The Challenges of Chinese and Japanese Searching

A. Researcher

September 3, 2025

Abstract

The process of tailoring website search functionality for Chinese and Japanese languages presents numerous challenges not commonly encountered with European languages. The way users interact with websites in China and Japan differs significantly, necessitating specialized adaptations for product and content search. The writing systems, which are based wholly or partly on Chinese characters, feature highly irregular orthography and a wide variety of language and character variants. This paper outlines the most significant linguistic quirks and idiosyncrasies that must be considered when implementing a language-aware full-text search. It covers critical aspects such as language detection, character and script variants, word segmentation, normalization, and handling of numerals, synonyms, and homophones. The objective is to provide a comprehensive overview of these issues and demonstrate how they can be addressed using modern information retrieval techniques.

Contents

Acknowledgements

A lot of thanks to my co-authors who helped me go through the linguistic quirks and idiosyncrasies, revise and extend this writing, and eventually having me come off looking like a pro. Thanks to **Timofey Klyubin** who is a guru in Japanese, and **Dmitry Antonov** who gave me valuable feedback, great tips, and pointers on Chinese.

1 Introduction

Three languages are traditionally considered together in the context of information retrieval, internationalization, and localization: Chinese, Japanese, and Korean. Their writing systems are based entirely or partly on Chinese characters. This research can be useful for internationalization, localization and information retrieval components and projects. Internationalization is mainly about support for multiple languages and cultures, while localization stands for adaptation of language, content, and design to specific countries, regions, or cultures. Cross-lingual information retrieval deals with documents in one or more different languages, and the techniques for indexing, searching, and retrieving information from large multi-language collections.

From the perspective of information retrieval, Chinese and Japanese present numerous challenges. The major issue is their highly irregular orthography and language variants. In this article, I collected the most important ones we need to take into account when implementing the language-aware full text search as well as how to address them.

2 Language Detection

When and where possible, the website should allow the user to specify unambiguously what language is going to be used for entering a search query and presenting the results. Normally, the users enter search queries in the same language as the website's interface is set to. However, our observations show that the customers use their native language if the website is advertised in their country even if the localized version of the website is not pre-selected, automatically or manually. If the first-level domain is from the local pool (.cn for China or .jp for Japan), the user's intent of using the native language is even stronger.

To address this case, there are AI and statistical techniques to determine the likely language. The automatic language detection is a very challenging task especially if the analyzed string is short. For example, if it has a mix of Latin and Chinese characters from the Japanese Kanji set, it may indicate that the text is either in Japanese or Chinese, which can be too abstract.

3 Language Variants

3.1 Dialects

There are many more dialects in Chinese than Japanese, both full of specificities interesting to us in regard to the topic. Chinese is not a single language; it is a family of spoken languages. China has a lot of dialects, but the most popular are Mandarin (or "Standard Chinese", over 1 billion speakers) and Cantonese (or Yue, over 100 million speakers). In Japan, there are two major types of the Japanese language: the Tokyo-type (or Eastern) and the Kyoto-Osaka type (or Western). The form that is considered the standard is called "Standard Japanese". Unlike Traditional and Simplified Chinese, standard Japanese has become prevalent nationwide.

3.2 Scripts

3.2.1 Japanese: Kana and Kanji

There are two typical Japanese scripts, Kana and Kanji.

- **Kanji** is logographic Chinese scripts, Chinese characters adapted to write Japanese words. There are thousands of kanji in Japanese.
- **Kana** is a collective term for Japanese syllabaries, Hiragana (46 characters) and Katakana (48 characters). They are derived by simplifying Chinese characters selected to represent syllables of Japanese.

The same Japanese word can be written in either kana or kanji.

Table 1: Japanese Script Variants for the word "fox".

English word	Japanese (Kanji)	Japanese (Katakana)	Japanese (Hiragana)
fox	狐	キツネ	きつね

This complexity is also illustrated by the sentence 金の卵を産む鶏 (“A hen that lays golden eggs”). The word ‘egg’ has four variants (卵, 玉子, たまご, タマゴ), ‘chicken’ has three (鶏, にわとり, ニワトリ) and ‘giving birth to’ has two (産む, 生む), which expands to 24 permutations. In many contexts only one option is correct.

Japanese has a large number of loan words or *gairaigo*. The considerable portion of them is derived from English. In written Japanese, gairaigos are usually written in katakana. Many gairaigos have native equivalents in Japanese. Sometimes a Japanese person can use either a native form or its English equivalent written in katakana. This is especially the case of proper names or science terms. If you are not familiar with the native variant, you will probably use a syllabic construct.

Table 2: Japanese Native Words and English Loan Equivalents.

English word	Japanese (native word)	Japanese (English loan word)
door	扉 /tobira/, 戸 /to/	ドア /doa/
mobile phone/cell phone	携帯 /keitai/, 携帯電話 /keitaidenwa/	モバイルフォン /mobairufon/, セルラー電話 /serur denwa/

School kids use hiragana more commonly since they might not have learned the kanji equivalents yet. Additionally, there is Romaji which uses Latin script to represent Japanese.

3.2.2 Chinese: Traditional and Simplified

Along with the sheer complexity and size of the character set, Chinese has several related language variants. In Taiwan, Hong Kong, and Macao, Traditional Chinese characters are predominant over the Simplified Chinese variant which is used mainly in Mainland China, Singapore, and Malaysia. Some traditional Chinese characters, or derivatives of them, are also found in Japanese writing. So there is a subset of characters common for different languages. These shared Chinese, Japanese, and Korean characters constitute a set named CJK Unified Ideographs. The CJK part of Unicode defines a total of 87,887 characters, though the characters needed for everyday use is much smaller. For search, queries can be in either traditional or simplified characters or a combination of the two; search results should contain all matching resources, whether traditional or simplified.

Below is a random text to demonstrate the differences between the writing systems. Characters highlighted by yellow marker have different spelling in Simplified (输入简体字点下面繁体字按钮进行在线转换) and Traditional (輸入簡體字點下面繁體字按鈕進行在線轉換) Chinese.

输	入	简	体	字	点	下	面	繁	体	字	按	钮	进	行	在	线	转	换
輸	入	簡	體	字	點	下	面	繁	體	字	按	鈕	進	行	在	線	轉	換

Figure 1: Comparison of a sentence in Simplified (top) and Traditional (bottom) Chinese.

4 Character Variants

Chinese and Japanese characters don’ t use upper or lower cases. They have only a single representation independent of context. The majority of letters are monospaced. There are no additional decorations for the letters as it is in Arabic, for example.

5 Conversion Between the Systems

The conversion is important when either a user or a document uses a mix of Chinese writing systems. For example, given a user query 舊小說 (‘old fiction’ in Traditional Chinese), the results should include matches for 舊小說 (traditional) and 旧小说 (simplified characters for ‘old fiction’). This means that conversion should be done at the query level. The accurate conversion between Simplified Chinese and Traditional Chinese is a deceptively simple but in fact extremely difficult computational task. If your search is used by millions, the system will be much more resource-intensive compared with the setup for European languages.

There are three methods of conversion:

- **Code conversion (codepoint-to-codepoint)**. This method is based on the mapping table and considered the most unreliable because of the numerous one-to-many mappings (in both directions). The rate of conversion failure is unacceptably high.
- **Orthographic conversion**. In this method, the meaningful linguistic units, especially compounds and phrases, are considered. Orthographic mapping tables enable conversion on the word or phrase level rather than the codepoint level. An excellent example is the Chinese word for “computer” .
- **Lexemic conversion**. A more sophisticated, and more challenging, approach to conversion. In this method, the mapping table contains lexemes that are semantically, rather than orthographically, equivalent. This is similar to the difference between *lorry* in British English and *truck* in American English. The complexity of this method lies in lexemic differences between Simplified and Traditional Chinese, especially in technical terms and proper nouns.

In Japanese, the kanji characters may or may not have the same-looking Chinese character.

It is generally believed that the top priority for Chinese discovery improvements is to equate Traditional characters with simplified characters. For Japanese, there is also a problem of equating Modern Kanji characters with Traditional Kanji characters, but it is not as strong as it is in Chinese where you deal with two different scripts. There is a priority for Japanese discovery improvements to equate all scripts used in the language: Kanji, Hiragana, Katakana, and Romaji. In Apache Solr, the only other relevant ICU script translation is a mapping between

Table 3: Examples of Simplified-to-Traditional Chinese Conversion.

Simplified Chinese	Traditional Chinese	Translation
干	幹 or 乾 or 榦	(dry, make, surname)
电话	電話	(telephone)
软件	軟體 (Taiwan)	(software)
计 算 机 (“calculating machine”)	電脑 (“electronic brain”)	(computer)

Table 4: Character Variation Between Chinese and Japanese.

Chinese (Simp.)	Chinese (Trad.)	Japanese (Kanji)	Japanese (Kana)	Translation
两	兩	両		(both)
龟	龜	亀	カメ	(tortoise)

Hiragana and Katakana. This is a straightforward one-to-one character mapping working in both directions .[1]

Consider making Simplified Chinese and traditional Chinese inter-searchable. If one searches for 计算机 (computer, Simplified) or 電脑 (computer, Traditional), the results should contain the records with both 计算机 and 電脑. At least measure how often each of these writing systems is used by your customers to make an educated decision on how to make search better.

6 Word Segmentation

Chinese and Japanese are written in a style that does not delimit word boundaries. Typical Chinese sentences include only Chinese characters, along with a select few punctuation marks and symbols. Typical Japanese sentences include mostly Japanese kana and some adopted Chinese characters that are used in the Japanese writing system. So, how does one decide how to break up the words when there are no separators in between?. As for spaces, they delineate words inconsistently and with variation among writers. Formally, there must always be a space between English words and Chinese words, but in fact this rule is not strict and many neglect it. There is no space between the Arabic numbers and Chinese characters.

There are different approaches for splitting the text into word units. The most common algorithms use dictionaries and, additionally, a set of rules. This topic is still an area of considerable research among the machine learning community. These are not perfect: this segmentation cannot be done unambiguously, but different methods show acceptable results for specific areas. For example, for scientific texts, the dictionary-based methods may show poorer results than statistical or machine-learning based ones.

For example, the word “中华人民共和国” (People’s Republic of China) is seven characters long and has smaller words within: “人民” (people) and “共和国” (republic country). The first two characters, “中华” are usually not be used as a word independently in modern Chinese, though it can be used as a word in ancient Chinese. Digging further, within the word “人民” (people), “人” is a word (human), but “民” (civilian or folk) is not a standalone word. As another example, while the proper segmentation of “中华人民共和国外交部” (Ministry of Foreign Affairs of the PRC) is “中华人民共和国 / 外交部”, another word, “国外” (overseas),

[1]Here I mentioned Apache Solr for the first time. For those who are not familiar with Solr, it is one of the most comprehensive opensource search engines. SAP Commerce Cloud uses Apache Solr for product and content search. One of the goals of this article is to give recommendations on how to configure Solr properly for Chinese and Japanese search).

could also be erroneously extracted. Consequently, a search for “国外” should most likely not match the string “中华人民共和国外交部” but a query for “外交部” should.

A group of characters might be segmented in different ways resulting in different meanings. The Chinese sentence 我喜欢新西兰花 can be interpreted in two ways, as shown in Table 5.

Table 5: Example of Word Segmentation Ambiguity.

Sentence: 我喜欢新西兰花				
Interpretation 1	我	喜欢	新西兰	花
Meaning	I	like	New Zealand	flower
Interpretation 2	我	喜欢	新	西兰花
Meaning	I	like	new	broccoli

Note: This ambiguity typically occurs in spoken language. A writer would use the particle 的 to clarify: 我喜欢新的西兰花 for interpretation 1, and 我喜欢新西兰的花 for interpretation 2.

The challenge is how to extract the meaningful units of knowledge from the text for indexing to return better results at the query phase. There are three approaches:

- **Unigrams**: treat individual Chinese characters as tokens.
- **Bigrams**: treat overlapping groups of two adjacent Chinese characters as tokens.
- **By part of speech or meaningful words**: performs word segmentation and indexes word units as tokens.

Table 6: Tokenization Approaches for “我是中国人”(“I’ m a Chinese”).

	Unigrams	**Bigrams**	**Word segmentation**
Token 1	我	我是	我 (“I”)
Token 2	是	是中	是 (“right”)
Token 3	中	中国	中国 (“China”)
Token 4	国	国人	人 (“man”)
Token 5	人		

6.1 Chinese Tokenizers for Apache Solr

For Chinese, Apache Solr supports various analyzers:

- **Standard Analyzer**: based on unigram indexing.
- **ChineseAnalyzer**: indexes unigrams.
- **CJKAnalyzer**: indexes bigrams.
- **SmartChineseAnalyzer**: indexes words based on a dictionary and heuristics for Simplified Chinese.
- **HanLPTokenizer**: A powerful tokenizer supporting multiple algorithms.
- **Paoding**: An older tokenizer that may have issues with recent Solr versions.

6.1.1 CJKAnalyzer

This analyzer has a simple bigram tokenizer. This is the fastest option, but the search recall will be the worst. Bigramming doesn't require any linguistic resources such as dictionaries or statistical tables. Every overlapping two-character sequence is placed into the index. There is a common practice is to index Chinese texts simultaneously as words and as overlapping bigrams, combining the methods in a weighted fashion to improve accuracy.

SF	text	我喜	喜欢	欢新	新西	西兰	兰花
	raw_bytes	[e6 88 91 e5 96 9c]	[e5 96 9c e6 ac a2]	[e6 ac a2 e6 96 b0]	[e6 96 b0 e8 a5 bf]	[e8 a5 bf e5 85 b0]	[e5 85 b0 e8 8a b1]
	start	0	1	2	3	4	5
	end	2	3	4	5	6	7
	positionLength	1	1	1	1	1	1
	type	<DOUBLE>	<DOUBLE>	<DOUBLE>	<DOUBLE>	<DOUBLE>	<DOUBLE>
	termFrequency	1	1	1	1	1	1
	position	1	2	3	4	5	6

Figure 2: Example of CJKAnalyzer tokenizing a Chinese sentence into bigrams.

6.1.2 Smart Chinese Analyzer

This analyzer uses *HMMChineseTokenizer* which uses probabilistic knowledge to find the optimal word segmentation for Simplified Chinese text. The text is first broken into sentences, then each sentence is segmented into words based on a Hidden Markov Model. It requires a dictionary to provide statistical data, which is included out-of-box from the ICTCLAS1.0 project. SmartChineseAnalyzer creates four terms (I + like + New Zealand (新西兰) + flower) for our example sentence.

SF	text	我	喜欢	新西兰	花
	raw_bytes	[e6 88 91]	[e5 96 9c e6 ac a2]	[e6 96 b0 e8 a5 bf e5 85 b0]	[e8 8a b1]
	start	0	1	3	6
	end	1	3	6	7
	positionLength	1	1	1	1
	type	word	word	word	word
	termFrequency	1	1	1	1
	keyword	false	false	false	false
	position	1	2	3	4

Figure 3: Example of SmartChineseAnalyzer performing word-based segmentation.

6.1.3 HanLPTokenizer

HanLPTokenizer supports multiple algorithms, with the Viterbi algorithm being the default, which offers a good balance of efficiency and effectiveness. Unlike SmartChineseAnalyzer, it supports Traditional Chinese. For our example, HanLPTokenizer creates six terms (I + like + New Zealand (新西兰) + Zealand(西兰) + flower).

HLPT	text	我	喜欢	新	西兰花	西兰	兰花
	raw_bytes	[e6 88 91]	[e5 96 9c e6 ac a2]	[e6 96 b0]	[e8 a5 bf e5 85 b0 e8 8a b1]	[e8 a5 bf e5 85 b0]	[e5 85 b0 e8 8a b1]
	start	0	1	3	4	4	5
	end	1	3	4	7	6	7
	positionLength	1	1	1	1	1	1
	type	r	v	a	nf	nrf	n
	termFrequency	1	1	1	1	1	1
	position	1	2	3	4	5	6

Figure 4: Example of HanLPTokenizer performing word-based segmentation.

6.2 Japanese Tokenizers for Apache Solr

For Japanese, Solr provides:

- **CJKAnalyzer**: indexes bigrams, same as for Chinese.
- **JapaneseTokenizer (Kuromoji)**: splits the text into word units using morphological analysis, and annotates each term with part-of-speech, base form (lemma), reading, and pronunciation.

6.2.1 JapaneseTokenizer (Kuromoji)

This morphological tokenizer, also known as the Kuromoji Japanese Morphological Analyzer, uses a rolling Viterbi search to find the least cost segmentation (path) of the incoming characters. For our test query “私は日本人です” (“I’m Japanese”), it returns four terms (“I + particle + Japanese + am).

JT	text	私	は	日本人	です
	raw_bytes	[e7 a7 81]	[e3 81 af]	[e6 97 a5 e6 9c ac e4 ba ba]	[e3 81 a7 e3 81 99]
	start	0	1	2	5
	end	1	2	5	7
	positionLength	1	1	1	1
	type	word	word	word	word
	termFrequency	1	1	1	1
	baseForm	名詞-代名詞-一般	助詞-係助詞	名詞-一般	助動詞
	partOfSpeech	noun-pronoun-misc	particle-dependency	noun-common	auxiliary-verb
	partOfSpeech (en)	ワタシ	ハ	ニッポンジン	デス
	reading	watashi	ha	nipponjin	desu
	reading (en)	ワタシ	ワ	ニッポンジン	デス
	pronunciation	watashi	wa	nipponjin	desu
	pronunciation (en)				特殊・デス
	inflectionType	1	2	3	special-desu
	inflectionType (en)				基本形
	inflectionForm				base
	inflectionForm (en)				4
	position				

Figure 5: Kuromoji tokenizing “私は日本人です”.

For a more complex sentence like “韓国に住んでいていい人に聞いた” (I asked a good person, who lives in South Korea), the tokenizer produces a more granular output, but the key point is that the base forms of core words are preserved, maintaining the meaning. Additional tokens like particles can be removed by a stop-words filter.

The tokenizer also supports different modes (*normal*, *search*, and *extended*) to handle compound nouns and provide additional segmentation useful for search. For example, ‘search‘ mode can split compounds like 関西国際空港 (Kansai International Airport) to allow a search for 空港 (airport) to match. It might be good to use search mode for indexing and normal mode for queries to increase precision.

7 Word Normalization

Word normalization refers to the process that maps a word to some canonical form. For example, in English the canonical form for “are”, “is”, and “being” is “be”. This normalization, performed at both index and query time, improves the accuracy of search results. Solr uses two approaches: stemming (reducing a word to its root) and lemmatization (identifying the dictionary form based on context).

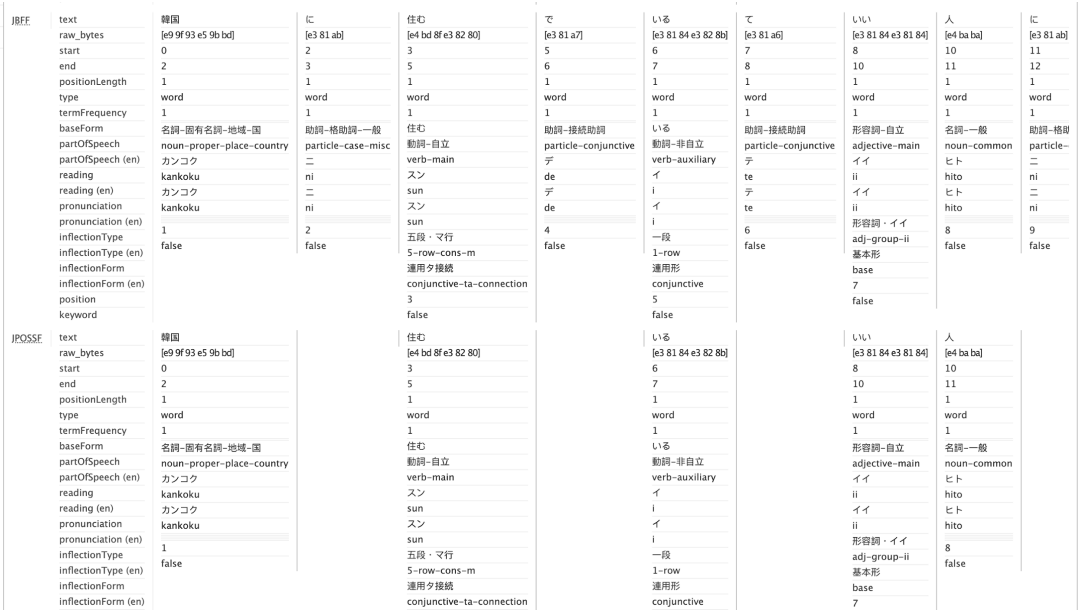

Figure 6: Kuromoji tokenizing a more complex Japanese sentence.

7.1 Solr Filters for Chinese and Japanese

7.1.1 Japanese Iteration Marks

For stemming in Japanese, Solr provides *JapaneseIterationMarkCharFilter* which normalizes horizontal iteration marks (々, odoriji) to their expanded form. These marks are used to represent a duplicated character in the same morpheme, like 人々 (hitobito, "people"), which is normalized to 人人. This is distinct from cases like 日日 (hinichi, "number of days"), where the character is duplicated because it represents different morphemes.

7.1.2 Half-Width/Full-Width Filter

By convention, 1/2 Em wide characters are called “half-width”; the others are called correspondingly “full-width” characters. *CJKWidthFilter* folds full-width ASCII variants into the equivalent basic Latin characters (“Ｉｊ Ｉ” -> “IjI”) and half-width Katakana variants into the equivalent standard Japanese kana (ｶ -> カ).

7.1.3 Japanese Base Form Filter

JapaneseBaseFormFilter reduces inflected Japanese verbs and adjectives to their base/dictionary forms. For example, for the phrase“それをください。”(That one, please.), ください (kudasai) is converted to its base form, くださる (kudasaru).

7.1.4 Japanese Non-meaningful Terms Removal Filter

JapanesePartOfSpeechStopFilterFactory removes tokens with certain part-of-speech tags. For example, the particle “を” (wo), which marks the direct object of a verb, is removed from the token stream.

7.1.5 Japanese Katakana Stemming

JapaneseKatakanaStemFilter normalizes common katakana spelling variations ending in a long sound character (ー) by removing it. Only katakana words longer than four characters are

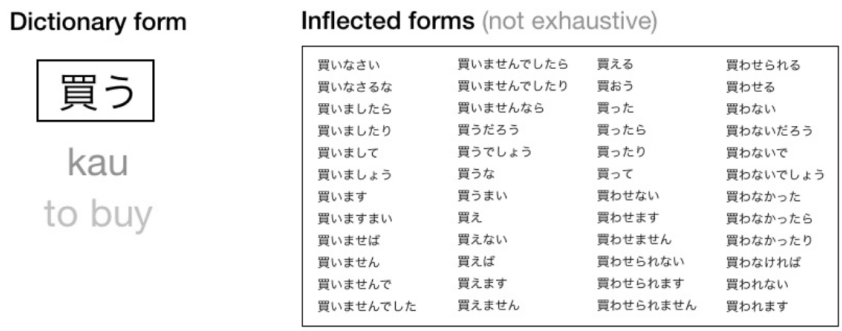

Figure 7: Example of the Japanese Base Form Filter in action.

processed. For example, パーティー (pt, "party") becomes パーティ (pti), but コピー ("copy") remains unchanged as it is too short.

8 Apache Solr Processing Flow Diagrams

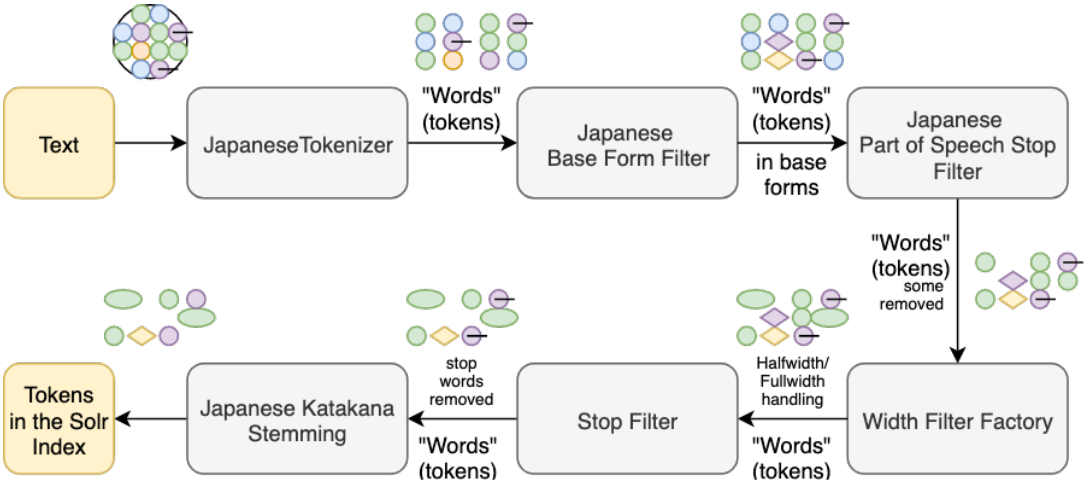

Figure 8: Apache Solr processing flow for Japanese.

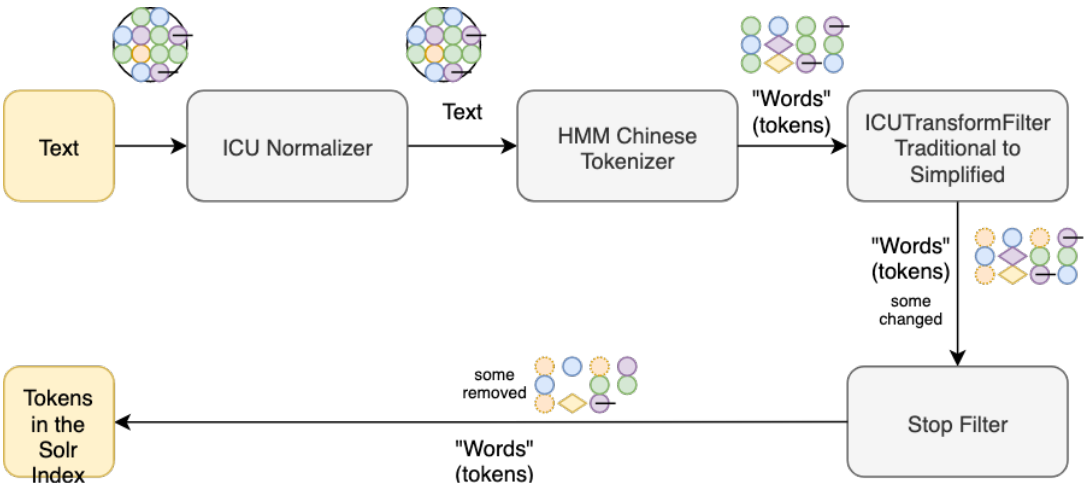

Figure 9: Apache Solr Processing Flow for Chinese.

9 Numerals

In Japan and China, most people and institutions primarily use Arabic numerals. However, Chinese and Japanese numerals are also used, and the system must support them. Japanese numerals are often written using a combination of kanji and Arabic numbers with various kinds of punctuation. For example, ３． ２千 means 3200.

Apache Solr comes with the *JapaneseNumberFilter*, which normalizes various Japanese number formats to regular Arabic decimal numbers. This allows a search for "3200" to match " ３． ２千" in the text.

Table 7: Examples of Japanese Number Normalization via JapaneseNumberFilter.

Before	**After**	**Comments**
○○七	7	○ (maru) is the same as numeral 0.
三千 2 百２十三	3223	
３． ２千	3200	千 means 1000. “．” is a double-byte point.
4,647.100	4647.1	Commas are ignored (removed).
2,500 万	25000000	万 means 10,000.

This filter may in some cases normalize tokens that are not numbers. For example, 田中京一 is a name, but out of context 京一 can also represent a large number. Formal numbers (daiji) and decimal fractions are currently not supported by the filter.

10 Synonyms and Homophones

10.1 Synonyms

Like other languages, Japanese has multiple words for the same concept. For example, there are many verbs meaning "to kill": 殺す, 殺害する, 暗殺する, etc. Apache Solr supports synonyms, but the dictionary of synonymous words must be user-defined.

10.2 Homophones

Homophones are words that are pronounced the same but differ in writing and meaning. Because of a smaller stock of phonemes in Japanese and Chinese, the number of homophones is very large. For many homophones, a universally-accepted orthography does not exist, and the choice of character is often governed by the personal preferences of the writer. A semantically classified database of homophones is needed to implement effective cross-homophone searching. The table below shows the many ways to write the verb *sasu*, depending on meaning.

11 Search by Pronunciation

In Japanese, the pronunciation is directly mapped to the written words. Users often type words in hiragana (phonetic script) and then convert them to kanji or katakana. A robust search system should be able to handle queries typed phonetically. For example, a search for “とうきょうえ” (tkye) should suggest or match “東京駅” (tkyeki, Tokyo Station), as their pronunciations are nearly identical.

Table 8: Orthographic Variants for the Japanese verb *sasu*.

English	Standard	Sometimes	Often also
to offer	差す	さす	
to hold up	差す	さす	
to pour into	差す	注す	さす
to color	差す	注す	さす
to shine on	差す	射す	さす
to aim at	指す	差す	
to point to	指す	さす	
to stab	刺す	さす	
to leave unfinished	さす	止す	

Figure 10: Google Suggest handling a phonetic Japanese query.

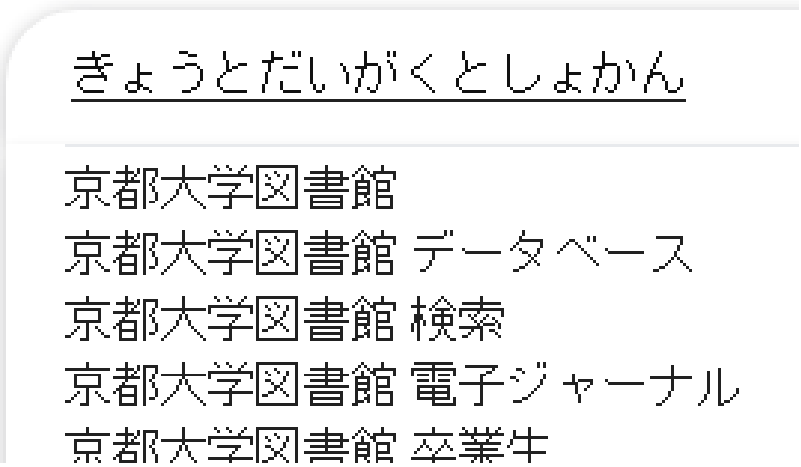

Figure 11: Search results for Kyoto University Library, likely initiated by a phonetic search.

12 Punctuation Marks

There are punctuation marks specific for Japanese and Chinese, some of which have similar-looking but non-interchangeable equivalents in European languages.

Table 9: Specific CJK Punctuation Marks.

Mark	Example	Explanation
〜	1〜2	Wavy dash, used for ranges.
。		Full stop (equivalent to ".").
、	a、b、c	Enumeration comma.
「」	「あいうえお」	Japanese equivalent of quotation marks.
・	ジョン・ドゥ	Nakaten, used to separate parts of foreign names.

13 Search UI Observations

Chinese and Japanese websites often have much less negative space, smaller images, and a higher density of information compared to Western sites. This layout style may be connected to a tendency for content efficiency, placing a maximum amount of content within a minimum space.

13.1 Input Methods

13.1.1 Chinese Input

Chinese text can be input in various ways, including Pinyin (Latin transcription), Wubi (stroke-based), handwriting, and voice recognition. Each method has its own user base and trade-offs.

- **Wubi** is the fastest but most challenging method, requiring memorization of a stroke-to-key mapping.
- **Pinyin** is slower due to the need to select from a list of homonyms, but is common among the younger generation and foreigners.
- **Handwriting** and **Stroke sequence** are used by those who don't know Latin alphabets.
- **Image recognition** is great for larger amounts of data.
- **Voice recognition** is popular but challenged by the numerous dialects.

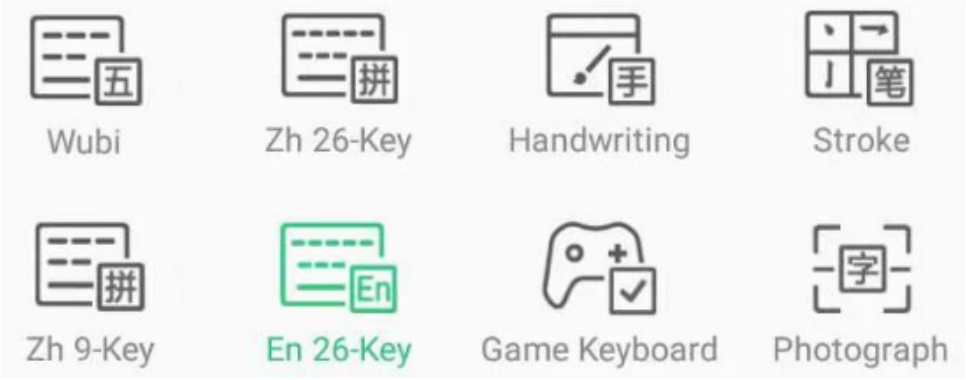

Figure 12: Chinese input methods available on a standard Android system.

13.1.2 Japanese Input

The primary input method is typing words by their reading in kana (phonetically) and then converting them to kanji. Because of many homonyms, a pop-up window often appears with a list of conversion variants. Japanese keyboards have special helper keys for conversion (変換), no conversion (無変換), and switching between kana modes.

Figure 13: A typical Japanese keyboard layout.

13.2 Less Search, More Navigation

It is common for Chinese and Japanese websites to de-emphasize the search field and prioritize navigation. With complex input methods, it is often faster for users to click through categories than to type a search query. Many apps use a "focus page" that appears when the search bar is clicked, offering tags and popular searches to guide the user and reduce the need for manual text entry.

Figure 14: The search UI on Suning.com, showing popular search tags below the search bar.

13.3 Other UI Trends

- **Voice Search:** Voice input for search queries is becoming increasingly common, especially for older audiences who may struggle with typing complex characters.

- **Context-aware Recommendations:** Many websites show recommended queries under the search bar that are based on user behavior and context.
- **Visual Search:** A growing number of e-commerce sites are implementing visual search, allowing users to find products by uploading an image.
- **Facets:** Facet panels are often arranged horizontally on Chinese websites due to the density of the script, while vertical facets remain more common on Japanese sites. All important facets are often expanded by default.

14 Recommendations

14.1 Implementation Questions

If your website is available in different language versions, you need to have answers to the following questions:

- What languages can be used on what language versions? Can I search in Chinese on the English website and vice versa?.
- Can we mix English and Chinese in the same query? This is especially important for brands and proper names (e.g., Sony / ソニー).
- What language variants (e.g., Simplified/Traditional Chinese) are supported?.

14.2 Web Typography

- **Line Length:** The optimal line length is 15-40 characters for desktop and 15-21 for mobile, about half that recommended for English.
- **Fonts:** Use standard classifications like Mincho (serif) and Gothic (sans-serif) for Japanese, and Song/Ming (serif) and Hei (sans-serif) for Chinese. Avoid embedding large custom font files, as the large character sets can significantly slow page loading.
- **Styling:** Do not use italics, as it skews characters and makes them unreadable.
- **Font Size:** Use a minimum font size of 12pt, and consider 16pt for websites targeting older users. It's best to set font size with relative units like "em" or "%" to respect user preferences.

15 Conclusions

The challenges of Japanese and Chinese searching are significant but addressable. Because of the complexities and irregularities of their writing systems, successful implementation requires not only computational linguistic tools like morphological analyzers but also rich lexical databases fine-tuned to specific project goals and content. Both analyzers and databases are constantly improving, and it is important to keep an eye on the latest breakthroughs in information retrieval to continue delivering a better user experience.