CJK Searching and Discovery: Recent Developments and Future Directions

CEAL Technology Forum 2018, Washington, D.C.

Brent Cook | Director of Product Management for Discovery
Kazuaki Maeda | Lead Software Developer, Summon
Introduction

• Summon is a multilingual search engine which supports language-specific search features (*native search features*) for 17 languages including Chinese, Japanese and Korean (CJK)
• Summon clients can access content in multiple languages; native search features are applied according to language of record
Challenges with CJK Searching and Discovery

• Lack of word boundary marks
  • Word boundaries are not marked with white spaces in Chinese and Japanese. In Korean, white spaces often mark larger syntactic units than words. (e.g., noun + postposition)
  • Examples:
    • Chinese: 科学首先指对应于自然领域的知识，经扩展、引用至社会、思维等领域，如社会学。
    • Japanese: 科学という語は文脈に応じて多様な意味をもつが、おおむね以下のような意味で用いられている。
    • Korean: 과학은 사물의 구조·성질·법칙 등을 관찰 가능한 방법으로 얻어진 체계적·이론적인 지식의 체계를 말한다.
Challenges with CJK Searching and Discovery (cont’d)

• Character variations – Same words can be spelled using different characters
  • Traditional vs. simplified Chinese characters
    • 科学 (“science” simplified) vs. 科學 (“science” traditional)
  • Chinese vs. Japanese variations
    • 經 (Chinese traditional) vs. 经 (Chinese simplified) vs. 経 (Japanese)
  • Other character variations
    • 研 (U+7814) vs. 研 (U+784F)
    • 拼 (U+62FC) vs. 拚 (U+62DA)
    • 郎 (U+90DE) vs. 郎 (U+90CE)
Challenges with CJK Searching and Discovery (cont’d)

• Other writing variations (spelling of foreign loan words, synonyms, etc.)
  • Japanese: コンピューター vs. コンピュータ (“computer”)
  • Japanese: ヴァージニア vs. バージニア (“Virginia”)
  • Chinese: 公交车 (“bus” – Mainland China) vs. 巴士 (“bus” – Singapore, Taiwan, Hong Kong, Macau)

• Use of mixed scripts and alternate scripts
  • Japanese mixed scripts: Kanji(漢字) + Katakana(カタカナ) + Hiragana(ひらがな)
  • Korean mixed scripts: Hanja(漢字) + Hangul(한글)
  • Romanization: Pinyin (Chinese), Romaji (Japanese)
Technology Advances in Recent (Past 10-15) Years

• Improved Unicode support by operating systems and web browsers
  • No/little need to deal with character encoding issues
• Availability of powerful open source search platforms, such as Solr and Elastic Search
  • Allows adding our own new features and improvements as plugins
• Advances in Natural Language Processing (NLP) research and availability of open source software
  • Chinese Word Segmentation
  • Japanese and Korean Morphological Analysis
  • And more
Relevance: “how well a retrieved document or set of documents meets the information need of the user. “

(W) Within-language) relevance
- Improve discoverability (recall) of items
- Maintain high precision, especially among top ranked results

Across-language relevance
- Example: searching for “心理学” could match Chinese, Japanese or Korean documents -> we need to ensure results from languages other than user’s primary language do not dominate top results
CJK Word Segmentation and Morphological Analysis

• Chinese:
  • Example: 梵文基础读本 -> 梵文 ("Sanskrit") + 基础 ("basic") + 读本 ("reader")

• Japanese:
  • Example: サクラを歌った -> サクラ (noun: "Sakura") + を (postposition) + 歌う (verb: "sing")

• Korean:
  • Example: 시애틀에서 -> 시애틀 ("Seattle") + 에서 ("from")

• Alternate approach
  • Tokenize every character (unigram tokenization)
  • Example: 梵文基础读本 -> 梵 + 文 + 基 + 础 + 读
  • Better for recall, but not for precision
  • In Summon, we use this to complement word segmentation for certain important fields such as Title
CJK Phrase Matching vs Non-phrase Matching

• Example: $q = \text{梵文 读本} \ (\text{梵文 “Sanskrit” + 读本 “reader”})$
  • Option 1: return only exact phrase matches to “梵文 读本”
  • Option 2: return matches to both “梵文 读本” and non-phrase matches such as “梵文 基础 读本” (梵文 “Sanskrit” + 基础 “basic” + 读本 “reader”)
  • Option 3: same as Option 2, but boost relevance ranking/scores of exact phrase matches over non-phrase matches => best approach
CJK Character Normalization

• Chinese traditional vs. simplified characters
  • By normalizing characters at index time and query time, we can make them cross-searchable
  • Example:
    • Chinese: 科學 (“science” traditional) -> 科学 (simplified)
    • Chinese: 經濟 (“economy” traditional) -> 经济 (simplified)

• Other character variations:
  • Examples:
    • Japanese: 经济 (Chinese simplified) -> 経済 (Japanese)
    • Japanese: 横濱 (Japanese archaic) -> 横浜 (Japanese modern)
    • Japanese: アイウエオ (half-width) -> アイウエオ (full-width)
Searching by Alternate Scripts of CJK

• Use of a morphological analyzer or dictionary-base converter allows supporting searching by alternate scripts
• Chinese: searching for Hanzi using Pinyin
  • Example: q=beijing daxue ("Peking University" Pinyin) => matches “北京大学” ("Peking University" Hanzi)
• Japanese: searching for Kanji or Katakana using Hiragana
  • Example: q=けいざいがく ("economics" Hiragana) => matches “経済学” ("economics" Kanji)
• Korean: searching for Hanja using Hangul
  • Example: q=해외 ("overseas" Hangul) => matches “海外” ("overseas" Hanja)
Verbatim Match Boosting for Better Precision

• Approaches such as character normalization and synonym mappings expend search results so that more items are discovered.
• However, this could cause non-relevant results being returned.
• Solution: “Verbatim Match Boosting” is an approach which boosts rankings/scores of items where the matching between the query and indexed string are verbatim.
  • Example: If query=科學 (“science” traditional), a match with “科學” gets a higher score than a match with “科学” (“science” simplified) if everything else is equal.
Search Suggestions for CJK – Did You Mean …?

• CJK supporting Search Suggestion feature based on language-specific dictionaries
• Example: q=山山大学 (non-existent/misspelled university name)
  • From Chinese UI: Summon suggests “山东大学” (“Shandong University” – existing university in China) using Chinese dictionary
  • From Japanese UI: Summon suggests “山形大学” (“Yamagata University” – existing university in Japan) using Japanese dictionary
More Advanced CJK Discovery Features

• Topic/subject exploration
Future Directions – from Multilingual Search to Multilingual Discovery

• Feedback from native-speakers is crucial in improving CJK search features and relevance
• Keeping abreast of latest technologies in Natural Language Processing (Information Extraction, Machine Translation, etc.) technologies for CJK languages
• Ideas for more innovative discovery features for all languages including CJK – e.g., expand metadata using NLP technologies, more advanced topic/subject/concept exploration features, query expansions, query suggestions, etc.
THANK YOU
Brent.Cook@exlibrisgroup.com
Kazuaki.Maeda@exlibrisgroup.com