Internationalizing Quantitative Text Analysis

27 June 2018

Kohei Watanabe (WIAS/LSE)
Oul Han (GESIS/SNU)
Visitor to docs.quanteda.io

Quanteda’s users are concentrated in the US and Europe
Indian and Japanese visitors

However, we are getting more Asian users.
Keys to internationalizing text analysis

They are mainly about tools but not necessarily.

- Tools (software packages, morphological analysis tools)
- Data (corpus, dictionary)
- Skills (knowledge of statistics and programing)
- Literature (text books, papers)
Tools
Text analysis tools

There aren’t a lot of text analysis packages that can handle Asian languages:

- **KH Coder (MySQL + Java + R + perl + Mecab)**
  - Officially supports Japanese, Chinese, Russian, Korean
  - Widely used in Japan but not in the overseas

- **Quanteda (R)**
  - Supports all languages in International Component for Unicode (ICU)
  - Use in the US and Europe, but less so in Asia

- **Homegrown tools (Python + morphological analysis tool)**
  - Python is not so good with Unicode
  - Usually heavily system dependent
<table>
<thead>
<tr>
<th>表層形</th>
<th>品詞</th>
<th>品詞細分類1</th>
<th>品詞細分類2</th>
<th>品詞細分類3</th>
<th>活用型</th>
<th>活用形</th>
<th>原形</th>
<th>読み</th>
<th>発音</th>
</tr>
</thead>
<tbody>
<tr>
<td>○</td>
<td>記号</td>
<td>一般</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>宮本</td>
<td>名詞</td>
<td>固有名詞</td>
<td>人名</td>
<td>姓</td>
<td>*</td>
<td>*</td>
<td>宮本</td>
<td>ミヤモト</td>
<td>ミヤモト</td>
</tr>
<tr>
<td>(</td>
<td>記号</td>
<td>括弧開</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>(</td>
<td>(</td>
<td>(</td>
</tr>
<tr>
<td>徹</td>
<td>名詞</td>
<td>固有名詞</td>
<td>人名</td>
<td>名</td>
<td>*</td>
<td>*</td>
<td>徹</td>
<td>トオル</td>
<td>トール</td>
</tr>
<tr>
<td>)</td>
<td>記号</td>
<td>括弧閉</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>)</td>
<td>)</td>
<td>)</td>
</tr>
<tr>
<td>委員</td>
<td>名詞</td>
<td>一般</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>委員</td>
<td>イイイン</td>
<td>イイイン</td>
</tr>
<tr>
<td>記号</td>
<td>空白</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>内容</td>
<td>名詞</td>
<td>一般</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>内容</td>
<td>ナイヨウ</td>
<td>ナイヨー</td>
</tr>
<tr>
<td>について</td>
<td>助詞</td>
<td>格助詞</td>
<td>連語</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>について</td>
<td>ニツイテ</td>
<td>ニツイテ</td>
</tr>
<tr>
<td>差し控え</td>
<td>動詞</td>
<td>自立</td>
<td>*</td>
<td>*</td>
<td>一段</td>
<td>末然形</td>
<td>差し控える</td>
<td>サシハカエ</td>
<td>サシハカエ</td>
</tr>
<tr>
<td>させ</td>
<td>動詞</td>
<td>接尾</td>
<td>*</td>
<td>*</td>
<td>一段</td>
<td>連用形</td>
<td>させる</td>
<td>サセ</td>
<td>サセ</td>
</tr>
<tr>
<td>て</td>
<td>助詞</td>
<td>接続助詞</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>て</td>
<td>テ</td>
<td>テ</td>
</tr>
<tr>
<td>じゃ</td>
<td>助詞</td>
<td>副助詞</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>じゃ</td>
<td>ジャ</td>
<td>ジャ</td>
</tr>
<tr>
<td>なく</td>
<td>助動詞</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>特殊・ナイ</td>
<td>連用＋接続</td>
<td>ない</td>
<td>ナク</td>
<td>ナク</td>
</tr>
<tr>
<td>て</td>
<td>助詞</td>
<td>接続助詞</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>て</td>
<td>テ</td>
<td>テ</td>
</tr>
</tbody>
</table>
Morphological analysis tools (1)

Morphological analysis tools are used in tokenization of Japanese.

- **Mecab (Japanese, Korean and Chinese)**
  - RMecab (not on CRAN)
    - Documentation is only available in Japanese
  - RcppMecab (on CRAN)
    - Internationalization of RMecabKo
    - Supports both Korean, Japanese and Chinese
    - Documented in English
  - mecab-python
    - No longer maintained

- **Rakuten MA (Japanese and Chinese)**
  - javascript library
Morphological analysis tools (2)

Morphological analysis tools are used in tokenization of Chinese and Korean.

- HanNanum, Kkma, Komoran, Twitter Korean Text, ... (Korean)
  - Java libraries, different performances, choose according to aim
  - Available all-in-one via KoNLP (on CRAN), KoNLPy (Python)
- jieba (Chinese)
  - Python module
  - jiebaR (on CRAN)
Challenges in tokenizing Korean

- We can “stem” Japanese but not Korean:

<table>
<thead>
<tr>
<th>하다</th>
<th>行う</th>
<th>do</th>
</tr>
</thead>
<tbody>
<tr>
<td>하였다</td>
<td>行った</td>
<td>did</td>
</tr>
<tr>
<td>했다</td>
<td>行った</td>
<td>did (shorter form)</td>
</tr>
<tr>
<td>했었다</td>
<td>行った</td>
<td>had done</td>
</tr>
</tbody>
</table>

- Two possible approaches:
  - Apply morphological analysis tools to identify lemma
  - Tokenize only nouns using noun extraction tools (HanNanum)
Part-of-speech (POS) tagging

<table>
<thead>
<tr>
<th>doc_id</th>
<th>sentence_id</th>
<th>token_id</th>
<th>token</th>
<th>lemma</th>
<th>pos</th>
<th>entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>text1</td>
<td>1</td>
<td>1</td>
<td>When</td>
<td>when</td>
<td>ADV</td>
<td></td>
</tr>
<tr>
<td>text1</td>
<td>1</td>
<td>2</td>
<td>I</td>
<td>-PRON-</td>
<td>PRON</td>
<td></td>
</tr>
<tr>
<td>text1</td>
<td>1</td>
<td>3</td>
<td>presented</td>
<td>present</td>
<td>VERB</td>
<td></td>
</tr>
<tr>
<td>text1</td>
<td>1</td>
<td>4</td>
<td>the</td>
<td>the</td>
<td>DET</td>
<td></td>
</tr>
<tr>
<td>text1</td>
<td>1</td>
<td>5</td>
<td>supplementary</td>
<td>supplementary</td>
<td>ADJ</td>
<td></td>
</tr>
<tr>
<td>text1</td>
<td>1</td>
<td>6</td>
<td>budget</td>
<td>budget</td>
<td>NOUN</td>
<td></td>
</tr>
<tr>
<td>text1</td>
<td>1</td>
<td>7</td>
<td>to</td>
<td>to</td>
<td>ADP</td>
<td></td>
</tr>
<tr>
<td>text1</td>
<td>1</td>
<td>8</td>
<td>this</td>
<td>this</td>
<td>DET</td>
<td></td>
</tr>
<tr>
<td>text1</td>
<td>1</td>
<td>9</td>
<td>House</td>
<td>house</td>
<td>PROPN</td>
<td>ORG_B</td>
</tr>
<tr>
<td>text1</td>
<td>1</td>
<td>10</td>
<td>last</td>
<td>last</td>
<td>ADJ</td>
<td>DATE_B</td>
</tr>
<tr>
<td>text1</td>
<td>1</td>
<td>11</td>
<td>April</td>
<td>april</td>
<td>PROPN</td>
<td>DATE_I</td>
</tr>
<tr>
<td>text1</td>
<td>1</td>
<td>12</td>
<td>,</td>
<td>,</td>
<td>PUNCT</td>
<td></td>
</tr>
<tr>
<td>text1</td>
<td>1</td>
<td>13</td>
<td>I</td>
<td>-PRON-</td>
<td>PRON</td>
<td></td>
</tr>
<tr>
<td>text1</td>
<td>1</td>
<td>14</td>
<td>said</td>
<td>say</td>
<td>VERB</td>
<td></td>
</tr>
<tr>
<td>text1</td>
<td>1</td>
<td>15</td>
<td>we</td>
<td>-PRON-</td>
<td>PRON</td>
<td></td>
</tr>
<tr>
<td>text1</td>
<td>1</td>
<td>16</td>
<td>could</td>
<td>could</td>
<td>VERB</td>
<td></td>
</tr>
</tbody>
</table>
Application of POS tools

Part-of-speech tagging tools can be adopted for Asian languages, but usually not as good as specialist tools

- **StanfordNLP** (Chinese)
  - Java tool for part-of-speech tagging
- **Spacy** (Chinese, Japanese, Thai, Vietnamese etc.)
  - Python tool for part-of-speech tagging
  - Currently only supports tokenization
- **Universal Dependencies** (60 languages including CJK)
  - udpipe (on CRAN)
Dictionary-based tokenization

Use of external tools is not necessary

- **International Component for Unicode (ICU)**
  - ICU is the IT infrastructure
    - Adobe, Amazon, Apple, DELL, Google, IBM etc.
    - Also part of Python and Java libraries
  - ICU defines word boundaries using dictionary
    - Dictionary for Chinese and Japanese are developed based on existing lexicon
      - IPA for Japanese
      - Libtabe for Chinese
    - However, dictionary is missing for Korea
  - The string package provides interface to ICU in R
Tokenization by ICU

Japanese segmentation without morphological analysis

```r
taxt_jp <- "政治とは社会に対して全体的な影響を及ぼし、社会で生きるひとりひとりの人の人
生にも様々な影響を及ぼす複雑な領域である。"
quanteda::tokens(txt_jp)
tokens from 1 document.
text1:
[1] "政治" "と" "は" "社会" "に対して"
[6] "全体" "的" "な" "影響" "を"
[11] "及" "ば" "し" "、" "社会"
[16] "で" "生きる" "ひとりひとり" "の" "人"
[21] "の" "人生" "に" "も" "様々"
[26] "な" "影響" "を" "及ぼす" "複雑"
[31] "な" "領域" "で" "ある" "."
```
Tokenization by ICU

Chinese segmentation without morphological analysis

```
> txt_cn <- "政治是各种团提进行集体决策的一个过程，也是各种团提或个人为了各自的领域所结成的特定关系，尤指对于某一政治实体的统治，例如统治一个国家，亦指对于一国内外事务之监督与管制。"
> quanteda::tokens(txt_cn)
tokens from 1 document.
```

<table>
<thead>
<tr>
<th>Text</th>
<th>Token</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;政治&quot;</td>
<td>&quot;政治&quot;</td>
</tr>
<tr>
<td>&quot;的&quot;</td>
<td>&quot;的&quot;</td>
</tr>
<tr>
<td>&quot;各种&quot;</td>
<td>&quot;各种&quot;</td>
</tr>
<tr>
<td>&quot;团&quot;</td>
<td>&quot;团&quot;</td>
</tr>
<tr>
<td>&quot;体&quot;</td>
<td>&quot;体&quot;</td>
</tr>
<tr>
<td>&quot;是&quot;</td>
<td>&quot;是&quot;</td>
</tr>
<tr>
<td>&quot;过程&quot;</td>
<td>&quot;过程&quot;</td>
</tr>
<tr>
<td>&quot;各&quot;</td>
<td>&quot;各&quot;</td>
</tr>
<tr>
<td>&quot;自&quot;</td>
<td>&quot;自&quot;</td>
</tr>
<tr>
<td>&quot;的&quot;</td>
<td>&quot;的&quot;</td>
</tr>
<tr>
<td>&quot;也是&quot;</td>
<td>&quot;也是&quot;</td>
</tr>
<tr>
<td>&quot;领域&quot;</td>
<td>&quot;领域&quot;</td>
</tr>
<tr>
<td>&quot;所&quot;</td>
<td>&quot;所&quot;</td>
</tr>
<tr>
<td>&quot;成&quot;</td>
<td>&quot;成&quot;</td>
</tr>
<tr>
<td>&quot;对于&quot;</td>
<td>&quot;对于&quot;</td>
</tr>
<tr>
<td>&quot;一&quot;</td>
<td>&quot;一&quot;</td>
</tr>
<tr>
<td>&quot;国家&quot;</td>
<td>&quot;国家&quot;</td>
</tr>
<tr>
<td>&quot;之&quot;</td>
<td>&quot;之&quot;</td>
</tr>
<tr>
<td>&quot;监督&quot;</td>
<td>&quot;监督&quot;</td>
</tr>
<tr>
<td>&quot;与&quot;</td>
<td>&quot;与&quot;</td>
</tr>
<tr>
<td>&quot;管制&quot;</td>
<td>&quot;管制&quot;</td>
</tr>
<tr>
<td>&quot;进行&quot;</td>
<td>&quot;进行&quot;</td>
</tr>
<tr>
<td>&quot;集体&quot;</td>
<td>&quot;集体&quot;</td>
</tr>
<tr>
<td>&quot;决策&quot;</td>
<td>&quot;决策&quot;</td>
</tr>
<tr>
<td>&quot;团&quot;</td>
<td>&quot;团&quot;</td>
</tr>
<tr>
<td>&quot;体&quot;</td>
<td>&quot;体&quot;</td>
</tr>
<tr>
<td>&quot;的&quot;</td>
<td>&quot;的&quot;</td>
</tr>
<tr>
<td>&quot;结成&quot;</td>
<td>&quot;结成&quot;</td>
</tr>
<tr>
<td>&quot;对于&quot;</td>
<td>&quot;对于&quot;</td>
</tr>
<tr>
<td>&quot;国内&quot;</td>
<td>&quot;国内&quot;</td>
</tr>
<tr>
<td>&quot;外&quot;</td>
<td>&quot;外&quot;</td>
</tr>
<tr>
<td>&quot;事务&quot;</td>
<td>&quot;事务&quot;</td>
</tr>
</tbody>
</table>
Data
Sources of textual data (Japanese)

- **Politics**
    - Accessible via the kaigiroku package ([https://github.com/amatsuo/kaigiroku/](https://github.com/amatsuo/kaigiroku/))
    - No API to download raw text yet

- **Online forum**
  - Yahoo Japan News

- **Mass media**
  - Asahi Shimbun (Kikuzo database)
  - Yomiuri Shimbun (Yomidas database)
  - Full newspaper corpus ([http://www.nichigai.co.jp/sales/corpus.html](http://www.nichigai.co.jp/sales/corpus.html))

- **More...**
Sources of textual data (Korean)

- Politics

- Online forum
  - Naver News

- Mass media
  - KINDS ([https://www.kinds.or.kr/](https://www.kinds.or.kr/))
  - Dow Jones Factiva

- More...
Skills
Necessary skills and knowledge

- **Data collection**
  - API
    - Knowledge of machine readable formats (XML, JSON etc.)
  - Scraping
    - Advanced skill in programming (R or Python)
    - Knowledge of HTML, JavaScript and Selenium

- **Analysis**
  - Basic knowledge of
    - Descriptive and inferential statistics (chi-square, t-test, regression analysis)
    - Network analysis
  - Basic skill in programming R or Python
    - However, KH Coder does not require programming skill

- **Research design**
  - Broad knowledge of social scientific text analysis is essential
Literature
Textbook

Many people need a textbook to teach themselves, but there aren’t any.

- There is no good textbook on the application of text analysis
  - Textbooks on text analysis are usually about computer programming
    - Ken Benoit should be writing a social scientific textbook
  - Textbooks in computer science are about complex algorithms

- CJK languages require special handling
  - Language specific textbooks are needed for the analysis of CJK languages
    - We have created a section for Japanese
      (https://tutorials.quanteda.io/language-specific/japanese/)
    - Add pages for Korean and Chinese
More papers using text analysis should be available, but there are obstacles:

- Papers on CJK texts might be difficult to publish in English journals
  - Weak incentives to use text analysis to study Asian countries
- Asian journals might be more conservative about methodology
  - Methodologically innovative papers are more difficult to publish
  - Comparative analysis in Asian languages and English is necessary
- There is no international and interdisciplinary venue for text analysis
  - Kyushu’s Multidisciplinary Text-mining Colloquium is too domestic
  - Asia PolMeth is good, but too narrowly focuses on political science
Summary (1)

● Tools
  ○ Tokenization of CJK text is becoming easier but not without problems
    ■ Cross-lingual morphological analysis tools (e.g. rcppMecab)
    ■ Tokenization without morphological analysis (e.g. ICU)
  ○ Requires international effort and demand stimulation to improve support of CJK languages
  ○ For the moment, there are pragmatic workarounds (e.g. noun stemming in Korean)

● Data
  ○ Political texts are becoming increasingly accessible in Asia
    ■ Comprehensive list of sources is needed
  ○ Media text is still difficult to collect
    ■ Scrapers is required for large scale data collection
Summary (2)

● Skill
  ○ Advanced quantitative text analysis requires programming skills
    ■ Waseda University’s Global Education Center will offer a statistics course using R from the next term
    ■ However, there is no course about programming itself
  ○

● Literature
  ○ Needs more opportunities and incentives to publish analysis of CJK languages
Conclusions

Actions to promote text analysis in CJK languages:

1. Develop more educational materials for CJK languages
   ○ Develop techniques to make Asian and European languages comparable
   ○ Translate materials into local languages

2. Improve accessibility to tools and datasets
   ○ Create a centralized list of tools (e.g. scrapers) and datasets

3. Expose students and researchers to the latest projects
   ○ Invite leading figures from overseas for research seminars
   ○ Fund participation to text conferences (Text-as-data, Manifesto Corpus, PolText etc.)

4. Establish interdisciplinary venue or medium for CJK text analysis
   ○ International and multidisciplinary events on text analysis should be organized
   ○ Launch an open-access journal on Asian text analysis