Lexicon-based Orthographic Disambiguation in CJK Intelligent Information Retrieval
面向中日韩文智能信息检索的基于词典的异形词排歧

Jack Halpern（春遍雀來）jack@cjki.org
The CJK Dictionary Institute（日中韓辭典研究所）
34-14, 2-chome, Tohoku, Niiza-shi, Saitama 352-0001, Japan
Presented at COLING 2002

Abstract

The orthographical complexity of Chinese, Japanese and Korean (CJK) poses a special challenge to the developers of computational linguistic tools, especially in the area of **intelligent information retrieval**. These difficulties are exacerbated by the lack of a standardized orthography in these languages, especially the highly irregular Japanese orthography. This paper focuses on the typology of CJK orthographic variation, provides a brief analysis of the linguistic issues, and discusses why lexical databases should play a central role in the disambiguation process.

1 Introduction

Various factors contribute to the difficulties of CJK information retrieval. To achieve truly "intelligent" retrieval many challenges must be overcome. Some of the major issues include:

1. The lack of a standard orthography. To process the extremely large number of orthographic variants (especially in Japanese) and character forms requires support for advanced IR technologies such as **cross-orthographic searching** (Halpern 2000).
2. The accurate conversion between Simplified Chinese (SC) and Traditional Chinese (TC), a deceptively simple but in fact extremely difficult computational task (Halpern and Kerman 1999).
3. The morphological complexity of Japanese and Korean poses a formidable challenge to the development of an accurate morphological analyzer. This performs such operations as canonicalization, *stemming* (removing inflectional endings) and *conflation* (reducing morphological variants to a single form) on the morphemic level.
4. The difficulty of performing accurate word segmentation, especially in Chinese and Japanese which are written without interword spacing. This involves identifying word boundaries by breaking a text stream into meaningful semantic units for dictionary lookup and indexing purposes. Good progress in this area is reported in Emerson (2000) and Yu et al. (2000).
5. Miscellaneous retrieval technologies such as lexeme-based retrieval (e.g. 'take off' + 'jacket' from 'took off his jacket'), identifying syntactic phrases (such as 研究する from 研究をした), synonym expansion, and cross-language information retrieval (CLIR) (Goto et al. 2001).
6. Miscellaneous technical requirements such as transcoding between multiple character sets and encodings, support for Unicode, and input method editors (IME). Most of these issues have been satisfactorily resolved, as reported in Lunde (1999).
7. Proper nouns pose special difficulties for IR tools, as they are extremely numerous, difficult to detect without a lexicon, and have an unstable orthography.
8. Automatic recognition of terms and their variants, a complex topic beyond the scope of this paper. It is described in detail for European languages in Jacquemin (2001), and we are currently investigating it for Chinese and Japanese.

Each of the above is a major issue that deserves a paper in its own right. Here, the focus is on **orthographic disambiguation,** which refers to

the detection, normalization and conversion of CJK orthographic variants. This paper summarizes the typology of CJK orthographic variation, briefly analyzes the linguistic issues, and discusses why lexical databases should play a central role in the disambiguation process.

2 Orthographic Variation in Chinese

2.1 One Language, Two Scripts

As a result of the postwar language reforms in the PRC, thousands of character forms underwent drastic simplifications (Zongbiao 1986). Chinese written in these simplified forms is called **Simplified Chinese** (SC). Taiwan, Hong Kong, and most overseas Chinese continue to use the old, complex forms, referred to as **Traditional Chinese** (TC).

The complexity of the Chinese writing system is well known. Some factors contributing to this are the large number of characters in common use, their complex forms, the major differences between TC and SC along various dimensions, the presence of numerous orthographic variants in TC, and others. The numerous variants and the difficulty of converting between SC and TC are of special importance to Chinese IR applications.

2.2 Chinese-to-Chinese Conversion

The process of automatically converting SC to/from TC, referred to as **C2C conversion**, is full of complexities and pitfalls. A detailed description of the linguistic issues can be found in Halpern and Kerman (1999), while technical issues related to encoding and character sets are described in Lunde (1999). The conversion can be implemented on three levels in increasing order of sophistication, briefly described below.

2.2.1 Code Conversion The easiest, but most unreliable, way to perform C2C conversion is on a codepoint-to-codepoint basis by looking the source up in a mapping table, such as the one shown below. This is referred to as **code conversion** or **transcoding.** Because of the numerous one-to-many ambiguities (which occur in both the SC-to-TC and the TC-to-SC directions), the rate of conversion failure is unacceptably high.

Table 1. Code Conversion

SC	TC1	TC2	TC3	TC4	Remarks
门	門				one-to-one
汤	湯				one-to-one
发	發	髮			one-to-many
暗	暗	闇			one-to-many
干	幹	乾	干	榦	one-to-many

2.2.2 Orthographic Conversion The next level of sophistication in C2C conversion is referred to as **orthographic conversion,** because the items being converted are orthographic units**,** rather than codepoints in a character set. That is, they are meaningful linguistic units, especially multi-character lexemes. While code conversion is ambiguous, orthographic conversion gives better results because the orthographic mapping tables enable conversion on the word level.

Table 2. Orthographic Conversion

English	SC	TC1	TC2	Incorrect	Comments
telephone	电话	電話			unambiguous
we	我们	我們			unambiguous
start-off	出发	出發		出髮 齣髮 齣發	one-to-many
dry	干燥	乾燥		干燥 幹燥 榦燥	one-to-many
	阴干	陰乾	陰干		depends on context

As can be seen, the ambiguities inherent in code conversion are resolved by using an orthographic mapping table, which avoids false conversions such as shown in the **Incorrect** column. Because of segmentation ambiguities, such conversion must be done with the aid of a morphological analyzer that can break the text stream into meaningful units (Emerson 2000).

2.2.3 Lexemic Conversion A more sophisticated, and far more challenging, approach to C2C conversion is called **lexemic conversion**, which maps SC and TC lexemes that are **semantically,** *not* 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, as demonstrated by Tsou (2000). For example, there are more than 10 variants for 'Osama bin Laden.' To complicate matters, the correct TC is sometimes locale-dependent. Lexemic conversion is the most difficult aspect of C2C conversion and can only be done with the help of mapping tables. Table 3 illustrates various patterns of cross-locale lexemic variation.

Table 3. Lexemic Conversion

English	SC	Taiwan TC	Hong Kong TC	Other TC	Incorrect TC (orthographic)
Software	软件	軟體	軟件		軟件
Taxi	出租汽车	計程車	的士	德士	出租汽車
Osama bin Laden	奥萨马本拉登	奧薩瑪賓拉登	奧薩瑪賓拉丹		奧薩馬本拉登
Oahu	瓦胡岛	歐胡島			瓦胡島

2.3 Traditional Chinese Variants

Traditional Chinese does not have a stable orthography. There are numerous TC variant forms, and much confusion prevails. To process TC (and to some extent SC) it is necessary to disambiguate these variants using mapping tables (Halpern 2001).

2.3.1 TC Variants in Taiwan and Hong Kong Traditional Chinese dictionaries often disagree on the choice of the standard TC form. TC variants can be classified into various types, as illustrated in Table 4.

Table 4. 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
沉	沈	sink; surname	partially interchangeable
泄	洩	leak; divulge	partially interchangeable

There are various reasons for the existence of TC variants, such as some TC forms are not being available in the Big Five character set, the occasional use of SC forms, and others.

2.3.2 Mainland vs. Taiwanese Variants To a limited extent, the TC forms are used in the PRC for some classical literature, newspapers for overseas Chinese, etc., based on a standard that maps the SC forms (GB 2312-80) to their corresponding TC forms (GB/T 12345-90). However, these mappings do not necessarily agree with those widely used in Taiwan. We will refer to the former as **"Simplified Traditional Chinese"** (STC), and to the latter as **"Traditional Traditional Chinese"** (TTC).

Table 5. STC vs. TTC Variants

Pinyin	SC	STC	TTC
xiàn	线	綫	線
bēng	绷	綳	繃
cè	厕	廁	廁

3 Orthographic Variation in Japanese

3.1 One Language, Four Scripts

The Japanese orthography is highly irregular. Because of the large number of orthographic variants and easily confused homophones, the Japanese writing system is significantly more complex than any other major language, including Chinese. A major factor is the complex interaction of the four scripts used to write Japanese, resulting in countless words that can be written in a variety of often unpredictable ways (Halpern 1990, 2000). Table 6 shows the orthographic variants of 取り扱い *toriatsukai* 'handling', illustrating a variety of variation patterns.

Table 6. Variants of *toriatsukai*

Toriatsukai	**Type of variant**
取り扱い	"standard" form
取扱い	okurigana variant
取扱	All kanji
とり扱い	replace kanji with hiragana
取りあつかい	replace kanji with hiragana
とりあつかい	All hiragana

An example of how difficult Japanese IR can be is the proverbial "A hen that lays golden eggs." The "standard" orthography would be 金の卵を産む鶏 (*Kin no tamago wo umu niwatori*). In reality, *tamago* 'egg' has four variants (卵, 玉子, たまご, タマゴ), *niwatori* 'chicken' three (鶏, にわとり, ニワトリ) and *umu* 'to lay' two (産む, 生む), which expands to 24 permutations like 金の卵を生むニワトリ, 金の玉子を産む鶏 etc. As can be easily verified by searching the web, these variants frequently occur in webpages. Clearly, the user has no hope of finding them unless the application supports orthographic disambiguation.

3.2 Okurigana Variants

One of the most common types of orthographic variation in Japanese occurs in kana endings, called 送り仮名 *okurigana*, that are attached to a kanji base or stem. Although it is possible to generate some okurigana variants algorithmically, such as nouns (飛出し) derived from verbs (飛出す), on the whole hard-coded tables are required. Because usage is often unpredictable and the variants are numerous, okurigana must play a major role in Japanese orthographic disambiguation.

Table 7. Okurigana Variants

English	**Reading**	**Standard**	**Variants**
publish	*kakiarawasu*	書き表す	書き表わす 書表わす 書表す
perform	*okonau*	行う	行なう
handling	*toriatsukai*	取り扱い	取扱い 取扱

3.3 Cross-Script Orthographic Variants

Japanese is written in a mixture of four scripts (Halpern 1990): **kanji** (Chinese characters), two syllabic scripts called **hiragana** and **katakana,** and **romaji** (the Latin alphabet). Orthographic variation across scripts, which should play a major role in Japanese IR, is extremely common and mostly unpredictable, so that the same word can be written in hiragana, katakana or kanji, or even in a mixture of two scripts. Table 8 shows the major cross-script variation patterns in Japanese.

Table 8. Cross-Script Variants

Kanji vs. Hiragana	大勢　おおぜい
Kanji vs. Katakana	硫黄　イオウ
Kanji vs. hiragana vs. katakana	猫　ねこ　ネコ
Katakana vs. hybrid	ワイシャツ　Yシャツ
Kanji vs. katakana vs. hybrid	皮膚　ヒフ　皮フ
Kanji vs. hybrid	彗星　すい星
Hiragana vs. katakana	ぴかぴか　ピカピカ

3.4 Kana Variants

Recent years have seen a sharp increase in the use of katakana, a syllabary used mostly to write loanwords. A major annoyance in Japanese IR is that katakana orthography is often irregular; it is quite common for the same word to be written in multiple, unpredictable ways which cannot be generated algorithmically. Hiragana is used

mostly to write grammatical elements and some native Japanese words. Although hiragana orthography is generally regular, a small number of irregularities persist. Some of the major types of kana variation are shown in Table 9.

Table 9. Katakana and Hiragana Variants

Type	English	Reading	Standard	Variants
Macron	computer	*konpyuuta* *konpyuutaa*	コンピュータ	コンピューター
Long vowels	maid	*meedo*	メード	メイド
Multiple kana	team	*chiimu* *tiimu*	チーム	ティーム
Traditional	big	*ookii*	おおきい	おうきい
づ vs. ず	continue	*tsuzuku*	つづく	つずく

The above is only a brief introduction to the most important types of kana variation. There are various others, including an optional middle dot (*nakaguro)* and small katakana variants (クォ vs. クオ), and the use of traditional (じ vs. ぢ) and historical (い vs. ゐ) kana.

3.5 Miscellaneous Variants

There are various other types of orthographic variants in Japanese, which are beyond the scope of this paper. Only a couple of the important ones are mentioned below. A detailed treatment can be found in Halpern (2000).

3.5.1 Kanji Variants Though the Japanese writing system underwent major reforms in the postwar period and the character forms have by now been standardized, there is still a significant number of variants in common use, such as abbreviated forms in contemporary Japanese (才 for 歳 and 巾 for 幅) and traditional forms in proper nouns and classical works (such as 嶋 for 島 and 發 for 発).

3.5.2 Kun Homophones An important factor that contributes to the complexity of the Japanese writing system is the existence of a large number of homophones (words pronounced the same but written differently) and their variable orthography (Halpern 2000). Not only can each kanji have many *kun* readings, but many *kun* words can be written in a bewildering variety of ways. The majority of *kun* homophones are often close or even identical in meaning and thus easily confused, i.e., *noboru* means 'go up' when written 上る but 'climb' when written 登る, while *yawarakai* 'soft' is written 柔らかい or 軟らかい with identical meanings.

4 Orthographic Variation in Korean

4.1 Irregular Orthography

The Korean orthography is not as regular as most people tend to believe. Though hangul is often described as "logical," the fact is that in modern Korean there is a significant amount of orthographic variation. This, combined with the morphological complexity of the language, poses a challenge to developers of IR tools. The major types of orthographic variation in Korean are described below.

4.2 Hangul Variants

The most important type of orthographic variation in Korean is the use of variant hangul spellings in the writing of loanwords. Another significant kind of variation is in the writing of non-Korean personal names, as shown in Table 10.

Table 10. Hangul Variants

cake	케이크 (*keikeu*)	케잌 (*keik*)
yellow	옐로우 (*yelrou*)	옐로 (*yelro*)
Mao Zedong	마오쩌뚱 (*maojjeottung*)	모택동 (*motaekdong*)
Clinton	클린턴 (*keulrinteon*)	클린톤 (*keulrinton*)

4.3 Cross-Script Orthographic Variants

A factor that contributes to the complexity of the Korean writing system is the use of multiple scripts. Korean is written in a mixture of three scripts: an alphabetic syllabary called **hangul,** Chinese characters called **hanja** (their use is declining) and the Latin alphabet called **romaja**. Orthographic variation across scripts is not uncommon. The major patterns of cross-script variation are shown Table 11.

Table 11. Cross-Script Orthographic Variants

Type of Variation	English	Var. 1	Var. 2	Var.3
Hanja vs. hangul	many people	大勢 (*daese*)	대세 (*daese*)	
Hangul vs. hybrid	shirt	와이셔츠(*wai-syeacheu*)	Y셔츠 (*wai-syeacheu*)	
Hangul vs. numeral vs. hanja	one o'clock	한시 (*hansi*)	1시 (*hansi*)	一時 (*hansi*)
English vs. hangul	sex	sex	섹스 (*sekseu*)	

4.4 Miscellaneous Variants

4.4.1 North vs. South Korea Another factor contributing to the irregularity of hangul orthography is the differences in spelling between South Korea (S.K.) and North Korea (N.K.). The major differences are in the writing of loanwords, a strong preference for native Korean words, and in the writing of non-Korean proper nouns. The major types are shown below.

1. **Place names:** N.K. 오사까 (*osakka*) vs. S.K. 오사카 (*osaka*) for 'Osaka'
2. **Personal names:** N.K. 부슈 (*busyu*) vs. S.K. 부시 (*busi*) for 'Bush'
3. **Loanwords:** N.K. 미싸일 (*missail*) vs. S.K. 미사일 (*misail*) for 'missile'
4. **Russian vs. English:** N.K. 그루빠 (*guruppa*) vs. S.K. 그룹 (*geurup*)
5. **Morphophonemic:** N.K. 람용 (*ramyong*) vs. S.K. 남용 (*namyong*)

4.4.2 New vs. Old Orthography The hangul script went through several reforms during its history, the latest one taking place as recently as 1988. Though the new orthography is now well established, the old orthography is still important because the affected words are of high frequency and their number is not insignificant. For example, the modern 일군 'worker' (*ilgun*) was written 일꾼 (*ilkkun*) before 1988, while 빛갈 'color' (*bitgal*) was written 빛깔 (*bitkkal*).

4.4.3 Hanja Variants Although language reforms in Korea did not include the simplification of the character forms, the Japanese occupation of Korea resulted in many simplified Japanese character forms coming into use, such as the Japanese form 発 to replace 發 (*bal*).

4.4.4 Miscellaneous Variants There are various other types of orthographic variation, which are beyond the scope of this paper. This includes the use of abbreviations and acronyms and variation in interword spacing in multiword compounds. For example, 'Caribbean Sea' (*karibeuhae*) may be written solid (카리브해) or open (카리브 해).

5 The Role of Lexical Databases

Because of the irregular orthography of CJK languages, 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 Brill (2001) and Goto et al. (2001), with some claiming performance equivalent to lexicon-based methods, while Kwok (1997) 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 and C2C conversion, Emerson (2000) 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 CJK Dictionary Institute (CJKI), which specializes in CJK computational lexicography, is engaged in an ongoing research and development effort to compile comprehensive CJK lexical databases (currently about 5.5 million entries), with special emphasis on orthographic disambiguation and proper nouns. Listed below are the principal components useful for intelligent IR tools and orthographic disambiguation.

1. **Chinese to Chinese conversion.** In 1996, CJKI launched a project to investigate C2C conversion issues in-depth, and to build comprehensive mapping tables (now at 1.3 million SC and 1.2 million TC items) whose goal is to achieve near 100% conversion accuracy. These include:
   a. SC-to/from-TC code-level mapping tables

b. SC-to/from-TC orthographic and lexemic mapping tables for general vocabulary
c. SC-to/from-TC orthographic mapping tables for proper nouns
d. Comprehensive SC-to/from-TC orthographic/lexemic mapping tables for technical terminology, especially IT terms

2. **TC orthographc normalization tables**
   a. TC normalization mapping tables
   b. STC-to/from-TTC character mapping tables
3. **Japanese orthographic variant databases**
   a. A comprehensive database of Japanese orthographic variants
   b. A database of semantically classified homophone groups
   c. Semantically classified synonym groups for synonym expansion (Japanese thesaurus)
   d. An English-Japanese lexicon for CLIR
   e. Rules for identifying unlisted variants

Conclusions

CJK IR tools have become increasingly important to information retrieval in particular and to information technology in general. As we have seen, because of the irregular orthography of the CJK writing systems, intelligent information retrieval requires not only sophisticated tools such as morphological analyzers, but also lexical databases fine-tuned to the needs of orthographic disambiguation.

Few if any CJK IR tools perform orthographic disambiguation. For truly "intelligent" IR to become a reality, not only must lexicon-based disambiguation be supported, but such emerging technologies as CLIR, synonym expansion and cross-homophone searching should also be implemented.

We are currently engaged in further developing the lexical resources required for building intelligent CJK information retrieval tools and for supporting accurate segmentation technology.

References

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.

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

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

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

Halpern, J. (1990) *Outline Of Japanese Writing System.* In "New Japanese-English Character Dictionary", 6th printing, Kenkyusha Ltd., Tokyo, Japan (www.kanji.org/kanji/japanese/writing/outline.htm)

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.

Halpern, J. (2000) *The Challenges of Intelligent Japanese Searching.* Working paper (www.cjk.org/cjk/joa/joapaper.htm), The CJK Dictionary Institute, Saitama, Japan.

Halpern, J. (2001) *Variation in Traditional Chinese Orthography.* Working paper (www.cjk.org/cjk/cjk/reference/chinvar.htm), The CJK Dictionary Institute, Saitama, Japan.

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

Lunde, Ken (1999) *CJKV Information Processing.* O'Reilly & Associates, Sebastopol, CA.

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.

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 .

Zongbiao (1986) 简化字总表 (*Jianhuazi zongbiao)* (Second Edition). 国家语言文字工作委员会, 语文出版社, China.