EXTRACTING TRANSLITERATION PAIRS FROM COMPARABLE CORPORA

Slaven Bilac and Hozumi Tanaka

Department of Computer Science
Tokyo Institute of Technology
Tokyo, Japan
{sbilac,tanaka}@cl.cs.titech.ac.jp

ABSTRACT

Transliterating words and names from one language to another is a frequent and highly productive phenomenon. For example, English word *cache* is transliterated in Japanese as キャッシュ "kyasshu". In many cases, recent transliterations are not recorded in machine readable dictionaries so it is impossible to rely on dictionary lookup to find transliteration equivalents.

In this paper we describe a method for extracting transliteration pairs from comparable corpora. Proposed method exploits the structure of comparable corpora to extract a large subset of similarly distributed English words for each Japanese transliteration and then relies on phonetic similarity (i.e. back-transliteration) to find the best match in this subset. Back-transliteration also produces a similarity score which can be used to order extracted pairs.

1. INTRODUCTION

Transliteration is a process of acquisition and assimilation of words from one language into the other. In the process, the words are adjusted to allow representation in the target language script and pronunciation by native speakers of the target language. Technical terms and proper names make up most of transliterations. For example, English word *cache* is transliterated in Japanese as キャッシュ "kyasshu". Furthermore, in Japanese transliterations are written in katakana and thus easily distinguishable. Since word assimilation is frequently occurring process, newly introduced transliterations are often not recorded in electronic dictionaries and thus represent a significant portion of out-of-vocabulary items (OOV). OOV pose a big problem for Machine Translation, where unsuccessful dictionary lookups can lead to translation failures and Cross Language Information Retrieval where inability to translate content words significantly reduces retrieval performance.

In order to reduce the OOV problem due to transliteration, we propose a novel method for transliteration pair extraction by combining the lexical knowledge acquisition techniques with a phonetic similarity selection process based on back-transliteration. First, given a katakana string for which we want to find an English pair, we use cross-language contextual distribution to select a relatively large number of back-transliteration candidates: given a transliteration candidate in katakana we extract a set of English terms which occur in similar contexts. Second, we apply back-transliteration to each katakana string to select the most likely English terms among the candidates selected based on contextual similarity. Initial katakana string and selected back-transliteration then form a transliteration pair candidate. Finally, normalized back-transliteration score is used to rank all the transliteration pair candidates and output only the highest ranking pairs.

2. RELATED RESEARCH

Automatic extraction of transliteration pairs from bilingual corpora has received a lot of attention from researchers. For example, [1] propose a minimum edit-distance based algorithm to match Japanese and English named entities. Their system relies heavily on heuristics (e.g. capitalization of English proper nouns) and manually specified mapping rules. On the other hand, [2] use web-search query logs as corpora. In this approach, novel English words appearing in the query logs are added to the system lexicon and then the closest fit is found by searching through a complete lexicon using an edit distance based measure. However, continually growing the lexicon size can greatly increase the ambiguity and eventually tax the system performance. [3] proposes a rule-based generative model to generate English equivalents. Candidate back-transliterations are exhaustively generated for each transliteration and then filtered using the [4] word alignment method. Besides the limitations due to manually determined mapping rules, this approach runs into efficiency problems due to combinatorial explosion in number of possible matches.

Extraction of bilingual terms without consideration of phonetic similarity (modeled through backward- or forward-transliteration) has also received attention. Mostly, the techniques are applied to aligned parallel or near parallel corpora [5, 6]. Extracting term pairs from comparable corpora is much more difficult since it is harder to exploit the text structure to limit the search space. However, other patterns such as cross-language context and usage can be used to extracted related terms [7].

Instead of relying only on contextual similarity or phonetic similarity several hybrid approaches try to look at both simultaneously. For example, [8] propose a model combining vector-based semantic similarity with a transliteration model. However, their method requires POS and named entity (NE) tagging. On the other hand, [9] propose a language modeling approach for measuring context similarity and combine it with a context-free transliteration model. However, their method cannot handle phrases although phrases make up a significant portion of NE and technical term pairs.

In order to improve on previous approaches, we propose a transliteration pair extraction method which exploits the distributional similarity of the terms in bilingual text to limit the number of candidates to consider and then applies back-transliteration to select the most likely transliteration equivalent from among those candidates. By using this two-step approach to pair extraction we try to draw on the strengths of both the statistical lexical knowledge acquisition methods and back-transliteration techniques to extract technical term pairs with high precision. Unlike previously proposed hybrid approaches our method does not require additional lexical tools (e.g. POS-tagger) and can handle phrases.

3. COMPARABLE CORPORA

A major requirement for extracting transliteration pairs is to have a bilingual corpus. Ideally, a sentence aligned bilingual corpora would be used since the potential difficulties of extracting transliteration pairs are greatly reduced for such corpora [10]. However, large sentence aligned corpora are not readily available so the expected coverage of transliterated pairs can hardly surpass a fraction of transliterations appearing in the language. More importantly, since sentence aligned corpora are unlikely to be continually updated, recent transliterations (i.e. OOV) will probably not be present in large numbers.

On the other hand, comparable corpora in form of electronically published newspapers [11] and bilingual conference abstracts are much more accessible and are likely to contain a significant number of transliterated named entities and technical terms. Here we concentrate on conference abstracts. It is a common practice for conferences to require an English abstract to be provided along with abstracts in original language and recently it has become a trend to make the abstracts available in electronic format. Note, however, that abstracts in English are not necessarily translations of the original, but rather paraphrases [12] and as such cannot be considered parallel corpora. Nonetheless, abstracts can be grouped by conference or by author and thus form a well structured corpora. Moreover, since topics covered at conferences tend to be technical and/or scientific, the abstracts are likely to include a large number of transliterations.

4. TRANSLITERATION PAIR EXTRACTION

Given the structured bilingual corpus similar to those just described, it is possible to exploit its structure to extract subsets with similar distribution (e.g. by-conference distribution of words) across languages. If word distributions are represented as vectors, distributional similarity can be calculated based on vector similarity measures [13].

For example, we create a distribution vector space so that each conference corresponds to a column (common for both languages) and each word appearing at any conference (unique for each language) corresponds to a row. Frequencies of each word at a conference populate each cell.[1] Note that column-wise dimensions are identical for both English and Japanese, thus giving us a common axis to leverage in comparing distribution across languages.

Given a katakana string (i.e. a suspected transliteration) that is not contained in the dictionary, we can retrieve its distribution vector $\overrightarrow{wq}$ and then extract n rows $\overrightarrow{wd}$ from the English half of the corpus with the most similar distribution using a Cosine similarity measure as given in equation (1).

$$sim(q,d) = \frac{\overrightarrow{wq} \cdot \overrightarrow{wd}}{|\overrightarrow{wq}| \cdot |\overrightarrow{wd}|} = \frac{\sum_{i=1}^{n} wq_i \times wd_i}{\sqrt{\sum_{i=1}^{n} wq_i^2} \times \sqrt{\sum_{i=1}^{n} wd_i^2}} \qquad (1)$$

The words corresponding to the most similar vectors are likely to contain English equivalents of the transliteration. However, rather than trying to determine a single candidate based on distribution similarity, we extract a significant number of candidates (e.g. 10,000), under assumption that later processing will be able to correctly disambiguate among them.

After distribution vectors for both sides of comparable corpora are created we process all the katakana strings not contained in the dictionary are processed as a batch. For each such katakana string we retrieve n English words with most similar distribution. Retrieved words are used to dynamically build a back-transliteration source (i.e. language) model unique for each katakana string. The back-transliteration is then calculated as described in [15]. Each back-transliteration produced has a probability $P(E|J)$ associated with it. However, the raw probability is not adequate as a measure of transliteration pair "plausibility", since longer strings naturally tend to have lower probabilities associated with them. Thus, we calculate the plausibility score as given by equation (2).

[1] Raw frequencies can easily be replaced with a different weighting schema many of which have been proposed for use in IR [14].

Table 1: Comparison between different language models

Highest n pairs	100	500	1000
FULL	100.00	96.20	93.30
FULL$_{SP}$	100.00	96.80	93.90
DYNS	98.00	95.80	91.40
DYNS$_{SP}$	100.00	96.60	96.90
DYNF	100.00	96.60	92.00
DYNF$_{SP}$	100.00	98.00	94.50

Table 2: Highest scoring pairs

Katakana	Extracted Pair
コンフォーメーション "koNfoomeeshoN"	*conformation*
ハイブリダイゼーション "haiburidaizeeshoN"	*hybridization*
トランスフェクション "toraNsufekushoN"	*transfection*
インターカレーション "iNtaakareeshoN"	*intercalation*
バイオレメディエーション "baioremedieeshoN"	*bioremediation*
キャラクタリゼーション "kyarakutarizeeshoN"	*characterization*
エレクトロポレーション "erekutoroporeeshoN"	*electroporation*
トランスコンダクタンス "toraNsukoNdakutaNsu"	*transconductance*
インプリンティング "iNpuriNtiNgu"	*imprinting*
コンボリューション "koNboryuushoN"	*convolution*

$$S(E, J) = \sqrt[|J|]{P(E|J)} \quad (2)$$

Here, $|J|$ is the length of input Japanese string in katakana characters. After all katakana strings in the corpus are processed, each katakana and obtained back-transliteration are output sorted by the plausibility score.

5. EVALUATION

We evaluate the proposed methodology on the NTCIR-2 data collection [12]. This data collection consists of English and Japanese conference abstracts. Although the abstracts were aligned when they were presented at respective conferences, in the data collections these alignments are not provided since the data set is intended for CLIR evaluation. Therefore, we use distribution by conference as the basis for similarity calculation. There are 616 different conferences in this collection with about 63,000 distinct katakana strings on the Japanese side and about 100,000 different words on English side (when punctuation and numbers are ignored).

Given this set we extract English equivalents of 4,400 katakana strings with frequency between 5 and 40 using the above described method and $n = 10,000$ most similar words in the dynamic language model. As a baseline, we extract pairs based on the full language model (as described in [15]). All 100,000 English words appearing in the corpus with weights reflecting the corpus frequencies are used. Since all the English words are considered and the search space is explored exhaustively this method is very computationally expensive. Back-transliteration module used in all models is also trained as described in [15].

Table 3: Lowest scoring pairs

Katakana	Extracted Pair	Correct English
カゴ "kago"	cod	*cage*
ヨシ "yoshi"	huse	*reed*
アワ "awa"	amur	*foam*
ヤシ "yashi"	hae	*palm-tree*
ネギ "negi"	none	*leek*

However, we do not segment the katakana strings before back-transliteration.

The results are given in Table 1 for manually evaluated 100, 500 and 1000 highest scoring extracted pairs. The pairs are deemed correct when the Japanese katakana matches the English equivalent in any of its inflected forms. Thus システム "shisutemu" matched with either *system* or *systems* would be deemed as correct. Besides the figures for the full language model (FULL) we give numbers for dynamic model with weights computed based on corpus frequencies (DYNF) and based on the vector similarity score (DYNS). Furthermore, since extracted pairs often contain spelling errors on the English side (e.g. ヒューマンインターフェイス "hyuumanintaafeisu" being erroneously matched with *humen inteface* instead of *human interface*), we give two different numbers for each model: one where spelling errors were considered as errors (e.g. DYNF) and the other where obvious spelling errors were ignored (e.g. DYNF$_{SP}$).

We can see that precision is high for all tested models and that DYNF and DYNS achieve similar precision to FULL model although only about one tenth of English vocabulary is considered for each katakana string. This shows that even a simple model of cross-language distribution can be used to reduce the back-transliteration search space.

Most erroneously extracted pairs are due to incorrectly matched transliterations of English phrases. For example, コミュニケーションツール "komyunikeeshontsuuru" is erroneously matched with *communicational* instead of *communication tool*. Many of these errors could be avoided by performing katakana string segmentation before looking for a back-transliteration. Another solution could be to add a bigram language model to bias the system toward likely English word sequences.[2]

Table 2 shows ten highest scoring transliteration pairs extracted obtained by DYNF. On the other hand, Table 3 shows five lowest scoring transliteration pairs for the same model. Katakana strings appearing in this table are not transliterations but Japanese words which were not derived from English and as such have no back-transliterations. Thus, it seems that, proposed plausibility score provides the desired ordering: high score for correct transliteration pairs and low

[2]We tried directly using English bigram distributions instead of unigram distributions when selecting most similarly distributed words but this significantly reduced the overall precision due to a high number of function words being selected.

score for incorrect ones. In the future, we hope to determine an adequate threshold for filtering incorrect pairings.

5.1. Discussion

The evaluation described above provides encouraging but still limited results. Proposed model relies on the notion that the original word and its transliteration will be appearing in comparable corpora with similar distribution but rather than choosing a single term (as was previously case in bilingual term extraction) we select a large subset of the lexicon with similar distribution and then rely on the back-transliteration module to select the most appropriate pairing.

Another assumption that we naively (yet deliberately) make is that all katakana strings appearing in Japanese texts are transliterations although this is clearly not the case (as can be seen in Table 3). However, proposed scoring schema takes care of this problem to a large extent.

Finally, in order to apply our methodology to other languages, initial requirement would be to identify likely transliteration candidates. While more difficult than in case of Japanese, this could be achieved by considering all words not contained in the system dictionary as possible transliterations or using character statistics to detect unusual character patterns [16] and treat them as likely transliterations.

6. CONCLUSION

In this paper we proposed a novel method for bilingual term extraction. For each possible transliteration a subset of similarly distributed English words is extracted and then back-transliteration is used to find the best match in the extracted subset. Preliminary evaluation on NTCIR-2 data collection shows that this approach can yield good results and therefore deserves further consideration.

7. REFERENCES

[1] N. Collier, A. Kumano, and H. Hirakawa, "Acquisition of English-Japanese proper nouns from noisy-parallel newswire articles using Katakana matching," in *Proc. of NLPRS97* , Phuket, Thailand, December 2–4 1997, pp. 309–314.

[2] E. Brill, G. Kacmarcik, and C. Brockett, "Automatically harvesting katakana-English term pairs from search engine query logs," in *Proc. of NLPRS2001*, Tokyo, Japan, 2001, pp. 393–399.

[3] Keita Tsuji, "Automatic extraction of translational Japanese-katakana and English word pairs from bilingual corpora," *International Journal of Computer Processing of Oriental Languages*, vol. 15, no. 3, pp. 261–279, 2002.

[4] I.D. Melamed, "Models of translational equivalence among words," *Computational Linguistics*, vol. 26, no. 2, pp. 221–49, 2000.

[5] A. Kumano and H. Hirakawa, "Building an MT dictionary from parallel texts based on linguistic and statistical information," in *Proc. of COLING1994*, 1994, pp. 76–81.

[6] P. Fung and K. W. Church, "K-vec : a new approach for aligning parallel texts," in *Proc. of the ACL1995*, 1994.

[7] P. Fung and L.Y. Yee, "An IR approach for translating new words from nonparallel, comparable texts," in *Proc. of the COLING/ACL-98*, 1998, pp. 414–20.

[8] Fei Huang, Stephan Vogel, and Alex Waibel, "Improving named entity translation combining phonetic and semantic similarities," in *HLT-NAACL 2004*, Boston, Massachusetts, USA, May 2 - May 7 2004, pp. 281–288.

[9] Li Shao and Hwee Tou Ng, "Mining new word translations from comparable corpora," in *Proceedings of Coling 2004*, Geneva, Switzerland, Aug 23–Aug 27 2004, pp. 618–624, COLING.

[10] Harold Somers, "Bilingual parallel corpora and language engineering," in *In Proc. of Workshop on Language Engineering for South-Asian languages*, 2001.

[11] Yusuke Shinyama and Satoshi Sekine, "Named entity discovery using comparable news articles," in *Proceedings of Coling 2004*, Geneva, Switzerland, Aug 23–Aug 27 2004, pp. 848–853, COLING.

[12] Noriko Kando, Kazuko Kuriyama, and Masaharu Yoshioka, "Overview of Japanese and English Information Retrieval Tasks (JEIR) at the Second NTCIR Wordshop," in *Proc. of NTCIR Workshop 2*, 2001.

[13] Yuji Matsumoto, "Lexical knowledge acquisition," in *The Oxford handbook of computational linguistics*, Ruslan Mitkov, Ed., pp. 395–413. Oxford University Press, 2003.

[14] T. Tokunaga, *Jouhou kensaku to gengo syori*, University of Tokyo Press, 1999, (in Japanese).

[15] Slaven Bilac and Hozumi Tanaka, "Direct combination of spelling and pronunciation information for robust back-transliteration," in *Proc. of the CICLING2005*, Mexico City, Mexico, 2005, (to appear).

[16] B. Kang and K. Choi, "Effective foreign word extraction for Korean information retrieval," *Information Processing and Management*, vol. 38, pp. 91–109, 2002.