Data-driven Language Independent Word Segmentation Using Character-Level Information

Dong-Hee Lim, Seung-Shik Kang
School of Computer Science, Kookmin Univerity
861-1 Chongnung-dong, Songbuk-gu, Seoul 136-702, Korea
{nlp, sskang}@cs.kookmin.ac.kr

Abstract

This paper presents a data-driven language independent word segmentation system that has been trained for Chinese corpus at the second Chinese word segmentation bakeoff. The system consists of a base segmentation algorithm and the refining procedures for the undecided character sequences. It does not use any lexicon and the base segmentation is simply done by character bigram and HMM-model is applied for the remaining character sequences. As a final step, high-frequency character trigram modifies the error-prone parts of the text.

1 Introduction

We participated in the closed track of the second Chinese word segmentation bakeoff for the training corpus of HK (City University of Hong Kong, PK (Beijing University), and MS (Microsoft Research). Our system is independent of the corpus or the language that we also registered for AS (Academia Sinica) track, but failed to generate a result because of the code set problem. AS uses two-byte space characters instead of a blank(0x20), and more 0x0A is used in AS that is regarded as EOF in Windows environment.

The result of our system is not a top-level system when compared to other systems. However, our approach is quite acceptable because the data-driven methods can contribute to improving the accuracy of other word segmentation systems because we did not performed a tuning the system to fix the frequently repeating error patterns.

This work was supported by the Korea Science and Engineering Foundation(KOSEF) through Advaned Information Technology Research Center(AITrc).

2 Bigram and trigram data

We extracted a character bigram data from the training corpus. In the previous studies, Shim(1996) and Kang(2001) constructed space generation probability for each adjacent two characters XY. They are inside probability "X_Y", left-side probability "_XY", and right-side probability "XY_".[2] That is, they ignored 'space information'. In our bigram data, inside and outside space information is extracted from the training corpus, together with the character pairs. We call it 'extended bigram data' and it has eight types of frequency data. For example, XY consists of the frequencies of "0X0Y0", "0X0Y1", "0X1Y0", "0X1Y1", "1X0Y0", "1X0Y1", "1X1Y0", and "1X1Y1".[3] From the frequencies of the extended bigram data, we compute the space generation probability of $P_{t=000}(C_iC_{i+1})$ and left/right/inside probabilities are also computed from the extended bigram data.

$$P_{t=000}(C_iC_{i+1}) = F_{t=000}(C_iC_{i+1}) / \sum_{t=000}^{t=111} Ft(CiCi+1)$$

Extended bigram data is more sophisticated than the basic bigram data that the accuracy is better than that of the basic bigram data.

3 Segmentation algorithm

3.1 Base Algorithm

The base segmentation algorithm is a HMM model together with the space-insertion probability. HMM model chooses the

[2] Lee(2002) used bigram and trigram data for HMM model which requires a more memory space.

[3] '0' is a non-space tag and '1' is a space tag.

segmentation with the higheset probability. Given a sentence of n characters, $S = c_1c_2...c_n$, has a segmentation of m words, then segmentation probability is estimated as $P(T,S) = P(t_{1,n}, c_{1,n}) \approx \prod_{i=0}^{i=n} p(t_i \mid c_{i-1}c_i, t_{i-2}t_{i-1})$. Our starting point of the word segmentation was the high precision ratio for the Korean language. We first tried to simply applying the extended bigram data with an appropriate threshold. However, it is supposed that there is a limitation of this approach because of the low recall ratio. It caused an adoption of HMM model together with the extended bigram data. Table 1 shows the results of HMM with extended bigram data.

Table 1. Results of HMM with extended bigram

Testing data	Recall	Precision	F-measure
HK	0.924	0.921	0.923
PK	0.902	0.919	0.910
MS	0.942	0.939	0.940

3.2 Postprocessing by trigram data

Extended bigram data in Section 2 consists of 2 adjacent characters and 3 space information (2C3S). In contrast, we may extract trigram data that is constructed by 3 characters and 2 space information(3C2S). This 3C2S trigram data has a form of "X0Y0Z", "X0Y1Z", "X1Y0Z", and "X1Y1Z". That is 3 character sequence XYZ has 4 frequency data. We supposed that "there are frequent 3-character sequences that are biased to one of the spacing pattern".

We verified this supposition by improving the accuracy of the word segmentation result. Table 2 shows the final result of the postprocessing. Postprocessing by trigram data got an increase of both recall and precision. When compared to the base segmentation results of Table 1, F-measures are increased by 0.3%, 0.4%, and 0.8%, respectively.

As an improvement of the system performance, character trigram data has been extracted from the training corpus.

Table 2. Final segmentation results[4]

Testing data	Recall	Precision	F-measure
HK	0.926	0.925	0.926
PK	0.904	0.925	0.914
MS	0.947	0.949	0.948

4 Pure data-driven method without using HMM

Only after submitting the results for bakeoff 2005, we noticed that the accuracy of HMM model is low. It is not clear what the problem is and there is a possibility of the implementation error. So, we looked for a pure data-driven method without using HMM model. The first step in the base segmentation is to apply extended bigram with no space information. In this step, only the spaces with high confidence are fixed and others are marked as 'undecided'.[5] In the second step, extended bigram with space information is applied. Two more postprocessing modules are added for refinements. One of them is to adopt the word-length feature by using the fact that average length of Chinese word is 1.6 characters. The other is to construct 'error dictionary' for the training data. Error dictionary is constructed by running training data and comparing the differences. The context information of error dictionary is four characters (left two and right two characters). The new approach got a better result than that of the final result of bakeoff 2005 as shown in Table 3.

Table 3. Pure data-driven method without HMM

Testng data	Recall	Precision	F-measure
HK	0.933	0.921	0.927
PK	0.912	0.929	0.920
MS	0.952	0.953	0.952

[4] The final results in Table 2 are a bit higher than the bakeoff 2005 results. F-measures of bakeoff 2005 results are 0.921, 0.912, and 0.947, respectively. The reason was not identified. Table 1 and Table 2 are computed by the evaluation program 'score.txt' in the website of SIGHAN bakeoff 2005.

[5] If space generation probability is higher than 0.7, space is inserted. With less than 0.3, space is not inserted, and 'undecided' mark for the range 0.3~0.7,

5 Conclusion

We presented our word segmentation method for the closed track of bakeoff 2005. Our approach is data-driven and language independent. That is, our method is purely statistical method that no language dependent features are applied for tuning or improving the accuracy. Word segmentation system for bakeoff 2005 applied HMM model together with extended bigram and trigram data. The results show that word segmentation problem can be solved with no lexicons or language-dependent resources.

One of the good point of our approach is that data-driven language independent approach is quite acceptable for the word segmentation problem. We also expect that our data-driven method would be a good solution for the enhancement of word segmentation systems as a postprocessing module.

References

Chen, A., Chinese Word Segmentation Using Minimal Linguistic Knowledge, SIGHAN 2003, pp.148-151, 2003.

Gao, J., M. Li, and C.N. Huang, Improved Source-Channel Models for Chinese Word Segmentation, ACL 2003.

Kang, S. S. and C. W. Woo, Automatic Segmentation of Words using Syllable Bigram Statistics, Proceedings of NLPRS'2001, pp.729-732, 2001.

Lee D. G, S. Z. Lee, and H. C. Rim, H. S. Lim, Automatic Word Spacing Using Hidden Markov Model for Refining Korean Text Corpora, Proc. of the 3rd Workshop on Asian Language Resources and International Standardization, pp.51-57, 2002.

Maosong, S., S. Dayang, and B. K. Tsou, Chinese Word Segmentation without Using Lexicon and Hand-crafted Training Data, Proceedings of the 17th International Conference on Computational Linguistics (Coling'98), pp.1265-1271, 1998.

Asahara, M., C. L. Go, X. Wang, and Y. Matsumoto, Combining Segmenter and Chunker for Chinese Word Segmentation, Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing, pp144-147, 2003.

Nakagawa, T., Chinese and Japanese Word Segmentation Using Word-Level and Character-Level Information, COLING'04., pp.466-472, 2004.

Ng, H.T. and J.K. Low, Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based, EMNLP'04.

Shim, K. S., Automated Word-Segmentation for Korean using Mutual Information of Syllables, Journal of KISS: Software and Applications, pp.991-1000, 1996.

Sproat, R. and T. Emerson, The First International Chinese Word Segmentation Bakeoff, SIGHAN 2003.