A Probabilistic Approach to Compound Noun Indexing in Korean Texts

Hyouk R. Park and **Young S. Han** and **Kang H. Lee**
Korea R&D Information Center/KIST
P.O. Box 122 YuSong Taejon, 305-600, Korea
{*hrpark,yshan,khlee*}*@stissbs.kordic.re.kr*

Key-Sun Choi
Computer Science Department KAIST
YuSong Taejon, 305-701, Korea
kschoi@world.kaist.ac.kr

Abstract

In this paper we address the problem of compound noun indexing that is about segmenting or decomposing compound nouns into promising index terms. Compound nouns as index terms that usually subscribe to specific notions tend to increase the precision of retrieval performance. The use of the component nouns of a compound noun as index terms, on the other hand, may improve the recall performance, but can decrease the precision.

Our proposed method to handle compound nouns with a goal to increase the recall while preserving the precision computes the relevance of the component nouns of a compound noun to the document content by comparing the document sets that are supported by the component nouns and the terms of the document. The operational content of a term is represented as the probabilistic distribution of the term over the document set.

Experiments with a set of 1,000 documents show that our method gains 33% increase of retrieval performance compared to the indexing method without compound noun analysis, and is as good as manual decomposition by human experts.

1 Introduction

Automatic indexing renders a form of document representation that visualizes the content of the document more explicitly. Indices that are carefully chosen to represent a document will bring about the improvement of retrieval performance in accuracy and time efficiency. The potential of a candidate index is often judged on the basis of its discriminating power over a document set as well as its linguistic significance in the document. Thus, a good index term should distinguish a certain class of documents from the rest of the documents and be relevant to the subject matters of the class of documents to be indexed by the term.

In general, automatic indexing consists of the identification of index terms and the assignment of weights to the terms (Salton 1983).

An index term can be either a simple noun or a compound noun composed of more than one simple nouns. Compound nouns tend to carry more specific contextual information than simple nouns, thus they are likely to contribute to the retrieval precision. Compound nouns may contain useful simple nouns that usually refer general contexts, and thus will boost the recall of retrieval. Processing compound nouns is decomposing them into simple nouns and evaluating the simple nouns as potential index terms. In both identifying and evaluating index terms, compound nouns require a different strategy from that for simple nouns. The identification of compound nouns involves a certain degree of linguistic or statistical analysis that varies from simple stemming to morphological analysis (Fagan 1989).

What makes it even more complicated to handle compound nouns in Korean documents lies in the convention of writing compound nouns. In Korean, it is allowed to write compound nouns with or without intervening blanks between constituent nouns. Arbitrarily long compound nouns are possible and not rare in real texts. The decomposition of a compound noun is particularly problematic because of the severe ambiguity of segmentations.

In this paper, we propose a method to identify and evaluate the candidate index terms from compound nouns. First, each possible decomposition of a compound noun is identified. To see the potential of the component nouns of the decomposition, we observe how the component nouns are distributed over the total document set, and

also examine how the simple and compound nouns of the current document are distributed over the same document set. The similarity of the two distributions implies how consistently the two term sets will behave given a query at retrieval time.

The proposed method assumes a dictionary of nouns that is automatically constructed from the document set. This is the practice that has never been tried in Korean document indexing, but has some important merits. A laborious work for the manual construction of nominal dictionaries is not needed. Since the noun dictionary contains only those in a document set, the ambiguity in analyzing words is greatly reduced.

Previous researches on the problem of compound noun indexing in Korean have been done in two directions. One approach adopts a full-scale morphological analysis to decompose a word into a sequence of the smallest morpheme units that are all treated as index terms. The other approach tries to avoid the complexity of the full scale analysis by using bigrams as in (Fujii 1993; Lee 1996; Ogawa 1993). Since these methods take all the components of compound nouns as index terms without evaluation, irrelevant terms can decrease retrieval precision.

Experiments on 1000 documents show that our evaluation scheme gave results closer to the human intuition and maintained the highest precision ratio of the existing methods.

In the following section, a brief review of related work on automatic indexing for Korean documents is made. Section 3 explains the proposed method in detail. The verification of the method through experiments is described in section 4. Section 5 concludes the paper.

2 Related Work

The previous approaches to compound noun indexing are based either on full scale morphological analysis (Kang 1995; Kim 1983; Lee 1995; Seo 1993) or on the syllabic patterns (Fujii 1993 ; Lee 1996; Ogawa 1993). Morphological analysis will return morphologically valid component words constituting a given compound word. Since this method does not exclude invalid or meaningless words, it can result in the degradation of precision. Besides the employment of full morphological analysis is often too expensive and requires costly maintenance.

Simpler methods segment compound nouns mechanically into unigram or bigram words that are all regarded as index terms (Lee 1996). Bigram indexes shows better precision than unigrams, but can suffer from big index size. In general, the existing methods for compound noun analysis have been focused mainly on recall performance with little attention to the precision. The work presented in this paper tries to achieve the improvement of recall without the deterioration of precision.

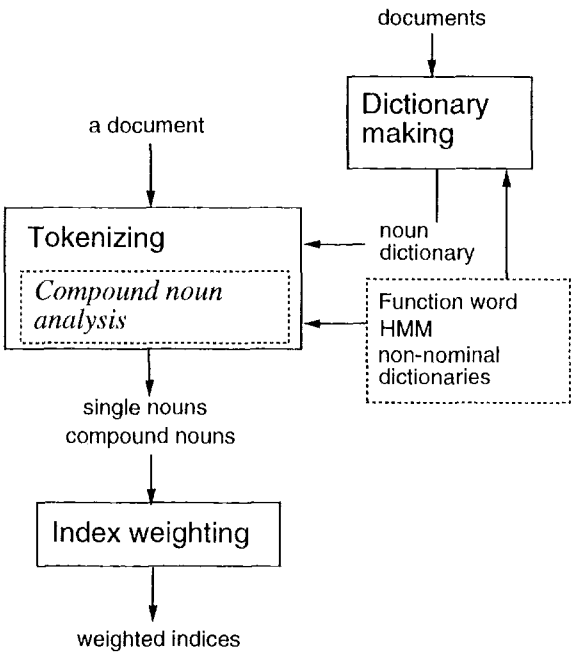

Figure 1: Compound noun indexing.

3 Probabilistic Compound Noun Indexing

In this section, we describe the algorithm to recognize and evaluate candidate index terms from compound nouns. Figure 1 summarizes the algorithm. The tokenizer produces a list of simple and compound nouns by utilizing the noun dictionary and the basic stemming rules. The noun dictionary is used to identify whether a noun is simple or compound, and the basic stemming rules are used to differentiate nominal words from others such as function words and verbs. The noun dictionary is automatically constructed from the observation on the document set. The compound noun analyzer investigates if the components of compound nouns are appropriate as indexes. The index terms that include simple nouns produced as a result of compound noun analysis are weighted, which finishes the indexing.

Let S and C denote the sets of simple and compound nouns, respectively. Simple nouns are, by definition, those that do not have any of their substrings as a noun according to the dictionary. Compound nouns are those one or more substrings of which are recognized as nouns. Let $T = \{T_1, T_2, \ldots, T_t\} = S \cup C$ be the set of all simple and compound nouns of a document set. Also, let $D = \{D_1, D_2, \ldots, D_d\}$ be the set of all documents. A document is represented as a list of term-weight (T_i, W_i) pairs.

For a compound noun C_i of a document, a *de-*

composition is a sequence of nouns $(T_1 T_2 \ldots T_k)$. In many cases, there are more than one decomposition, but only a few of them are sensible with respect to the context of the document. Indiscreet use of the component nouns may bring about the improvement of recall, but can lead to the significant decrease of precision. In the following discussions, we describe the details of the algorithm to select useful component nouns from compound nouns.

3.1 Dictionary buildup

It is very difficult to provide an IR system with the sufficient list of nouns. Because the nominals outnumber and grow faster than other categories of words, it is more efficient to handle non-nominal words manually. We consider building noun dictionary by identifying the remaining string as a noun after eliminating non-nominal part of a word. The non-nominals are verbs, adverbs, adjectives, prefixes, and suffixes.

The words in non-nominal dictionaries do not include those that can also be used as nouns, which is not a problem since unlike in English, the multi-categorial words in Korean tend to be invariant of meaning. The non-nominal dictionaries are made usually by manual work.

Those recognized as non-nominal words but not as function words are regarded as nouns. There can be multiple interpretations in segmenting a word due to the ambiguity of function words as illustrated in the following example.

wencalo	$\longrightarrow$	wencalo (reactor),
	$\longrightarrow$	wenca+lo (with atom)
		atom+INSTRUMENTAL

One way to deal with the problem is to use the probability of each function word and choose the one with the highest value. More accurate measure would be made using a Hidden Markov Model that is about a stochastic process of function words. The function words are classified into 32 groups according to their roles and position in sentences. In particular, each segmentation of a word is evaluated as follows.

$$P(C_i|C_{i-1})P(n)P(f|n).$$

$P(C_i|C_{i-1})$ is the probability of the function category of current word given the category of the previous word. $P(n)$ is the probability of candidate noun and $P(f|n)$ is the probability of a function word given the candidate noun. The best sequence of these segmentations for a sentence can be obtained. The candidate nouns n of the best sequence are then added to the noun dictionary.

3.2 Tokenizing and compound noun analysis

Tokenizing aims at recognizing simple and compound nouns from a text and reporting them as the the final index terms. The method for dictionary making is also used for tokenizing. Since the dictionary making method gives a list of candidate nouns, we only need to check if a candidate is a compound noun and judge if the components of the candidate compound noun are consistent with the content of the document.

To deal with the notion of consistency, we have to define the meaning of a term or a set of terms. It is a well recognized practice to regard the discriminating power of a term as the value of the term. The quality of the discriminating power is the distribution of the term over a document set. We define the distribution of a term as the meaning of the term. Similarly the meaning of a set of terms is the distribution of terms on the document set.

Let M be the distribution of a term T_i over a document set $D = D_1 \cdots D_n$ such that

$$\sum_{j}^{n} M(T_i, D_j) = 1.$$

One definition of $M(\cdot)$ may be as follows.

$$M(T_i, D_j) = \frac{freq(T_i, D_j)}{\sum_k freq(T_i, D_k)}.$$

For the case of multiple terms,

$$M(\{T_1 \cdots T_i\}, D_j) = \frac{\sum_{k=1}^{i} M(T_k, D_j)}{i}.$$

The similarity between two terms (or sets of terms) can be defined as any of vector similarity measures. The measurement of relative information of the two distributions corresponding to the two terms gives the distance between the distributions. Given two distributions M_i *and* M_j for T_i *and* T_j respectively, the discrimination $L()$ is defined as follows (Blahut, 1988).

$$L(M_i, M_j) = \sum_{i=0}^{I-1} M_i \log \frac{M_i}{M_j}.$$

Since we want the dissimilarity between two distributions, *divergence* that is a symmetric version of discrimination is more appropriate for our case. It is defined as follows (Blahut, 1988).

$$\overline{L}(M_i, M_j) = L(M_i, M_j) + L(M_j, M_i).$$

Figure 2 illustrates the different distributions of terms over the same document set suggesting the usefulness of the distributions as the representation of the terms. The divergence $\overline{L}(\cdot)$ gives about the information (uncertainty) of the two distributions as compared with each other, and has the following characteristics.

- The more uniform the distribution is, the larger $\overline{L}(\cdot)$ will be.
- The more the two distributions agree, the less $\overline{L}(\cdot)$ will be.

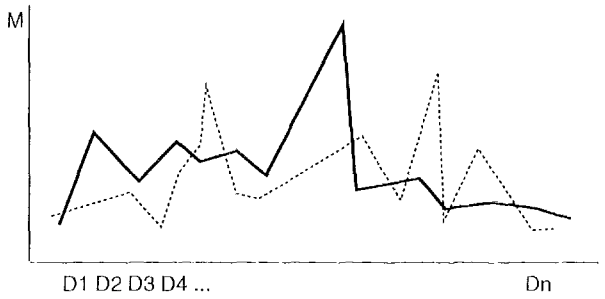

Figure 2: Illustration of term distributions over the same document set.

The characteristics are useful because good index terms should be less uniform and share similar contexts with other terms in a document. In this respect, information theoretic measure is more concrete and thus possibly more accurate than vector similarity measures.

For each decomposition $(T_i, \cdots, T_j)$ of a compound noun C_k, what we want to see is how different the decomposed terms and the document terms are. That is, $\overline{L}(\{T_i, \cdots, T_j\}, D_k)$ becomes the score of the particular decomposition. What we select here is one decomposition with the lowest divergence. Letting τ and τ' denote a decomposition and the best decomposition respectively,

$$\tau' = \arg\min_{\tau} \overline{L}(\tau, D_k).$$

The following summarizes the procedure of extracting simple nouns from compound nouns.

1. Remove non-nominal words using the method for dictionary making.
2. Identify compound nouns using nominal dictionary.
3. For each decomposition τ_i of a compound noun C_i, compute $\overline{L}(\tau_i, D)$.
4. Select τ_i with the lowest $\overline{L}(\tau_i, D)$.

3.3 Index weighting

There are three well known methods for weighting index terms. They are based on the information of inverse document frequency, discrimination value, and probabilistic value (Salton 1988). It turned out that these methods lead to similar performance, but inverse document frequency is by far the simplest of them in terms of time complexity and required resources (Salton 1988; Harmann 1992).

Inverse document frequency method is also shown to work with little performance variation across different domains. For this reason, we adopted inverse document frequency in the experiments. It is defined as follows.

$$w_{ij} = tf_{ij} \times \log(\frac{1}{df_i})$$

Table 1: The proportion of compound nouns in the 1000 science abstract. About 9% of nouns are compound nouns.

no. of components	nouns	proportion
1	49639	90.55 %
2	4665	8.50 %
3	469	.85 %
4	53	.09 %
5	6	.01 %

where w_{ij} is the weight of the i'th term in the j'th document, t_{ij} is the number of occurrences of the i'th term in the j'th document, and df_i is the number of documents in which the i'th term occurs

4 Experiments

The goal of experiments is to validate the proposed algorithm for analyzing compound nouns by comparing it with the manual analysis and the bigram method.

The test data set consists of 1000 science abstracts written in Korean (Kim 1994). All nominals are manually identified and compound nouns were decomposed into appropriate simple nouns by an expert indexer. In the first experiment, our proposed algorithm is asked to do the same thing over the test data, and retrieval performances on the two different outcomes (manually indexed and automatically indexed abstracts) are compared. In the second experiments, the performances of the proposed method and bigram method are compared to observe how the precision is affected.

As is shown at table 1, the portion of compound nouns is about 9% of total nouns found in the test set, but can make critical effects on the retrieval performance because often compound nouns carrying more specific information become a more accurate index to the documents.

Figure 3 and Table 2 summarize the performance of the indexing methods: manual analysis, the proposed probabilistic method, and the bigram method. The proposed method showed a slightly better performance (around 3% - 4%) than manual indexing or bigram indexing. However, our method has was more efficient than bigram indexing in terms of the number of index terms and the average number of retrieved documents per a query.

The average ambiguity of a compound noun is 1.43, and this low ambiguity must have contributed to the high agreement ratio of the proposed indexing method with manual indexing. The low ambiguity is partly attributed to the noun dictionary that has no unnecessary entries

Recall	Man.	Prob.	Big.	No Anal.
0.00	0.8719	0.8579	0.8406	0.7957
0.10	0.7719	0.7587	0.7841	0.6455
0.20	0.7122	0.6981	0.6812	0.5894
0.30	0.5895	0.6312	0.5939	0.4931
0.40	0.5458	0.5854	0.5637	0.4103
0.50	0.4957	0.5287	0.5240	0.3646
0.60	0.4272	0.4438	0.4370	0.2844
0.70	0.3304	0.3665	0.3322	0.2311
0.80	0.2552	0.2876	0.2569	0.1695
0.90	0.2102	0.2280	0.2028	0.0900
1.00	0.1428	0.1724	0.1600	0.0514

Table 2: Performance of Manual, Prob., and Bigram Indexing

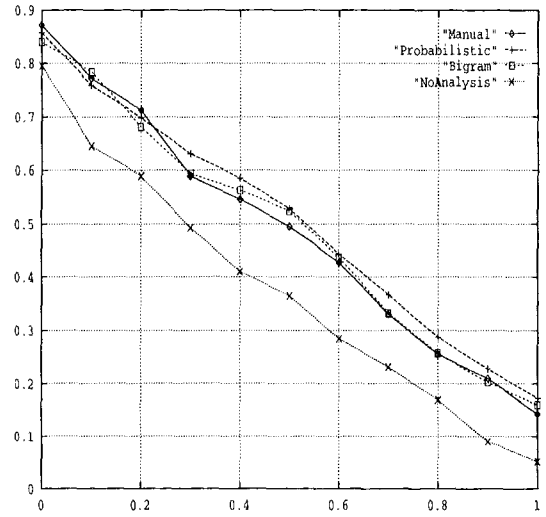

Figure 3: Recall-Precison curve of indexing methods

not found at the documents.

5 Conclusion

The compound analysis in automatic indexing aims at the improvement of recall performance by extracting useful component nouns from compound nouns. The task for Korean texts requires extra efforts due to the complexity of inflections. The proposed method gives better potential of sustaining the precision while improving the recall than other approaches by making use of probabilistic distributions of terms as the representation of meaning of the terms.

The proposed method to evaluate the components of compound nouns is unique in that it defines and uses term representation, which explains the superiority of the method to other methods. The method requires little human involvement and is very promising for the implementation of practical systems by achieving efficiency and accuracy at the same time.

References

Blahut, Richard E. (1987). *Principles and Practice of Information Theory.* Addison-Wesley.

Fagan, J. L. (1989). The effectiveness of a Nonsyntactic Approach to Automatic Phrase Indexing for Document Retrieval, *Journal of American Society for Information Science*, Vol. 40, No. 2.

Harman, D. (1992). "Ranking Algorithms" in *Information Retrieval: Data Structure and Algorithms*, (Frakes, W. B., and Baeza-Yates, R. ed.) Prentice Hall.

Fujii, H., and Croft, W. B. (1993). "A comparison of indexing techniques for Japanese text retrieval," In Proceedings of 16'th ACM SIGIR Conference.

Kang, S. S. (1995). "Role of Morphological Analysis for Korean Automatic Indexing,", In Proceedings of the 22nd Korea Information Science Society Conference.

Kim, Y. H. (1983). *Automatic Indexing System of Korean Texts mixed with Chinese and English* M.S. Thesis, Dept. of Computer Science, Korea Advanced Institute of Science and Technology.

Kim, S. H. (1994). A Development of the Test Collection for Estimating the Retrieval Performance of an Automatic Indexer, *Journal of Korea Information Management Society*, Vol. 11, No. 1.

Lee, J. H. (1996). "n-Gram-Based Indexing for Effective Retrieval of Korean Texts," In Proceedings of 1st Australian Document Computing Symposium 1996

Lee, H. A. (1995). "Implementation of an Indexing System Based on Korean Morpheme Structural Rules,", In Proceedings of Spring Conference of Korea Information Science Society.

Ogawa Y. (1993). "Simple word strings as compound keywords: An indexing and ranking method for Japanese texts,", In Proceedings of 16'th ACM SIGIR Conference.

Salton, G., and McGill M. J. (1983). *Introduction to Modern Information Retrieval* McGraw-Hill Inc.

Salton, G., and Buckey, C. (1988). Term Weighting Approaches in Automatic Text Retrieval, *Information Processing & and Management*, Vol. 24, No. 5.

Seo, E. K. (1993). An Experiment in Automatic Indexing with Korean Texts: A Comparison of Syntactico-Statistical and Manual Methods, *Journal of Korea Information Management Society*, Vol. 10, No. 1.