Constructing Transliteration Lexicons from Web Corpora

Jin-Shea Kuo[1, 2]
[1]Chung-Hwa Telecommunication Laboratories, Taiwan, R. O. C., 326
jskuo@cht.com.tw

Ying-Kuei Yang[2]
[2]E. E. Dept., National Taiwan University of Science and Technology, Taiwan, R.O.C., 106
ykyang@mouse.ee.ntust.edu.tw

Abstract

This paper proposes a novel approach to automating the construction of transliterated-term lexicons. A simple syllable alignment algorithm is used to construct confusion matrices for cross-language syllable-phoneme conversion. Each row in the confusion matrix consists of a set of syllables in the source language that are (correctly or erroneously) matched phonetically and statistically to a syllable in the target language. Two conversions using phoneme-to-phoneme and text-to-phoneme syllabification algorithms are automatically deduced from a training corpus of paired terms and are used to calculate the degree of similarity between phonemes for transliterated-term extraction. In a large-scale experiment using this automated learning process for conversions, more than 200,000 transliterated-term pairs were successfully extracted by analyzing query results from Internet search engines. Experimental results indicate the proposed approach shows promise in transliterated-term extraction.

1 Introduction

Machine transliteration plays an important role in machine translation. The importance of term transliteration can be realized from our analysis of the terms used in 200 qualifying sentences that were randomly selected from English-Chinese mixed news pages. Each qualifying sentence contained at least one English word. Analysis showed that 17.43% of the English terms were transliterated, and that most of them were content words (words that carry essential meaning, as opposed to grammatical function words such as conjunctions, prepositions, and auxiliary verbs).

In general, a transliteration process starts by first examining a pre-compiled lexicon which contains many transliterated-term pairs collected manually or automatically. If a term is not found in the lexicon, the transliteration system then deals with this out-of-vocabulary (OOV) term to try to generate a transliterated-term via a sequence of pipelined conversions (Knight, 1998). Before this issue can be dealt with, a large quantity of transliterated-term pairs are required to train conversion models.

Preparing a lexicon composed of transliterated term pairs is time- and labor-intensive. Constructing such a lexicon automatically is the most important goal of this paper. The problem is how to collect transliterated-term pairs from text resources.

Query logs recorded by Internet search engines reveal users' intentions and contain much information about users' behaviors. (Brill, 2001) proposed an interactive process that used query logs for extracting English-Japanese transliterated-terms. Under this method, a large initial number of term pairs were compiled manually. It is time-consuming to prepare such an initial training set, and the resource used is not publicly accessible.

The Internet is one of the largest distributed databases in the world. It comprises various kinds of data and at the same time is growing rapidly. Though the World Wide Web is not systematically organized, much invaluable information can be obtained from this large text corpus. Many researchers dealing with natural language processing, machine translation, and information retrieval have focused on exploiting such non-parallel Web data (Al-Onaizan, 2002; Fung, 1998;). Also, online texts contain the latest terms that may not be found in existing dictionaries. Regularly exploring Web corpora is a good way to update dictionaries.

Transliterated-term extraction using non-parallel corpora has also been conducted (Kuo, 2003). Automated speech recognition-generated confusion matrices (AGCM) have been used successfully to bootstrap term extraction from Web pages collected by a software spider.

AGCM were used successfully not only to alleviate pronunciation variation, especially the sociolinguistic causes, but also to construct a method for cross-language syllable-phoneme conversion (CLSPC). This is a mapping from a source-language syllable into its target-language counterpart. The problem is how to produce such conversions if AGCM are not available for the targeted language pair. To generate confusion matrices from automated speech recognition requires the effort of collecting many speech corpora for model training, costing time and labor. Automatically constructing a CLSPC without AGCM is the other main focus of this paper.

Web pages, which are dynamically updated and publicly accessible, are important to many researchers. However, if many personally guided spiders were simultaneously collecting Web pages, they might cause a network traffic jam. Internet search engines, which update their data periodically, provide search services that are also publicly accessible. A user can select only the pages of interest from Internet search engines; this mitigates the possibility that a network traffic jam will be caused by many personally guided spiders.

Possibly aligned candidate strings in two languages, which may belong to two completely different language families, are selected using local context analysis from non-parallel corpora (Kuo, 2003). In order to determine the degree of similarity between possible candidate strings, a method for converting such aligned terms cross-linguistically into the same representation in syllables is needed. A syllable is the basic pronunciation unit used in this paper. The tasks discussed in this paper are first to align syllables cross-linguistically, then to construct a cross-linguistic relation, and third to use the trained relation to extract transliterated-term pairs.

The remainder of the paper is organized as follows: Section 2 describes how English-Chinese transliterated-term pairs can be extracted automatically. Experimental results are presented in Section 3. Section 4 analyzes on the performance achieved by the extraction. Conclusions are drawn in Section 5.

2. The Proposed Approach

An algorithm based on minimizing the edit distance between words with the same representation has been proposed (Brill, 2001). However, the mapping between cross-linguistic phonemes is obtained only after the cross-linguistic relation is constructed. Such a relation is not available at the very beginning.

A simple and fast approach is proposed here to overcome this problem. Initially, 200 verified correct English-Chinese transliterated-term pairs are collected manually. One of the most important attributes of these term pairs is that the numbers of syllables in the source-language term and the target-language term are equal. The syllables of both languages can also be decomposed further into phonemes. The algorithm that adopts equal syllable numbers to align syllables and phonemes cross-linguistically is called the simple syllable alignment algorithm (SSAA). This algorithm generates syllable and phoneme mapping tables between the source and target languages. These two mapping tables can be used to calculate similarity between candidate strings in transliterated-term extraction. With the mapping, the transliterated-term pairs can be extracted. The obtained term pairs can be selected according to the criterion of equal syllable segments. These qualified term pairs can then be merged with the previous set to form a larger set of qualified term pairs. The new set of qualified term pairs can be used again to construct a new cross-linguistic mapping for the next term extraction. This process iterates until no more new term pairs are produced or until other criteria are met. The conversions used in the last round of the training phase are then used to extract large-scale transliterated-term pairs from query results.

Two types of cross-linguistic relations, phoneme-to-phoneme (PP) and text-to-phoneme (TP), can be used depending on whether a source-language letter-to-sound system is available or not.

2.1 Construction of a Relation Using Phoneme-to-Phoneme Mapping

If a letter-to-phoneme system is available, a phoneme-based syllabification algorithm (PSA) is used for constructing a cross-linguistic relation, then a phoneme-to-phoneme (PP) mapping is selected. Each word in the located English string is converted into phonemes using MBRDICO (Pagel, 1998). In order to compare English terms with Chinese terms in syllables, the generated English phonemes are syllabified into consonant-vowel pairs. Each consonant-vowel pair is then converted into a Chinese syllable. The PSA used here is basically the same as the classical one (Jurafsky, 2000), but has some minor modifications. Traditionally, an English syllable is composed of an initial consonant cluster followed by a vowel and then a final consonant cluster. However, in order to convert English syllables to Chinese ones, the final consonant cluster is appended only when it is a nasal. The other consonants in the final consonant cluster are then segmented into isolated consonants. Such a syllable may be viewed as the basic pronunciation unit in transliterated-term extraction.

After English phonemes are grouped into syllables, the English syllables can be converted into Chinese ones according to the results produced by using SSAA. The accuracy of the conversion can improve progressively if the cross-linguistic relation is deduced from a large quantity of transliterated-term pairs.

Take the word “polder” as an example. First, it is converted into /poldə/ using the letter-to-phoneme system, and then according to the phoneme-based syllabification algorithm (PSA), it is divided into /po/, /l/, and /də/, where /l/ is an isolated consonant. Second, these English syllables are then converted into Chinese syllables using the trained cross-

linguistic relation; for example, /po/, /l/, and /də/ are converted into /po/, /er/, and /de/ (in Pin-yin), respectively. /l/ is a syllable with only an isolated consonant. A final is appended to its converted Chinese syllable in order to make it complete because not all Chinese initials are legal syllables. The other point worth noting is that /l/, a consonant in English, is converted into its Chinese equivalent, /er/, but, /er/ is a final (a kind of complex vowel) in Chinese.

2.2 Construction of a Relation Using Text-to-Phoneme Mapping

If a source language letter-to-phoneme system is not available, a simple text-based syllabification algorithm (TSA) is used and a text-to-phoneme (TP) mapping is selected. An English word is frequently composed of multiple syllables; whereas, every Chinese character is a monosyllable. First, each English character in an English term is identified as a consonant, a vowel or a nasal. For example, the characters "a", "b" and "n" are viewed as a vowel, a consonant and a nasal, respectively. Second, consecutive characters of the same attribute form a cluster. However, some characters, such as "ch", "ng" and "ph", always combine together to form complex consonants. Such complex consonants are also taken into account in the syllabification process. A Chinese syllable is composed of an initial and a final. An initial is similar to a consonant in English, and a final is analogous to a vowel or a combination of a vowel and a nasal. Using the proposed simple syllable alignment algorithm, a conversion using TP mapping can be produced. The conversion can also be used in transliterated-term extraction from non-parallel web corpora.

The automated construction of a cross-linguistic mapping eliminates the dependency on AGCM reported in (Kuo, 2003) and makes transliterated-term extraction for other language pairs possible. The cross-linguistic relation constructed using TSA and TP is called CTP; on the other hand, the cross-linguistic relation using PSA and PP is called CPP.

3 The Experimental Results

3.1 Training Cross-language Syllable-phoneme Conversions

An English-Chinese text corpus of 500MB in 15,822,984 pages, which was collected from the Internet using a web spider and was converted to plain text, was used as a training set. This corpus is called SET1. From SET1, 80,094 qualifying sentences that occupied 5MB were extracted. A qualifying sentence was a sentence composed of at least one English string.

Two experiments were conducted using either CPP or CTP on SET1. Figure 1 shows the progress of extracting transliterated-term pairs achieved using CPP mapping. A noteworthy phenomenon was that phoneme conversion produced more term pairs than syllable conversion did at the very beginning of training. This is because, initially, the quality of the syllable combinations is not good enough. The phonemes exerted finer-grained control than syllables did. However, when the generated syllable combinations improved in quality, the situation changed. Finally, extraction performed using syllable conversion outperformed that achieved using phoneme conversion. Note also that the results produced by using phonemes quickly approached the saturation state. This is because the English phoneme set is small. When phonemes were used independently to perform term extraction, fewer extracted term pairs were produced than were produced using syllables or a combination of syllables and phonemes.

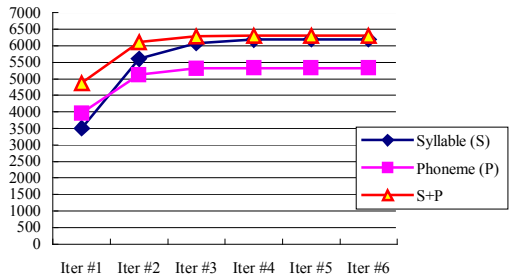

Figure 1. The progress of extracting transliterated-term pairs using CPP conversion

Figure 2 shows the progress of extracting transliterated-term pairs using CTP. The same situation also occurred at the very beginning of training. Comparing the results generated using CPP and CTP, CPP outperformed CTP in terms of the quantity of extracted term pairs because the combinations obtained using TSA are larger than those obtained using PSA. This is also revealed by the results generated at iteration 1 and shown in Figures 1 and 2.

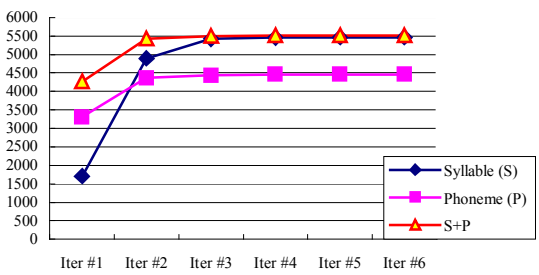

Figure 2. The progress of extracting transliterated-term pairs using CTP conversion.

3.2 Transliterated-term Extraction

The Web is growing rapidly. It is a rich information source for many researchers. Internet search engines have collected a huge number of Web pages for public searching (Brin, 1998). Submitting queries to these search engines and analyzing the results can help researchers to understand the usages of transliterated-term pairs.

Query results are text snippets shown in a page returned from an Internet search engine in response to a query. These text snippets may be composed of texts that are extracted from the beginning of pages or from the texts around the keywords matched in the pages. Though a snippet presents only a portion of the full text, it provides an alternative way to summarize the pages matched.

Initially, 200 personal names were randomly selected from the names in the 1990 census conducted by the US Census Bureau[1] as queries to be submitted to Internet search engines. CPP and CTP were obtained in the last round of the training phase. The estimated numbers of distinct qualifying term pairs (EDQTP) obtained by analyzing query results and by using CPP and CTP mappings for 7 days are shown in Table 1. A qualifying term pair means a term pair that is verified manually to be correct. EDQTP are term pairs that are not verified manually but are estimated according to the precision achieved during the training phase.

Finally, a text corpus called SET2 was obtained by iteratively submitting queries to search engines. SET2 occupies 3.17GB and is composed of 67,944 pages in total. The term pairs extracted using CTP were much fewer in number than those extracted using CPP. This is because the TSA used in this study, though effective, is very simple and rudimentary. A finer-grained syllabification algorithm would improve performance.

	CPP	CTP
EDQTP	201,732	110,295

Table 1. The term pairs extracted from Internet search engines using PP and TP mappings.

4 Discussion

Comparing the performances achieved by CPP and CTP, the results obtained by using CPP were better than those with CTP. The reason is that TSA is very simple. A better TSA would produce better results. Though TSA is simple, it is still effective in automatically extracting a large quantity of term pairs. Also, TSA has an advantage over PSA is that no letter-to-phoneme system is required. It could be helpful when applying the proposed approach to other language pairs, where such a mapping may not be available.

[1] http://www.census.gov/genealogy/names/

5 Conclusions

An approach to constructing transliterated-term lexicons has been presented in this paper. A simple alignment algorithm has been used to automatically construct confusion matrices for cross-language syllable-phoneme conversion using phoneme-to-phoneme (PP) and text-to-phoneme (TP) syllabification algorithms. The proposed approach not only reduces the need for using automated speech recognition-generated confusion matrices, but also eliminates the need for a letter-to-phoneme system for source-language terms if TP is used to construct a cross-language syllable-phoneme conversion and to successfully extract transliterated-term pairs from query results returned by Internet search engines. The performance achieved using PP and TP has been compared and discussed. The overall experimental results show that this approach is very promising for transliterated-term extraction.

References

Al-Onaizan Y. and Knight K. 2002. Machine Transliteration of Names in Arabic Text, In *Proceedings of ACL Workshop on Computational Approaches to Semitic Languages, pp. 34-46.*

Brill E., Kacmarcik G., Brockett C. 2001. Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Logs, In *Proceedings of Natural Language Processing Pacific Rim Symposium*, pp. 393-399.

Brin S. and Page L. 1998. The Anatomy of a Large-scale Hypertextual Web Search Engine, In *Proceedings of 7th International World Wide Web Conference,* pp. 107-117.

Fung P. and Yee L.-Y. 1998. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. *In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 7th International Conference on Computational Linguistics,* pp. 414-420.

Jurafsky D. and Martin J. H. 2000. Speech and Language Processing, pp. 102-120, Prentice-Hall, New Jersey.

Knight K. and Graehl J. 1998. Machine Transliteration, Computational Linguistics, Vol. 24, No. 4, pp.599-612.

Kuo J. S. and Yang Y. K. 2003. Automatic Transliterated-term Extraction Using Confusion Matrix from Non-parallel Corpora, In *Proceedings of ROCLING XV Computational Linguistics Conference*, pp.17-32.

Pagel V., Lenzo K., and Black A. 1998. Letter to Sound Rules for Accented Lexicon Compression, In *Proceedings of ICSLP*, pp. 2015-2020.