Crawling English-Japanese Person-Name Transliterations from the Web

Satoshi Sato
Graduate School of Engineering
Nagoya University
Furo-cho, Chikusa-ku
Nagoya, Japan
ssato@nuee.nagoya-u.ac.jp

ABSTRACT

Automatic compilation of lexicon is a dream of lexicon compilers as well as lexicon users. This paper proposes a system that crawls English-Japanese person-name transliterations from the Web, which works a back-end collector for automatic compilation of bilingual person-name lexicon. Our crawler collected 561K transliterations in five months. From them, an English-Japanese person-name lexicon with 406K entries has been compiled by an automatic post processing. This lexicon is much larger than other similar resources including English-Japanese lexicon of HeiNER obtained from Wikipedia.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Design, Experimentation.

Keywords

Mining transliteration pairs, automatic lexicon compilation, person name.

1. INTRODUCTION

Frequent update and enhancement of lexicon is always welcomed by lexicon users. However, lexicons are rarely updated because it requires a lot of human labor. A solution is replacement of human labor with a computer program. Automatic compilation of lexicon is a dream of lexicon compilers as well as lexicon users.

In this paper, we report an attempt of automatic compilation of a bilingual lexicon from the Web. We have selected *English-Japanese (Latin-Katakana) person-name lexicon* as a target of automatic compilation because of the following reasons.

1. Between languages that use different alphabets, person name translation is a serious problem for human translators. In English-Japanese translation, person names written in Latin script are transliterated into one in Katakana script according to their pronunciations. English-Japanese transliteration of person name is difficult because of several reasons, such as limited coverage of existing bilingual lexicons, non-English (e.g., French and German) person names appeared in English texts, and spelling variants in Katakana script.

2. There is a possibility that we can compile a large English-Japanese person-name lexicon from the Web, because a lot of transliteration instances of person names exist on the Web. Actually, human translators use the Web as a virtual low-quality bilingual lexicon.

3. New person names are produced; new person-name transliterations are produced in every day. Human translators hope frequent update of bilingual person-name lexicon.

This paper proposes a system that crawls English-Japanese person-name transliterations from the Web, which works as a back-end collector for automatic lexicon compilation. From collected transliterations, a bilingual person-name lexicon is produced by an automatic post processing. This attempt of automatic lexicon compilation can be viewed as a conversion from a virtual low-quality bilingual lexicon (i.e., the Web) to a real high-quality bilingual lexicon.

2. SYSTEM

Figure 1 shows an overview of the system, *Tsumugi Crawler*. The system consists of four components: (1) *person-name pool*, which keeps monolingual (English or Japanese) person names to be searched. (2) *Tsumugi Finder*, which searches transliterations (or back-transliterations) of a monolingual person name. (3) *PN-filter*, a person-name filter, which selects person-names from candidates. (4) *transliteration database*, which keeps collected transliteration pairs.

The central component of the system is the Tsumugi Finder. First, the Finder picks a monolingual person name from the person-name pool, and sends it to Yahoo Japan!'s search engine and obtains the search results (100 snippets in Japanese). Because Japanese texts frequently include *partially bilingual texts*, as Nagata et al.[2] reported, it is highly expected that the obtained snippets include (back-)transliterations of the person name of the search query.

From the obtained snippets, S-filter extracts strings that satisfy a predefined *style* requirement of person name. We have defined a person name as a term that consists of a first

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

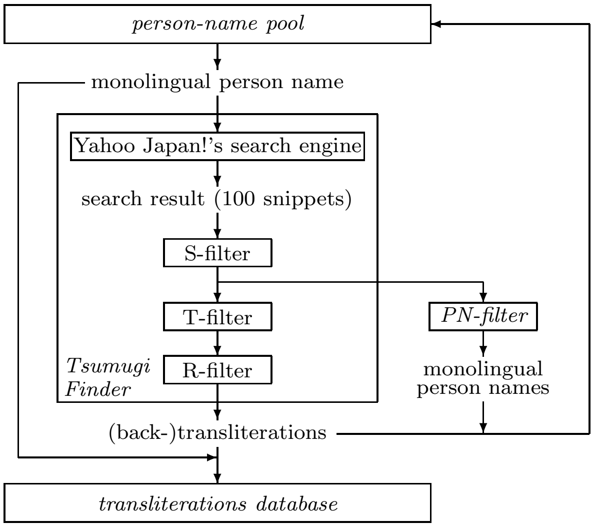

Figure 1: Overview of *Tsumugi Crawler*

name and a last name, and optionally a middle name. The style requirement is implemented by two regular expressions: one for English, the other for Japanese.

T-filter checks transliteration correspondence between the person name (input) and every candidate produced by S-filter. T-filter uses *transliteration score* between an English term and a Japanese term [3], which is an estimated pronunciation similarity of two terms. T-filter removes a *hopeless* candidates whose score is less than the predetermined threshold.

R-filter sorts remaining candidates based on their frequencies in the snippets. It always passes the first-ranked candidates; it also passes a few *hopeful* candidates as spelling variants of the first-ranked candidate. Finally, the obtained transliteration pairs are stored into the transliteration database.

The snippets obtained in the above process are also used as a source from which the system collects new monolingual person names. Monolingual person names are obtained by applying PN-filter [1] to the candidates produced by S-filter and they are stored into the person-name pool.

3. COMPILED LEXICON

Tsumugi Crawler collected 561K transliterations in five months run. The final bilingual lexicon was produced by the following post processing.

1. Applying the newest version of S-filter and T-filter to the collected transliterations. (561K → 558K)

2. Assign a forward rank and a backward rank to each transliteration. The forward rank of a transliteration pair $t_i = \langle en, ja_i \rangle$ is a rank among the competitive set $\bigcup_i t_i$; the rank is determined by the estimated frequency calculated from the hit counts of "*en* and ja_i". When t_i is in the first rank, ja_i is the best (frequently used) transliteration of *en*. A backward rank is calculated by the same way but reverse direction.

3. Select transliterations in the following types and filter out other transliterations. (558K → 454K)

   Type-1 (standard transliteration) transliteration that has the first forward rank and the first backward rank.

   Type-2 (transliteration variant) transliteration that has the first forward rank or the first backward rank (except type-1).

4. Filter out every transliteration whose estimated frequency is 0. (454K → 441K)

5. Apply the newest version of PN-filter to the remaining transliterations. (441K → 406K)

The size of the produced lexicon is 406,416 (type-1 entries = 286,047, type-2 entries = 120,369; English spellings = 316,259, Japanese spellings = 379,504).

Typical size of existing book-style English-Japanese person-name dictionary is around 20,000. HeiNER[4] provides bilingual lexicon of named entities, which was compiled from inter-language links of Wikipedia. The size of English-Japanese lexicon of HeiNER is 125,053, which includes other types of name entities such as place, product, and company. The size of the largest database of English-Japanese person-name that we know is 175,000, which is sold by a database company. Our lexicon is much larger than these existing resources.

Because of the large size, evaluation of our lexicon is difficult and time-consuming. We have conducted a preliminary evaluation of precision, where 420 type-1 entries have been examined by a human evaluator. We have obtained the result that 390 (93%) entries are correct and 30 (7%) entries are incorrect (they are correct transliteration pairs of other types of named entities).

Evaluation of recall of our lexicon is much difficult because we do not know the total size of English-Japanese person-name transliterations on the Web. We have conducted a rough estimation by using a set of known transliterations extracted from four English books and their Japanese translations. Among 658 known transliterations that can be found by Yahoo Japan!'s search engine, 431 (65.5%) transliterations have been already recorded in our lexicon.

4. ACKNOWLEDGMENTS

This research has been supported by the Kayamori Foundation of Informational Science Advancement.

5. REFERENCES

[1] S. Kaide and S. Sato. A person-name classifier by using probability difference (in Japanese). In *Proc. of NLP-09*, 2009.

[2] M. Nagata, T. Saito, and K. Suzuki. Using the Web as a bilingual dictionary. In *Proc. of the workshop on Data-driven methods in machine translation*, pages 1–8, 2001.

[3] Y. Sakakibara and S. Sato. Automatic compilation of a bilingual person-name lexicon (in Japanese). In *Proc. of NLP-07*, pages 879–882, 2007.

[4] W. Wentland, J. Knopp, C. Silberer, and M. Hartung. Building a multilingual lexical resource for named entity disambiguation, translation and transliteration. In *Proc. of LREC-08*, 2008.