this process is word with the sentences described the concept meaning of the word. The figure 2 shows that this word consists of 2 meanings, collected from 2 text groups.

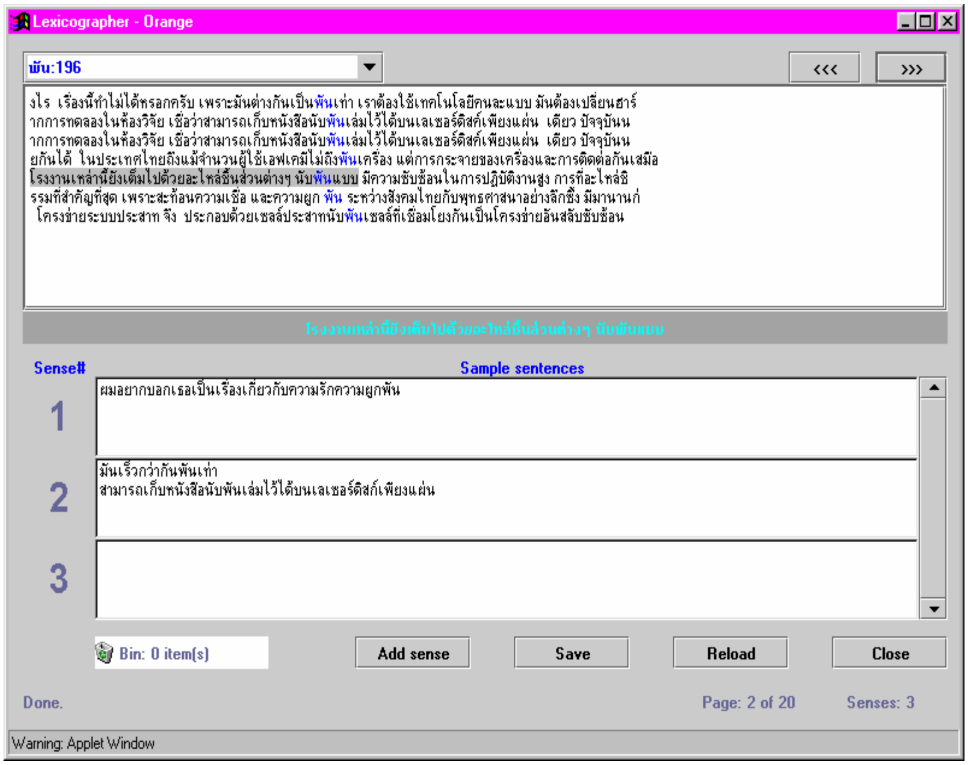
4. Conclusion

Aiming at constructing a dictionary network for Asian Languages, we start with Thai language. The semi-automatic model of dictionary construction is proposed as a common model for non-word-boundary languages. The word list is automatic extracted from the WWW, while the classified text group as the representative meaning of the word is prepared automatically as well. With our proposed method, the corpus-based dictionary is automatically generated from WWW. Native speaker will involve in the final step, to select or verify the word concept or sense classification. Any other Asian languages similar to Thai can also use the same procedure in constructing their own machine readable dictionary.

References

[1] Hotho, A., Mädche, A., Staab, S. (2001) Ontology-based Text Clustering, Workshop "Text Learning: Beyond Supervision", IJCAI 2001.

[2] Letnattee, V. and Theeramunkong, T. (2002) Improving Centroid-based Text Classification Using Term-Distribution-Based Weighting and Feature Selection. Proceedings of The 2nd International Conference on



พัน:196

งไร เรื่องนี้ทำไม่ได้หรอกครับ เพราะมันต่างกันเป็นพันเท่า เราต้องใช้เทคโนโลยีคนละแบบ มันต้องเปลี่ยนฮาร์
าการทดลองในห้องวิจัย เชื่อว่าสามารถเก็บหนังสือนับพันเล่มไว้ได้บนเลเซอร์ดิสค์เพียงแผ่น เดียว ปัจจุบันน
าการทดลองในห้องวิจัย เชื่อว่าสามารถเก็บหนังสือนับพันเล่มไว้ได้บนเลเซอร์ดิสค์เพียงแผ่น เดียว ปัจจุบันน
ยกันได้ ในประเทศไทยถึงแม้จำนวนผู้ใช้เอฟเคมีไม่ถึงพันเครื่อง แต่การกระจายของเครื่องและการติดต่อกันเสมือ
โรงงานเหล่านี้ยังเต็มไปด้วยอะไหล่ชิ้นส่วนต่างๆ นับพันแบบ มีความซับซ้อนในการปฏิบัติงานสูง การที่อะไหล่ชิ
รรมที่สำคัญที่สุด เพราะสะท้อนความเชื่อ และความผูก พัน ระหว่างสังคมไทยกับพุทธศาสนาอย่างลึกซึ้ง มีมานานก่
โครงข่ายระบบประสาท จึง ประกอบด้วยเซลล์ประสาทนับพันเซลล์ที่เชื่อมโยงกันเป็นโครงข่ายอันสลับซับซ้อน

โรงงานเหล่านี้ยังเต็มไปด้วยอะไหล่ชิ้นส่วนต่างๆ นับพันแบบ

Sense#	Sample sentences
1	ผมอยากบอกเธอเป็นเรื่องเกี่ยวกับความรักความผูกพัน
2	มันเร็วกว่ากันพันเท่า สามารถเก็บหนังสือนับพันเล่มไว้ได้บนเลเซอร์ดิสก์เพียงแผ่น
3	

Figure 2 Editor for meaning selection

Intelligent Technologies (InTech'2001) pp.349-355

[3] McCallum, A. and Nigam, K. (1999) Text Classification by Bootstrapping with Keywords, EM and Shrinkage. Proceeding of ACL'99, pp. 52-58

[4] Sornlertlamvanich, V., Potipiti, T. and Charoenporn, T. (2000) Automatic Corpus-based Thai Word Extraction with the C4.5 Learning Algorithm. Proceedings of COLING 2000, pp.802-807