Transliterated Named Entity Recognition Based on Chinese Word Sketch

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

Chu-Ren Huang
Institute of Linguistics
Academia Sinica,Taiwan
churen@gate.sinica.edu.tw

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

Jia-Fei Hong
Institute of Linguistics
Academia Sinica, Taiwan
jiafei@gate.sinica.edu.tw

Abstract

One of the unique challenges to Chinese Language Processing is cross-strait named entity recognition. Due to the adoption of different transliteration strategies, foreign name transliterations can vary greatly between PRC and Taiwan. This situation poses a serious problem for NLP tasks: including data mining, translation and information retrieval. In this paper, we introduce a novel approach to automatic extraction of divergent transliterations of foreign named entities by bootstrapping co-occurrence statistics from tagged Chinese corpora. In this study, we use Chinese Word Sketch The automatically bootstrapped transliteration pairs are further screened based on phonetic similarity. The precision is evaluated to be more than 90% against manually corrected transliteration pairs.

1 Introduction

Named entity recognition (NER) is one of the most difficult problems in Natural Language Processing and Document Understanding (MUC, 1998). In the field of Chinese NER, several approaches have been proposed to recognize personal names, date/time expressions, monetary and percentage expressions (Chen and Lee, 1996; Chen, Ding and Tsai, 1998).

However, the discovery of transliteration variations has not been well-studied in Chinese NER. This is perhaps due to the fact that the transliterations 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.

The most difficult, and potentially very rewarding type of transliteration variants to discover are the variations between different Mandarin Chinese regional variants. Identification of such variants will allow various knowledge engineering applications, such as search engine and information retrieval applications, to greatly expand searchable domain while able to identity information on the same named entity. Of all possible variations of transliterated named entities, the cross-strait difference between PRC and Taiwan is the most prevalent and most challenging. While there is centralized transliteration in PRC, a large diversity is prevalent in Taiwan[1].

In order to achieve this goal, two main difficulties must be resolved: First, transliterated names in Chinese do not typically occur with their original language. Hence there is no easy anchor to link different variants or to establish the identity of each transliteration. Second, there is no parallel corpus for different varieties of Mandarin Chinese. Since these are variants and not different languages, there is no need for translation. Even given some subcor-

[1] For instance, there are at least 14 different attested transliterations for *Lewinsky* in Taiwan, 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.

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

Table 1: Contrasting Transliteration of 'Clinton'

pora of PRC and Taiwan, transliteration variants do not necessarily occur only in one of the sub-corpus (due to citations or cross-strait borrowing), hence some variants do overlap. Table 1 illustrates this phenomenon, where CNA stands for Central News Agency in Taiwan, XIN stands for Xinhua News Agency in PRC, respectively. Lastly, in terms of NLP, this means that there is no parallel corpus for extraction of paired variants.

With the availability of Chinese Gigaword corpus and Chinese Word Sketch (Huang et al., 2005), we have an opportunity face the problem of contrasting transliterations of identical foreign named entities in Chinese text from PRC and Taiwan. In general, our approach depends on information about concurrence of word pairs in certain grammatical relation that are then compared phonologically.

In sum, our research issue is how to discover transliteration variants without identifying the original name and without mapping to the original term. Instead of rule-based or statistics-based model (Virga and Khudanpur, 2003; Wan and Verspoor, 1998), we propose that such variants can be discovered by extracting terms from similar linguistic context. In particular, when we compare the context of a pair of known transliteration variants, we can identify those transliterations that are uniquely co-occur with one of the variants to be a candidate for a new transliteration variant. Our study shows that resolution of transliterated NE variations can be fully automated. This will have strong implications for online information retrieval for cross-lingual and multi-lingual informational retrieval.

The rest of this paper is organized as follows: first we briefly introduce our resources: Chinese Word Sketch and Chinese Gigaword Corpus, then we will provide an overview of the extraction and decision procedures and finally we conclude with and experiment and an evaluation. Finally, we discuss possible applications and future work needed.

2 Resources

Two important resources, Chinese Gigaword Corpus and Chinese Word Sketch (or similar), are vital for the present method.

Chinese Gigaword Second Edition. Chinese Gigaword Second Edition corpus (CGSE) (Graff el al., 2005) contains data from Central News Agency (CNA, Taiwan), Xinhua News Agency (XIN, PRC) and Zaobao Newspaper (ZBN, Singapore) and therefore covers lexical data from three Chinese communities. Given the nature of the sources of the texts, CGSE provides rich resource of journalistic document style textual data. This corpus has been segmented and tagged automatically (Huang et al., 1997). Also the non-Taiwanese components of corpus have been converted to traditional characters to allow for simultaneous searches.

We focus on CNS and XIN components of CGSE. Taiwanese CNA component provides cca 700 million words and PRC XIN components cca 400 million words. This allows us to commence contrastive study of lexical differences between the two Chinese speaking communities.

We didn't include Singapore data in our study.

Chinese Word Sketch. A new corpus management tool, Sketch Engine, introduced in (Kilgarriff, 2004) is used to manage large Chinese Gigaword corpus. This tool can be found as Chinese Word Sketch at *http://wordsketch.ling.sinica.edu.tw/* (Huang et al., 2005). Apart from typical functionality of corpus managers like KWIC displays, frequency and co-occurence statistics, Sketch Engine (SkE), and naturally also Chinese Word Sketch (CWS), also provides grammar-wise co-occurrence statistics, i.e. it enables us to study words in co-occur in certain grammatical relation.

This grammar-wise co-occurrence information is provided via so called Word Sketches, which are triples of $\langle lemma_1, relation, lemma_2 \rangle$, where $lemma_1$ is a keyword of a query, displayed in tables for each relation. Thus we can obtain tables words that occur in relation Object, Subject, Modifier etc. with $lemma_1$. The tables also contain frequency of co-occurrence and salience of the collocate pair with

Object	40340	3.7
敗仗	452	75.68
藥	1843	74.03
晚飯	361	73.27
飯	834	70.05
…	…	…
早餐	493	62.12
頓飯	104	61.72
狗肉	204	61.68
…	…	…

Table 2: Word Sketch for 吃 (*chī*, to eat) and it's frequent objects. The second column indicates frequency of the item in the first column, i.e. the first row: *chī* appears 40340 times with any object; the second row: *bàizhàn* appears 452 times as an object of *chī*.

克林頓 and/or	1197	3940	1.9	1.5
江澤民	80	889	34.2	64.8
布萊爾	70	100	44.7	42.8
…	…	…	…	…
穆巴拉克	36	31	36.4	28.2
阿拉法特	45	76	33.0	33.2
拉賓	17	15	26.6	20.3
…	…	…	…	…

Table 3: Example of Word Sketch difference: ”克林頓” and ”柯林頓” common patterns

respect to the relation in question (Kilgarriff, 2004). Example is shown in Table 2.

Another vital function of SkE is *Word Sketch difference*. This function provides side-by-side comparison of two keywords in the same fashion as Word Sketch, i.e. as $\langle lemma_1, relation, lemma_2\rangle$ triples, in three different displays: *common* for collocates that appear with both keywords and exclusive collocates specific for each keyword. Table 3 shows *common* patterns, Figure 1 show patterns for that are special for keyword ”克林頓” and ”柯林頓” respectively.

3 Bootstrapping transliteration pairs

The main idea of our method is to collect collocates using a certain grammatical relation as a constraint and then compare phonological properties of these items. The bootstrapping process starts with few seed transliteration pairs that are selected so they cover different parts of the corpus.

Figure 1: Exclusive Sketch Difference for ”克林頓” and ”柯林頓”

克林頓		
and/or	1197	1.9
葉利欽	169	60.9
戈爾	57	51.8
巴拉克	48	46.0
布什	86	43.8
阿薩德	32	41.5
希拉克	38	39.1
萊温斯基	11	34.9
科爾	22	33.5
侯賽因	18	31.9
…	…	…

柯林頓		
and/or	3940	1.5
呂茵斯基	211	68.7
高爾	311	65.9
葉爾勤	317	59.2
布希	357	52.5
希拉蕊	86	50.1
巴瑞克	62	45.8
宋嘉斯	26	42.9
阿塞德	38	40.5
裴洛	39	40.0
…	…	…

Our method also exploits following properties of textual data:

- *frequent co-occurrence of named entities within certain stigmatic 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

- *presence and co-occurrence of named entities in a text is dependent on a text type* – journalistic style of Chinese Gigaword corpus accumulates many foreign named entities in close relations

- *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

- *many named entities will occur in different domains* – famous person can be mentioned together with someone from politician, musical, artistic or sport domain. This allows us to make leaps during the bootstrapping process 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 realised 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 or 穆巴拉 vs. 穆巴拉克. The ending consonant does not have to be represented. XIN transliterations tend towards longer (i.e. phonologically more complete) forms, whereas CNA data shows tendency towards shorter transliterations that blend in the Chinese text better.
- *difference between feminine/masculine choice of characters*: 萊絲 莉 vs. 萊斯利 or 克莉絲特 vs. 克莉斯特. Both versions can refer to either sex.
- *slightly diverging representations based on choice what to transcribe, e.g. both given and surname vs. only surname*: 大威廉絲 / 維・ 威廉絲.
- *difference in phonological interpretation, resulting in different choice of initial or final*: 賴夫特 vs. 拉夫特 or 康諾斯 vs. 康那斯.
- *non-English names transcribed according to English pronunciation, i.e. the commonly heard one, or a native pronunciation (native in respect to the named entity in question)*: 艾斯庫德 vs. 伊斯庫德 (Escudero) or 費德洛 vs. 費德勒 (Federer).

3.1 Data collection

All transliteration pairs were collected using Word Sketch difference queries. A query for seed pair $\langle w^{XIN}, w^{CNA}\rangle$ returns patterns that are common for both words and also patterns that are specific for each of them. We concentrate on the latter, but we neglect the frequency and salience information. Therefore, after each Word Sketch difference query, we end up with a list of collocates for each keyword. Since we are interested only in NE, we use only relation and/or, because only this relation is expected to yield named entities Relation and/or is defined as a relation of two nouns separated either by a conjunction or by an *IDEOGRAPHIC COMMA* "、".

Due to the fact that Chinese Gigaword corpus is composed of texts of journalistic style, it is expected that queries for peoples names will yield mostly names of persons, maybe institutions, etc. As we can see in the examples in Table 3 and Figure 1, this is really the case.

General formulation of the data collection phase is as follows:

- have a set of seed pairs $S = \{S_1, S_2, ..., S_n\}$, where $S_i = \langle w_{XIN}, w_{CNA}\rangle$.
- for each seed pair, retrieve Word Sketch difference for `and/or` relation, thus have two word lists, $L = \langle W_i^{XIN}, W_i^{CNA}\rangle$, where W_i^k is an unordered list of words.
- process each list of candidates L with the *pairs extraction algorithm*.

3.2 Pairs extraction

Using the wordlists of collocates specific for each of the keywords from the seed pair, we then compare phonological representation of each form specific to the first keyword to each form specific to the second keyword. Similarity of each candidate pair is scored.

General version of our extraction algorithm can be formulated as follows:

- given two lists of words W^{XIN} and W^{CNA} that potentially contain several identical NEs, although transliterated differently
- for each word $w_i^{XIN} \in W^{XIN}$ find matching counterpart(s) $w_j^{CNA} \in W^{CNA}$. Comparison is done using simple phonological rules, viz. 3.3
- use newly extracted pairs as new seeds (original seeds are stored as confirmed pairs and not queried any more)
- loop until there are no new pairs

3.3 Phonological comparison

All word forms are converted from Chinese script into a phonological representation[2] 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. We neglect the tone altogether in our study. When comparing syllables such as 裴[pei,fei] and 斐[fei], 裴 will be represented as [fei], since it's the best matching counterpart for 斐. In case of pairs such as 葉爾欽 [ye er qin] and 葉爾侵 [ye er qin],

[2] Using Unihan database: http://unicode.org/charts/unihan.html

which have syllables with multiple pronunciations (葉 has three possible pronunciations: *yè, xié, shè*) and thus multiple representations, a representation can be chosen randomly. However, since these syllables are represented by similar Chinese characters, we, naturally, skip the phonological comparison at all and select the first pronunciation[3]. Also note that as we loop through the list W^{XIN}, each $w_i^{XIN} \in W^{XIN}$ can have several different representations, because at each step j, the representation of w_i^{XIN} is influenced by $w_j^{CNA} \in W^{CNA}$. The major rule is: try to find the best matching representation. Therefore, when character 葉 is matched against 社 *shè*, it will be represented as *shè*.

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:

Initials

- distinctive features for initials are evaluated as similar: *g:k*,*d:t*, *b:p*. E.g. 高 [gao] (高爾) vs. 科 [ke] (科爾)

- *r:l* 瑞 [rui] (柯吉瑞夫) 列 [lie] (科濟列夫) is added to distinctive feature set based on observation.

Finals

- pair *ei:ui* is evaluated as equivalent

- opposition of non-/nasalised final is considered dissimilar.

3.4 Extraction algorithm

Our method will potentially exhaust the whole corpus, i.e. find all named entities that cooccur with at least few other names entities in the and/or, but only if seeds are chosen wisely and cover different domains[4]. However, concurrence of NE within some domains might be sparser then in other, e.g. politicians tend to be mentioned together more often then novelists. Nature of the corpus also plays important role. It is likely to retrieve more `and/or` related NEs from journalistic style. This is one of reasons why we chose Chinese Gigaword Corpus for this task.

Motivation. Our goal is to retrieve two lists of words that might contain indentical NEs that are transliterated differently. Given a news report stating: *Clinton and Yeltsin discussed. . .* , we can expect similar news to appear in both XIN and CNA news. Names of Clinton and Yeltsin, will be, however, transliterated as 克林頓 and 葉利欽 in XIN news and as 柯林頓 and 葉爾勤 in CNA news. Therefore, if we choose Clinton, i.e. (克林頓, 柯林頓), as our seed, we will obtain two lists of names coocuring with Clinton in `and/or`; two transliterations for Yeltsin will be present in respective lists. Our task is to identify indentical NEs. We cannot adhere to the original representation, Chinese script, but we can exploit the mechanism how these transliterations are created, i.e. using phonological similarity.

Each NE from the two lists, retrieved by $WSDiff$, is converted to phonological representation. Since each NE can have several phonological representations (viz. 3.3), from this many-to-many mapping we have to select the best matching pair. The similarity of this pair is then scored. If the score falls below a $threshold = 1.6$, which has been found experimentally, the score is stored as the $best$ and the loop continues.

During the phonological correction phase, perform several weighting step, such as:

1. phonologically very close pairs, such as *g*:*k*: -0.6

2. nasalization, *en*:*er*: different

3. pinyin incoherence for finals, `if` *ui*:*ei*, `then` -0.3, `else` +0.6

[3] Please note that it's irrelevant whether we choose "correct" pronunciation or not, as long as we choose consistently among the two compared syllables.

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

XIN	CNA	English
克林頓	柯林頓	Clinton
巴赫	巴哈	Bach
喬丹	喬登	Jordan
達文西	達芬奇	Da Vinci
畢卡索	畢加索	Picasso
碧咸	貝克漢	Beckham
萊温斯基	呂茵斯	Lewinsky
阿斯平	亞斯平	Aspen
侯賽因	胡笙	Hussain
卡斯特羅	卡斯楚	Castro
萊昂納多迪卡普里奧	李奧 納多迪卡皮歐	Leonardo DiCaprio

Table 4: Seed pairs

The score that results from these corrections is then compared agains our threshold.

The processing of one seed ends when we have compared all NEs from list L^{XIN} with all NEs from list L^{CNA}. At this point we might have zero or more new pairs of transliteration pairs.

```
Data: seed pairs $S = \{S_1, S_2, \ldots, S_n\}$
Result: transliteration pairs
while seeds $S$ contains unprocessed pairs do
    for each $\langle w^{XIN}, w^{CNA} \rangle \in S_i$ in seeds do
        retrieve $L^{XIN}$,$L^{CNA}$ = $WSDiff(w^{XIN}, w^{CNA})$;
        for $w_i^{XIN} \in L^{XIN}$ do
            candidates $cands$ = [];
            for $w_j^{CNA} \in L^{CNA}$ do
                get phonological representations, $R^{XIN}$, $R^{CNA}$;
                if size of $R^{XIN}$ or $(R^{CNA} > 1$ then
                    find the best matching pair of representations:
                    $r^{XIN}$ and $r^{CNA}$;
                end
                apply phonological corrections 3.3;
            end
            $score = levenstein(r^{XIN}, r^{CNA})$;
            if $score < threshold$ then
                $Best = 99$;                              /* dummy */
                if $score < Best$ then
                    $Best = score$;
                    $cands_{w^{XIN}}$ = $[w^{CNA}]$;
                end
                else if $score == Best$ then
                    add $w^{CNA}$ to $cands_{w^{XIN}}$;
                end
            end
        end
        remove processed seed $s_i$ from seeds $S$ and add it to transliteration
        pairs;
        add retrieved pair(s) to seeds $S$;
    end
end
```

Algorithm 1: Extraction algorithm

4 Experiment

We have tested our method on the Chinese Gigaword Second Edition corpus with 11 manually selected seeds. Table 4 shows our default 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.

CNA	XIN
威恩	貝恩
裴利	斐斯
柏格	比克
卡塞	卡特
腓力普	費利佩

Table 5: Example of common errors

CNA	XIN	Latinized form
安德瑞奧帝	安德烈奧蒂	Giulio Andreotti
阿塞德	阿薩德	Hafez Al-Assad
卡多索	卡多佐	Cardozo
德維勒班	德維爾潘	Dominique G. de Villepin
培瑞斯	佩雷斯	Shimon Peres
葉爾辛	葉爾勤	Boris Yeltsin
杜馬	迪馬	Dumas
魯賓	拉賓	Rabin Michael
霍斯	胡斯	Robert Huth
庫克	科克	Cork
加斯	夏斯	Tommy Haas
克羅西亞	克羅地亞	Croatia
季德	基德	Tracy Kidder

Table 6: Example of correct matches

Preliminary results are evaluated manually and shows that this approach is competitive against other approaches reported in previous studies. Performance of our algorithms is calculated in terms of precision rate with 90.01%. Table 5 shows example of common errors and Table 6 example of correctly identified NEs with their common form in Latin alphabet.

Please note, that out method also retrieves multiple forms within the tranliteration pairs, so e.g. for NE *Boris Yeltsin*, we would retrieve: 葉爾勤, 葉爾辛, 葉爾侵, 葉爾欽 , because all these forms are sufficiently close in their phonological representation.

5 Conclusion and Future work

We have shown that it is possible to retrieve transliterated named entities in Chinese with quite high efficiency. Our method has important implications in field of information retrieval and data mining of Chinese data.

As for the future directions: more refined phonological analysis might be needed to improve precision of the task. Also our present experiment was limited to comparison of lists limited to each seed pair. Extension to global comparison might yield better results.

References

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

Hsin-Hsi Chen and Jen-Chang Lee. 1996. *Identification and Classification of Proper Nouns in Chinese Texts*. Proceedings of 16th International Conference on Computational Linguistics. pp. 222-229.

Hsin-Hsi Chen and Yung-Wei Ding and Shih-Chung Tsai. 1998. *Named Entity Extraction for Information Retrieval*. Computer Processing of Oriental Languages, Special Issue on Information Retrieval on Oriental Languages, 12(1), pp. 75-85.

Jia-Fei Hong and Chu-Ren Huang. 2006. *WordNet Based Comparison of Language Variation - A study based on CCD and CWN Global WordNet*. Jeju Island, Korea.

Huang, Chu-Ren and Adam Kilgarriff and Yicing Wu and Chih-Min Chiu and Simon Smith and Pavel Rychlý and Ming-Hong Bai and Keh-Jiann Chen. 2005. *Chinese Sketch Engine and the Extraction of Collocations*. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, 48-55. Jeju Island, Korea.

Huang, Chu-Ren et al. 1997. *Segmentation Standard for Chinese Natural Language Processing*. Computational Linguistics and Chinese Language Processing. 2(2), pp. 47-62.

MUC. 1998. Proceedings of 7th Message Understanding Conference.

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.

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