Memory-Efficient Katakana Compound Segmentation Using Conditional Random Fields

KRAUCHANKA Siarhei[1,2] *ARTSYMENIA Artsiom*[3]
(1) MINSK STATE LINGUISTIC UNIVERSITY, Belarus, Minsk, Zakharova st. 21
(2) IHS Inc., Japan, Tokyo, Minato-ku, Toranomon 5-13-1
(3) IHS Inc., Belarus, Minsk, Starovilenskaya st. 131
Sergey.Kravchenko@ihs.com, Artsiom.Artsymenia@ihs.com

ABSTRACT

The absence of explicit word boundary delimiters, such as spaces, in Japanese texts causes all kinds of troubles for Japanese morphological analysis systems. Particularly, out-of-vocabulary words represent a very serious problem for the systems which rely on dictionary data to establish word boundaries. In this paper we present a solution for decompounding of katakana sequences (one of the main sources of the out-of-vocabulary words) using a discriminative model based on Conditional Random Fields. One of the notable features of the proposed approach is its simplicity and memory efficiency.

KEYWORDS : Japanese language, Tokenization, Katakana Compound Segmentation, Conditional Random Fields

Proceedings of COLING 2012: Posters, pages 1131–1140,
COLING 2012, Mumbai, December 2012.

1 Introduction

It is well known that in Japanese language the absence of word boundary delimiters, such as spaces, adds to the morphological ambiguity, which, in turn, causes many difficulties for the language processing systems, which rely on precise tokenization and POS tagging. This problem is especially grave in compound nouns of Japanese and foreign origin. A number of morphological analysis systems were developed to resolve this problem and a number of such systems show some relatively good results. It is reported, however, that one of the biggest problems in most of these researches is related to the out-of-vocabulary (OOV) words. In some cases, an OOV word can be a sign of an insufficient core lexicon, which simply needs to be updated to correspond to the requirements of the processing system, while in other cases, an OOV word can be related to the constantly changing and growing peripheral lexicon, which cannot be reflected in any existing dictionary, and, therefore, it needs to be identified by non-vocabulary means.

In relation to Japanese language, this duality has its own specifics. In general, the Japanese texts consist of different types of writings – kanji, hiragana, katakana and a small amount of non-Japanese characters (for words and abbreviations coming from foreign languages). Most of the words of Japanese origin are written using kanji and hiragana, while katakana is usually used to transcribe the words of a foreign language (mostly, English) origin. When it comes to compound nouns, each of these two types of writings have their own specifics of formation of new expressions. For Japanese compound nouns, the most common way of formation is the concatenation of simple nouns (which can also be accompanied by abbreviation):

磁気共鳴画像 (Magnetic Resonance Imaging, MRI)

where 磁気 - “magnetism”, 共鳴 - “resonance, sympathy”, 画像 - “image, picture”.

The compound nouns of non-Japanese origin are formed by transliteration of foreign words using katakana syllables and then concatenation of the elements:

プリントダイポールアレイアンテナ (printed dipole array antenna)

where プリント - “print” (here, “printed”), ダイポール - “dipole”, アレイ - “array”, アンテナ - “antenna”.

These two are by far the richest sources of OOV words and, consequently, of the problems for many morphological analysis systems. Therefore, by solving the problem of correct identification of words in these expressions, it is possible to significantly reduce the bad influence of OOV words on the results of morphological analysis in general.

The method described in this paper focuses primarily on katakana expressions, and the possibility of applying a similar approach to the kanji-based expression segmentation will be left for another research.

2 Related work

Our research can be characterized as a discriminative approach to katakana compound segmentation. It can be viewed as both – as a specific task aimed at a narrow problematic area of the Japanese morphological analysis, and as a part of the Japanese text tokenization problem in general.

The specific problem of katakana compound segmentation has received an extra attention in the works of (Nakazawa, 2005) and (Kaji, 2011). Both of these researches present a number of techniques, such as dictionary and corpus validation, back-transliteration, web search, which, when combined, do a really good work of identifying the words in compound expressions. It should be noted, though, that in order to de-compound katakana expressions, these systems rely on large external resources like large vocabulary, “huge” English or Japanese corpora, parallel corpora or Web search results. Because of that, these approaches can be very efficient for the task of extraction of new vocabulary data, but it is doubtful, that they can be efficiently implemented in a morphological analysis system to solve the problem of constantly appearing OOV words.

As for the Japanese text tokenization problem in general, a number of researches have been performed for many years (Kurohashi, 1994; Kudo, 2004; Asahara, Matsumoto, 2000), some of them culminating in working morphological analysis systems such as Juman, Chasen or Mecab. It has been noted on numerous occasions that one of the weaknesses of these systems is their very tight connection to their respective dictionaries, which results in poor performance when it comes to the OOV words. Traditional vocabulary-based approaches using Hidden Markov Model (such as Chasen) preform quite well on texts with fewer OOV words. More advanced approaches using Maximum Entropy Markov Models (Uchimoto, 2001) or Conditional Random Fields (Kudo, 2004) take better care of OOV words, but still underperform when it comes to processing the texts out of training domain. As opposed to sequential vocabulary-based approaches, advances in pointwise modeling for Japanese morphological analysis have been made recently. The method described in (Neubig, Nakata, Mori, 2011) combines character context and dictionary information in one model to get good results on both – vocabulary and OOV words. Our research is very closely related to these works, but in our paper we focus exclusively on the problem of word segmentation, because POS tagging of katakana expressions does not require highly complex models. By doing that, we concentrate on one of the most problematic areas of Japanese morphological analysis and try to solve this problem using minimal resources with maximum efficiency. In perspective, it could mean that even using a less complicated (and, therefore, less resource demanding) model for such task as morphological analysis, it might be possible to achieve superior results by solving some of the problems with their own dedicated methods.

3 Research environment

For the general morphological analysis we have been using our original system, whose resource base includes a POS dictionary, which consists of more than 210000 unique entries (out of which, at least 21000 entries are katakana words), and a morphologically annotated corpus of technical and scientific documents with 33,134 sentences, which contains 6,909 unique katakana single- and multi-word expressions. Here, morphological annotation means word boundary and POS tag assignment. For the Japanese POS tagset, we use our own original formalism which is close to that of the well-known Penn Treebank project.

Additionally, in order to train and test our model, we needed a much larger and more representative corpus of katakana expressions with explicitly indicated word boundaries. For this purpose, we have developed a simple method of katakana expression extraction based on back-transliteration and employing parallel sentence-aligned Japanese-English text corpora. We have managed to automatically acquire a corpus of 4,977,790 non-unique katakana (both, single- and multi-word) expressions matched to their English translations with explicitly indicated word boundaries. We used this result for both - training (4,000,000 katakana expressions which after unification turned into 80,550 expressions) and testing (977,790 non-unique katakana expressions).

4 CRF model for katakana decompounding

For splitting of katakana compounds into words we use a method where each katakana character is treated as a separate token with a label assigned to it depending on the position of the character in the word it belongs to, similar to that of (Ng, Low, 2004). The following labels are used for marking the word boundaries:

F - First character of the word

M - Middle character of the word

L - Last character of the word

S - Single character word

This way the task of katakana compound segmentation can be defined as a label tagging problem. For parameter training we use Conditional Random Fields as described in (Lafferty, 2001). For each katakana character in the sequence with each possible label a number of features are assigned based on the dictionary and context information about the word the character may belong to. The set of features is provided in table 1, where T_i is one of the labels (F, M, L, S) for the i-th character (katakana syllable) in the compound.

Feature form	**Description**
$T_i\_C_i$	Character itself
$T_i\_C_{i-1}$	Previous character in the sequence
$T_i\_C_{i+1}$	Next character in the sequence
$T_i\_C_{i-2}\_C_{i-1}$	Context character pairs
$T_i\_C_{i-1}\_C_i$	-
$T_i\_C_i\_C_{i+1}$	-
$T_i\_C_{i+1}\_C_{i+2}$	-
$T_i\_C_{i-3}\_C_{i-2}\_C_{i-1}$	Context character triples
$T_i\_C_{i-2}\_C_{i-1}\_C_i$	-
$T_i\_C_{i-1}\_C_i\_C_{i+1}$	-
$T_i\_C_i\_C_{i+1}\_C_{i+2}$	-
$T_i\_C_{i+1}\_C_{i+2}\_C_{i+3}$	-
T_i_D	Dictionary information
$T_{i-1}\_ T_i$	Label bi-gram feature, which takes into account the label of the previous character to avoid impossible label sequences (F F, M F, L L and so on)

TABLE 1 – Feature set.

The dictionary information feature T_i_D is assigned to a character based on pre-splitting of the katakana sequence using some dictionary-based heuristic. For example the greedy algorithm may be used where, at first, the longest dictionary word is selected starting at the first syllable of katakana expression, then the longest dictionary word is chosen from the position next to the first word and so on. Any other dictionary-based method of word segmentation can also be used. After splitting of the word with heuristics each character gets a feature $T_i\_WL\_WP$ where WP – position of the character in the word, WL – length of the word.

For example, the word キャンセレーション which is absent from our system dictionary will be split by the greedy algorithm into キャン + セレーション. So the dictionary feature T_i_D will be assigned the following way during the training procedure:

キ → F_3_0

ャ → M_3_1

ン → M_3_2

セ → M_6_0

レ → M_6_1

ー → M_6_2

シ → M_6_3

ョ → M_6_4

ン → L_6_5

During the testing, the features M_3_2 and L_3_2, M_6_0 and F_6_0 will compete with each other and together with other features will vote towards M M or L F chain of labels. T_{i-1}_T_i feature will also add the weight of M M and L F transitions to get the final decision whether the word should be split into two or not.

5 Experiments and results

The training data for our experiments included the following resources (all described in section 3):

- approx. 21,000 katakana words from our POS dictionary to train the dictionary feature;
- 6,909 unique katakana expressions from our POS corpus, and 80,550 unique automatically extracted katakana expressions from English-Japanese parallel corpora (originally - 4,000,000 non-uniqued expressions) to train the context features;

One of the characteristics which allowed us to achieve a high memory-efficiency in our approach is the usage of regularization techinque described in (Vail, Lafferty, Veloso, 2007). All of the training for our model was performed at L1 regularization. L2 regularization shows slightly better results, but it also produces a very large number of features (45,484 features for L1 against 638,472 for L2), which we did not consider as efficient implementation.

In order to test our method we used the following corpora:

- automatically extracted katakana compounds from parallel English-Japanese texts of the same domain as the training corpus (PAR) containing 977,790 non-unique compounds with 1,160,770 words, out of which 54% (631,168) are OOV words;
- out-of-domain manually annotated corpus of general newspaper articles (NEWS) containing 4,986 non-unique katakana expressions with 5,338 words, out of which 44% (2,390) are OOV words.

While the results received using our approach were satisfactory for our original task (improvement of OOV words processing in katakana compounds), we also needed to compare them with those of other existing systems. We compared our approach with the most common methods of Japanese morphological analysis – a simple dictionary-based Hidden Markov Model approach (HMM), and a more sophisticated Conditional Random Fields approach from Mecab morphological analysis system (MECAB). For training of HMM and MECAB we used the same training data which was used for our system. The results of comparison are presented in the table 2.

System	**PAR**	**NEWS**
HMM	84.01%	82.73%
MECAB	87.50%	91.45%
Our approach	98.27%	96.67%

TABLE 2 – Comparison with other systems (F-Measure).

The results show that our approach heavily outperforms some of the most popular morphological analysis methods used in katakana compound segmentation task. The reason for that is the usage of character-based feature assignment and boundary labelling, which, instead of dictionary data, relies on more robust syllable sequence data. Because of that, the influence of OOV words is significantly reduced. However, as it was mentioned earlier, our approach also uses the dictionary information, which gives it additional domain specific training data. The influence of dictionary data (D-feature) on performance of our model is explored in table 3.

Model variation	**PAR**	**NEWS**
Model without D-feature	98.10%	95.89%
D-feature, greedy	98.21%	96.49%
D-feature, smart	98.27%	96.67%

TABLE 3 – Influence of the dictionary feature on the decompounding results (F-Measure).

In the table, "D-feature, greedy" relates to the greedy dictionary-based algorithm of compound pre-splitting, "D-feature, smart" relates to the dictionary-based pre-splitting algorithm which chooses the splitting variant with as few splittings as possible (longest words from the dictionary). As we can see from the results, the influence of D-feature is very small, which suggests that it is possible to employ our method without using any dictionary data at all, and still reach high level of performance, thus reducing the total amount of features required for this task, and contributing to the memory-efficiency of the system. The usage of D-feature would most certainly improve performance on texts "familiar" to the dictionary, which have very few OOV words. The difference in results between greedy and smart way of using dictionary is not significant – a slight advantage goes to the latter.

Finally, we decided to evaluate the impact of employing our approach for katakana compound segmentation (KAT) within a general morphological analysis system based on a first order Hidden Markov Model (HMM). The testing was performed on the NEWS corpus, which was used in the out-of-domain testing of katakana compound segmentation earlier. The results are presented in table 4.

System configuration	**Segmentation F-Measure**	**Segmentation +Tagging F-Measure**
HMM only	93.24%	90.12%
HMM + KAT	93.64%	90.44%

TABLE 4 – The impact of the proposed katakana compound segmentation approach on the performance of a morphological analysis system.

As seen in the table, the overall performance of a morphological analysis system shows a small but notable improvement in overall performance after implementation of our approach for the processing of katakana expressions. While the overall performance itself might not be so high due to the simplicity of the test model (first order dictionary-based HMM), the difference of approx. 0.35% gained on katakana expressions rich with OOV words, cannot be discounted.

Conclusions

In this paper we have presented a new solution for katakana compound segmentation problem. Such characteristics as limited lexicon (only 50 syllables of katakana, instead of thousands of vocabulary words), possibility to implement the model without using any vocabulary data at all (without D-Feature), and a small number of resulting features due to a corresponding regularization technique make our approach very memory-efficient. This simplicity is a great advantage to other existing approaches especially considering the gravity of such problem as katakana decompounding in the overall performance of a morphological analysis system. As a result, our approach can be implemented as a dedicated solution for katakana expression tokenization within a general morphological analysis system of any complexity.

For the future work, we plan to explore the possibility of applying a similar dedicated approach for kanji-based multi-word expression tokenization and POS-tagging.

References

T. Nakazawa, D. Kawahara, and S. Kurohashi. (2005). Automatic acquisition of basic Katakana lexicon from a given corpus. In Proceedings of IJCNLP, pages 682–693.

N. Kaji, M. Kitsuregawa. (2011). Splitting noun compounds via monolingual and bilingual paraphrasing: a study on Japanese katakana words. In Proceedings of EMNLP, pages 959–969.

S. Kurohashi, M. Nagao. (1994). Improvements of Japanese morphological analyzer JUMAN. In Proceedings of the International Workshop on Sharable Natural Language Resources, pages 22–38.

T. Kudo, K. Yamamoto, and Y. Matsumoto. (2004). Applying conditional random fields to Japanese morphological analysis. In Proceedings of EMNLP, pages 230–237.

K. Uchimoto, S. Sekine, and H. Isahara. (2001). The unknown word problem: a morphological analysis of Japanese using maximum entropy aided by a dictionary. In Proceedings Of EMNLP, pages 91–99.

M. Asahara and Y. Matsumoto. (2000). Extended models and tools for high-performance part-of-speech tagger. In Proceedings of COLING, pages 21–27.

G. Neubig, Y. Nakata, S. Mori. (2011). Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011), pages 529-533.

J. Lafferty, A. McCallum, and F. Pereira. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Intl. Conf. on Machine Learning.

H. T. Ng, J. K. Low. (2004). Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based? In EMNLP 9.

D. Vail, J. Lafferty, and M. Veloso. (2007). Feature Selection in Conditional Random Fields for Activity Recognition. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).