A Joint Statistical Model for Simultaneous Word Spacing and Spelling Error Correction for Korean

Hyungjong Noh* Jeong-Won Cha** Gary Geunbae Lee*
*Department of Computer Science and Engineering
Pohang University of Science & Technology (POSTECH)
San 31, Hyoja-Dong, Pohang, 790-784, Republic of Korea

** Changwon National University
Department of Computer information & Communication
9 Sarim-dong, Changwon Gyeongnam, Korea 641-773

nohhj@postech.ac.kr jcha@changwon.ac.kr gblee@postech.ac.kr

Abstract

This paper presents noisy-channel based Korean preprocessor system, which corrects word spacing and typographical errors. The proposed algorithm corrects both errors simultaneously. Using Eojeol transition pattern dictionary and statistical data such as Eumjeol n-gram and Jaso transition probabilities, the algorithm minimizes the usage of huge word dictionaries.

1 Introduction

With increasing usages of messenger and SMS, we need an efficient text normalizer that processes colloquial style sentences. As in the case of general literary sentences, correcting word spacing error and spelling error is the very essential problem with colloquial style sentences.

In order to correct word spacing errors, many algorithms were used, which can be divided into statistical algorithms and rule-based algorithms. Statistical algorithms generally use character n-gram (Eojeol[1] or Eumjeol[2] n-gram in Korean) (Kang and Woo, 2001; Kwon, 2002) or noisy-channel model (Gao et. al., 2003). Rule-based algorithms are mostly heuristic algorithms that reflect linguistic knowledge (Yang et al., 2005) to solve word spacing problem. Word spacing problem is treated especially in Japanese or Chinese, which does not use word boundary, or Korean, which is normally segmented into Eojeols, not into words or morphemes.

The previous algorithms for spelling error correction basically use a word dictionary. Each word in a sentence is compared to word dictionary entries, and if the word is not in the dictionary, then the system assumes that the word has spelling errors. Then corrected candidate words are suggested by the system from the word dictionary, according to some metric to measure the similarity between the target word and its candidate word, such as edit-distance (Kashyap and Oommen, 1984; Mays et al., 1991).

But these previous algorithms have a critical limitation: They all corrected word spacing errors and spelling errors separately. Word spacing algorithms define the problem as a task for determining whether to insert the delimiter between characters or not. Since the determination is made according to the characters, the algorithms cannot work if the characters have spelling errors. Likewise, algorithms for solving spelling error problem cannot work well with word spacing errors.

To cope with the limitation, there is an algorithm proposed for Japanese (Nagata, 1996). Japanese sentence cannot be divided into words, but into chunks (bunsetsu in Japanese), like Eojeol in Korean. The proposed system is for sentences recognized by OCR, and it uses character transition probabilities and POS (part of speech) tag n-gram. However it needs a word dictionary and takes long time for searching many character combinations.

[1] Eojeol is a Korean spacing unit which consists of one or more Eumjeols (morphemes).
[2] Eumjeol is a Korean syllable.

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 61–64,
Prague, June 2007.

We propose a new algorithm which can correct both word spacing error and spelling error simultaneously for Korean. This algorithm is based on noisy-channel model, which uses Jaso[3] transition probabilities and Eojeol transition probabilities to create spelling correction candidates. Candidates are increased in number by inserting the blank characters on the created candidates, which cover the spacing error correction candidates. We find the best candidate sentence from the networks of Jaso/Eojeol candidates. This method decreases the size of Eojeol transition pattern dictionary and corrects the patterns which are not in the dictionary.

The remainder of this paper is as follows: Section 2 describes why we use Jaso transition probability for Korean. Section 3 describes the proposed model in detail. Section 4 provides the experiment results and analyses. Finally, section 5 presents our conclusion.

2 Spelling Error Correction with Jaso Transition[4] Probabilities

We can use Eumjeol transition probabilities or Jaso transition probabilities for spelling error correction for Korean. We choose Jaso transition probabilities because there are several advantages. Since an Eumjeol is a combination of 3 Jasos, the number of all possible Eumjeols is much larger than that of all possible Jasos. In other words, Jaso-based language model is smaller than Eumjeol-based language model. Various errors in Eumjeol (even if they do not appear as an Eumjeol pattern in a training corpus) can be corrected by correction in Jaso unit. Also, Jaso transition probabilities can be extracted from relatively small corpus. This merit is very important since we do not normally have such a huge corpus which is very hard to collect, since we have to pair the spelling errors with corresponding corrections.

We obtain probabilities differently for each case: single Jaso transition case, two Jaso's transition case, and more than two Jasos transition case.

In single Jaso transition case, the spelling errors are corrected by only one Jaso transition (e.g. 같애요→같아요 / ㅐ→ㅏ). The case of correcting by deleting Jaso is also one of the single Jaso transition case (나와욧→나와요 / ㅅ→X[5]). The Jaso transition probabilities are calculated by counting the transition frequencies in a training corpus.

In two Jaso's transition case, the spelling errors are corrected by adjacent two Jasos transition (촙오→초보 / ㅂㅇ→Xㅂ). In this case, we treat two Jaso's as one transition unit. The transition probability calculation is the same as above.

In more than two Jaso's transition case, the spelling errors cannot be corrected only by Jaso transition (걍→그냥). In this case, we treat the whole Eojeols as one transition unit, and build an Eojeol transition pattern dictionary for these special cases.

3 A Joint Statistical Model for Word Spacing and Spelling Error Correction

3.1 Problem Definition

Given a sentence T which includes both word spacing errors and spelling errors, we create correction candidates C from T, and find the best candidate C' that has the highest transition probability from C.

$$C' = \arg\max_C P(C \mid T). \qquad (1)$$

3.2 Model Description

A given sentence T and candidates C consist of Eumjeol s_i and the blank character b_i.

$$T = s_1 b_1 s_2 b_2 s_3 b_3 \ldots s_n b_n .$$
$$C = s_1 b_1 s_2 b_2 s_3 b_3 \ldots s_n b_n . \qquad (2)$$

(n is the number of Eumjeols)

Eumjeol s_i consists of 3 Jasos, Choseong (onset), Jungseong (nucleus), and Jongseong (coda). The empty Jaso is defined as 'X'. b_i is 'B' when the blank exists, and 'Φ' when the blank does not exist.

$$s_i = j_{i1} j_{i2} j_{i3} . \qquad (3)$$

(j_{i1} : Choseong, j_{i2} : Jungseong, j_{i3} : Jongseong)

Now we apply Bayes' Rule for C':

$$\begin{aligned} C' &= \arg\max_C P(C \mid T) \\ &= \arg\max_C P(T \mid C) P(C) / P(T) \\ &= \arg\max_C P(T \mid C) P(C). \end{aligned} \qquad (4)$$

[3] Jaso is a Korean character.

[4] 'Transition' means the correct character is changed to other character due to some causes, such as typographical errors.

5 'X' indicates that there is no Jaso in that position.

$P(C)$ can be obtained using trigrams of Eumjeols (with the blank character) that C includes.

$$P(C) = \prod_{i=1}^{n} P(c_i \mid c_{i-1} c_{i-2}), \; c = s \text{ or } b. \quad (5)$$

And $P(T \mid C)$ can be written as multiplication of each Jaso transition probability and the blank character transition probability.

$$P(T \mid C) = \prod_{i=1}^{n} P(s_i \mid s_i^{'})$$
$$= \prod_{i=1}^{n} [P(j_{i1} \mid j_{i1}^{'}) P(j_{i2} \mid j_{i2}^{'}) P(j_{i3} \mid j_{i3}^{'}) P(b_i \mid b_i^{'})]. \quad (6)$$

We use logarithm of $P(C \mid T)$ in implementation. Figure 1 shows how the system creates the Jaso candidates network.

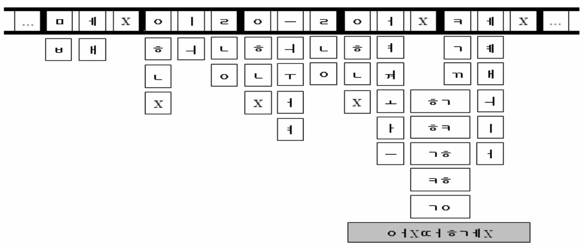

Figure 1: An example[6] of Jaso candidate network.

In Figure 1, the topmost line is the sequence of Jasos of the input sentence. Each Eumjeol in the sentence is decomposed into 3 Jasos as above, and each Jaso has its own correction candidates. For example, Jaso 'ㅇ' at 4th column has its candidates 'ㅎ', 'ㄴ' and 'X'. And two jaso's 'Xㅋ' at 13th and 14th column has its candidates 'ㅎㄱ', 'ㅎㅋ', 'ㄱㅎ', 'ㅋㅎ', and 'ㄱㅇ'. The undermost gray square is an Eojeol (which is decomposed into Jasos) candidate 'ㅇㅓXㄸㅓㅎㄱㅔX' created from 'ㅇㅓXㅋㅔX'. Each jaso candidate has its own transition probability, $\log P(j_{ik} \mid j_{ik}^{'})$[7], that is used for calculating $P(C \mid T)$.

In order to calculate $P(C)$, we need Eumjeol-based candidate network. Hence, we convert the above Jaso candidate network into Eumjeol/Eojeol candidate network. Figure 2 shows part of the final network briefly. At this time, the blank characters 'B' and 'Φ' are inserted into each Eumjeol/Eojeol candidates. To find the best path from the candidates, we conduct viterbi-search from leftmost node corresponding to the beginning of the sentence. When Eumjeol/Eojeol candidates are selected, the algorithm prunes the candidates according to the accumulated probabilities, doing beam search. Once the best path is found, the sentence corrected by both spacing and spelling errors is extracted by backtracking the path. In Figure 2, thick squares represent the nodes selected by the best path.

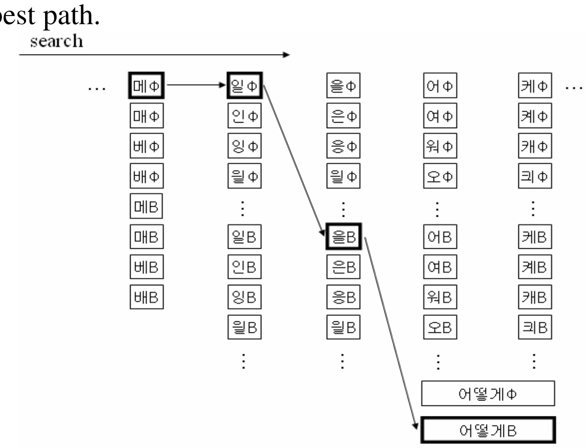

Figure 2: A final Eumjeol/Eojeol candidate network[8]

4 Experiments and Analyses

4.1 Corpus Information

	Training	Test
Sentences	60076	6006
Eojeols	302397	30376
Error Sentences (%)	15335 (25.53)	1512 (25.17)
Error Eojeols (%)	31297 (10.35)	3111 (10.24)

Table 1: Corpus information

Table 1 shows the information of corpus which is used for experiments. All corpora are obtained from Korean web chatting site log. Each corpus has pair of sentences, sentences containing errors and sentences with those errors corrected. Jaso transition patterns and Eojeol transition patterns are extracted from training corpus. Also, Eumjeol n-grams are also obtained as a language model.

[6] The example sentence is "데체메일을어케보내는거지".

[7] In real implementation, we used "a*log$P(j_{ik}|j'_{ik})$ + b" by determining constants a and b with parameter optimization (a = 1.0, b = 3.0).

[8] The final corrected sentence is "대체 메일을 어떻게 보내는 거지".

4.2 Experiment Results and Analyses

We used two separate Eumjeol n-grams as language models for experiments. N-gram A is obtained from only training corpus and n-gram B is obtained from all training and test corpora. All accuracies are measured based on Eojeol unit.

Table 2 shows the results of word spacing error correction only for the test corpus.

	n-gram A	n-gram B
Accuracy	91.03%	96.00%

Table 2: The word spacing error correction results

The results of both word spacing error and spelling error correction are shown in Table 3. Error containing test corpus (the blank characters are all deleted) was applied to this evaluation.

System	n-gram A	n-gram B
Basic joint model	88.34%	93.83%

Table 3: The joint model results

Table 4 shows the results of the same experiment, without deleting the blank characters in the test corpus. The experiment shows that our joint model has a flexibility of utilizing already existing blanks (spacing) in the input sentence.

System	n-gram A	n-gram B
Baseline	89.35%	89.35%
Basic joint model with keeping the blank characters	90.35%	95.25%

Table 4: The joint model results without deleting the exist spaces

As shown above, the performance is dependent of the language model (n-gram) performance. Jaso transition probabilities can be obtained easily from small corpus because the number of Jaso is very small, under 100, in contrast with Eumjeol.

Using the existing blank information is also an important factor. If test sentences have no or few blank characters, then we simply use joint algorithm to correct both errors. But when the test sentences already have some blank characters, we can use the information since some of the spacing can be given by the user. By keeping the blank characters, we can get better accuracy because blank insertion errors are generally fewer than the blank deletion errors in the corpus.

5 Conclusions

We proposed a joint text preprocessing model that can correct both word spacing and spelling errors simultaneously for Korean. To our best knowledge, this is the first model which can handle inter-related errors between spacing and spelling in Korean. The usage and size of the word dictionaries are decreased by using Jaso statistical probabilities effectively.

6 Acknowledgement

This work was supported in part by MIC & IITA through IT Leading R&D Support Project.

References

Jianfeng Gao, Mu Li and Chang-Ning Huang. 2003. *Improved Source-Channel Models for Chinese Word Segmentation*. Proceedings of the 41st Annual Meeting of the ACL, pp. 272-279

Seung-Shik Kang and Chong-Woo Woo. 2001. *Automatic Segmentation of Words Using Syllable Bigram Statistics*. Proceedings of 6th Natural Language Processing Pacific Rim Symposium, pp. 729-732

R. L Kashyap, B. J. Oommen. 1984. *Spelling Correction Using Probabilistic Methods*. Pattern Recognition Letters, pp. 147-154

Oh-Wook Kwon. 2002. *Korean Word Segmentation and Compound-noun Decomposition Using Markov Chain and Syllable N-gram*. The Journal of the Acoustical Society of Korea, pp. 274-283.

Mu Li, Muhua Zhu, Yang Zhang and Ming Zhou. 2006. *Exploring Distributional Similarity Based Models for Query Spelling Correction*. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pp. 1025-1032

Eric Mays, Fred J. Damerau and Robert L. Mercer. 1991. *Context Based Spelling Correction*. IP&M, pp. 517-522.

Masaaki Nagata. 1996. *Context-Based Spelling Correction for Japanese OCR*. Proceedings of the 16th conference on Computational Linguistics, pp. 806-811

Christoper C. Yang and K. W. Li. 2005. *A Heuristic Method Based on a Statistical Approach for Chinese Text Segmentation*. Journal of the American Society for Information Science and Technology, pp. 1438-1447.