Identification of Neologisms in Japanese by Corpus Analysis

James Breen
Monash University, Australia

Abstract

In Japanese and other languages that do not use spaces or other markers between words, the identification and extraction of neologisms and other unrecorded words presents some particular challenges. In this paper we discuss the problems encountered with neologism identification and describe and discuss some of the methods that have been employed to overcome these problems.

Keywords : Japanese neologism kanji hiragana katakana segmentation corpus n-gram

1. Introduction

In "The Oxford Guide to Practical Lexicography" (Atkins and Rundell, 2008) we find the unqualified statement "It's easy for computer programs to spot completely new words". The authors must have been thinking of European languages, where modern orthographical practice has each word separated by spaces. The quoted statement clearly does not apply languages such as Japanese or Chinese where apart from punctuation there is no clear marker between words, and where the very concept of "word" is often debated.

In this paper we describe recent and planned work to extend some techniques reported earlier to identify and extract neologisms from Japanese texts.(Breen 2004a; Breen 2005; Kaji, Uno and Katsuregawa 2009) The purpose of the work is to extend the recorded lexicon of Japanese, both in free and commercial dictionaries.

2. Overview of Japanese Orthography

Modern Japanese is written in a mixture scripts:

a. *kanji* (Chinese characters), which are used mainly for nouns and the roots of verb s, adjectives, etc. Approximately 2,000 *kanji* are in common use, although the full set available is estimated to be around 80,000. Most nouns written with *kanji* use two or more characters, whereas verbs typically use a single *kanji.*

b. the *hiragana* syllabary (46 symbols plus diacritics: あいうえおかきくけこ, etc.) In modern Japanese *hiragana* is used mainly for particles, verb and adjective inflections, conjunctions, etc.
c. the *katakana* syllabary (also 46 symbols plus diacritics: アイウエオカキクケコ, etc.) *Katakana* are currently used for loanwords, scientific names, transcriptions of foreign names, etc.

An illustration of the use of the scripts can be seen in the sentence スーパーで食品を買いました *su-pa- de shokuhin o kaimashita*: [I] bought some food at [a/the] supermarket). Here *kanji* are used for the noun 食品 (foodstuffs) and to root of the verb 買う (*kau* to buy), *hiragana* are used for the particles で and を and the polite past-tense inflection of the verb (いました), and *katakana is used for the abbreviated form of the loanword* スーパーマーケット (*su-pa-ma-ketto* supermarket).

3. Neologisms in Japanese

Despite having a rich lexicon, the Japanese language has a noted tendency to adopt and create new words (Lee 2002; Tsujimura 2006;). While the reasons for adopting new words are varied, there are number of processes associated with the Japanese language which tend to encourage neologism creation:

a. the readiness to accept loanwords. Unlike some countries, which attempt to restrict loanword usage, Japan has placed no formal restriction on their use. Estimates of the number of loanwords used in Japanese range as high as 80,000. Most of these words have been borrowed directly from English, however a significant number, known as *wasei eigo* (Japanese-made English) have been assembled from English words or word fragments.
b. the accepted morphological process of creating words by combining two or more *kanji* (Chinese characters) chosen for their semantic properties. This process was used extensively in the mid-19th century when Japan re-engaged with the rest of the world and needed an expanded lexicon to handle the technological, cultural, etc. information flowing into the country. This process has continued. A broadly similar process is used to create compound verbs.
c. the tendency to create abbreviations, particularly from compound nouns and long loanwords. For example, the formal term for "student discount" in Japanese is *gakusei waribiki* (学生割引), however the common term is *gakuwari* (学割) formed from the first *kanji* in each of the two constituent nouns. A similar process is applied to loanwords, resulting in words such as *sekuhara* (セクハラ) for "sexual harassment" (a contraction of *sekushuaru harasumento).*

Many neologisms find their way eventually into published dictionaries, and there are several special neologism dictionaries *(shingo jiten, gendaiyôgo jiten),* however many abbreviations, compound verbs and loanwords are less well lexicalized as native speakers can usually recognize them as such and recognize the pronunciation and meaning.

Traditional techniques for identifying neologisms involve extracting lexemes and comparing them with a lexical database. This process can have problems in Japanese as the orthography does not use any separators between words. As described below, text segmentation software packages for Japanese typically use extensive lexicons to enable word segments to be identified, but behave unpredictable when out-of-lexicon strings are encountered.

4. Word Segmentation in Japanese

Computerized segmentation of Japanese text was once considered a very difficult task; some writers in the 1980s thought it impossible. Since the 1990s several good systems have emerged, e.g. the open-source research-oriented Juman (Kyoto University) and Chasen and MeCab (Nara Institute of Science and Technology), and commercial and in-house systems from Basis Technology, NTT and Google. All of these combine artificial intelligence techniques with large lexicons (which implies that for correct operation the words must be known already).

As an example of such segmentation software consider the sentence "その教師は講堂に学生を集めた。" (that teacher assembled the students in the auditorium) when processed by the Chasen system. Table 1 shows the results of the segmentation.

Word Segment	**Reading**	**Lexical Form**	**POS Information**
その	ソノ	その	連体詞
教師	キョウシ	教師	名詞-一般
は	ハ	は	助詞-係助詞
講堂	コウドウ	講堂	名詞-一般
に	ニ	に	助詞-格助詞-一般
学生	ガクセイ	学生	名詞-一般
を	ヲ	を	助詞-格助詞-一般

集め	アツメ	集める	動詞-自立 一段 連用形
た	タ	た	助動詞 特殊・タ 基本形
。	。	。	記号-句点

Table 1. Example of CHASEN Text Segmentation

The sentence has been correctly segmented, and the 集めた has been correctly identified as the た (past tense) inflection of 集める.

Such segmentation software usually outputs unassociated strings of characters when words are encountered which are not in their lexicons. In Table 2 we illustrate this substituting some unknown words (全堂 instead of 講堂 and 凩毘 instead of 学生) which results in the following segmentation.

Word Segment	Reading	Lexical Form	POS Information
全	ゼン	全	接頭詞−名詞接続
堂	ドウ	堂	名詞−一般
凩	−	−	未知語
毘	−	−	未知語

Table 2. Example of CHASEN Parsing Unknown Words

The 全堂 has been identified as a prefix-noun combination, which is plausible, but the *kanji* in 凩毘 have been flagged 未知語 (*michigo:* unknown word). The tendency of these software systems to output unassociated strings of characters when words are encountered which are not in their lexicons is well known. Some work has been carried out on reconstructing these "unknown words", but usually in the context of part-of-speech tagging and dependency analysis. (Asahara and Matsumoto 2004; Uchimoto, Sekine and Isahara 2001; Utsuro, Shime, Tsuchiya, Matsuyoshi and Sato 2007)

5. Approaches to Finding New Words in Japanese Texts

Three broad approaches are proposed for identifying neologisms and other unlexicalized Japanese words:

a. scanning texts and other corpora for possible "new" words, typically by processing the texts through segmentation software and dealing with the "out-of-lexicon" problem;
b. mimicking Japanese morphological processes to generate possible words, then testing to for the presence of the "words" in corpora;
c. application of machine learning techniques in which software has been trained to identify the language constructs typically associated with the introduction and discussion of new or rare words.

These approaches are discussed more detail below.

6. Scanning Texts for Neologisms and Unlexicalized Words

The general approach is as follows:

a. process texts through segmentation software to extract lexemes. Ideally the lexicons used by the software should be extended to include as many known words as possible;
b. detect and analyze the cases where the analysis has failed. This will involve considerable post-processing, including careful profiling of any affixes identified, as Japanese is an agglutinative language which makes considerable use of highly productive single-character affixes;
c. extraction of possible unrecorded words;
d. examination of the words in the original textual contexts;
e. development of the reading (i.e. pronunciation) and the meaning of the words.

As reported in (Breen 2005), an initial trial of this was carried out in which 500 articles from the Asahi Shimbun newspaper were analyzed. The process concentrated on isolated unlexicalized *kanji* pairs. A number of hitherto unrecorded words were identified, e.g.

- previously unrecorded names e.g. 武示 (Takeshi), 晃毅 (Kouki), 潔重 (Yukishige);
- newly-arrived terms, e.g. 米紙 (American press/newspapers) and 軍歴 (military service record)
- many abbreviations, e.g. 日歯連 (from 日本歯科医師連盟 - Japan Dentists Federation)

- newspaper-style formations such as 中韓 (Chinese-Korean) and 仏誌 (French publication)
- several apparently new formations such as 入境 (border crossing or border entry) and 公助 (public assistance)

We can draw on the fact that loanwords in Japanese are are written in the *katakana* syllabary, thus enabling relatively straightforward extraction and comparison. The study also harvested unrecorded words written in *katakana.* Approximately 20% of the words in *katakana* were "new", and contained:

- many transcribed names (esp. Chinese and Korean);
- Japanese flora/fauna terms;
- many variants of common loanwords e.g. プロフィル (profile) instead of the more common プロフィール.
- a number of words and expressions worth adding to the lexicon, e.g. ピープルパワー (people-power) and ゼロメートル (zero metre, which in Japanese means sea level).

7. Generation of Possible Words

In this approach we mimic Japanese morphological processes to synthesize potential words, then test if the "word" exists in the lexicon or is in use in corpora.

Early trials used the WWW as a test corpus, with accesses via a programmed interface to a search engine (in this case the Google API.) A new WWW-derived resource for such testing is the Google Japanese Web N-gram Corpus (Kudo and Kazawa 2007). This corpus uses text extracted from a one-month WWW snapshot taken in July 2007. Text strings were processed through MeCab, and all 1-gram to 7-gram sequences occurring more than 20 times were counted and recorded. The resulting n-grams are published as a set of files containing form 2.5M 1-grams to 570M 7-grams (over 1.7M of the 1-grams are *katakana* words or compounds). This corpus has huge potential in corpus linguistics research and will be a very important resource in neologism detection and extraction.

A trial of the technique was carried out using synthesized *kanji* abbreviations based on the above-mentioned 4-*kanji* to 2-*kanji* pattern (e.g. 学生割引 being abbreviated to 学割). (Breen 2004a) Approximately 8,000 4-*kanji* compound verbs were extracted from the JMdict lexicon (Breen 2004b), 2-*kanji* abbreviations formed, and those that were not already lexicalized were tested against WWW pages. As *kanji* pairs can occur in many contexts, the text in which the potential abbreviations appeared was analyzed and classified according to the location of the *kanji* pair, surrounding kana, kanji, punctuation, etc.) and WWW page hits.

Approximately 700 potential abbreviations were identified for deeper analysis, and a large number of abbreviations established.

A further study was carried out using synthesized compound verbs (Breen and Baldwin 2009). (In Japanese compound verbs, formed from two (or more) verbs and acting as a single verb, are very common and highly productive. For example 歌い始める (to start singing) is formed from 歌う (to sing) and 始める (to start or begin). 2,900 compound verbs were selected from the JMdict lexicon, the two verb portions extracted (700 and 600 respectively) and 420,000 potential compound verbs generated. These were tested in the three most common inflections against the Google n-gram corpus, and approximately 22,800 were found to be in use (of these 4,800 were recorded in a range of lexicons.) Samples of the 22,800 were examined in detail, indicating that over 90% precision was being achieved.

8. Direct Scan of the N-gram Corpus

The availability of the Japanese n-gram corpus has opened the possibility of searching it directly for unlexicalized words. For example with 2,000 common *kanji* the possible 2-*kanji* compounds is only 4 million, and it is possible to scan the n-gram corpus for occurrences of such compounds in suitable textual contexts such as *kana-kanji-kanji-kana* sequences.

A direct scan approach was also used as an extension of the compound verb extraction mentioned above. The n-gram corpus was scanned using a symbolic template of a compound verb and the selected candidates filtered for valid inflectional values. Approximately 80,000 possible compound verbs were detected (of which 6,200 were in the range of lexicons), and sampling indicated that a precision of approximately 60% was achieved.

9. Machine Learning

As noted above, the Japanese language has a tendency to adopt and create new words. As a result there is considerable discussion of new words in Japanese newspapers, WWW pages, etc., and there are several WWW sites in Japan devoted to such discussions. Discussion of word meanings associated with neologisms, etc. tend to follow particular linguistic patterns, for example a passage discussing the neologism オタ芸 has
"オタ芸（オタげい・ヲタげい）とは、アイドルや声優などのコンサートや. ..". In this the pronunciation is parenthesized after the word, and followed by the "とは" particle which is typically used to flag an explication of a term. There are a number of such linguistic patterns, and research is under way to train text

classification software to detect documents containing such passages, this enabling a focussed analysis on documents likely to contain neologisms.

10. Derivation of Readings

The pronunciation or reading of an unrecorded word will be a function of the *kanji* with which it is written. (Words written in *hiragana* or *katakana* will have established pronunciations.) There are two issues that need to be dealt with in establishing the pronunciation of such words:

a. unlike Chinese, where each character typically has a single pronunciation, Japanese usually has several pronunciations for each character. Some pronunciations are more common than others and it is possible to generate most probable pronunciations for later testing;
b. a number of character pronunciations are not voiced when occurring at the start of a word, but voiced within a word, e.g. 所 is *tokoro* in initial positions, but usually *dokoro* elsewhere. The rules for this process are complex and not complete, for example 島 (island) is pronounced both *shima* and *jima* in identical contexts.

There is a tendency to write the pronunciation of unusual words in parentheses after its first occurrence in a text. This enables testing of candidate pronunciations by search for a collocation of a word with its possible pronunciation.

11. Derivation of Meanings

This is, of course, traditionally the most intensive and time-consuming part of lexicography. In our corpus-based processes we have been working towards automatic derivation of candidate meanings followed by human checking and verification.

With regard to automatic derivation of meanings, some general observations can be made:

a. for abbreviations the process is relatively straightforward as the abbreviation almost always carries the meaning of source word or expression;
b. for compound verbs there has been considerable success combining semantic and lexical information associated with the component verbs. In this area the English n-gram corpus is also proving useful in identifying most likely candidates;
c. for multi-word expressions it is often possible to get good results by testing combinations of the meanings of constituent words, e.g. 海底電線 →

(undersea, submarine) (electric, telephone, line, cable, wire) leading to "undersea cable" or "submarine cable" as the most likely candidate;

d. loanwords written in *katakana* can be a challenge. While there is some success in back-translation into English, especially when combined with checking for collocations on WWW pages, there is a persistent problem with pseudo-loanwords constructed from foreign words or word-fragments, and from non-English loanwords (Korean, French, German, etc.);
e. compound loanword nouns/expressions can be handled similarly to others. e.g. スパイスライス could be parsed as spice+rice or spy+slice, however checking against English n-grams indicates that former is the correct translation.

12. Conclusion

The nature of Japanese orthography makes neologism detection more difficult than in many other languages.

Modern computational linguistics has techniques and resources to assist in both identification of Japanese neologisms and other unrecorded words, and in deriving readings and meanings. This is a major research area, and a lot more work remains to be done.

References

ASAHARA M and MATSUMOTO Y. (2004) Japanese Unknown Word Identification by Character-based Chunking. COLING 2004, Geneva.

BREEN J. (2004a). Expanding the Lexicon: the Search for Abbreviations. Papillon Multi-lingual Dictionary Project Workshop, Grenoble, 2004.

BREEN J. (2004b). JMdict: a Japanese-Multilingual Dictionary. COLING Multilingual Linguistic Resources Workshop, Geneva, August 2004.

BREEN J. (2005). Expanding the Lexicon: Harvesting Neologisms in Japanese, Papillon Multi-lingual Dictionary Project Workshop, Chiang Rai, Thailand, 2005

BREEN J. and BALDWIN T. (2009). Corpus-based Extraction of Japanese Compound Verbs. Australasian Language Technology Workshop (ALTW2009), Sydney, December 2009.

KAJI N., UNO N. and KITSUREGAWA M. (2009). Mining Neologisms from a Large Diachronic Web Archive for Supporting Linguistic Research. Data Engineering and Information Management (DEIM2009), Tokyo, Japan.(in Japanese)

KUDO T. and KAZAWA H. (2007). Japanese Web N-gram Corpus Version 1, Google/Linguistic Data Consortium, http://www.ldc.upenn.edu/

LEE S.C. (2002). Lexical Neologisms in Japanese. Australian Association for Research in Education Conference, Brisbane, 2002.

NAKAZAWA T., KAWAHARA D. and KUROHASHI S. (2005). Automatic acquisition of basic katakana lexicon from a given corpus. IJCNLP, 2005.

TSUJIMURA N. (2006). An Introduction to Japanese Linguistics. Blackwell, 2nd Edition, 2006.
UCHIMOTO K., SEKINE S. and ISAHARA H. (2001). The unknown word problem: a morphological analysis of Japanese using maximum entropy aided by a dictionary. EMNLP 2001
UCHIYAMA K., BALDWIN T. and ISHIZAKI S. (2005). Disambiguating Japanese Compound Verbs Computer Speech & Language, Volume 19, Issue 4, October 2005, (Special issue on Multiword Expression)
UTSURU T., SHIME T., TSUCHIYA M., MATSUYOSHI S, and SATO S. (2007). Chunking and Dependency Analysis of Japanese Compound Functional Expressions by Machine Learning Text, Speech and Dialogue: 10th International Conference, TSD 2007, Plzen, Czech Republic