A Japanese Word Segmentation Proposal

Stalin Aguirre
National Polytechnic School of Ecuador
Faculty of Systems Engineering
stalin.aguirre@epn.edu.ec

Josafá de Jesus Aguiar Pontes
National Polytechnic School of Ecuador
Faculty of Systems Engineering
josafa.aguiar@epn.edu.ec

Abstract

Current Japanese word segmentation methods, that use a morpheme-based approach, may produce different segmentations for the same strings. This occurs when these strings appear in different sentences. The cause is the influence of different contexts around these strings affecting the probabilistic models used in segmentation algorithms. This paper presents an alternative to the current morpheme-based scheme for Japanese word segmentation. The proposed scheme focuses on segmenting inflections as single words instead of separating the auxiliary verbs and other morphemes from the stems. Some morphological segmentation rules are presented for each type of word and these rules are implemented in a program which is properly described. The program is used to generate a segmentation of a sentence corpus, whose consistency is calculated and compared with the current morpheme-based segmentation of the same corpus. The experiments show that this method produces a much more consistent segmentation than the morpheme-based one.

1 Introduction

In computational linguistics, the first step in text-processing tasks is segmenting an input text into words. Most languages make use of white spaces as word boundaries, facilitating this segmentation step. However, Japanese is one of the few languages that does not use a word delimiter. This particular problem has been the focus of many researchers because its solution is key to subsequent processing tasks, such as Part-of-Speech (PoS) tagging, machine translation or file indexing.

Segmenting a text requires the definition of a segmentation unit (Indurkhya and Damerau, 2010). This unit must be strictly defined to describe all the elements in a language. But languages are not perfect and have changed abruptly throughout the years, making it difficult or nearly impossible to define such a unit. The consensus has been that the unit to be used was the *word*, because it defines the majority of the elements in a language, elements that have a meaning and can stand by themselves (Katamba, 1994).

Even though there still are some constructions that do not fit in the word definition (Bauer, 1983), this segmentation unit is useful in languages that use spaces because they separate the majority of words in a text. For Japanese, however, this is not the case. It is a language that does not use spaces in its written form.

私たちの性格はまったく異なる。
(Our personalities are completely different.)

Furthermore, Japanese is an agglutinative language, which means that some constructions (specially inflected words) are formed by consecutively attaching morphemes to a stem (Kamermans, 2010). These long words are very important because they can work as full sentences without the need to add context that was previously stated, as illustrated in the following example:

待たされていました。
(I have been kept waiting.)

待つ (*wait*)
待たされる (*be kept waiting*)
待たされている (*being kept waiting*)
待たされています (*being kept waiting*) (*P)[1]
待たされていました (*been kept waiting*) (*P)

Given the nature of the language and the lack of a need for native speakers to explicitly separate words, there is no standard on how to segment a written text. Because of this, the segmentation unit and rules for text processing tasks are set by each researcher, although most of them have

[1] *P: Polite form

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 429–435
Florence, Italy, July 28 - August 2, 2019.

chosen a morpheme-based approach (Matsumoto et al., 1991; Kudo, 2005; Matsumoto et al., 2007).

The downside about this morpheme approach is that, in many cases, there is no consistency when segmenting the same string. The cause seems to be the influence of different contexts on the probabilistic models used in segmentation algorithms. In other words, by producing short morphemes as candidates, there are many segmentation possibilities from which the final one may change due to the context. This inconsistency problem is visible in n-gram data produced by Kudo and Kazawa (2009) and Yata (2010). Within these files, there are various entries of the same string as result of different segmentations, as shown in Table 1. This problem directly affects any later processing task that relies on the resulting segmentation. In machine translation, for example, different segmentations for the same word would produce different incorrect translations.

File	N-gram	Frequency
3gm-0034	行き ま し た	2681486
4gm-0056	行 き ま し た	290
4gm-0056	行き ま し た	384

Table 1: The word 行きました (*went*) (*P) as found in n-gram data files produced by Yata (2010).

Inflected words follow a limited set of rules. These rules properly define all possible inflections (Kamermans, 2010). As such, they can only lead to one possible correct segmentation. Taking this premise, the Proposed approach aims for a more consistent segmentation by focusing on the treatment of inflected words to limit their segmentation possibilities to a single one in all cases. Thus, reducing word segmentation inconsistency errors.

The present work is structured as follows: Section 2 describes the rules that lead the Proposed segmentation method. Section 3 describes the implementation of the algorithm that applies these rules; Section 4 introduces the evaluation parameters and the results obtained with the Proposed method, a comparison of these metrics with a morpheme-based method and discussion of the results; and Section 5 presents the conclusions of the work.

2 Segmentation Definition

Most Japanese constructions are created by directly connecting an affix to a word. A few of these constructions are considered separated words when translating them into English, such as: 日本式 (*Japan style*, from 日本 *Japan* and 式 *style*) or 外国語 (*foreign language*, from 外国 *foreign* and 語 *language*). Other constructions have a single word as translation, such as: 日本語 (*Japanese* (language) from 日本 *Japan* and 語 *language*). However, this word construction rule in Japanese can be found in "basic" words as well. For example: 大人 (*adult*, from 大 *great* and 人 *person*) or 女子 (*girl*, from 女 *woman* and 子 *child*), which are kept as single words in both languages. This means that we cannot generalize how these constructions should always get segmented, either as single words or multiple words.

As such, we have established a few segmentation rules where some affixes were connected to the words they modify by the use of symbols, regardless of the number of words it forms when translating them into English. For prefix concatenation, the backtick symbol (`) was used, on the other hand, for suffix concatenation, the hyphen symbol (-) was used. In general, the words that were considered for these affix concatenation rules were nouns and verbs.

Overall, the base of the segmentation for this work was the IPADIC (Asahara and Matsumoto, 2003) dictionary. This means that what was considered as *word* was each entry in this dictionary, with the exception of the verb and adjective entries. For these inflectional words, what was considered as word were the union of the inflectional word stems and the morphemes or auxiliary verbs that form the inflection, as described by Pontes (2013). In the case of the inflectional words, only the most common inflections are shown below [2].

Based on this word definition, the following rules were set:

2.1 Nouns

Most regular nouns were kept as single words according to the entries that belonged to this tag within the IPADIC. Some PoS included in this category were: common nouns, proper names, pronouns, pronoun contractions, adverbial nouns, verbal nouns (nouns that can be followed by する or related verbs), adjectival nouns (nouns that can be followed by な), Arabic numbers (wide and short length), counters and Chinese numbers.

[2]When mentioning an inflection based on an auxiliary verb, it also refers to all the inflections of such auxiliary.

- Names and pronouns were connected to personal suffixes.

  和子-さん (*Ms. Kazuko*)

- Common nouns were connected to non-inflectional affixes.

  フィリピン-人 (*Filipino*) (demonym)
  貧困-者 (*Pauper*)
  ドイツ-語 (*German*) (language)
  数`年 (*Several years*)

- Pronouns were connected to plural suffixes.

  私-たち (*We*)
  彼-等 (*They*)

- Nouns were connected to honorific prefixes.

  ご`注文 (*Order*) (*P)

- Nouns were connected to more than one affix when it was the case. The main noun was always right after the backtick or just before the first hyphen.

  お`手伝い-さん-たち (*Servants*) (*P)

- Nouns connected to the 的 character to adjectivize them were treated as suffixes. For the possible inflections of the 的 character, refer to Section 2.4.

  世界-的な (*Global*)

2.2 Verbs

Verbs are the most varied words in terms of inflections. This is due to the possibility of concatenating many auxiliary verbs to a single stem, producing really long words. In current segmentation methods, each inflectional word gets segmented by its stem and by each auxiliary verb. This scheme can be found in linguistic texts but it might not be the best way to segment these words because of the inconsistency problem illustrated in Table 1. For this method, the inflections were treated as single words.

- Present affirmative.

  会う (*Meet*)

- Negative form with auxiliary ない.

  合わない (*Do not meet*)

- Polite form with auxiliary ます.

  会います (*Meet*) (*P)

- Past form with auxiliary た.

  描いた (*Drew*)

- Continuative form with auxiliary て.

  話して (*Talk and ...*)

- Continuative form using the auxiliary いる.

  走っている (*Is running*)

- Desire with auxiliary たい.

  買いたい (*Would like to buy*)

- Hypothetical with auxiliary ば.

  読めば (*Should (you) read ...*)

- Passive with the auxiliary される.

  待たされる (*Be kept waiting*)

2.3 Adjectival Verbs

Adjectival verbs, also called *i-adjectives*, work the same as verbs. They keep a static stem while their suffixes change. These suffixes are formed by inflected auxiliary verbs from which a few are the same as the ones for verbs. Just like verbs, the inflections are treated as single words.

- Attributive form with い.

  欲しい (*Wanted, Desired*)

- Adverbial form with く.

  楽しく (*Happily*)

- Past form with かった.

  寒かった (*Was cold*)

- Negative form with auxiliary ない.

  面白くない (*Not interesting*)

2.4 Adjectival Nouns

Adjectival nouns, usually known as *na-adjectives*, can be connected to just three morphemes which are directly connected to the stem in this method.

- Copula な to directly modify a noun.

  大変な (*Terrible*)

- Continuative particle で to chain adjectives.

  知的で (*Intelligent and ...*)

- Nominalising particle さ.

  深刻さ (*Seriousness*)

3 Implementation

3.1 The Dictionaries

The dictionaries needed for this work were a list of non-inflectional words and various lists of inflectional word stems:

The **Non-Inflectional Word Dictionary (NIWD)** was formed by unifying the IPADIC files into a single list; omitting verb, adjectival verb, adjectival noun and symbol files. For the final dictionary file, a column with frequency counts was added. These counts were obtained from n-gram data produced by Yata (2010) and assigned to each entry. For this, the n-gram data was first cleaned by removing the white spaces separating the n-gram tokens, and the counts of the repeated entries were summed. Once cleaned, the whole n-gram data was sorted. Then, each entry of the dictionary was searched within the cleaned data to extract its frequency count.

The **Inflectional Word Dictionary (IWD)** included lists of the stems for all the inflectional words obtained by Pontes (2013). This dictionary was divided in two sets. The first set contained the adjectives classified in: noun adjectives (*na*, な), verbal adjectives (*i*, い) and irregular adjectives (*ii*, いい). The second set covered all verbs classified in eleven groups: *u* (う), *bu* (ぶ), *gu* (ぐ), *ku* (く), *mu* (む), *nu* (ぬ), *ru* (る), *su* (す) and *tsu* (つ) for first group verbs, and *ichidan* (える, いる) for second group verbs. One additional group was added for the honorific verbs that end in *aru* (ある).

3.2 The Inflection Automaton

Due to the large number of possible inflected words in the Japanese language, as shown by Pontes (2013), it was not practical to store them all in memory. Instead, a Deterministic Finite Automaton (DFA) was built to validate them.

The objective of the DFA was identifying whether an input string corresponds to an inflected word or not. This was done by checking if the string was formed by a stem (by making use of the IWD) and a correct inflectional suffix. The inflectional suffixes that were implemented in the DFA involved treating each character as a transition to a new state. The states were final if all the previous transitions formed a valid inflection. The transitions and states were created following the inflectional patterns obtained by (Pontes, 2013).

The process that the DFA implemented was:

1. Receive a string, set *position* to the start of the string.
2. Take a substring from *position* to *i*, where *i* grows by 1 in each iteration.
3. Look for the substring in the IWD. If not found, take the next substring and repeat this step. If there is no next substring, the string is not an inflected word, as it does not contain a stem.
4. If the substring is found, set the initial state to the corresponding stem group and move *position* to the end of the substring within the original string.
5. Read each next character from *position* and use it as a transition to a new state. If the next character is not a valid transition, go to step 3.
6. When there are no more characters left and the last state reached is an accepting state, the string is considered an inflected word. If the last state reached is not an accepting state, go to step 3.

Given that irregular verbs (suru する and kuru くる) do not have static stems, the method started at step 5 by setting the initial state to *suru* and, if not found, to *kuru*.

If the string was not recognized as an inflected word by the DFA, it verified if the first character of the string was a honorific prefix (お, ご or 御), and if so, a substring starting from the second character was sent to step 1.

3.3 The Segmentation Program

The main program implemented the NIWD and the DFA for word and inflection recognition respectively, which were part of a unigram language model for word segmentation (Jurafsky and Martin, 2000). This probabilistic model was accompanied with a few grammar rules for overriding the final segmentation decision. To refine the program, 2,500 sentences from the Tatoeba (Ho and Simon, 2006) sentence corpus were used as validation data.

The steps that were followed by the program were:

1. Split the input text into phrases by using delimiter symbols such as parenthesis, punctuation, etc., as separators.

2. For each phrase, take substrings of all sizes > 0, and from all positions $<$ phrase length.

3. For each substring, verify if it is an inflected word with the DFA. If it is not, look for it in the NIWD. If it is not found, verify if it is a number or a foreign word in Katakana or other alphabets. If it is not, repeat this step with the next substring.

4. If the substring was verified or found in step 3, save it as a candidate word and assign it a frequency count by looking for its value within n-gram data like the one produced by Yata (2010), and accumulate the frequency count in a variable.

5. Once all candidate words are available, calculate their score by taking the negative logarithm of the frequency count assigned to each word, divided to the accumulated frequency count.

6. Create a graph where its nodes represent the positions between each character of the phrase.

7. For each node, select all the candidate words whose last character position is right before the node. Check if any of the grammar rules apply to them in order to directly choose one or remove them. If no rules applied, choose the one with the least score value.

8. Set the chosen word as the edge that connects the node before the position of the first character of the candidate word, and the current node.

9. When all the edges are set, obtain each previous edge that connects the current node, starting from the last one and going backwards.

10. Return the obtained edges in order while adding a separation symbol between them such as a backtick (‘) or hyphen (-) for affixes or a white space for other words.

4 Evaluation

Two evaluations were established for the Proposed method. The first one checked how correctly it segments a test corpus in comparison with a gold segmentation. The second one compared the consistency of its segmentation of a corpus with MeCab's (Kudo, 2005) for the same corpus.

4.1 Segmentation Evaluation

For this evaluation, we used 1,000 sentences from the Tatoeba (Ho and Simon, 2006) sentence corpus, which were manually segmented to create a gold segmentation corpus.

A Baseline method by Pontes (2013), that segments words based on longest string matches, was used for comparison. Both methods' outputs are comparable given that the Baseline also uses the IWD for inflected word segmentation. MeCab, on the other hand, is not. It produces a morpheme-based segmentation for inflected words.

The metrics used for evaluating both segmentation methods were: recall, precision, and f-measure (Wong et al., 2009). From these metrics, the following abbreviations were considered: number of correctly segmented words (CW), total number of words in gold corpus (GW), total number of segmented words (SW). The obtained results are shown in Table 2.

Table 2: Results from evaluating the segmentation methods.

Method	**GW**	**SW**	**CW**
Proposed	9757	9798	9656
Baseline	9757	9190	8069
Method	**Recall**	**Precision**	**F-Measure**
Proposed	98.96%	98.55%	98.76%
Baseline	82.70%	87.80%	85.17%

The Proposed method outperforms the Baseline method. This score is apparently high, but notice that it is not statistically significant, as the time allowed us to manually prepare and revise only 1,000 sentences. Definitely, a larger corpus is necessary in order to provide a higher confidence level on the evaluation of our Japanese word segmentation method. For the next evaluation, however, we do count with a larger corpus for testing as explained below.

4.2 Consistency Evaluation

To evaluate the consistency of the segmentation method, a corpus of 185,393 sentences from the Tatoeba (Ho and Simon, 2006) sentence corpus was used. This corpus was segmented with four segmentation methods which were: the Proposed method that attaches affixes to words as defined in Section 2.1 (PMA), the Proposed method that does not attach affixes to words (PM), the Baseline method (BM) and MeCab.

PMA, PM and the Baseline method consider inflected words as single words, which means, all the auxiliary verbs that forms the inflections are directly connected to the stems.

For each corpus segmentation, the following process was applied:

Generate the n-gram data. Generate up to 7-gram data of a corpus segmentation by using the SRILM toolkit (Stolcke, 2004).

Clean the n-gram data. Remove the white spaces that separate the n-gram tokens and sort the whole n-gram data.

Create a list of repeated entries. Extract the repeated entries (RE) from the clean n-gram data by the use of regular expressions and produce a *repetition list*. Count the number of RE to calculate the inconsistency.

Count the RE that contain inflected words. Apply the DFA on the repetition list in order to obtain a subset of entries that contain inflected words and count them.

Due to the large amount of entries and the lack of context in n-gram data, it was not reasonable to say that the inflected words detected were the correct words in the corpus. Therefore, we made three different sets for inflected word count approximation: inflected words of more than one character within the entry (IW1), inflected words of more than two characters within the entry (IW2) and inflected words found as the whole entry (IWW). Table 3 shows an example of each set.

Table 3: Repeated n-gram entries, generated from MeCab's segmentation, that contain inflected words as found by the DFA.

Set	Clean N-gram Entry	Inflected Word Found
IW1	にいるか	いる
IW2	は思ったより	思った
IWW	変われる	変われる

In order to calculate the inconsistency of each method, the entries of the RE, IW1, IW2 and IWW lists of all the methods were summed. The share of the inconsistency from each method is shown in Table 4.

The evaluation of the four methods shows that both Proposed methods produce the least RE, which means that they are more consistent overall. Regarding the RE that contain inflected words, the Baseline method has the least inconsistency.

Table 4: Number of n-gram entries and inconsistency error distribution for each method.

Method	N-gram Entries
PMA	5,803,353
PM	5,808,828
BM	5,717,438
MeCab	6,245,224

Method	RE	IW1	IW2	IWW
PMA	14.58%	16.46%	17.42%	19.94%
PM	15.40%	17.20%	18.12%	20.66%
BM	16.23%	13.25%	13.46%	8.12%
MeCab	53.79%	53.09%	51.00%	51.28%
Total	100%	100%	100%	100%

The total number of n-gram entries produced from MeCab's segmentation is approximately 8% higher than the one produced by the second higher (PM). However, such a rate is insignificant compared to the rate of RE within the n-gram data, in which MeCab is around 300% more inconsistent than each one of the other three methods.

5 Conclusion

We have demonstrated that by considering inflectional words (with all their auxiliary verbs) as single words, the number of possible segmentations for those words in different contexts gets reduced. Therefore, the resulting segmentation is more consistent and more accurate. Tasks that use word segmentation would also see an improvement, such as language models and machine translation systems.

This approach relies on the fact that it is possible to define all the inflectional rules of the Japanese language. The same method could be applied to other words that can be defined by rules, or to other unsegmented languages whose rules can be defined the same way.

Acknowledgments

We give our most sincere thanks to the Japanese professor at Catholic University of Ecuador 出麻樹子 (Makiko Ide) for her very valuable contribution. She voluntarily helped us prepare the gold segmentation corpus which was a very important part of the project.

References

Masayuki Asahara and Yuji Matsumoto. 2003. Ipadic version 2.7.0 user's manual. [online] Open Source Development Network. Available at: https://ja.osdn.net/projects/ipadic/docs/ipadic-2.7.0-manual-en.pdf [Accessed 25 Apr. 2019].

Laurie Bauer. 1983. *Some basic concepts*, Cambridge Textbooks in Linguistics, page 7–41. Cambridge University Press.

Trang Ho and Allan Simon. 2006. Tatoeba. [online] Tatoeba: Collection of Sentences and Translations. Available at: https://tatoeba.org [Accessed 25 Apr. 2019].

Nitin Indurkhya and Fred J. Damerau. 2010. *Handbook of Natural Language Processing*, 2nd edition. Chapman & Hall/CRC.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st edition. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Michiel Kamermans. 2010. *An Introduction to Japanese - Syntax, Grammar and Language*. SJGR Publishing.

Francis Katamba. 1994. *English Words*. Routledge.

Taisei Kudo. 2005. Mecab : Yet another part-of-speech and morphological analyzer.

Taku Kudo and Hideto Kazawa. 2009. Japanese web n-gram version 1 ldc2009t08. [online] Linguistic Data Consortium. Available at: https://catalog.ldc.upenn.edu/LDC2009T08 [Accessed 25 Apr. 2019].

Yuji Matsumoto, Sadao Kurohashi, Yutaka Nyoki, Hitoshi Shinho, and Makoto Nagao. 1991. User's guide for the juman system, a user-extensible morphological analyzer for Japanese. *Nagao Laboratory, Kyoto University*.

Yuji Matsumoto, Kazuma Takaoka, and Masayuki Asahara. 2007. Chasen morphological analyzer version 2.4. 0 user's manual. *Nara Institute of Science and Technology*.

Josafá Pontes. 2013. A corpus of inflected japanese verbs and adjectives. [Unpublished].

Andreas Stolcke. 2004. Srilm — an extensible language modeling toolkit. *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, 2.

Kam-Fai Wong, Wenjie Li, Ruifeng Xu, and Zhengsheng Zhang. 2009. *Introduction to Chinese Natural Language Processing*, volume 2.

Susumu Yata. 2010. N-gram corpus - japanese web corpus 2010. [online] S-yata.jp. Available at: http://www.s-yata.jp/corpus/nwc2010/ngrams/ [Accessed 25 Apr. 2019].