International Journal of Fuzzy Logic and Intelligent Systems, vol. 4, no. 3, December 2004 pp. 306-310

Design and Implementation of the Compound Noun Segmentation Algorithm Based on Statistical Information

Chang-Geun Kim*, Han-Ho Tack**

*** Department of Computer Science, Jinju National University**

**** Department of Electronic Engineering, Jinju National University**

Abstract

This paper suggests a reverse segmentation algorithm using affix information and some preference pattern information of Korean compound nouns. The structure of Korean compound nouns is mostly derived from Chinese characters, and it includes some preference patterns utilized as a segmentation rule in this paper. To evaluate the accuracy of the proposed algorithm, an experiment was performed with 36,061 compound nouns.

The experiment resulted in getting 99.3% of correct segmentation and showed excellent satisfactory results from the comparative experimentation with other algorithms. Especially, most of the four-syllable or five-syllable compound nouns were successfully segmented without fail.

Key words : reverse segmentation algorithm, five-syllable compound nouns

I. Introduction

A compound noun in Korean is a noun where two or more unit nouns are connected without leaving space. It is all right in everyday life but causes serious problems in the information retrieval system or the machine translation system dependent on dictionaries. That's why it needs to be properly processed.

Two ways are possible. Either each compound noun is individually registered or it can be segmented at the level of unit nouns. However, the former is realistically impossible due to its unlimited creation. So, various algorithms have been suggested as to the segmentation methods[6, 7, 8, 9, 10, 11, 12].

In segmenting compound nouns, polysemy and unregistered words should be handled. This phenomenon also appears in Japanese and other languages in the countries of the bloc of the Chinese-character culture.

Chinese needs the process of making the borders between words. Correct word separation from the input sentence is important[1, 2]. Also, in Japanese, compound noun segmentation is needed and diverse semantic and syntactic data are used to reveal the structure of compound nouns. But a great problem is how to efficiently deal with the unregistered words nonexistent in dictionaries[3, 4].

In this paper, an algorithm to utilize statistical information is suggested as most Korean compound nouns come from Chinese characters composed of the unions of favorite syllables. Here, a reverse segmentation algorithm is utilized to apply collected statistics.

Manuscript received Nov. 20, 2004; Revised Dec. 11, 2004.

II. Related Research

Mostly heuristic, the compound noun segmentation methods are divided into those based on dictionaries[7, 11] and statistical processing[5, 8, 12]. The former is segmenting the nouns into a few unit nouns and using empirical processing to defeat polysemy. The latter is finding each word's appearance rate by way of a corpus and obtaining statistical information to parse compound nouns.

Yi Hyeon-min[6] with a heuristic method and Yun Bo-hyeon[12] with statistical information and preference rules show comparatively superior results. But they reveal comparatively low processing results as to polysemous segmentation and unregistered nouns. Considerations to solve polysemy are also somewhat lacking in other papers.

III. Segmentation Algorithm

How to process polysemy and unregistered words matters. First, in polysemous segmentation, it is important to choose a right union among some possible noun unions. For example, '특기적성교육' can be either '특기+적성+교육' (special-skill & aptitude education) or '특기적+성교육' (special-skill sex education). But the first one is selected.

Second, as in '시운전속력' (test run speed), '운전' (run) and '속력' (speed) are in the dictionary. But '시운전+속력' is the correct way of segmentation.

3.1 Suggestions for Compound Noun Segmentation

When all possible unions of unit nouns are put in the dictionary, the same words can increase multiple meanings. If '시운전속력' is divided into '시운전' and '전속력' (full speed), we are at a loss what to choose. To solve this problem of right segmentation, the following ways are proposed.

3.1.1 Use of a Noun Dictionary Except Monosyllable Nouns

Meaningful words in Korean are hardly composed of monosyllables, but made up of 2 or 3 Chinese characters. Thus a noun dictionary excluding monosyllables is utilized.

3.1.2 Priority to the Most Agreeable Nouns

For instance, '민주주의' (democracy) as well as '민주' (democratic) and '주의' (ideology) exist in the dictionary. But '민주주의' is chosen as it is the most agreeable noun.

3.1.3 Reverse Segmentation Using Affix Information

When the most agreeable noun is not found, each syllable is removed from the reverse direction. Suffixes and prefixes are checked in this order because most unregistered words are derivatives by affixation.

3.1.4 Segmentation by Preference Patterns

Compound nouns are mostly grounded on Chinese characters, holding preferred length of syllables. Korean has such pronunciation rules as an acronym rule, consonant shuffle, forward assimilation, and backward assimilation. Thus, statistical preference syllable information based on Korean grammar is used. For example, '수신제가치국평천하' can be segmented into '수신'+'제가'+'치국'+'평천하' (training myself, controlling my family, governing the country, and ruling the world).

In this paper, reverse segmentation, segmentation through affixes and favorite patterns, and segmentation via statistical information are applied by turns.

3.2 Dictionary Composition & Statistical nformation

3.2.1 Unit Noun Dictionary

Based on the single noun dictionary, the dependent noun dictionary, the compound noun dictionary, and the affix dictionary compiled by the National Korean Research Institute, unit nouns were produced to use here as a dictionary. As stated earlier, monosyllables were excluded as they increase a polysemous phenomenon.

3.2.2 Affix Dictionary

On the basis of the affix dictionary made by the Project Sejong 21 of the National Korean Research Institute, a dictionary of affixes was created, including prefixes, suffixes, prefixes/suffixes, and preferred first/last syllables. Table 1 shows those affixes and affix-acting syllables with the frequency of 5 or more times to appear first or last.

3.2.3 Segmentation-Preference Syllable Patterns

As the statistical data of syllable length, three highest patterns were drawn from the head words of the compound dictionary (National Korean Research Institute) and 6 encyclopedias, and summed up in Table 2.

3.3 Segmentation Algorithm Using Statistical Information

Fig. 1 depicts the overall flow of compound noun segmentation by using the preference rules suggested in this paper.

Table 1. Category of affixes

Division	Affix classes
Prefixes	가, 고, 과, 당, 대, 명, 무, 미, 반, 부, 불, 비, 생, 소, 신, 역, 재, 저, 전, 정, 주, 준, 초, 총, 최, 타, 탈, 피, 한, 항, 헛
	The prefix is regarded as a syllable to correspond to the head sound >= the tail sound * 5
Suffixes	가, 각, 간, 계, 고, 곡, 관, 구, 국, 권, 금, 기, 군, 꾼, 네, 님, 답, 대, 댁, 도, 력, 령, 로, 록, 론, 료, 류, 률, 만, 망, 물, 미, 민, 방, 배, 법, 보, 복, 부, 비, 사, 산, 상, 서, 석, 선, 설, 성, 소, 수, 술, 식, 실, 액, 어, 용, 째, 적, 제, 차, 축, 풍, 학, 해, 행, 형, 호, 화
	The suffix is regarded as a syllable to correspond to the head sound *5 <= the tail sound

What matters is that the algorithm is not sure whether the present method of segmenting compound nouns into unit nouns is right or not. So, in this study, statistical data for preference rules, consonantal frequency information, and segmentation-preference sy lable patterns are used to solve the matters of polysemy and unregistered words.

Table 2. Preference patterns

Number of affixes	Preference pattern with high frequency
3	1+2/ 2+1, Null
4	2+2/ 1+3/ 3+1
5	2+3/ 3+2/ 1+4
6	3+3/ 2+4/ 2+2+2
7	2+2+3/ 3+2+2/ 2+3+2
8	2+3+3/ 3+3+2/ 3+2+3
9	2+2+2+3/ 3+2+2+2/ 2+3+2+2
10	2+2+2+2+2/ 3+2+2+3/ 2+2+3+3
11 or more	Odd number: First pattern of 2+2+…+3/ 3+2+…+2+2/ n-2 number Even number: First pattern of 2+2+…+2/ 3+2+…+2+3/ n-2 number

The work of the preferred syllable processing routine implies the existence of unregistered words. The current algorithm tries segmentation according to the syllable information patterns in Table 2. When the segmented words are all in the unit noun dictionary, segmentation is completed. If even one unregistered word is seen, the forced segmentation

routine is practiced to divide it into 2 or 3 syllables.

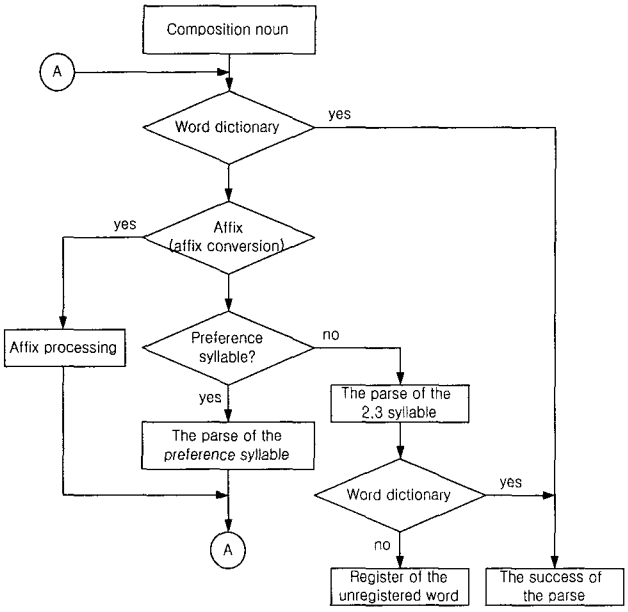

Fig. 1. Proposed algorithm

IV. Experiments & Analysis

4.1 Experimental Data

In order to assess the algorithm capacity suggested here, 36,061 compound nouns were extracted from the entries in 6 main encyclopedias at home. Those deriving from historic incidents and proper nouns were excluded if possible. Table 3 indicates the syllable rates of the collected compound nouns.

Table 3. Length of compound nouns

Number of syllables	Number of compound nouns	Structure rate (%)
3	114	0.32
4	23,564	65.29
5	7,817	21.66
6	2,833	7.85
7	1,137	3.15
8	371	1.03
9	161	0.45
10	47	0.13
11 or more	48	0.13
Total	36,092	100.00

4.2 Experiments

When the prefix- or suffix-related rates were processed from 2 to 10, the highest segmentation was shown at 5. Therefore, it was chosen as the rate of affixation.

Table 4. Comparison of segmentation rates with affixes

Suffix \ Prefix	3	4	5	6	7	8	9	10
3	97.873	98.239	98.948	98.750	98.576	98.011	98.113	97.295
4	97.899	98.246	99.053	99.008	98.641	98.322	98.254	97.776
5	98.121	98.451	99.343	99.177	98.767	98.579	98.423	98.285
6	98.033	98.484	99.301	99.020	98.761	98.577	98.409	98.285
7	98.001	98.413	99.301	98.985	98.655	98.437	98.403	98.128
8	98.001	98.402	99.005	98.985	98.548	98.229	98.402	98.006
9	97.998	98.376	99.000	98.944	98.333	98.100	97.878	97.994
10	97.939	98.374	98.998	98.926	98.239	98.046	97.874	97.866

Reverse segmentation obtained 98.2% of successful parsing, while the proposed algorithm earned 99.3%. Table 5 shows the results of the syllable numbers of the experimented compound nouns. Fig. 2 reveals the comparison of the algorithm results.

Table 5. Results of segmentation

Number of syllables	Number of compound nouns	Number of parsing successes	Rate
3	114	109	0.96
4	23,564	23,562	0.999
5	7,817	7,780	0.995
6	2,833	2,724	0.96
7	1,137	1,092	0.96
8	371	338	0.91
9	161	151	0.94
10	47	44	0.94
11 or more	48	42	0.88
Total	36,092	35,842	99.3

4.3 Analysis

As Fig. 2 indicates, the proposed algorithm using the preference rules had higher precision rate than reverse segmentation. In particular, the parsing rate was near 100% in the 4th and 5th syllables.

However, like the established methods, the proposed algorithm here showed a phenomenon of decreasing parsing rates according to the increase of syllables. The data for the nouns with 10 or more syllables were hard to find, and a few failures greatly affected the results. Most of the compound nouns had proper nouns and unregistered words before and after unit nouns. This kind of error took place in the compound nouns of other length.

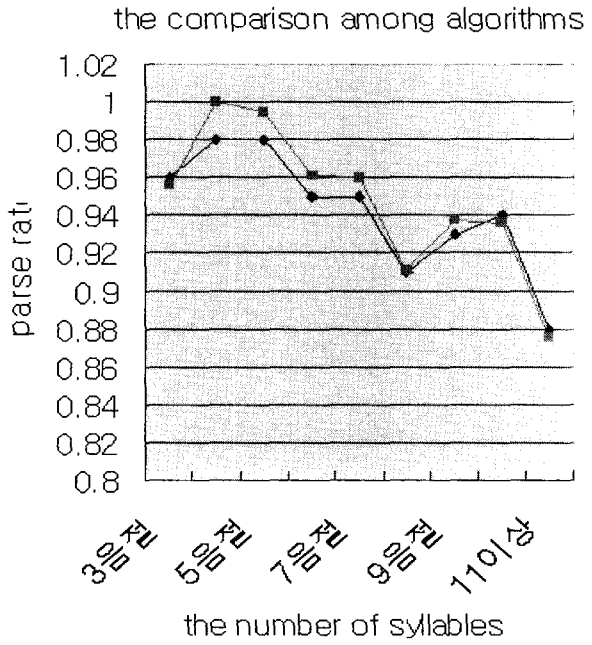

Fig. 2. Comparison of algorithms

V. Conclusion

Based on statistical information, this paper suggested and experimented compound noun segmentation algorithm. Along with dictionary retrieval, a dictionary of unit nouns with 2 or more syllables was used. To prevent polysemy, monosyllables were avoided. To deal with prefixes and suffixes, a dictionary of affixation was utilized.

With the compound nouns drawn from the Korean dictionary, encyclopedias, and Internet retrieval, about 99.3% of parsing rate was earned. The unregistered words among the compound nouns were affix derivatives, and the proposed segmentation methods produced rather high parsing precision.

Further research is needed for foreign-imported words. When registered nouns exist in the middle of unregistered words, this should be ignored. But the parsing rate dropped when the current algorithm was applied to disregard this. So, some additional research is necessary for capacity improvement.

References

[1] K. J. Chen & S. H. Liu, "Word Identification for Mandarin Chinese Sentences," *Proceedings of the 14th International Conference on Computational Linguistics*, pp. 101~107, 1992.

[2] R. Sproat, C. Shih, W. Gale & N. Chang, "A Stochastic Finite-state Word Segmentation Algorithm for Chinese," *Proceedings of ACL*, 1994.

[3] K. Yosiyuki & T. Hozumi, "Analysis of Japanese Compound Nouns Using Collocation Information," *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 865~869, 1994.

[4] T. Hisamitsu & Y. Nitta, "Analysis of Japanese Compound Nouns by Direct Text Scanning," *Proceedings of the 16th International Conference on Computational Linguistics*, pp. 550~555, 1996.

[5] Bo-Hyun Yun, Ho Lee & Hae-Chang Rim, "Analysis of Korean Compound Nouns Using Statistical Information," *Proceedings of the 1995 International Conference on the Computer Processing of Oriental Languages*, pp. 76~79, 1995.

[6] Yi Hyeon-min & Pak Hyeok-ro, "Reverse Segmentation Algorithm of Compound Nouns," *Korea Information Processing Society, Journal B*, No. 8-B, Vol. 4, 2001.

[7] Kang Seung-sik, "Segmentation Algorithm of Korean Compound Nouns," *Korea Information Science Society, Journal B*, Vol. 25-1, pp. 172~182, 1998.

[8] Sim Gwang-seop, "Parsing of Compound Nouns by Using the Composed Mutual Information," *Korea Information Science Society, Journal B*, Vol. 24-117, pp. 1307~1317, 1997.

[9] Pak Hyeok-ro & Shin Jung-ho, "Analysis of Korean Compound Nouns by Using biverty Learning Algorithm," *Korea Information Science Society*, 1997.

[10] Sim Gwang-seop, "Automatic Korean Spacing by Using Mutual Syllable Information," *Korea Information Science Society, Journal B*, Vol. 23-9, pp. 991~1000, 1996.

[11] Choe Jae-hyeok, "Parsing of Korean Compound Nouns According to Syllable Numbers," *8th Hangeul and Korean Information Processing Seminar Synopsis*, pp. 262~267, 1996.

[12] Yun Bo-hyeon, Jo Jeong-min & Im Hae-chang, "Parsing of Korean Compound Nouns by Using Statistical Information and Preference Rules," *Korea Information Science Society, Journal B*, Vol. 24-8, pp. 925~928, 1995.

Chang-Geun Kim

Chang-Geun Kim was born November. 20. 1962. He received the B.S. degree in Department of Computer Statistic from Gyeongsang National University, Jinju, Korea, in 1985. He received the M..S. degree in Department of Computer Science Gyeongnam University, Masan, Korea, in 1991. He received Ph. D. degree in Department of Computer Science from Gyeongnam University, Masan, Korea, in 1999. Since 1995, he has been a faculty member of the Department of Computer Science at the Jinju National University, where he is currently a associate Professor. His research interests are Neural Network, Fuzzy System, Computer networks, and Multimedia Communication etc. He is a member of KIMISC, KMS, and KIEE.

Phone : +82-55-751-3324
Fax : +82-55-751-3329
E-mail : cgkim@jinju.ac.kr

Han-Ho Tack

Han-Ho Tack was born july 6. 1959. He received the B.S. degree in Department of Electronic Engineering from Pukyung National University, Busan, Korea, in 1987 He received the M..S. degree in Department of Electronic Engineering from Dong-A University, Busan, Korea, in 1992. He received Ph. D. degree in Department of Electronic & Communication Engineering from the Korea Maritime University, Busan, Korea, in 1998. From 1987 to 1989, he was a Researcher at the Laboratory of Hung Chang Co. Ltd. Since 1991, he has been a faculty member of the Electronic Engineering at the Jinju National University, where he is currently a Professor. His research interests are Neural Network, Fuzzy System, Robotics, Factory Automation, Mechanical Vibration, Transfortation, and Multimedia System etc. He is a member of IEEE, KIMISC, KMS, KIEE, and KFIS.

Phone : +82-55-751-3332
Fax : +82-55-751-3339
E-mail : fmtack@jinju.ac.kr