An Internet-Based Collaborative Dictionary Development Project: SAIKAM

Vuthichai Ampornaramveth, Akiko Aizawa, Keizo Oyama
Research and Development Department
National Center for Science Information Systems
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo 101-8430, Japan
vuthi@rd.nacsis.ac.jp, akiko@rd.nacsis.ac.jp, oyama@rd.nacsis.ac.jp

Tasanee Methapisit
Thammasat University, Bangkok, Thailand
tasmetha@hotmail.com

Abstract

In this paper, an on–line Japanese↔Thai dictionary development project called “Saikam” is introduced. Saikam provides an on–line integrated environment to support collaborative development of Japaese–Thai dictionary on the Internet. Dictionary developers from all over the World can connect to the centralized dictionary database and update the content anytime at their convenience using standard web browsing tools. Some technical efforts have been made to enable trilingual data entry on the existing WWW tools. Beside basic word lookup and editing features, Saikam also provides Japanese word usage navigator which can extract ist of frequently–used words or sentences matching the specified patterns from large–scale Japanese text corpus to assist students of Japanese language.

1 Introduction to Saikam

Saikam project was initiated by the Association of Thai Professionals in Japan (ATPIJ) [6] with supports from a number of individuals and organizations including ATPIJ volunteer staffs, National Electronics and Computer Technology Center (NECTEC/Thailand), National Center for Science Information Systems (NACSIS/Japan), and the university of Electro–Communication. The primary objective of Saikam is to support development of Thai–Japanese dictionary database by employing the Internet and multi–lingual computing environment to provide an on–line collaborative working platform.

Saikam was inspired by the believe that many of the obstacles such as time limitation, and geographical barrier which detained the progress of similar Thai–Japanese dictionary developing efforts in the past would be overcome by utilizing these recent technologies. In other words, the technical challenge is to create a system which allows anybody with a computer and Internet access to contribute easily from his/her office or home at anytime through SAIKAM centralized website (figure 1).

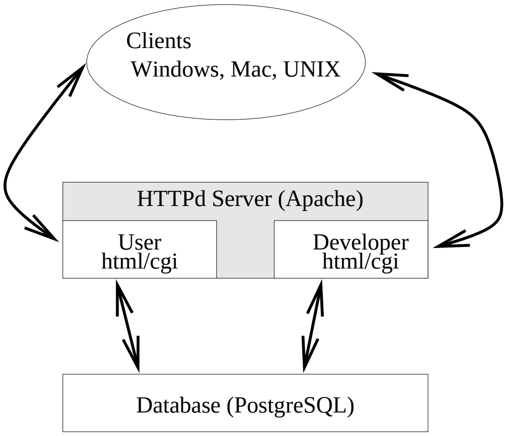


Figure 1: Saikam: Internet–based dictionary development system

Concurrently, the most recent dictionary data is also made available for anybody to access. Japanese words, for example, can be searched from a number of criteria including word pronunciation (reading), level of difficulty, frequency of occurrence, number of strokes, etc. The search result displays corresponding Thai words, descriptive meaning in Thai, some usage examples, synonyms, etc.

Saikam website has been made available for public access since the beginning of November 1998 at the URL

`http://thaigate.nacsis.ac.jp:8888/` or
`http://come.to/saikam` (alias)

2 Saikam Services

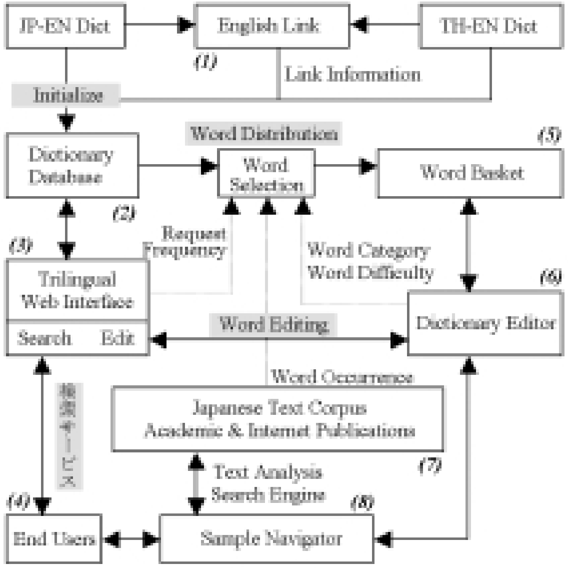


Figure 2: Saikam system components and services

Saikam system consists of a number of components and services as shown in figure 2. At the beginning of the project, the dictionary database(2) is initialized with list of Japanese words, Thai words and links between Thai and Japanese words. The link information was automatically generated by cross referencing the English definitions of Japanese and Thai words in a Japanese–English and a Thai–English dictionaries(1). Saikam users can access the database via the trilingual web interface(3).

2.1 Word Lookup

For the dictionary end–users(4), Saikam provides three types of word lookup service :–

1. "Thai→Japanese" word lookup. On a non–Thai platforms, Thai words can be entered using Thai input JAVA applet [5] on the browser with JAVA support.
2. "Japanese→Thai" word lookup. Japanese word with known pronunciation can be typed either in `KANA` or `ROMAJI` (on non–Japanese platforms). Search for words containing a specific kanji character requires kana–kanji input support from the operating system.
3. "Kanji→Japanese→Thai" word lookup. Kanji can be searched by specifying the number of strokes, level of difficulty, frequency of use, or kanji reading. Similary, the reading may be entered as `KANA` or `ROMAJI`. Frequency and difficulty searches allow students of Japanese languages to obtain list of most–frequently–used words which is useful for improving his/her own Japanese vacabulary. Directl jump from a kanji character to list of Japanese words containing that kanji character is also possible.

The information given by Saikam system is similar to those found on ordinary dictionary. For example, result of a Japanese word lookup contains the word reading, part of speech, descriptive meaning in Thai and English, together with list of corresponding Thai words. Since search result contains both Japanese and Thai texts, user may instruct the system to display Japanese/Thai texts as GIF images for best visualization result on all platforms.

2.2 Word Editing

Anybody with a valid email address can register with the system and become Saikam dictionary developer. Upon entering login name and password, developers can log in to the system and edit the dictionary content by following the process below

1. Adding words into developer's word basket(5). Each developer is assigned a word basket which is a logical collection of words being edited by the developer. The concept of word basket is introduced to maintain the consistency of words in the database. A word may not belong to more than one word basket at the same time. Thus, only a single developer can work on any particular words.

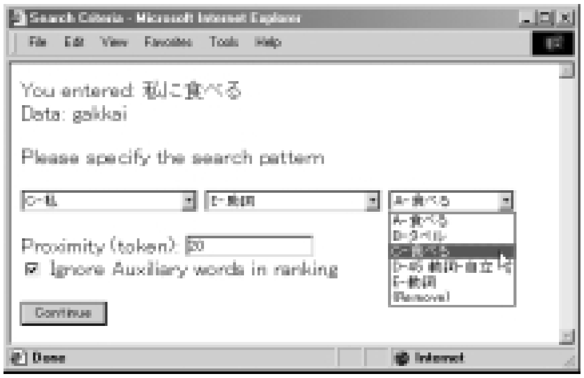
   Since it is desirable that most–frequently used words are updated first, the system also provides an automatic word assignment feature which can insert into the basket words with high occurency using statistical data from Japanese text corpus(7).
2. Pick up a word from the basket and edit the word definition and usage samples by com-

pleting the form in figure 3. Usually the developer has to provide descriptive meaning of that Japanese word in Thai, and edit the list of corresponding Thai words. This form contains Java applets to assist in editing Thai meaning and Thai word list.

3. Release words from word basket. Other developers may review or further improve the definition of this word at a later time.

2.3 Sample Navigator

Students of Japanese language may also search for sample sentences from the Japanese text corpus. Saikam allows high–level search criteria in which the search pattern may be specified as sequential occurences of words derived from the given word stem, or having the given part of speech as shown in figure 4. This is accomplished by first creating indices of all word stems and part of speeches using a Japanese morphological analysis [9] and full text indexing [3] tools. Sample sentences containing the specified pattern are sorted by level of difficulty which is calculated from length of the sentence and the difficulty of kanji characters it contains.



You entered: 私に食べる
Data: gakkai

Please specify the search pattern

C-私 | E-動詞 | A-食べる

A-食べる
B-タベル
C-食べる
D-助動詞-自立
E-動詞
(Remove)

Proximity (token): 20
☑ Ignore Auxiliary words in ranking

Continue

Figure 4: Sample Search Criteria

3 Implementation

In this section, some issues regarding the implementation of Saikam system is briefly discussed.

3.1 Multi–Lingual Web Interface

Due to multi–lingual nature of the dictionary, it is necessary that SAIKAM's WWW interface supports displaying and data input of Thai and Japanese text at the same time. Also the following issues should be considered when implementing the web interface.

- Platforms independency: The interface should behave similarly regardless of the operating system used (Windows, UNIX, or Macintosh) so that Saikam can reach the widest area of audience.
- Ease of use : Data input in SAIKAM should obey the standard data input methods for the corresponding language :– direct keyboard input for Thai and English, `ROMAJI` or `kana` input with kana–kanji front–end processor for Japanese. This is to encourage a quick acceptance among Saikam users.

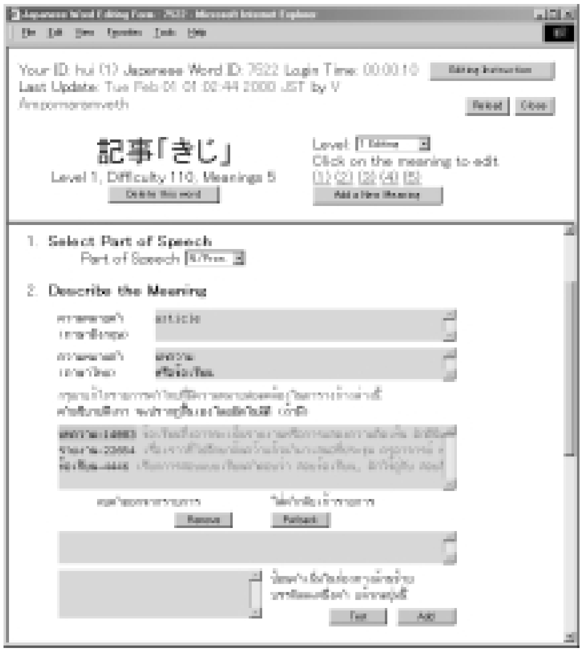
After investigating a number of multilingual tools for developing web application, the development team found that one of the best solution [7] is to 1) install Japanese font on user system, display Japanese text as plain text, 2) let the server converts Thai text to GIF image, 3) install Global–IME [2] and use IE browser for Japanese input on non–Japanese platforms, and 4) use JAVA applet for Thai input [5]. However, the system also allows users to customize the interface for optimal result on his/her own system by enabling/disabling any or all of these text-to-GIF convertors, and JAVA applet.

3.2 Database Initialization

Saikam database is initalized using word list from a Japanese–English [1] and a Thai–English [4] dictionaries. Initial links between Japannese words and Thai words are generated by calculating score measuring the similarity between their respective English definitions [8].

Initially, it turned out that there are more than 2.09 millions links with score higher than 1 (has some form of relationship). These links tie 20,468(93%) Japanese words with 43,739(85%) Thai words. However, after a careful investigation of the result, it turned out that those links with score between 1–33 are too weak and should be removed from the initial database. As a result, only 214,122 links with score higher than 33 tying 18,743(85%) Japanese words with 35,737(69%) Thai words make up the initial database.

By providing some automatically generated initial link like this, workload of the developer can be reduced. Most of the time, developers are asked to remove incorrect links from the system. This is much easier than adding a new link into the system, a task which requires rich knowledge of language vocabulary.


Japanese Word Editing Form - 7522 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Editing Instruction
Reload
Close
Delete this word
Add a New Meaning
Remove
Refresh
Test
Add

Your ID: hui (1) Japanese Word ID: 7522 Login Time: 00:00:10
Last Update: Tue Feb 01 01:02:44 2000 JST by V Ampornaramveth

記事「きじ」
Level 1, Difficulty 110, Meanings 5

Level: 1 Editing
Click on the meaning to edit
(1) (2) (3) (4) (5)

1. Select Part of Speech
Part of Speech N/Pron.

2. Describe the Meaning
article

Figure 3: Japanese word editing form

4 Conclusion

By playing a major role in information delivery, Internet has penetrated far into everybody's daily life as a communication tool of the new century. Saikam is an example of project taking advantage of this Internet infrastructure to provide a collaborative environment for doing a labor–intensive task, developing a dictionary.

Among the community of Thais in Japan, Saikam project represents an unprecedented co-ordination effort led by ATPIJ toward the collaborative development of the long awaited Japanese ↔ Thai dictionaries. Currently more than 60 developers have signed up to participate in developing the Japanese→Thai part of the Saikam dictionary. Data from a survey confirms the necessity of the dictionary development project and, at the same time, indicates the suitability of Internet as a new approach for dictionary development. However, in order to achieve a steady growth of this project, it is necessary identify the optimal working model for this kind of task. Also, close co–operation between the technical development team and the dictionary editors must be encouraged and maintained.

For futher inquiries on Saikam project, please contact saikam@fedu.uec.ac.jp

References

[1] Edict: Japanese–english dictionary. http://www.dgs.monash.edu.au/%7eejwb/japanese.html.

[2] Microsoft global input method editor. http://www.asia.microsoft.com/windows/ie/intlhome.htm.

[3] Pat reference manual. OpenText Corp.

[4] Resource of thai language processing. http://www.links.nectec.or.th/thaires/.

[5] Thai java applets. http://thaigate.nacsis.ac.jp/refer/thaijava/.

[6] V. Ampornaramveth. Saikam: An online dictionary development project. *Proc. of the 4th Intl. Workshop on Academic Information Networks and Systems, February 1998, NACSIS Seminar House, Karuizawa, Japan. http://thaigate.nacsis.ac.jp:8888/.*

[7] V. Ampornaramveth. Trilingual www interface to saikam dictionary project. *Proc. of the 5th Intl. Workshop on Academic Information Networks and Systems, December 1998, AIT, Thailand.*

[8] V. Ampornaramveth, A. Aizawa, K. Oyama. インターネット上の協調対訳辞書構築プロジェクト saikam. *Digital Libraries*, (16):121--128, November 1999.

[9] 松本 裕治 他. 日本語形態素解析システム「茶筌」使用説明書. 奈良先端科学技術大学院大学, *1997.*