The Company They Keep: Extracting Japanese Neologisms Using Language Patterns

James Breen
University of Melbourne, Australia
jimbreen@gmail.com

Timothy Baldwin
University of Melbourne, Australia
tb@ldwin.net

Francis Bond
Nanyang Technological University, Singapore
bond@ieee.org

December 24, 2017

Searching for Japanese Neologisms

- Part of a major project to develop and test techniques for extracting neologisms from Japanese text.
- The challenge with Japanese is the lack of word boundaries:
  - Japanese segmenters use large lexicons of known terms
  - unknown words are usually treated as unsructured sequences of basic morphemes
- Current investigation involves:
  - searching for terms that are highlighted as noteworthy by adjacent text
  - extracting those not currently in lexicons.

Quick Overview of Japanese Orthography

- Japanese is written in a mixture of scripts:
  - *kanji* (Chinese characters), e.g. 猫, 犬, 鳥, 牛, etc., used mainly for nouns and roots of verbs, adjectives, etc. Approx. 2,500 in common use.
  - most nouns in *kanji* use 2 or more characters, verbs and adjectives typically use one *kanji* for the non-inflecting root part
  - the *hiragana* syllabary (46 symbols plus diacritics: あいうえお かきくけこ, etc.), used mainly for particles, inflections, conjunctives, etc.
    - Texts aimed at children are initially only in *hiragana*, sometimes with spacing between terms
  - the *katakana* syllabary (アイウエオカキクケコ, etc.) used for loanwords, foreign names, scientific names, etc.)
  - Latin alphabetics - in text mainly used for initials, acronyms, etc (*USB*, *bps*, etc.) or product names (*iPhone*, *Windows*, etc.)
- E.g.: スーパーで食品を買いました。

Finding Neologism using Language Patterns

- Project genesis is common use in Japanese of phrases highlighting terms, e.g.
  - ⟨term⟩ というのは *to iu no wa* “as for that which is said ⟨term⟩”
  - ⟨term⟩ とは *to wa* “as for ⟨term⟩”
- Translators will often Google for ⟨term⟩ とは when encountering an unknown term.
- Aim to identify and test a repertoire of such patterns/phrases

Identification of Language Patterns

- Finding possible patterns
  - sampled WWW passages containing new terms recently added to dictionaries
    - no useful patterns
    - showed terms often used within parentheses e.g.「...」, "...", etc.
  - sampling using constructs such as という造語 *toiuzōgo* "thus said neologism" and という新語 *toiushingo* "thus said new word" detected several cases of new terms in use
  - search made for terms likely to be used with new words
    - workshopped with native speakers
    - set of 37 phrases developed, e.g.
    — という言葉 *to iu kotoba* "thus said word",
    — xx という不思議な *to iu fushigi na* "the said xx is strange/curious"
    — 最近流行の *saikin ryūkō no* "recent vogue ..."

Initial Tests

- Initially used the Kyoto University WWW Corpus
  - 500M ex-WWW sentences from 2004
- extracted all passages from the Corpus containing the 37 phrases (280,000)
  - examined a sample of 20 of each
  - classified: discussing term or not, new term or known term
- some patterns had high precision, but did not occur very often
- only about 0.06% of passages in the Corpus were collected
- about half the identified terms were parenthesized
- issues using a 12-year-old corpus - many “new” terms discussed were no longer new.
- decided to extend the examination to include 870M Twitter passages from 2014/2015.

Detailed Investigation

- constructed an extraction system using the 18 most productive patterns (97% of useful extractions came from two patterns)
  - used fast tree-based text matching (patterns start with こ, と, 近 and 最)
  - extracted possible terms following or preceding patterns
    - parenthesized terms extracted (10 parenth. types)
    - non-parenthesized terms based on restricted morpheme patterns: noun-noun, adjective-noun, etc.
- 235k terms extracted from WWW corpus - 53% parenthesized - 68k were unrecorded
- 108k terms extracted from Twitter data - 34% parenthesized - 48k were unrecorded
- 76% were single occurrences; up to 55 multiples (WWW)

Evaluation of Samples

- Samples of extracted terms were examined
  - the 50 most commonly recurring
  - a sample of 20 occurring 5 times each
  - a sample of 50 occurring once
- The samples were evaluated and classified:
  - A - known term but a variant form
  - B - known but in an inflected form
  - C - valid and of interest
  - D - valid, but not of interest
  - E - invalid

WWW Corpus Results

Categ.	Top 50	5 Times (20)	Once (50)
A	15	2	0
B	6	6	1
C	18	10	3
D	8	2	46
E	3	0	0

Examples:

- A - ガイジン *gaijin*: *katakana* form of 外人 “foreigner”
- B - 愛している *aishiteiru*: from the verb 愛する and meaning “to be in love”
- C - ゲーム性 *gēmusei* “quality of a video game; game rating”
- C - 共創 *kyōsō* “growing together; joint development”
- D - シンプルイズベスト *shinpuru izu besuto* (“Simple Is Best”: pop song name)

Twitter Results (Not Retweets)

Categ.	Top 50	5 Times (20)	Once (20)
A	5	3	0
B	2	6	1
C	21	5	3
D	20	10	16
E	3	0	0

Category C Examples:

- 放射脳 *hōshanō* “obsession with the effects of radiation”
- クリぼっち *kuribotchi* “spending Christmas alone”
- アホノミクス *ahonimikusu* “Ahonomics” (idiot economics: play on “Abenomics”)
- パイスラ *paisura* woman with a diagonal shoulder strap between her breasts
- アラサーメンズ *arasāmenzu* fashions for men over 30

Twitter Issues

Retweets

- Very common in Twitter (now a UI function)
- Can significantly skew term frequencies, BUT could also signal a useful term
- Often difficult to identify (e.g. added/modified text)
- Analysis showed no particular advantage for terms in retweets

Burstiness

- Twitter metadata allowed detection of time/date of term usage
- No advantage detected for repeated terms in bursts compared with other repeats

Advantages of Multiple Occurrences

- Clear that terms that occurred multiple times were more likely to be useful
- Noted that useful singly-occurring terms usually had high *n*-gram counts
  - selected 2,000 singly-occurring terms and added *n*-gram counts
  - tested samples for usefulness
  - strong correlation between higher counts and usefulness
- Combining the process with an *n*-gram corpus would enhance precision

Pattern-based Term Extraction

- clearly effective for highlighting useful unrecorded terms
    - multiple occurrences a useful signal
    - can be boosted using an n-gram corpus (enhances precision)
    - sorting out re-tweets is a pain
- only lightly skimming texts
    - assessing recall a challenge
      – successful in finding most occurrences
    - interesting future work
- obvious application to monitoring real-time text flows: Twitter, RSS, etc.