

Character Categorization via Latent Dirichlet Allocation for Kana Sequence Segmentation with Conditional Random Fields

Tomonari Masada^a

^a*Department of Computer and Information Sciences, Nagasaki University, Japan*
masada@cis.nagasaki-u.ac.jp

Abstract: We propose an efficient Kana sequence segmentation as a component of faster and easier interfaces for e-learning systems. We assign categories to Kana characters via latent Dirichlet allocation (LDA) and use the categories to compose additional features for conditional random fields (CRF). We compare the categories our method gives and those manually prepared by their efficiency in Kana sequence segmentation.

Keywords: sequence segmentation, latent Dirichlet allocation, conditional random fields

Introduction

Input methods as a component of e-learning interfaces in Japanese are our main topic. Researches on Japanese input methods mainly focus on Kana-Kanji conversion. However, persons with disabilities may experience difficulty in typing for conversion. Therefore, we focus on Kana sequence segmentation [5] for communication without Kanji characters, because the meaning of a Kana sequence is clear as long as the sequence is appropriately segmented into words. We propose a character categorization via latent Dirichlet allocation (LDA) [2] [7] and use the categories to compose additional features for conditional random fields (CRF) [4]. While we can manually assign categories to characters based on our knowledge, we aim to go beyond such off-the-shelf categories by our method.

1. Our Method

Kana sequence segmentation can be described as an assignment of 0/1 labels to characters. For example, “a-shi-ta-no-te-n-ki-ha-do-u” is labeled as “1-0-0-1-1-0-0-1-1-0”, where “1” means the first character of a word. We use linear chain CRF [6] for segmentation. Let x_i be a random variable for i th character in a sequence, and y_i be a random variable for the label of i th character. Our basic features in CRF are $f_u(x_i = c, y_i = l)$, $f_v(x_i = c, y_i = l, y_{i-1} = l')$, and $f_b(x_i = c, x_{i-1} = c', y_i = l, y_{i-1} = l')$, where c and c' (resp. l and l') denote observed characters (resp. labels assigned to characters). Further, we use character categories to compose additional features, $f_u(t_i = s, y_i = l)$, $f_v(t_i = s, y_i = l, y_{i-1} = l')$, and $f_b(t_i = s, t_{i-1} = s', y_i = l, y_{i-1} = l')$, where t_i is a random variable for the category of i th character. Each feature is equal to 1 when the corresponding condition holds, and is equal to 0 otherwise. Our method can provide character categories for CRF as follows. We regard characters in our problem as words in LDA document modeling and compute a topic frequency distribution for each

distinct character. We use this frequency distribution as a feature vector of each distinct character and obtain categories of characters by Gaussian mixture clustering [1] of these feature vectors. We interpret resulting clusters as character categories.

2. Evaluation Experiment

Ten character categories, i.e., ten clusters, our method provides are compared with the seven manually prepared categories: 1) Kana characters which never appear as a particle, 2) other Kana characters, 3) digits, 4) uppercase and 5) lowercase alphabets, 6) punctuations, and 7) others. Training and test sets include 21,704 and 19,427 newswire articles from <http://japan.internet.com>, respectively. Since a Japanese morphological analyzer *MeCab* [3] gives pronunciations of segmented words in Kata-Kana, we regard the pronunciations as the gold standard for Kana sequence segmentation. After we apply MeCab, training and test set turn out to be sequences of 23,720,438 and of 22,719,333 characters, respectively. Note that Kana sequence segmentation is different from Japanese morphological analysis, because Kana sequences provide fewer clues to CRF learning processes than the sequences where Kanji characters are also observable. Both of training and test sets include 270 distinct characters among which 85 are Kana. Table 1 presents the percentages of correctly labeled characters among 22,719,333 test set characters. Character categories automatically generated by our method lead to better results. Manually prepared categories result in over-segmentation, i.e., large number of mistakes of type “0→1”, where a character which is not the first character of a word is incorrectly labeled as the first character of a word.

Table 1. Evaluation Results.

manually prepared categories	93.2 % (1→0: 635,965; 0→1: 908,592)
categories obtained by our method	95.3 % (1→0: 580,263; 0→1: 475,022)

3. Conclusion

We provide an efficient Kana sequence segmentation. Our contribution is a proposal of automatic character categorization via LDA to compose additional features for CRF. We now plan to implement an input method using character categories given by our method for more realistic evaluations.

References

- [1] Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Chapter 9. Springer.
- [2] Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol.3 (pp.993-1022).
- [3] Kudo, T., Yamamoto, K., & Matsumoto, Y. (2004). Applying Conditional Random Fields to Japanese Morphological Analysis. *Proceedings of EMNLP'04* (pp. 230-237).
- [4] Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of ICML'01* (pp.282-289).
- [5] Makino, H., & Kizawa, M. (1980). An Automatic Translation System of Non-Segmented Kana Sentences into Kanji-Kana Sentences. *Proceedings of COLING'80* (pp.295-302).
- [6] Sutton, C., & McCallum, A. (2007). An Introduction to Conditional Random Fields for Relational Learning. *Introduction to Statistical Relational Learning*. MIT Press.
- [7] Teh, Y.W., Newman, D., & Welling, M. (2006). A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems*.