Identification of Neologisms in Japanese Corpora using Synthesis

James Breen[1], Timothy Baldwin[2], Francis Bond[3]
[1]Monash University, Melbourne, Australia
[2]University Of Melbourne, Melbourne, Australia
[3]Nanyang Technological University, Singapore

Abstract

We report on the investigation of a neologism detection approach involving the synthesis of possible Japanese words by mimicking Japanese morphological processes, followed by testing for the presence of candidate words in Japanese corpora. A 2-*kanji* compound generation and classification technique resulted in the detection of significant numbers of unrecorded terms.

Keywords: Japanese, term synthesis; *n*-gram corpora, neology

Introduction

This paper reports on part of a major study into the extraction of neologisms from Japanese corpora. In the study three main approaches were explored:

a) analysis of morpheme sequences in Japanese texts to determine the presence of potential new or unrecorded terms. The processes included processing the texts with a morphological analyzer to produce sequences of tagged morphemes, tagging of the morphemes with features derived from combinatory data derived from large lexicons and corpora, and processing the tagged morphemes with rule-based and machine-learning-based chunkers to assemble candidate words and expressions.
b) analysis of language patterns which are often used in Japanese in association with new and emerging terms. These patterns are usually associated with discussions or explanations of new terms. (Breen et al., 2018)
c) synthesis of possible Japanese terms by mimicking Japanese morphological processes, followed by testing for the presence of candidate terms in Japanese corpora.

In this paper we report on the third of these, based on the synthesis of possible Japanese words. (Another component, covering compound verbs, has been reported separately. (Breen and Baldwin, 2009))

A central issue when dealing with neologisms in Japanese is the nature of the orthography, with its use of multiple scripts, primarily *kanji* (Chinese characters), e.g. 猫, 犬, 鳥, 牛, etc., of which approximately 2,500 are in common use and which are used mainly for nouns and the roots of verbs, adjectives, etc.; and the *hiragana* and *katakana* syllabaries, each of 46 symbols plus diacritics. A major issue is the absence of any indication of the boundaries between the syntactic elements in texts. Automated text-processing in Japanese usually relies on morphological analysis software such as *MeCab* (Kudo, 2008) which employ large morpheme lexicons such as *UniDic* (Den et al., 2007), however unrecorded terms will (by definition) usually not be found in these lexicons.

Lexicographical Aspects

The study reported here has concentrated on the *identification* of neologisms with a view to possibly including them in a dictionary, and indeed many of those identified have been added to the online *JMdict* Japanese dictionary (Breen, 2004). The process of assessing potential lexical items for such inclusion is central to lexicography, and along with the establishment of an accurate translation into the target language (English in this case) involves a range of processes and issues in determining their suitability. In our study we have aimed identifying neologisms which are suitable for inclusion in both *coding* and *decoding* dictionaries, an important issue as Japanese dictionaries compiled for native speakers do not typically contain terms such as 凹状 (*ôjô* – concavity) as it is seen as a prefix-plus-noun (concave shape), whereas the term should, and does, occur in dictionaries intended for non-native speakers.

The techniques used in this study focussed on terms that occur in corpora in significant enough quantities to warrant further investigation. In addition, several of the techniques identified 2-*kanji* terms which are typically nouns, and thus would be strong candidates for lexicalization. There remain questions such as whether identified compound nouns or multi-word expressions have meanings which are idiomatic, novel or non-intuitive enough to warrant lexicalization. This is widely recognized as one of the major challenges in lexicography (Atkins and Rundell, 2008). Some have expressed the view that with the rise of electronic dictionaries, which do not have the size limitations of printed dictionaries, there is little harm in relaxing this evaluation and including larger numbers of non-idiomatic compounds, expressions, etc. In the case of Japanese there are additional issues such as multiple surface-forms, reading variations, the extensive use of pseudo-English constructions, etc. to take into account. A detailed analysis of the lexicographic issues associated with Japanese dictionary entries, including the criteria for lexicalization, is reported elsewhere (Breen, 2017).

Resources

As the experimentation with synthesized terms takes the form of create-and-test, a key requirement is access to appropriate large-scale corpora to test for the presence and usage patterns of the terms.

The main accessible Japanese corpus for this type of testing is the Google *n*-gram Corpus (Kudo and Kazawa, 2007), based on approximately 20 billion text segments extracted from WWW pages, and is provided in the form of sets of 1-grams to 7-grams with counts of the numbers of occurrences. As the Google corpus only reports *n*-grams which occur 20 or more times a second *n*-gram corpus was assembled using the smaller Kyoto University WWW Corpus, containing about 500 million text segments.

Investigation Approach

Three types of synthesized term formation were investigated:

a) **Abbreviation/Clipping**. This is a very common and productive process in Japanese, wherein the (usually) leading character of each of the components of a composite are taken to form an abbreviated compound. An example of this is 学割 *gakuwari* "student discount" from the full compound 学生割引 *gakuseiwaribiki*. The terms produced by this process are typically nouns or adjectival nouns, and hence if valid would be clear candidates for lexicalization.

b) **Affixation**. The addition of prefixes and suffixes, often written with a single *kanji*, is a very common morphological process in Japanese (Tsujimura, 2006). Vance (1991) describes 63 single-*kanji* affixes commonly employed. The process is very productive and the resulting terms are not usually lexicalized unless they have an idiomatic meaning or unusual reading. Most of the terms arising from this process are nouns, e.g. those arising from the affixation of 化 (*ka*: -ization). Some others, such as 的 (*teki:* -like, -ical, -ish, etc.) form adjectives.

c) **Compounding**. As in many languages, the formation of terms by combining two or more words or morphemes is very common. The components can be independent words, as in 秋空 *akizora* "autumn sky" where both 秋 *aki* and 空 *sora* can be used independently, or bound morphemes as in 警告 *keikoku* "warning" where neither component can be used alone. See Tanaka (2002) and Baldwin and Tanaka (2004) for earlier work in this area.) Two types of synthesized compounds were investigated:
   i. 2-kanji compounds, as in the examples above;
   ii. composites formed by aggregating known 2-kanji compounds, for example combining 警告 (above) with 射撃 *shageki* "firing, shooting" forms a composite 警告射撃 *keikoku shageki* meaning "warning shot".

Synthesized terms which were not already recorded in a reference lexicon were considered if they occurred more than 100 times in the corpus. The evaluation of such terms was carried out by examining their occurrences in a combination of syntactic contexts typically associated with nouns. Japanese uses a large number of particles which are usually written in the *hiragana* syllabary. Counts of occurrences were extracted for the terms encapsulated in 37 combinations of the following common pre/postpended particles.

pre: は (*wa*), が (*ga*), に (*ni*), の (*no*), な (*na*), て (*te*), や (*ya*)
post: を (*wo*), が (*ga*), に (*ni*), の (*no*), な (*na*), や (*ya*)

Two types of evaluation were carried out on the sets of counts:

a) a machine-learning analysis using support-vector machine (SVM) models trained on the patterns of counts from a range of established terms (Chang and Lin, 2011);
b) a heuristic approach using rules based on the numbers of encapsulations.

Investigation Outcome

a) **Abbreviation/Clipping**. Of 33,000 synthesized abbreviations 7,900 were evaluated of which 162 were identified as potential new nouns (2.0%). Hand-checking a sample revealed that few were actually valid abbreviations. Most were other types of collocations.
b) **Affixation**. Initial testing of potential terms generated by this technique resulted in large numbers which were clearly in regular use. It quickly emerged, however, that they almost always had quite predictable meanings and were unlikely to be included in a dictionary unless quite relaxed lexicalization criteria were applied. For example the noun 衒衒 *gengaku* "pedantry" can take suffixes such as the personalizing suffix 者 *sha* to form 衒学者 *gengakusha* "pedant" or the adjective-forming suffix 的 *teki* (mentioned earlier) to form 衒学的 *gengakuteki* "pedantic".
c) **Compounding**.
    i. 2-*kanji* compounds. As there are over 6 million combinations of the most common 2,500 *kanji*, two samples each of 40,000 compounds were generated from two ranges of *kanji* from which were excluded *kanji* for numerics, common affixes, etc. The unlexicalized compounds were tested against the two corpora, resulting in 200-500 compounds being classified from each sample. Hand-checking selections of these revealed that most were valid terms, with about 60% being proper names. Examples of such terms include 移弦 *igen* "string-crossing" (violin, etc. technique), 春苗 *shunbyô* "spring seedlings" (also a girl's name), and 母珠 *moshu* "large bead(s) in a Buddhist rosary". The precision of the technique, i.e. the proportion of classified terms which proved to be valid, was quite high. As the compounds are typically simple nouns most of the non-name terms were clear candidates for lexicalization.
    ii. 4-*kanji* compounds. The potential numbers of 4-*kanji* compounds which can be generated from known 2-*kanji* compounds is very large, however few record sufficient counts in the *n*-gram corpora to be considered further. Several batches of 1 million compounds were generated, with 400-500 from each being accepted for further analysis, and 30-50 being flagged by the models as potential terms. Again hand-checking confirmed their validity, with terms such as 英国王室 *eikokuôshitsu* "British royal family", and 欧州遠征 *ôshûensei* "European campaign" (esp. with sporting teams) being identified. A point to note is that many of the accepted terms have meanings which are readily apparent from the components, and hence are unlikely to be included in a general dictionary unless the usual criteria were relaxed. It was noted that compounds formed from components which were polysemous were more likely to have non-intuitive meanings, which indicates a possibly fruitful area of future study.

As mentioned earlier, the synthesized compounds were tested using both machine-learning and heuristic models. It was noted that in general the heuristic models performed more effectively than the machine-learning approach.

Conclusion

In this paper we describe the investigation of a neologism detection approach involving the synthesis of possible Japanese terms by mimicking Japanese morphological processes, followed by testing for the presence of candidate terms in Japanese corpora in appropriate syntactic contexts. Synthesized terms which passed this evaluation were assessed by regular lexicographic processes to determine their suitability for lexicalization. Of the techniques tested: abbreviation, affixation and compounding, the latter showed particular promise, with the 2-*kanji* compound generation and classification resulting significant numbers of unrecorded terms deemed suitable for inclusion in dictionaries.

References

B.T. Sue Atkins and Michael Rundell. 2008. T*he Oxford Guide to Practical Lexicography.* Oxford University Press, Oxford, UK.

Timothy Baldwin and Takaaki Tanaka. 2004. Translation by machine of compound nominals: Getting it right. In *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing.*

James Breen. 2004. JMdict: a Japanese-Multilingual dictionary. In *Proceedings of COLING-2004 Workshop on Multilingual Resources.*

James Breen and Timothy Baldwin. 2009. Corpus-based Extraction of Japanese Compound Verbs. In *Proceedings of the Australasian Language Technology Workshop (ALTW 2009).*

James Breen. 2017. Extraction of Neologisms from Japanese Corpora *(PhD Thesis, The University of Melbourne)* Chapter: "Neologisms: Lexicographic Issues and Terminology" . https://minerva-access.unimelb.edu.au/handle/11343/211675

James Breen, Timothy Baldwin, and Francis Bond. 2018. The Company They Keep: Extracting Japanese Neologisms Using Language Patterns. In *Proceedings of the Global Wordnet Conference*.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Yasuharu Den, Toshinobu Ogiso, Hideki Ogura, Atsushi Yamada, Nobuaki Minematsu, Kiyotaka Uchimoto, and Hanae Koiso. 2007. The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics (in Japanese). *Japanese Linguistics*, 22:101-123.
Taku Kudo. 2008. *MeCab*: *Yet Another Part-of-Speech and Morphological Analyzer*. http://mecab.sourceforge.net/.

Taku Kudo and Hideto Kazawa. 2007. Japanese Web N-gram Corpus version 1. http://www.ldc.upenn.edu/Catalog/docs/LDC2009T08/.

Takaaki Tanaka. 2002. Measuring the similarity between compound nouns in different languages using non-parallel corpora. In *Proceedings of Coling 2002*.

Natsuko Tsujimura. 2006. *An Introduction to Japanese Linguistics*. Blackwell, Oxford, UK, second edition.

Timothy J Vance. 1991. *Instant Vocabulary through Prefixes and Suffixes*. Kodansha International, Tokyo.