Combined Word-Spacing Method for Disambiguating Korean Texts

Mi-young Kang, Aesun Yoon, and Hyuk-chul Kwon

Korean Language Processing Lab., School of Electrical & Computer Engineering,
Pusan National University, San 30, Jangjeon-dong,
609-735, Busan, Korea
{kmyoung, asyoon, hckwon}@pusan.ac.kr
http://klpl.re.pusan.ac.kr

Abstract. In this paper, we propose an automatic word-spacing method for a Korean text preprocessing system in resolving the problem of context-dependent word-spacing. The current method combines the stochastic-based method and partial parsing. First, the stochastic method splits an input sentence into a candidate-word sequence using word unigrams and syllable bigrams. Second, the system engages a partial parsing module based on the asymmetric relation between the candidate-words. The partial parsing module manages the governing relationship using words which are incorporated into the knowledge base as having a high-probability of spacing-error words. These elements serve as parsing trigger points based on their linguistic information, and they determine the parsing direction as well as the parsing scope. Combining the stochastic- and linguistic-based methods, the current automatic word-spacing system becomes robust against the problem of context-dependant word-spacing. An average 8.98% amelioration of the total error rate is obtained for inner and external data.

1 Introduction

Compared with Chinese and Japanese which do not use any spaces or other delimiters excepting punctuation marks within a sentence, Western European languages are somewhat easy to break into meaningful semantic units because 'words'[1] are delimited by spaces.[10], [11] Korean words are delimited by spaces like words in Western European languages. A Korean word can be composed of one or several concatenated morphemes of different linguistic features which are equivalent to a phrase in English. This spacing unit is called a 'word', 'eo-jeol', or 'morpheme cluster' in Korean linguistic literature. In this paper, we adopt the term 'word' in order to refer to 'an alphanumeric cluster of morphemes` located between two blanks in Korean. Korean

[1] The concept of word is not easy to define. Generally, words are linguistically-considered atomic elements: they are the indivisible building blocks of syntax.

G.I. Webb and Xinghuo Yu (Eds.): AI 2004, LNAI 3339, pp. 562–573, 2004.

normative grammar prescribes word-spacing rules. [2] The word-spacing error in Korean is the second-most frequent[3] among other errors that we encounter while processing Korean texts. It is due to the ignorance of word-spacing rules or simple omission of a space when typing. Sometimes, writers commit word-spacing errors on purpose. (For example, users commit word-spacing errors on purpose due to word-count limits from opinion pages in a daily newspaper which should be written with less than 100 words). Violation of those word-spacing constraints in Korean induces linguistic errors and ambiguities in the lexical interpretation of parts of speech, because there are many homographs of different parts-of-speech (POS) in Korean which can be disambiguated only by their spacing status. Resolving those ambiguities created by word-spacing errors, therefore, is crucial in Korean-language-processing application domains such as information retrieval, Optical Character Recognition (OCR), Text-to-Speech Synthesis (TTS), Korean text editing, among others. In information retrieval systems, a morphological analysis of query words or phrases fails when there is a word-spacing error. Also, a correct recognition of word boundary cannot be expected if target documents are not processed with regard to word-spacing. Besides, the correct conversion of Korean text to phonemes in developing TTS is impossible when there are word-spacing errors in the text.

Several studies have proposed automatic word-spacing methods for Korean text. Among these previous studies on word-spacing, we can distinguish (a) rule-/knowledge-based approaches and (b) the stochastic approach. Rule-/knowledge-based approaches use morphological analysis and heuristic linguistic knowledge. [6], [9] Nevertheless, a disadvantage is the amount of language that has to be treated without the possibility of expanding this knowledge base and applying it to natural language processing. Contrary to rule/ knowledge-based methods, the stochastic approach has advantages in set-up time and cost savings and the capability of coping with unregistered words. [4], [5], [7], [10], [11] However, the stochastic method shows a strong training-data-dependency and data sparseness. Data sparseness becomes more serious while processing agglutinative languages such as Korean. In such languages, using syllable statistics at the right-hand boundary of a word exacerbates the data sparseness problem. Therefore, word-spacing errors in Korean need to be processed while taking into consideration Korean language particularity.

For the implementation of an efficient word-spacing system for Korean-text pre-processing, this paper proposes an automatic word-spacing method to resolve the

[2] The following rules are found in The Revised Korean Spelling System and Korean Standard Language (officially announced by the Ministry of Education and which came into effect in March, 1989.):

- a postposition is attached to a preceding noun;
- a dependent noun appears with a space in a sentence;
- a space is placed in every four-digit number in the Korean numeric system; and others.

In addition, the following word-spacing rules are commonly accepted by Korean normative grammar even though they are not prescribed in RKSSKSL:

- a determiner is used with a space on both sides;`
- an adverb appears in a sentence with a space on both sides; and others.

Furthermore, some word-spacing rules in Korean are facultative.

[3] See **Table 3** in the text.

problem of context-dependent word-spacing in Korean. Our current method combines (a) the stochastic-based method which splits an input sentence into a candidate-word sequence using word unigrams and syllable bigrams; (b) the method of dynamic extension of a candidate-word list using the 'longest match strategy' based on the viable-prefix; and (c) the partial parsing module based on the asymmetric relation between candidate words. To achieve this aim, the current paper is composed of five sections. Following this introduction, Section 2 presents our ongoing automatic word-spacing system which uses syllable bi-grams, word stochastic information, and the dynamic extension of a candidate-word list using the 'longest match strategy' based on the viable-prefix and the dynamic selection of candidate words. Section 3 discusses an automatic word-spacing algorithm based on partial parsing with a heuristic linguistic knowledge base and the combined word-spacing system. Section 4 presents the results of our experiments. Finally, in section 5, we give concluding comments and suggestions for future studies.

2 Korean Word-Spacing Based on Word Unigrams and Syllable Bigrams (Ongoing Work)

Our first attempt to construct an automatic word-spacing method for Korean text preprocessing is based on the stochastic information extracted from a large training database which was composed of articles of two different newspaper companies (Corpus A and B) and of three years' worth of news broadcasting scripts (Corpus C). Arabic numerals, Roman-alphabet letters, symbols, among others, were all included in the word count. Seven patterns of Arabic numerals were extracted. First, we extracted five patterns according to one-digit numbers, two-digit numbers, three-digit numbers, four-digit numbers and more than five-digit numbers. Second, every figure located on both sides of one period were grouped in one pattern in order to consider float. Finally, IP addresses, dates (year•month•day), subsection numbering (chapter•section•subsection), times (hour•minute•second) and others used with more than two periods were grouped in one pattern. Each of those patterns was treated as one word with regard to word-spacing.

Table 1. Words and Disyllables in the Training Data (Corpus A, B and C)

Total N° of different word unigrams	1,950,068
Total N° of word unigrams	33,643,884
Total N° of different syllable bigrams	391,732
Total N° of syllable bigrams	90,235,529

Word-Spacing Based on Word Unigrams and Syllable Bigrams. Word-spacing probabilities are estimated, in our current word-spacing system, by using a maximum likelihood estimator with two parameters: word probability *P(W)*, which means in this paper the probability that a sequence could be a possible word estimated by relative word frequencies; odds favoring the inner-spacing probability (P_{innerS}) of a disyllable at the current k$^{\text{th}}$ possible word boundary compared to the rate of no-inner-spacing probability. When an input sentence is given, the most probable word sequence is selected among candidate-word sequences applying the following estimator.[4]

$$\begin{aligned}&\textit{The optimal sentence}\\ &= \arg\max_{S} \prod_{k=1}^{n} P(W_k)\frac{P_{innerS}(LS\ of\ W_k, FS\ of\ W_{k+1})}{1-P_{innerS}(LS\ of\ W_k, FS\ of\ W_{k+1})}\end{aligned} \qquad \text{(Eq. 1)}$$

$$\begin{aligned}&P_{innerS}(LS\ of\ W_k, FS\ of\ W_{k+1})\\ &= \frac{freq(LS\ of\ W_k \# FS\ of\ W_{k+1})}{freq(LS\ of\ W_k \# FS\ of\ W_{k+1}) + freq(LS\ of\ W_k \phi FS\ of\ W_{k+1})}\end{aligned} \qquad \text{(Eq. 2)}$$

If $k = n$, then $P_{inneS}(LS\ of\ W_k,\ FS\ of\ W_{k+1}) = 0.5$; $W_k = k^{th}$ word; FS = First syllable; LS = Last syllable; #: spacing (word boundary); $\emptyset$: absence of spacing.

Smoothing Method Based on Syllable Bigrams. In order to mitigate data sparseness, we use a stochastic smoothing approach using a 'longest match strategy' based on the syllable bigram statistics. This smoothing method estimates the inner-spacing probability of each disyllable starting from the last syllable of a stochastic candidate word, attaches each syllable while the inner spacing probability of each successive disyllable, $P_{innerS}(x_1, x_2)$, $P_{innerS}(x_2,x_3)$,...and $P_{innerS}(x_{n-1},x_n)$, is over a threshold, and selects the longest word among the unseen candidate-words. The inner data varies according to the training corpus. It is extracted at the same distribution ratio as a given corpus in the whole training corpora.

Comparison of Stochastic Models. We can compare our stochastic model's performance with other stochastic-based studies on Korean word-spacing. Lee D.G. et al. (2003) treated word-spacing problems such as POS tagging using a hidden Markov model (HMM), and found the most likely sequence of word-spacing tags T = (t1, t2, …, tn) for a given sentence of syllables S = (s1, s2, …, sn) with the equation: $\arg\max_{T} p(T|S)$. The best result is given by the model using syllable tri-grams: it shows a 93.06% word-unit precision with POS-tagged corpus by ETRI. And a syllable bigram-based model of Kang, S.S. and Woo C.W (2001) shows a 71.22% word-unit precision according to the experiment by Lee D.G. et al. (2003). Compared with these stochastic models, our method using word unigrams and syllable bigrams shows a better performance, a 93.39% word-unit precision, with the test data equivalent to that used by Lee D.G. et al. (2003) (i.e. ExT 2 in Table 4). Some would consider that our system would use more system memory with a stochastic candidate-words list than that using tri-grams. However, as shown in Table 2, our model using word unigrams and syllable bigrams with smoothing based on syllable bigrams requires a relatively small memory size compared with the model using syllable trigrams.

[4] We observed that the computation of logarithms avoids underflow, and that the multiplication of odds favoring the inner spacing probability of a disyllable by the exponent of a power m produces the best performance.

Table 2. Memory Comparison of Stochastic Models

Stochastic Models	Total memory size of the total N° of different words
Syllable Bigram	**4.1**MB
Syllable Bigram + Word Unigram	4.1MB + 25.1MB = **29.2**MB
Syllable Trigram	**63.7**MB

3 Combined Word-Spacing Method Rule- and Knowledge-Based Word-Spacing Module

Our basic system provides a list of possible words based on word unigram information. The word-unigram-based model is an intuitively natural approach to the word-spacing problem and requires a relatively small memory size compared with the model using syllable trigrams. However, the model naturally induces data sparseness because of the agglutinative morphology of Korean and semantic and syntactic ambiguities which can be removed only by considering the candidate-words' contexts. In this section, we propose a method combining the stochastic model and the rule-/knowledge-based model.

Agglutinative Morphology and Linguistic Ambiguity. In Korean, sequences of suffixes are productively and successively attached to the ends of word stems and determine most of the grammatical relations. In the following examples, various suffixes successively attach to the stem *namgi*.[5, 6]

(1) a. namgi-nim-i # nam-gi-si-eoss-da
Namgi-H-Nom /to stay-CS-H-Past-E "Mr. Namgi left something"
b. nam-gi-si-eoss-da-lago # ha-go # us-da
to stay-{CS|*Nol}-H-Past-E-QS /to deliver (a speech)-Conj /to smile-E "to smile saying that (somebody) left (something)"
c. nam-gi to stay-Nol "the staying"

The form *namgi* can be disambiguated only by considering its context. It can be analyzed as a noun considering noun suffixes such as *-nim* <Hon>, *-ga* <Nom>, among others: (1-a) ***namgi***-*nim-i*; as a causative verb stem derived from another verb stem *nam-* considering verb endings or verb pre-endings (verb suffixes) attached to it: (1-a) ***nam-gi***-*si-eoss-da* and (1-b) ***nam-gi***-*si-eoss-da-lago*; and as a derived noun from a verb with nominal suffix *-gi* (1-c). Moreover, we can find some sub-chains of the whole used as spaced words: *#nam-gi-si-eoss-da-lago#*; *#nam-gi-si-eoss-da#*; *#nam-gi#*. This variability of a word boundary produces a higher difficulty in processing the word-spacing of Korean than of English, in which fewer morphemes for inflec-

[5] Throughout this paper, we adopt the Revised Romanization of Korean, released on July 4, 2000 by the South Korean Ministry of Culture and Tourism.

[6] Symbols and Abbreviations: |: separation; -: morphological boundary; *: unacceptable form; { }: alternative elements; W: word; Acc: accusative; Auxvb: auxiliary verb; Adv: adverb; CS: causative suffix; Conj: conjunction; D: determiner; E: verbal ending; H: honorific suffix; N: noun; Nom: nominative; Nol: nominal suffix; Past: past tense; Post: postposition; Pfx: Prefix; QS: quotation suffix used in indirect narrative; Sfx: suffix;

tion simultaneously encode several meanings. [1] A large number of conjugated forms of each Korean verb are possible considering all the possible combinations among verbal stems, verbal pre-endings and verbal endings.

Word-Spacing Errors and Linguistic Ambiguities. There is a word-spacing problem due to linguistic ambiguities that the statistical method can hardly overcome and can only be resolved by considering enough of the context of the word being checked.

(2) Ambiguity between a postposition and a noun
a. banghag{-|*#}nae # don-eul # da # sseo{-|#}beoli-da
vacation-{all through<Sfx>|*inside<N> /money-Acc /all<Adv> /to spend-Conj{-|#}Auxvb-E “Somebody spent all his money all through the vacation.”
b. haggyo # nae.
school /inside<N> “the inside of school”
(3) Ambiguity between a prefix and a noun
a. geum{-|*#}segi # choego{-|#}haengsa
{this<Pfx>|*gold<N>}-century /the greatest<N> /event<N> “The greatest event of the century”
b. geum # paljji gold /wristlet “a gold wristlet”
(4) Ambiguity between a determiner and a part of a noun
a. chongal # su{*-|#}bal-eul # sso-da
ball /{*care|many<D>}/round-acc /to shot-E “to shot many rounds of shot”
b. hwanja # subal invalid /care “care for an invalid”

The same morpheme, *nae* in (2), can be interpreted as a postposition ‘all through’ when it is used without a space on its left side, or as a noun ‘inside’ when it is used with a space. The same morpheme *geum* in (3) can prefix ‘this’ without a space on its right side or the noun ‘gold’ with a space. The forms *subal* and *su#bal* in (4) differ only regarding the inner space. They are analyzed as a noun and a sequence of a determiner ‘many rounds of shot’ and its determined noun.

Constructing Spacing-Error Checking Rules. In order to mitigate data-sparseness due to the agglutinative morphology of Korean and remove linguistic ambiguities, it is necessary to understand the factors that influence spacing-errors found in real texts. Table 3 shows the summary of frequencies for each error type identified by our recent experiment on an earlier system which was carried out with Corpus A of approximately 19 million words (see Table 1). The result shows that 1,404,777 words were rejected by running a Korean grammar checker.[7]

Table 3. Error Types in Web-Documents

Error Types	N° of Erroneous Words	Frequency (%)
Grammatical Errors	222,516	15.84
Spacing Errors	**334,899**	**23.84**
Spelling Errors	846,940	60.29
Verbal Conjugation Errors	421	0.03
Total	1,404,777	100.00

[7] The Korean Grammar Checker is used in Pusan National University [3] .

According to the results in Table 3, the word-spacing error in Korean is the second-most frequent among other errors that we encounter while processing Korean texts. Based on statistical results and heuristic evidence obtained while treating various word-spacing error types in the texts, error analysis makes possible the building of an adequate knowledge base. These linguistic conditions can be used to make correction rules based on a morph-syntactical analysis or on collocation and anti-collocation, which are then incorporated into the knowledge base. Each rule is composed of a Word Potentially Involved in Word-spacing Error (WPI) and its one or more targets' information. The WPI triggers partial parsing. While the singularity of the WPI should be respected, the target can be the subject of several rules: an eventual dependent could be a target of a different WPI. Our current system provides about 800 words or patterns as WPIs which trigger partial parsing.

Rule (number) =
{Word Potentially Involved in Word-spacing Error, Parsing Direction;
Negligible Words' Linguistic Information;
Compatible or Incompatible Words' Linguistic Conditions;
Correction (Splitting, Attachment)}

Fig. 1. Knowledgebase of Word-spacing Rules

There are approximately 1,541 rules in our system that are related to context-dependant word-spacing error correction. And, 9,407 words or patterns and 15,679 rules are dedicated to compound nouns.

Word-Spacing Error Correction Based on Partial Parsing Method. We engage the partial parsing method based on the rule-/knowledge-based approach in order to resolve context-dependant word-spacing problems. The orientation of the parsing to detect and correct grammatical and semantic errors in our system is constructed with respect to a WPI [8]. The partial parsing module is triggered when this WPI is detected. Partial parsing proceeds from a selected WPI until its target (i.e. a word that forms a collocation or anti-collocation with it), or no other possible targets, are found. The word-spacing module implemented with partial parsing provides three possible checking directions a posteriori: right-hand parsing, left-hand parsing, and conditional parsing.

[8] Our partial parsing method deals with knowledge based on dependency grammar. Nevertheless, this is rather far from classical dependency grammar. Whereas dependency grammar describes order and restrictions in the same formula, our system tolerates different descriptions for each case. But in the current word-spacing module, the governor in our system is not conceived in a linguistic sense but as a word potentially involved in word-spacing error (WPI) that the system identifies when checking a sentence and which selection as governor doesn't burden the spacing system since the system always has to have the fewest rules possible in order to avoid lowering the efficiency of the system. We refer reader to [3] for further discussion on the parsing trigger and parsing direction.

Consider first the WPI having a ‘right-hand headed relation’ with their targets. For example, we can find many spacing errors between the POS of a homograph, *nae* in (2-a). It can be attached only to a noun [+time] as a suffix. Otherwise, it should be spaced from its left-hand noun as another noun. The construction of Korean compound nouns such as *choego{- | #}haengsa* in (3-a) is endocentric, with the possible exception of coordinated structures. It means that they have a head on the right-hand side in a construction. This linguistic aspect is reflected as such in the parsing direction of our system. Our system controls the semantic and lexical scope of each noun belonging to a compound noun in order to check if the collocation relation is established or not. The form *geum* in (3-a) is often mis-split in real texts, producing semantic ambiguity. The form is included as a WPI, in the knowledge base, having a ‘left-hand headed relation’ with its targets. When it is attached to a noun [+time], there is compatibility between the WPI and the target. Otherwise, the morpheme is interpreted as a noun. Finally, we have a ‘conditional WPI’. We call a ‘conditional WPI’ a syntactic entity that can govern its target regarding the semantic or morpho-syntactic state of the element situated on its other wing. Many Korean writers confuse *subal* with *su#bal*. Thus *subal* in (4-a) triggers parsing as a WPI and selects a verb *sso-da* ‘to shoot’ as its target. However, the WPI needs to satisfy the co-occurrence condition with the item on its other wing, *chongal* ‘ball’. *subal* is not compatible with *chongal*. Thus correction of *su#bal* by splitting proceeds.

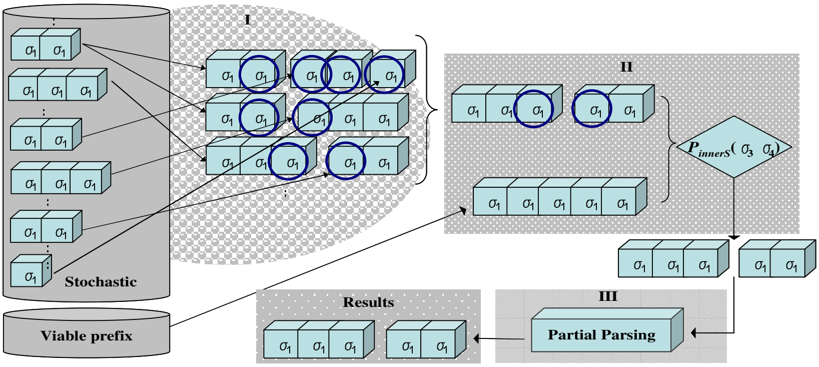

Fig. 2. Combined Word-spacing Module

Combined Word-Spacing Model. The current combined word-spacing model provides: (I) the stochastic-based method which splits an input sentence into a candidate-word sequence using word unigrams and syllable bigrams; (II) the method of dynamic extension of the candidate-word list using the ‘longest match strategy’ based on the viable-prefix which searches the longest candidate-word which can be analyzed morphologically. The extension does not proceed using the stochastic method based on syllable bigrams in the combined model, in the interests of the efficiency of partial parsing. The ‘longest match strategy’ based on the viable-prefix suggests dynamically possible words and includes them among possible kth candidate words, and the system dynamically selects a candidate-word by estimating the inner-spacing probability of disyllables located at the boundary of stochastic-based words (i.e. σ_3 and σ_4 at the step II in the Figure 2). If this value is under the threshold, the system selects the word from the stochastic word list and, if the value is over the threshold, the system selects the longest-radix word; and (III) the partial parsing method based on the asymmetric relation between candidate words selected by applying methods (I) and (II). The word-spacing module based on statistic and partial parsing is depicted as follows:

4 Experimentation

For the test of our stochastic word-spacing model's performance, two types of test data were provided: (a) the inner test data extracted at the same distribution ratio as a given corpus over the whole training data, Corpus A, B, and C, and (b) three external test data, each extracted from the Sejong Project's processed and balanced corpus (ExT 1), the POS-tagged corpus by ETRI(Electronics and Telecommunications Research Institute) (ExT 2): and from special opinion pages in a daily newspaper on the web of less than 100 words (Ext 3).

Table 4. Test Data Suite

	N° of Sentences	N° of Words	N° of Syllables
Inner	2,000	25,020	103,196
ExT 1	2,000	13,971	40,353
ExT 2	2,000	17,191	52,688
ExT 3	2,000	12,504	40,088

Table 5. Experimental Results (PP: Partial parsing)

		Stochastic Model Without Smoothing	Knowledge-based Dynamic Extension	With Partial Parsing
InT	*Pw*	98.45%	98.46%	98.52%
	Rw	98.19%	98.00%	97.98%
ExT 1	*Pw*	90.91%	97.81%	98.07%
	Rw	93.63%	97.77%	97.95%
ExT 2	*Pw*	90.43%	98.88%	99.02%
	Rw	94.61%	99.01%	99.12%
ExT 3	*Pw*	86.88%	93.60%	94.09%
	Rw	90.89%	94.73%	95.01%

The system test was preceded by removing spaces from the input test data and selecting according to the following two evaluation measures: (a) correctly spaced words compared to the total number of words created by the system (word-unit precision, *Pw*); and (b) correctly spaced words compared to the total number of words in the test document (word-unit recall, *Rw*). The result of stochastic automatic word-spacing with dynamic expansion of the candidate-word list using stochastic smoothing is shown in the following Table.

The system thus becomes robust against ambiguity that would be encountered because of word-spacing errors while processing Korean texts. And a similar performance is provided for inner and balanced standard external data (ExT 1) (98.52% and 98.07% respective word-unit precisions). The performance observed with text from opinion pages in a daily newspaper (ExT 3) is lower than that with other test data. This kind of text is especially intricate because it contains stylistic errors that require complete reconstitution of sentences. In these types of text, due to the constraint of the number of words, users commit spacing errors purposely and produce especially intricate text. Therefore, even though we could obtain only a 94.09% word-unit precision, which is lower than with other test data, it is very encouraging in preprocessing Korean texts.

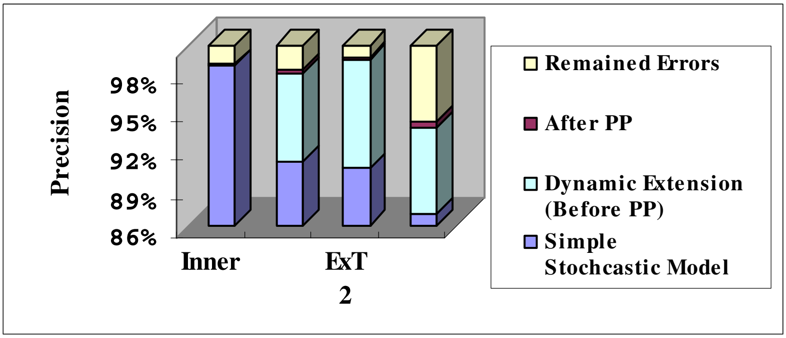

Fig. 3. Amelioration of total error rate

By smoothing with the dynamic extension of the candidate-word list based on the viable prefix, a high amelioration of performance was observed for the inner and the external data. The system thus compensates for the data sparseness problem in the simple stochastic method. Figure 3 shows that, according to different test data, about 2.81% of words are not processed correctly after applying methods (I) and (II). Most of those errors are context-dependant word-spacing problems. Implementing the combined word-spacing method using partial parsing, context-dependant word-spacing problems are resolved and 3.90%, 11.87%, 12.50% and 7.66% amelioration is obtained on the total error rate for Inner, ExT 1, ExT 2 and ExT 3 respectively as depicted in Figure 3.

6 Conclusions and Further Research

In this paper we have implemented an automatic word-spacing system which combines (a) a stochastic-based method based on word unigram and syllable bigram, (b) a normalization of data sparseness in providing a candidate-words list extension using the 'longest match strategy' based on the viable-prefix and (c) the rule-/knowledge-based method using partial parsing. This combined method efficiently (a) normalizes data sparseness due to Korean agglutinative morphology where chains of suffixes are commonly attached to the ends of stems, and therefore there is considerable risk in using the syllable statistics for the right-hand boundary of a word, and (b) resolves semantic and syntactic ambiguities concerning word-spacing which remain after applying only stochastic and dynamic candidate-words list extension. Using partial parsing, semantic and syntactic ambiguities are removed and an average 8.98% amelioration of the total error rate is obtained for inner and external data. Thus our method becomes robust against homographs, the meanings of which can only be disambiguated by spacing.

Our further research will develop a predictive algorithm for unseen words in order to extract linguistic categorical information from a training corpus of different types of texts so that it can be applied to Korean text preprocessing, and to define the optimal combining algorithm between the statistical spacing method and the rule-based spacing method.

Acknowledgements

This work was supported by National Research Laboratory Program (Contract Number: M1-0412-00-0028-04-J00-00-014-00).

References

1. Comrie, B.: Language Universals and Linguistic Typology, Blackwell (1989)
2. Grefenstette, G.: What is a Word, What is a Sentence, Problems of Tokenization Proceedings of the conference on computational lexicography and text research (1994)
3. Kang, M.Y., Yoon, A.S., Kwon, H.C.: Improving Partial Parsing Based on Error Pattern Analysis for Korean Grammar Checker, TALIP Volume 2, Issue 4, ACM (2003) 301-323
4. Kang, M.Y., Choi, S.W., Kwon, H.C.: A Hybrid Approach to Automatic Word-spacing in Korean, LNCS Vol.3029 (2004) 284 - 294
5. Kang, S.S., Woo C.W.: Automatic Segmentation of Words Using Syllable Bigram Statistics. Proceedings of 6th Natural Language Processing Pacific Rim Symposium (2001) 729-732
6. Kang, S.S.: Korean Morphological Analysis and Information Retrieval, Hongleunggwahag Publisher, Seoul (2002)
7. Lee, D.G., Lee, S.Z., Lim, H.S., Rim, H.CH.: Two Statistical Models for Automatic Word Spacing of Korean Sentences, Journal of KISS(B): Software and Applications, Vol. 30. 4, (2003) 358~370

8. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processin The MIT Press, Cambridge London (2001)
9. Sim, CH.M., Kwon, H.CH.: Implementation of a Korean Spelling Checker Based on Collocation of Words, Journal of KISS(B): Software and Applications, Vol. 23. 7 (1996) 776-785
10. Teahan, W. J., McNab R., Wen, Y., Witten, I. H.: A compression-based algorithm for Chinese word segmentation, Computational Linguistics, Vol 26. 3 (2000) 375 – 393
11. Tsutsumi, J., Nitta, T., Ono, K., Jiang, S.D., Nakaishi, M.: Segmenting a Sentence into Morphemes using Statistic Information between Words, Proceedings of Coling'94 (1994) 227-233
12. Korean Standard Pronunciation Dictionary, Edited by Kim, S.T et al., Eomungak (1993)