A Multi-language Translation Example Browser

Isao Goto, Naoto Kato, and Noriyoshi Uratani

NHK Science and Technical Research Laboratories
1-11-10 Kinuta, Setagaya, Tokyo, Japan
{goto.i-es, katou.n-ga, uratani.n-fc}@nhk.or.jp

Terumasa Ehara[1]

Tokyo University of Science, Suwa
5000-1, Toyohira, Chino, Nagano, Japan
eharate@rs.suwa.tus.ac.jp

Tadashi Kumano[1] **and Hideki Tanaka**[1]

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Keihanna Science City, Kyoto, Japan
{tadashi.kumano, hideki.tanaka}@atr.co.jp

Abstract

This paper describes a Multi-language Translation Example Browser, a type of translation memory system. The system is able to retrieve translation examples from bilingual news databases, which consist of news transcripts of past broadcasts. We put a Japanese-English system to practical use and undertook trial operations of a system of eight language-pairs.

1 Introduction

NHK, Japan Broadcasting Corporation, provides news programs to foreign countries via radio broadcasting services in 22 languages.[2] For these programs, news transcripts are produced by translating source news transcripts. Broadcast news translation requires considerable skill and experience, due to the following factors:

- Lack of time for translation

The news scripts must be translated by the time news programs are broadcast, with news airing several times a day. Translators thus have very short time to translate.

- Difficulty in translating proper nouns

Proper nouns are commonplace in news scripts. In translating these words, translators must consult the appropriate organizations for the correct translation, since they are often not registered in extant bilingual dictionaries.

- Need to produce understandable translations

The translators must produce scripts suitable for foreign listeners who are unfamiliar with Japanese affairs. Thus, background information needs to be added to translated news scripts. The translated scripts must be liberal rather than literal translations.

To support such translations, we have been developing a Multi-language Translation Example Browser, a type of translation memory system. The system retrieves translation examples from bilingual databases consisting of scripts of past news broadcasts. Translators enter expressions in the source language they want to translate from, and then obtain the corresponding scripts in a target language that contain the translated expressions.

This system provides a translation aid of both speed and ease of use. Users can easily find translations of proper nouns from past translation examples. Since the system displays an entire script, users can examine the context of an interesting expression and consult examples of translations of background information from past translated scripts. We put a Japanese-English system to practical use (Kumano et al. 2002) while undertaking trial operations of a system of eight language-pairs.

---

[1] This work was done when these authors were working for NHK.

[2] Arabic, Bengali, Burmese, Chinese, English, French, German, Hindi, Indonesian, Italian, Japanese, Korean, Malay, Persian, Portuguese, Russian, Spanish, Swahili, Swedish, Thai, Urdu, and Vietnamese.

This paper describes our Multi-language Translation Example Browser. Translators access our system server via a web browser. The system server has four main components: a search engine, a multi-language database, CGI programs running on a web server, and a morphological analyzer. The search engine plays the primary role in our system, retrieving translation examples from all of our databases. The database currently consists of ten language databases. Our system can be easily extended by adding other language databases. Section 2 of this paper presents the system design, while Section 3 describes the multi-language database used in the system.

2 System design

2.1 An overview of our system

Figure 1 gives an overview of our Multi-language Translation Example Browser. The system uses a Web-based approach to eliminate the need for special software on terminals and to reduce maintenance overhead. In Figure 1, the web browser functions as the user interface for translators, accepting the entry of problematic expressions and returning translation examples as results. The system server consists of the following four modules:

- Web server

The web server is the system server manager, whose functions are performed by CGI programs on the web server. The CGI applications send the expressions to the Morphological analyzer to obtain keywords from the analyzer. It then forwards the keywords to the Search engine and obtains the translation examples. Finally, the results are returned by the server.

- Morphological analyzer

The Morphological analyzer finds keywords from the expressions by extracting content words. The current analyzer handles expressions in their native language for Japanese, English, and Korean. In other languages, it returns the original expressions themselves.

- Search engine

The Search engine retrieves translation examples from the news database. Even if the expressions do not find a perfect match in the database, the Search engine can retrieve similar expressions, since it looks for them based on keywords extracted by the Morphological analyzer. If the search engine receives multiple keywords, it considers the number of matched keywords, the sequence of the

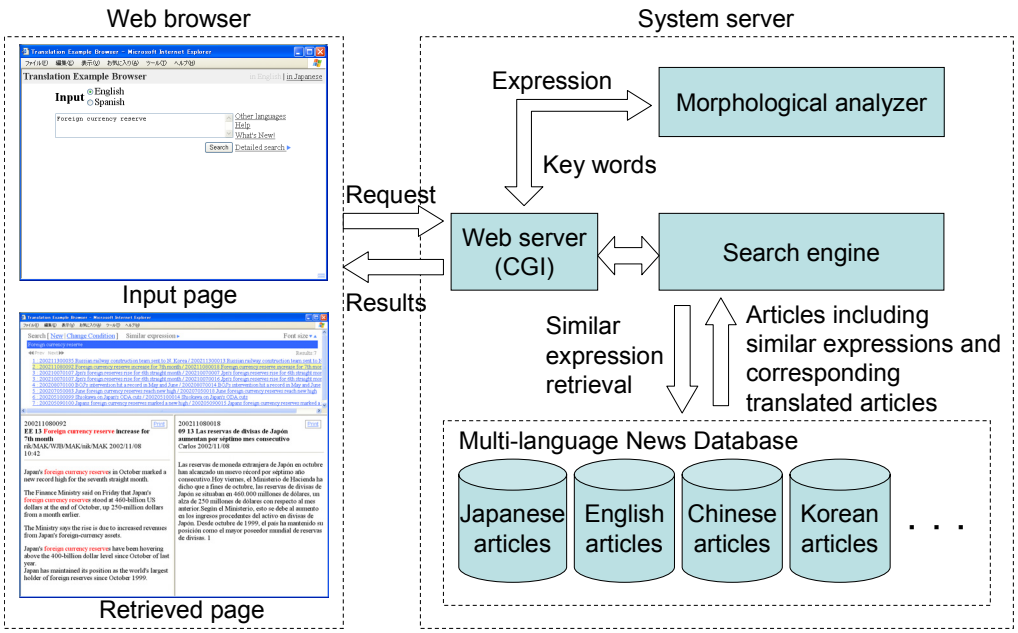

Figure 1: An overview of our Translation Example Browser

keywords in the original search key, and the distance between the keywords to find similar expressions. This similar expression search algorithm is effective especially for long Japanese news expressions. For a detailed account of our search algorithm, see Tanaka et al. (1999).

The search engine uses multithreaded processing to handle requests from multiple users at the same time. It also caches search results to accelerate searches based on the same search key. The engine is not language-specific and can be easily applied to many languages.

- Multi-language News database

For detailed information on Multi-language News databases, see Section 3.

2.2 An example of retrieved translation examples

Figure 2-4 shows the system user interface. First, a translator selects a language pair using select language page in figure 2. In this case, the English-Spanish database is selected.

Next, the translator selects entering language in the input page in figure 3. If the translator wishes to obtain a translation of the English expression "foreign currency reserve," the translator enters the expression into the input page and pushes the Search button. The system then retrieves the database and returns the results in retrieved page.

The retrieved page in figure 4 contains examples of translation examples. The page has three parts. The upper half of the page is a list of titles of search results. The translator can refer to other scripts when selecting subsequent titles. The lower left side displays a script in the source language (English). Similar expressions for which the search was made in the script on the lower left side are indicated by red character strings. The lower right side displays the corresponding script in the target language (Spanish). The translator can obtain the required expression for a translation from the script, or check for expressions used within a specific context.

Translators can change the required level of similarity between an input expression and retrieved expressions by changing the number of keywords to be matched. Pressing the matching level button in figure 4 modifies this value.

Figure 2: Select Language page

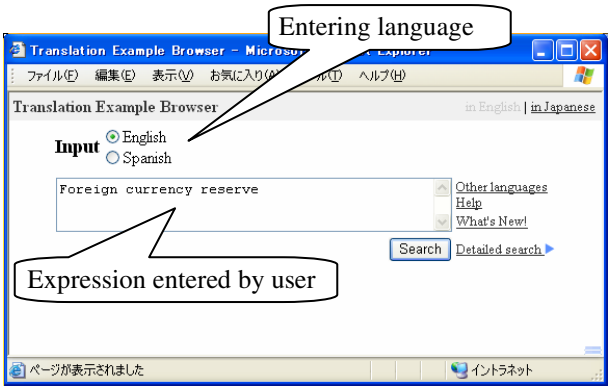

Figure 3: Input page

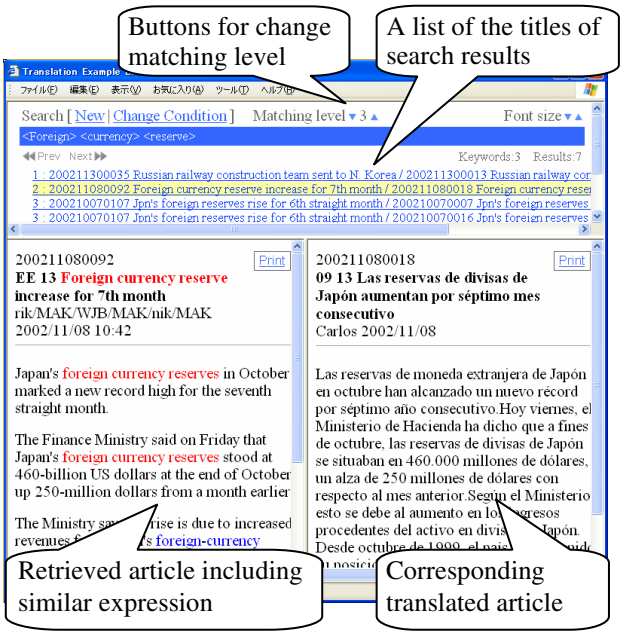

Figure 4: Retrieved page

3 The multi-language database used in the system

Our current system includes 10 language databases. The languages are Chinese, English, Hindi, Indonesian, Japanese, Korean, Malay, Russian, Spanish, and Urdu.

Translated news scripts are produced by human translation from a source language into a target language as follows: First, Japanese scripts are written for domestic and international news programs. English, Chinese, and Korean versions are translated directly from the Japanese. Other language versions are created from the English version. Figure 5 shows the translation sequence. When translators translate scripts, they insert the number of the script in a source language to the script in a target language as header information. This makes it possible to match scripts in a source language and in a target language using the header information. The link information is used to build bilingual databases by combining mono-lingual databases.

Our system treats the multi-language database as multi-bilingual databases. The language pairs of our bilingual databases correspond to the relations of the source and target languages in the translation sequence. We plan to extend our multi-language database to encompass the 22 languages in which NHK broadcasts news programs. For a detailed discussion of our news database, see Goto et al. (2001).

Since news is broadcast daily, news scripts for them continue to be written throughout the day. Thus, our databases expand as time passes. The number of news scripts is approximately 300/day in Japanese, approximately 90/day in English, and approximately 10-20/day in the other languages.

Our databases currently hold 585,000 scripts in Japanese for an eight-year period, 157,000 scripts in English for an eight-year period, and 5,000 to 12,000 scripts in the other eight languages for a two-year period.

4 Conclusion

This paper describes our Multi-language Translation Example Browser. This simple system provides effective support for new translations. We plan to extend this system to translation examples in additional languages. For languages not handled by our morphological analyzer, we will develop a word-delimiting function without dictionaries based on a statistical method. This function will allow our system to retrieve similar expressions in all languages. We will also examine the issue of sentence-to-sentence or word-to-word automatic alignment.

5 Bibliographical References

Goto, Isao, Naoto Kato, and Terumasa Ehara. (2001). *A multilingual news database and its application to a translation memory system*. 6th Natural Language Processing Pacific Rim Symposium Post-Conference Workshop on Language Resources in Asia, pp.1-6.

Kumano, Tadashi, Isao Goto, Hideki Tanaka, Noriyoshi Uratani, and Terumasa Ehara. (2002). *A Translation Aid System by Retrieving Bilingual News Database*. Systems and Computers in Japan, Vol.33, No.8, pp.19-29.

Tanaka, Hideki, Tadashi Kumano, Noriyoshi Uratani, and Terumasa Ehara. (1999). *An Efficient Way of Gauging Similarity between Long Japanese News Expressions*. Journal of Natural Language Processing, Vol.6, No.5, pp.93-116. (in Japanese)

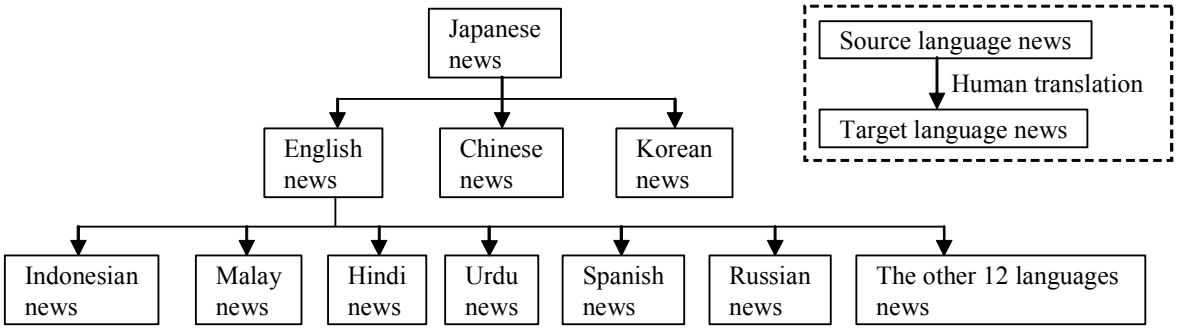

Figure 5: The translation sequence at NHK