Automatic Discovery of Named Entity Variants
– Grammar-driven Approaches to Non-alphabetical Transliterations

Chu-Ren Huang
Institute of Linguistics
Academia Sinica, Taiwan
churenhuang@gmail.com

Petr Šimon
Institute of Linguistics
Academia Sinica, Taiwan
sim@klubko.net

Shu-Kai Hsieh
DoFLAL
NIU, Taiwan
shukai@gmail.com

Abstract

Identification of transliterated names is a particularly difficult task of Named Entity Recognition (NER), especially in the Chinese context. Of all possible variations of transliterated named entities, the difference between PRC and Taiwan is the most prevalent and most challenging. In this paper, we introduce a novel approach to the automatic extraction of diverging transliterations of foreign named entities by bootstrapping co-occurrence statistics from tagged and segmented Chinese corpus. Preliminary experiment yields promising results and shows its potential in NLP applications.

1 Introduction

Named Entity Recognition (NER) is one of the most difficult problems in NLP and Document Understanding. In the field of Chinese NER, several approaches have been proposed to recognize personal names, date/time expressions, monetary and percentage expressions. However, the discovery of transliteration variations has not been well-studied in Chinese NER. This is perhaps due to the fact that the transliteration forms in a non-alphabetic language such as Chinese are opaque and not easy to compare. On the hand, there is often more than one way to transliterate a foreign name. On the other hand, dialectal difference as well as different transliteration strategies often lead to the same named entity to be transliterated differently in different Chinese speaking communities.

Corpus	Example (Clinton)	Frequency
XIN	克林頓	24382
CNA	克林頓	150
XIN	柯林頓	0
CNA	柯林頓	120842

Table 1: Distribution of two transliteration variants for ”Clinton” in two sub-corpora

Of all possible variations, the cross-strait difference between PRC and Taiwan is the most prevalent and most challenging.[1]The main reason may lie in the lack of suitable corpus.

Even given some subcorpora of PRC and Taiwan variants of Chinese, a simple contrastive approach is still not possible. It is because: (1) some variants might overlap and (2) there are more variants used in each corpus due to citations or borrowing cross-strait. Table 1 illustrates this phenomenon, where CNA stands for Central News Agency in Taiwan, XIN stands for Xinhua News Agency in PRC, respectively.

With the availability of Chinese Gigaword Corpus (CGC) and Word Sketch Engine (WSE) Tools (Kilgarriff, 2004). We propose a novel approach towards discovery of transliteration variants by utilizing a full range of grammatical information augmented with phonological analysis.

Existing literatures on processing of transliteration concentrate on the identification of either the transliterated term or the original term, given knowledge of the other (e.g. (Virga and Khudanpur,

[1] For instance, we found at least 14 transliteration variants for Lewinsky,such as 呂茵斯基，呂文絲基，呂茵斯，陸文斯基，陸茵斯 基，柳思基，陸雯絲姬，陸文斯基，呂茵斯基，露文斯基，李文斯基，露温 斯基，蘿恩斯基，李雯斯基 and so on.

2003)). These studies are typically either rule-based or statistics-based, and specific to a language pair with a fixed direction (e.g. (Wan and Verspoor, 1998; Jiang et al., 2007)). To the best of our knowledge, ours is the first attempt to discover transliterated NE's without assuming prior knowledge of the entities. In particular, we propose that transliteration variants can be discovered by extracting and comparing terms from similar linguistic context based on CGC and WSE tools. This proposal has great potential of increasing robustness of future NER work by enabling discovery of new and unknown transliterated NE's.

Our study shows that resolution of transliterated NE variations can be fully automated. This will have strong and positive implications for cross-lingual and multi-lingual informational retrieval.

2 Bootstrapping transliteration pairs

The current study is based on Chinese Gigaword Corpus (CGC) (Graff el al., 2005), a large corpus contains with 1.1 billion Chinese characters containing data from Central News Agency of Taiwan (ca. 700 million characters), Xinhua News Agency of PRC (ca. 400 million characters). These two sub-corpora represent news dispatches from roughly the same period of time, i.e. 1990-2002. Hence the two sub-corpora can be expected to have reasonably parallel contents for comparative studies.[2]

The premises of our proposal are that transliterated NE's are likely to collocate with other transliterated NE's, and that collocates of a pair of transliteration variants may form contrasting pairs and are potential variants. In particular, since the transliteration variations that we are interested in are those between PRC and Taiwan Mandarin, we will start with known contrasting pairs of these two language variants and mine potential variant pairs from their collocates. These potential variant pairs are then checked for their phonological similarity to determine whether they are true variants or not. In order to effectively select collocates from specific grammatical constructions, the Chinese Word Sketch[3] is adopted. In particular, we use the Word Sketch difference (WSDiff) function to pick the grammatical contexts as well as contrasting pairs. It is important to bear in mind that Chinese texts are composed of Chinese characters, hence it is impossible to compare a transliterated NE with the alphabetical form in its original language. The following characteristics of a transliterated NE's in CGC are exploited to allow discovery of transliteration variations without referring to original NE.

- *frequent co-occurrence of named entities within certain syntagmatic relations* – named entities frequently co-occur in relations such as AND or OR and this fact can be used to collect and score mutual predictability.

- *foreign named entities are typically transliterated phonetically* – transliterations of the same name entity using different characters can be matched by using simple heuristics to map their phonological value.

- *presence and co-occurrence of named entities in a text is dependent on a text type* – journalistic style cumulates many foreign named entities in close relations.

- *many entities will occur in different domains* – famous person can be mentioned together with someone from politician, musician, artist or athlete. Thus allows us to make leaps from one domain to another.

There are, however, several problems with the phonological representation of foreign named entities in Chinese. Due to the nature of Chinese script, NE transliterations can be realized very differently. The following is a summary of several problems that have to be taken into account:

- *word ending*: 阿拉法 vs.阿拉法特 ”Arafat” or 穆巴拉 vs.穆巴拉克 ”Mubarak”. The final consonant is not always transliterated. XIN transliterations tend to try to represent all phonemes and often add vowels to a final consonant to form a new syllable, whereas CNA transliteration tends to be shorter and may simply leave out a final consonant.

- *gender dependent choice of characters*: 萊絲 莉 ”Leslie” vs.萊斯利 ”Chris” or 克莉絲特 vs. 克莉斯

[2] To facilitate processing, the complete CGC was segmented and POS tagged using the Academia Sinica segmentation and tagging system (Ma and Huang, 2006).

[3] http://wordsketch.ling.sinica.edu.tw

特. Some occidental names are gender neutral. However, the choice of characters in a personal name in Chinese is often gender sensitive. So these names are likely to be transliterated differently depending on the gender of its referent.

- *divergent representations caused by scope of transliteration, e.g. both given and surname vs. only surname*: 大威廉絲 / 維・威廉絲 ”Venus Williams”.
- *difference in phonological interpretation*: 賴夫特 vs. 拉夫特 ”Rafter” or 康諾斯 vs. 康那斯 ”Connors”.
- *native vs. non-native pronunciation*: 艾斯庫 德 vs. 伊斯庫德 ”Escudero” or 費德洛 vs. 費德勒 ”Federer”.

2.1 Data collection

All data were collected from Chinese Gigaword Corpus using Chinese Sketch Engine with `WSDiff` function, which provides side-by-side syntagmatic comparison of Word Sketches for two different words. `WSDiff` query for w_i and w_j returns patterns that are common for both words and also patterns that are particular for each of them. Three data sets are thus provided. We neglect the common patterns set and concentrate only on the wordlists specific for each word.

2.2 Pairs extraction

Transliteration pairs are extracted from the two sets, A and B, collected with `WSDiff` using default set of seed pairs :

- for each seed pair in seeds retrieve `WSDiff` for `and/or` relation, thus have pairs of word lists, $< A_i, B_i >$
- for each word $w_{ii} \in A_i$ find best matching counterpart(s) $w_{ij} \in B_i$. Comparison is done using simple phonological rules, viz. 2.3
- use newly extracted pairs as new seeds (original seeds are stored as good pairs and not queried any more)
- loop until there are no new pairs

Notice that even though substantial proportion of borrowing among different communities, there is no mixing in the local context of collocation, which means, local collocation could be the most reliable way to detect language variants with known variants.

2.3 Phonological comparison

All word forms are converted from Chinese script into a phonological representation[4] during the pairs extraction phase and then these representations are compared and similarity scores are given to all pair candidates.

A lot of Chinese characters have multiple pronunciations and thus multiple representations are derived. In case of multiple pronunciations for certain syllable, this syllable is commpared to its counterpart from the other set. E.g. (葉 has three pronunciations: *yè, xié, shè*. When comparing syllables such as 裴[pei,fei] and 斐[fei], 裴 will be represented as [fei]. In case of pairs such as 葉爾欽 [ye er qin] and 葉爾侵 [ye er qin], which have syllables with multiple pronunciations and this multiple representations. However, since these two potential variants share the first two characters (out of three), they are considered as variants without superfluous phonological checking.

Phonological representations of whole words are then compared by Levenstein algorithm, which is widely used to measure the similarity between two strings. First, each syllable is split into initial and final components: *gao*:*g*+*ao*. In case of syllables without initials like *er*, an ’ is inserted before the syllable, thus *er*:*’+er*.

Before we ran the Levenstein measure, we also apply phonological corrections on each pair of candidate representations. Rules used for these corrections are derived from phonological features of Mandarin Chinese and extended with few rules from observation of the data: (1) For **Initials**, (a): voiced/voiceless stop contrasts are considered as similar for initials: *g:k*, e.g. 高 [gao] (高爾) vs. 科 [ke] (科爾),*d:t*, *b:p*, (b): *r:l* 瑞 [rui] (柯吉瑞夫) 列 [lie] (科濟列夫) is added to distinctive feature set based on observation. (2). For **Finals**, (a): pair *ei:ui* is evaluated as equivalent.[5] (b): oppositions of nasalised final is evaluated as dissimilar.

[4] http://unicode.org/charts/unihan.html

[5] *Pinyin* representation of phonology of Mandarin Chinese does not follow the phonological reality exactly: [ui] = [uei] etc.

2.4 Extraction algorithm

Our algorithm will potentially exhaust the whole corpus, i.e. find most of the named entities that occur with at least few other names entities, but only if seeds are chosen wisely and cover different domains[6]. However, some domains might not overlap at all, that is, members of those domains never appear in the corpus in relation `and/or`. And concurrence of members within some domains might be sparser than in other, e.g. politicians tend to be mentioned together more often than novelists. Nature of the corpus also plays important role. It is likely to retrieve more `and/or` related names from journalistic style. This is one of the reasons why we chose Chinese Gigaword Corpus for this task.

3 Experiment and evaluation

We have tested our method on the Chinese Gigaword Second Edition corpus with 11 manually selected seeds Apart from the selection of the starter seeds, the whole process is fully automatic. For this task we have collected data from syntagmatic relation `and/or`, which contains words co-occurring frequently with our seed words. When we make a query for peoples names, it is expected that most of the retrieved items will also be names, perhaps also names of locations, organizations etc.

The whole experiment took 505 iterations in which 494 pairs were extracted.

Our complete experiment with 11 pre-selected transliteration pairs as seed took 505 iterations to end. The iterations identified 494 effective transliteration variant pairs (i.e. those which were not among the seeds or pairs identified by earlier iteration.) All the 494 candidate pairs were manually evaluated 445 of them are found to be actual contrast pairs, a precision of 90.01%. In addition, the number of new transliteration pairs yielded is 4,045%, a very productive yield for NE discovery.

Preliminary results show that this approach is competitive against other approaches reported in previous studies. Performances of our algorithms is calculated in terms of precision rate with 90.01%.

4 Conclusion and Future work

In this paper, we have shown that it is possible to identify NE's without having prior knowledge of them. We also showed that, applying WSE to restrict grammatical context and saliency of collocation, we are able to effectively extract transliteration variants in a language where transliteration is not explicitly represented. We also show that a small set of seeds is all it needs for the proposed method to identify hundreds of transliteration variants. This proposed method has important applications in information retrieval and data mining in Chinese data.

In the future, we will be experimenting with a different set of seeds in a different domain to test the robustness of this approach, as well as to discover transliteration variants in our fields. We will also be focusing on more refined phonological analysis. In addition, we would like to explore the possibility of extending this proposal to other language pairs.

References

Jiang, L. and M.Zhou and L.f. Chien. 2007. *Named Entity Discovery based on Transliteration and WWW* [In Chinese]. Journal of the Chinese Information Processing Society. 2007 no.1. pp.23-29.

Graff, David et al. 2005. *Chinese Gigaword Second Edition*. Linguistic Data Consortium, Philadelphia.

Ma, Wei-Yun and Huang, Chu-Ren. 2006. *Uniform and Effective Tagging of a Heterogeneous Giga-word Corpus*. Presented at the 5th International Conference on Language Resources and Evaluation (LREC2006), 24-28 May. Genoa, Italy.

Kilgarriff, Adam et al. 2004. *The Sketch Engine*. Proceedings of EURALEX 2004. Lorient, France.

Paola Virga and Sanjeev Khudanpur. 2003. *Transliteration of proper names in cross-lingual information retrieval*. In Proc. of the ACL Workshop on Multilingual Named Entity Recognition, pp.57-64.

Wan, Stephen and Cornelia Verspoor. 1998. *Automatic English-Chinese Name Transliteration for Development of Multiple Resources*. In Proc. of COLING/ACL, pp.1352-1356.

[6] The term domain refers to *politics,music,sport, film* etc.