Very Large-scale Lexical Resources to Enhance Chinese and Japanese IR and NLP

Extended Abstract[†]

ABSTRACT

A major issue in IR and other NLP applications such as machine translation is the recognition and translation of named entities.This is especially true for Chinese and Japanese, whose scripts present linguistic and algorithmic challenges not found in other languages. This paper focuses on the linguistic issues related to orthographic variation, shows how **Very Large-scale Lexical Resources** (VLSLR) can significantly enhance the accuracy of NLP tools, with focus on information retrieval (IR) and named entity recognition (NER) and named entity translation (NET).

1 INTRODUCTION

This paper summarizes some of the major linguistic issues related to Chinese and Japaneses named entities, including orthographic variation, and introduces several Very Large Scale Lexical Resources (VLSLR) consisting of millions of CJK named entities, such as a multilingual database of Japanese POIs (points of interest, such as schools, highways, hotels, etc) and personal names, and a very comprehensive multilingual database of Chinese personal names.

1.1 **Major Issues**

Some of the major factors that contribute to the difficulties of Chinese and Japanese NLP in general, and of IR and NER in particular, include:

1. Since the Japanese orthography is highly irregular, identifying, disambiguating and normalizing the large number of orthographic variants requires support for advanced IR capabilities such as cross-orthographic searching [6].
2. The morphological complexity of Japanese requires the use of a robust morphological analyzer, rather than a simple n-gram tokenizer, to perform such operations as segmentation, lemmatization, and decompounding.
3. The accurate conversion between Simplified Chinese (SC) and Traditional Chinese (TC), a deceptively simple but in fact extremely difficult computational task [5].
4. The difficulty of performing accurate segmentation of Japanese and Chinese texts [3,10], which requires identifying word boundaries by breaking a text stream into meaningful semantic units for information retrieval, dictionary lookup and indexing.
5. Miscellaneous retrieval technologies such as lexeme-based retrieval and the detection of discontinuous MWEs (e.g. extracting 'take off' + 'jacket' from 'he took his jacket off'), synonym expansion, and cross-language information retrieval (CLIR) [3].
6. Proper nouns and POIs pose special difficulties, as they are extremely numerous, difficult to detect without a lexicon, and have an unstable orthography.
7. Recognition of technical terms and term variants as described Jacquemin [4].

Each of the above is a major issue in itself. This paper will focus is on (1) summarizing the typology of orthographic variation, (2) the challenges of processing named entities, and (3) how large-scale lexical resources can significantly contribute to the disambiguation, identification, and translation of named entities.

1.2 **The role of Lexical Resources**

There is no question the lexical resources, such as dictionary databases and terminology glossaries, should play an important role in IR and MT. Even the advanced corpus-based technology used in state-of-the-art NMT (neural machine translation) systems and sophisticated NER systems often fail to recognize and accurately process entities such as Japanese proper nouns, especially POIs. As an informal experiment, we conducted some spot tests using Google Translate, which uses NMT technology, on several Japanese POIs. This resulted in a failure rate of 55%, a snippet of which is shown below.

Table 1: Translating POIs by NMT

Japanese	Google NMT	Correct translation
海の中道線	Midair line of the sea	Umi-no-Nakamichi Line
十和田観光電鉄線	Towada Shimbun photoelectric wire	Towada Kanko Electric Railway Line
神津島空港	Kozu Island airport	Kozushima Airport

Because of the irregular orthography of Japanese, lexeme-based procedures such as orthographic disambiguation cannot be based on probabilistic methods (e.g. bigramming) alone. Many attempts have been made along these lines, as for example in Brill [1] and Goto [3], with some claiming performance equivalent to lexicon-based methods, while Kwok [7] reports good results with only a small lexicon and simple segmentor.

These methods may be satisfactory for pure IR (relevant document retrieval), but for orthographic disambiguation, Emerson [2] and others have shown that a robust morphological analyzer capable of processing lexemes, rather than bigrams or n-grams, must be supported by a large-scale computational lexicon (even 100,000 entries is much too small).

The fundamental problem is that statistical methods, even when based on large-scale corpora, often fail to achieve

the high accuracy required for robust NLP applications unless they are supported by up-to-date, large-scale lexica. It has been shown in various studies [3,9] that MT systems and morphological analyzers capable of processing lexemes, rather than bigrams or n-grams, must be supported by large-scale computational lexica (often referred to as the hybrid approach). However, as is well known, such resources are expensive to build and time-consuming to maintain.

2 ORTHOGRAPHIC VARIATION IN JAPANESE

2.1 Irregular Orthography

One reason that the Japanese script is difficult to process by NLP tools is its highly irregular orthography. The numerous orthographic variants result from, among other factors, the unpredictable interaction between the four scripts used in Japanese; namely, kanji, hiragana, katakana and the Latin script [6]. This can be illustrated by the sentence 金の卵を産む鶏 *Kin no tamago wo umu niwatori* 'A hen that lays golden eggs.' . *Tamago* 'egg' has four variants (卵, 玉子, たまご, タマゴ), *niwatori* 'chicken' has three (鶏, にわとり, ニワトリ) and *umu* 'give birth to' has two (産む, 生む), which expands to 24 permutations. Since these variants occur frequently, the user has no hope of retrieving all instances of this sentence unless the IR application supports orthographic disambiguation.

2.2 Variant Typology

There are eight types of Japanese orthographical variation. The three most important ones are described below.

1. *Okurigana variants.* This refers to kana endings attached to a kanji base or stem, such as *okonau* 'perform', written 行う or 行なう, whereas *atarihazue* can be written in the six ways shown in the table below. Identifying and normalizing okurigana variants, which are numerous and unpredictable, is a major issue [6]. An effective solution is to use an orthographic variants lexicon, a solution adopted by some major search engine portals.

Table 2: Variants of *atarihazure*

Atarihazure	Type of variant
当たり外れ	"standard" form
当り外れ	okurigana variant
当外れ	okurigana variant
当外	all kanji
当たりはずれ	replace kanji with hiragana
あたり外れ	replace kanji with hiragana

2. *Cross-script variants.* This refers to variation across the four Japanese scripts in Japanese, including hybrid words written in multiple scripts, as shown below.

Table 3: Cross-Script Variation

Kanji	Hiragana	Katakana	Latin	Hybrid	English
人参	にんじん	ニンジン			carrot
		オープン	OPEN		open
硫黄		イオウ			sulfur
		ワイシャツ		Yシャツ	shirt
皮膚		ヒフ		皮フ	skin

3. Cross-script variants, which are also common and unpredictable, negatively impact recall and pose a major headache to NLP applications, including IR.
4. *Katakana variants.* The use of katakana loanwords is a major annoyance to NLP since katakana words are very numerous and their orthography is often irregular. It is common for the same word to be written in multiple, unpredictable ways, as shown below:

Table 4: Katakana Variants

Type	English	Standard	Variants
Macron	computer	コンピュータ	コンピューター
Long vowels	maid	メード	メイド
Multiple kana	team	チーム	ティーム

3 ORTHOGRAPHIC VARIATION IN CHINESE

Below is a brief description of the major issues in Chinese orthographical variation.

3.1 Simplified vs. Traditional

In mainland China and Singapore, the characters are written in simplified forms called Simplified Chinese (SC), whereas Taiwan, Hong Kong, Macau and most overseas Chinese communities continue to use the old, complex forms referred to as Traditional Chinese (TC) (Zongbiao 1986). Several factors contribute to the complexity of the Chinese script: (1) the large number of characters, (2) the major differences between SC and TC along various dimensions (graphemic, semantic and phonemic), (3) the many orthographic variants in TC, and (4) the difficulty of accurately converting between SC and TC.

3.2 Chinese-to-Chinese Conversion

The process of automatically converting SC to/from TC, referred to as C2C, is fraught with pitfalls. A detailed description of the linguistic issues can be found in [5], while the technical issues related to encoding and character sets are described in Lunde [8]. The conversion can be implemented on three levels in increasing order of sophistication.

3.2.1 *Code conversion.* The most unreliable way to perform C2C conversion is on a codepoint-to-codepoint basis by looking up in a mapping table, such as the one below. Because of the numerous one-to-many mappings (which occur in both the SC-to-TC and the TC-to-SC directions), the rate of conversion failure is unacceptably high.

Table 5: Code Conversion

SC	TC1	TC2	TC3	TC4
门	門			
发	發	髮		
干	幹	乾	干	榦

3.2.2 *Orthographic conversion.* The next level of sophistication in C2C is to convert orthographic units: that is, meaningful linguistic units, especially compounds and phrases that match on a one-to-one SC character to TC character basis. This gives better results because the orthographic mapping tables enable conversion on the word or phrase level rather than the codepoint level.

Table 6: Orthographic Conversion

English	SC	TC1
telephone	电话	電話
dry	干燥	乾燥
set out	出发	出發

The ambiguities inherent in code conversion can be resolved by using orthographic mapping table like the above, but because there are no word boundaries ambiguities must be done with the aid of a segmenor that can break the text stream into meaningful units [2].

3.2.3 *Lexemic conversion.* A more sophisticated, and more challenging, approach to C2C conversion is to map SC to TC lexemes that are semantically, rather than orthographically, equivalent. For example, SC 信息 (*xìnxī*) 'information' is converted to the semantically equivalent TC 資訊 (*zīxùn*). This is similar to the difference between *lorry* in British English and *truck* in American English.

There are numerous lexemic differences between SC and TC, especially in technical terms and proper nouns [11] . To complicate matters, the correct TC is sometimes locale-dependent, as shown below. Lexemic conversion is the most difficult aspect of C2C conversion and can only be done with the help of mapping tables.

Table 7: Lexemic Conversion

English	SC	Taiwan TC	HK TC	Other TC
software	软件	軟體	軟件	軟件
taxi	出租汽车	計程車	的士	德士
Osama Bin Laden	奥萨马本拉登	奥薩瑪賓拉登	奥薩瑪賓拉丹	

3.3 **Traditional Chinese Variants**

Traditional Chinese does not have a stable orthography. There are numerous TC variant forms, and some confusion prevails, as shown below:

Table 8: TC Variants

Var. 1	Var. 2	English	Comment
裏	裡	inside	100% interchangeable
教	教	teach	100% interchangeable
著	着	particle	variant 2 not in Big5
為	爲	for	variant 2 not in Big5

To some extent, TC forms are also used in the PRC, as in classical literature and newspapers for overseas Chinese. These are based on a standard that maps the SC forms in the character set GB 2312-80 to their corresponding TC forms in GB/T 12345-90. However, these mappings do not always agree with those used in Taiwan, as shown below:

Table 9: Mainland TC vs. Taiwan TC

Pinyin	SC	Mainland TC	Taiwan TC
xiàn	线	綫	線
cè	厕	厠	廁

There are various reasons for the existence of TC variants, such as that some TC forms were not available in the original Big Five character set used in Taiwan, the occasional use of SC forms, and others. To process TC texts it is necessary to disambiguate these variants using mapping tables [6].

4 LEXICAL RESOURCES

4.1 **Large-scale Resources**

There is no question that large-scale lexical resources can dramatically improve that accuracy of NLP tools. Though attempts at algorithmic solutions for some tasks, such as processing katakana loanwords, have been made [1], such major portals as Yahoo have adopted the most practical solution, namely using a hard-coded lexical databases.

Our institute, which specializes in CJK and Arabic computational lexicography, has for decades been engaged in research and development to compile comprehensive lexical databases with special emphasis on orthographic disambiguation, named entities, and technical terminology. Below are the principal lexical resources designed enhance the accuracy of IR and NLP applications.

4.2 **Japanese Resources**

Below is subset of some of the databases included in our VLSLR for Japanese.

1. The Japanese Personal Names Database covers over five million entries, including hiragana readings, numerous romanized variants (sometimes over 100 per name) and their English, SC, TC, and Korean equivalents.
2. The Japanese Lexical/Orthographic Database covers about 400,000 entries, including okurigana, kanji, and kana variants for orthographic disambiguation and grammar codes for morphological analysis.
3. The Comprehensive Database of Japanese POIs and Place Names, which covers about 3.1 million entries in 14 languages along with hiragana and romanized variants,.
4. The Database of Katakana Loanwords covers about 50,000 entries.
5. The Database of Japanese Companies and Organizations covers about 600,000 entries

4.3 **Chinese Resources**

Below is subset of some of the databases included in our VLSLR for Chinese.

1. The Comprehensive Chinese to Chinese Mapping Tables (C2C) exceeds 2.5 million entries and is constantly expanding. This covers general words, named entities and technical terms mapped to their TC equivalents, including such attributes as POS codes and type codes, and supports all three conversion levels, namely code, orthographic and lexemic conversion.
2. The Database of 100 Million Chinese Personal Names, an extremely comprehensive resource, covers Chinese personal names, their romanized variants, dialectical variants for Cantonese, Hokkien and Hakka, multilingual coverage for English, Japanese, Korean, and Vietnamese.
3. Database of Chinese Full Names: covers 4 million Chinese full names of real people, including celebrities.
4. Miscellaneous mapping tables such TC orthographic normalization tables, large scale pinyin databases showing the difference between SC and TC pronunciation, and others.

5 CONCLUSIONS

With computer memory inexpensive and virtually unlimited, it is no longer necessary for NLP to over-rely on corpora and algorithmic solutions. The time has come to leverage the power of VLSLR to significantly enhance the accuracy of IR, NER, MT and other NLP applications.

REFERENCES

[1] Brill, E. and Kacmarick, G. and Brocket, C. (2001) Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Logs. Microsoft Research, Proc. of the Sixth Natural Language Processing Pacific Rim Symposium, Tokyo, Japan.

[2] Emerson, T. (2000) Segmenting Chinese in Unicode. Proc. of the 16th International Unicode Conference, Amsterdam

[3] Goto, I., Uratani, N. and Ehara T. (2001) Cross-Language Information Retrieval of Proper Nouns using Context Information. NHK Science and Technical Research Laboratories. Proc. of the Sixth Natural Language Processing Pacific Rim Symposium, Tokyo, Japan

[4] Jacquemin, C. (2001) Spotting and Discovering Terms through Natural Language Processing. The MIT Press, Cambridge, MA

[5] Halpern, J. and Kerman J. (1999) The Pitfalls and Complexities of Chinese to Chinese Conversion. Proc. of the Fourteenth International Unicode Conference in Cambridge, MA.

[6] Halpern, J. (2008) Exploiting Lexical Resources for Disambiguating Orthographic CJK and Arabic Orthographic Variants. Proceedings of LREC 2008 in Marrakesh, Morocco

[7] Kwok, K.L. (1997) Lexicon Effects on Chinese Information Retrieval. Proc. of 2nd Conf. on Empirical Methods in NLP. ACL. pp.141-8.

[8] Lunde, Ken (2008) CJKV Information Processing. O'Reilly & Associates, Sebastopol, CA.

[9] Nakagawa, Tetsuji Chinese and Japanese Word Segmentation Using Word-Level and Character-Level Information. COLING '04 in Geneva, Switzerland. Proceedings of the 20th international conference on Computational Linguistics

[10] Yu, Shiwen, Zhu, Xue-feng and Wang, Hui (2000) New Progress of the Grammatical Knowledge-base of Contemporary Chinese. Journal of Chinese Information Processing, Institute of Computational Linguistics, Peking University, Vol.15 No.1.

[11] Tsou, B.K., Tsoi, W.F., Lai, T.B.Y. Hu, J., and Chan S.W.K. (2000) LIVAC, a Chinese synchronous corpus, and some applications. In "2000 International Conference on Chinese Language ComputingICCLC2000", Chicago.