퍼지 및 지능시스템학회 논문지 2001, Vol. 11, No. 6, pp. 549-557

A Machine Learning Approach to Korean Language Stemming

Se-hyeong Cho

MyongJi University

Abstract

Morphological analysis and POS tagging require a dictionary for the language at hand. In this fashion, though, it is impossible to analyze a language without a dictionary. We also have difficulty if significant portion of the vocabulary is new or unknown. This paper explores the possibility of learning morphology of an agglutinative language, in particular, Korean language, without any prior lexical knowledge of the language. We use unsupervised learning, in that there is no instructor to guide the outcome of the learner, nor any tagged corpus. Here are the main characteristics of the approach: First, we use only raw corpus without any tags attached or any dictionary. Second, unlike many heuristics that are theoretically ungrounded, this method is based on statistical methods, which are widely accepted. The method is currently applied only to Korean language, but since it is essentially language-neutral, it can easily be adapted to other agglutinative languages.

Key Words : machine learning, natural language, morphology, unsupervised learning

1. Introduction

Morphological analysis and part-of-speech tagging usually require a dictionary. Therefore it is not possible to tag sentences in a language for which there is no dictionary. We also have difficulty if a significant portion of the vocabulary is unknown. For instance, North Korea and South Korea share the same language, but during the last 50 years or so of her separation, so many new words have been "invented" or changed their meanings. For another example, virtually all youngster who chat over the Internet use "여" [y uh][1] instead of "요" [y ao] as the concluding eomi[2]. A morphological analyzer or a POS-tagger will mark them as errors, or unknown words, unless they are equipped with a brand new dictionary that encompasses all such morphemes. Unfortunately, constructing a new dictionary or even updating a dictionary is an extremely time-consuming task.

This paper describes a novel method of automatically constructing a lexicon given only a raw corpus using machine learning. The purpose of the research is two-fold. The first is to construct a lexicon-free stemmer that is to be used in an Internet search engine. By not requiring a human-made dictionary, we are freed from the problem of intellectual property rights. Also, machine learning makes the process of constructing a lexicon extremely fast, and therefore accelerates the adaptation to the constantly evolving language. The second is to explore the possibility of unsupervised machine learning of natural languages. Of morphology learning, grammar learning, and semantics learning, we focus only on morphology learning in this paper.

Machine learning can be classified into unsupervised learning and supervised learning. A learning process is regarded as being supervised if there exists a means of telling the right from the wrong. It is regarded as being unsupervised otherwise. In morphology learning, if the corpus used for learning is POS-tagged, then it is regarded as supervised; it is regarded as unsupervised if the corpus is untagged. The proposed method uses unsupervised learning, since the corpus that is available will probably be a collection of sentences in a "new" language. Section 2 discusses related work. In section 3, we discuss the characteristics of Korean morphology, together with its statistical manifestations we can observe from a corpus. We also examine if such statistical manifestations can be used as criteria for identifying individual morphemes. In section 4, we propose a multi-stage method that overcomes the limitations of such simple-minded statistical solutions. Section 5 discusses the results and future work.

접수일자 : 2001년 9월 11일
완료일자 : 2001년 11월 20일

1) Throughout the paper, examples in Korean are first written in Korean. If necessary, it will be followed by ARPAbet[20] symbols in square brackets. The translations are in round parentheses, where each eojul is represented by the meaning part optionally followed by the role of grammatical morphemes written between curly braces. For instance, the sentence "He went" in Korean is "그가 갔다", and will be represented in this paper as "그가 갔다" [g ux g aa g aa t aa](he{subject} go{past}).

2) Literally, an eomi is a suffix. However, in Korean language, it is a class of grammatical morpheme that determines the tense, mood, voice, or honor.

2. Related work

Recently, there have been quite a number of research results on morphology learning. Unfortunately, most of them consider only inflectional languages, such as French or English.

2.1 Morphology learning in inflectional languages

Studies in morphology learning and lexicon-free stemmers in inflectional languages have been there for quite a time[9]. Early work used pre-constructed suffix list and/or rules concerning the stems. More recent work shows attempts in language-neutral morphology learning schemes[10] [11].

Marquez[12] developed a machine-learning method for part-of-speech tagging, but it requires a dictionary and only learns knowledge for ambiguity resolution. Porter[16] developed a simple lexicon free stemmer. Porter's algorithm uses knowledge about commonly used suffixes and deletes the suffixes from inflected or derived words. Such a simple algorithm works because English does not have so many suffixes. Also the performance (in terms of correct stemming) is not known to be very good because of its simplicity. Gaussier[13] proposed an unsupervised learner that uses an inflectional lexicon to learn inflectional operations. He proposed a method of classifying words with common prefixes as candidates, and used clustering to group words into families. However, the "certain length" is totally arbitrary, and also it requires inflectional lexicon, making the method language-dependent.

Goldsmith[11] and DeJean[14] also developed morphology learners, but they are restricted to inflectional language. They use statistical method to find out candidates for stems, and searches for appropriate inflectional suffixes. There are cases where proper suffixes are inappropriately applied. Also, ambiguity problem remains. When selecting candidate affixes, Goldsmith[11], Gaussier[13], and Schone[10] all used p-similarity. Two terms are p-similar if they share the first p letters. The affixes to p(or more)-similar words are collected first, and K most frequent affixes are selected as candidates. The value of p is arbitrary, and depends on the language or the knowledge of the researcher. The value of K is also arbitrary: Goldsmith[11] chose the top 100, and Shone[10] chose 200. No theoretical ground has been suggested on the determination of the values. Schone[10] added to Goldsmith's and Gaussier's methods the technique of singular value decomposition to reduce the dimension. This method is used to find out the hidden similarity to resolve ambiguity problem. This method may be compatible with the learning algorithm suggested in this paper.

2.2 Korean morphology analysis technology

Korean language morphological analyzers and POS taggers are reported to have a high quality, enough for use in commercial products. According to Shin[1], the accuracy of Korean POS taggers range from 89 to 97 percents. Since even the tags assigned by an expert's manual work (which are used as the "gold standard") usually have a few percents errors[21], this percentage can be considered near perfect. Unfortunately, these taggers rely on manually constructed lexicons. Nam[2] used statistical method for constructing information base for noun-derived suffixes. The suffix list, however, is from an already constructed dictionary. Kang[3] used the characteristics of Korean language syllables for morphological analysis. For instance, an eojul (the unit of writing separated by white spaces or punctuation marks) ending with the syllable "은" [uh n] or "는" [n uh n] is unlikely to be a single word. Such knowledge is used as heuristic rules. Since we aim at using no lexicographic knowledge, this method is not suitable for our purpose. Other work in Korean language morphology include construction of efficient dictionary[4][5], use of language-dependent knowledge for enhancing the accuracy of morphological analysis and POS-tagging[3][6], and combining statistical method and rule-based method[1]. In [7], mutual information is used for automatic segmentation of incorrectly unsegmented eojuls. It differs from our work in that it aims at separation of eojuls, while ours aim at separation of morphemes. Despite the many pieces of work on Korean morphology, no work that the authors have reviewed has attempted morphology learning.

3. Characteristics of Korean morphology and some statistical observations

In this section, we discuss some characteristics of Korean language morphology that can be of help in learning the morphology, and also discuss some statistics that such characteristics manifest. Although such statistics are quite interesting in that they can be used in identifying some morpheme boundaries, none of them are sufficient to be used as criteria for morpheme segmentation by themselves, as we shall see later in this section.

3.1. Characteristics of Korean language morphology

Korean language is an agglutinative language, and affixes that indicate the case, mood, tense, or honor are agglutinated to words, or content morphemes. This may look similar to inflection or derivation in that affixes are concatenated to word stems. However, agglutinated morphemes play more important grammatical and semantic roles. One of the categories, josa, is the set of words that are attached to nominals to determine the case. Thus "영희는 고양이를 좋아한다." (YungHee{subject} cat{object} like{present}) and "고양이를 영희는 좋아한다." (cat{object} YungHee{subject} like{present})

are two sentences with the same meaning despite the change of positions of the eojuls in the sentence. Another important class of grammatical morphemes is eomi. Eomis are further divided into leading eomis and trailing eomis. Eomis are used to indicate honor, tense, mood, and voice. These are concatenated into one eojul, and we must be able to identify the morpheme boundaries before we proceed to POS-tag, analyze syntax, or identify meanings of sentences. The content morphemes are usually open classes, and therefore new words are constantly added and are large in numbers. Grammatical morphemes are usually closed classes and small in numbers. We can then easily conjecture that at the end of eojuls same patterns will occur repeatedly.

3.2. Statistics manifested by the combination of content-grammatical morphemes

One important property of grammatical morphemes is that they appear at the end of eojuls. They are smaller in numbers compared to content morphemes, and thus will appear more often. This leads us to deduce (or abduce, to be more precise) that suffixes that appear "often" are grammatical morphemes. Whether they appeared "often" can be judged by different statistical measures. The first is the absolute frequencies. Second, it can be relative frequencies. That is, the ratio of the number of occurrences of a string at the end of eojuls to the number of occurrences of the same string regardless of the positions. The third is the number of different prefixes that are combined with the suffix to form an eojul. Fourth, there are the t-test values that tell us whether given samples are from a population of a certain probability distribution. We will compare these four statistical measures observed in the same corpus. To compare the relative merits and disadvantages, we plotted the accuracy against the recall.

3.2.1. Absolute frequencies

The conclusive eomi "다" [d ax] is used in a majority of sentences. "다" is roughly comparable to Japanese "です" [d eh s uh]. If an average sentence has 15 eojuls and 90% of sentences end with "다", then at least 6% of all eojuls will end with the syllable. Considering that there are around 2,000 syllables in Korean, the average probability of occurrence of a syllable at the end of eojuls is around 0.05%. 6% is unusually higher than 0.05%, and therefore we may conclude that "다" is a grammatical morpheme. Table 1 shows the rankings of some syllables in terms of the number of occurrences among 19,000 eojuls/59,000 syllables in Dong-A Ilbo newspaper of Jan. 25th, 2001. Rank 1 through 25 are found to be grammatical morphemes, while "시" [sh ih] in rank 26 is not. We found two reasons for this high frequency. One is that the syllable "시" does not appear often only at the end of eojuls, but it also appears in the middle of eojuls frequently, too. Another is that the high frequency is due to a few frequently used words ("다시" [d aa sh ih] and "당시" [d aa ng sh ih]) that happened to end with the same syllable "시" [sh ih].

Table 1. syllables ranked by frequencies.

rank	syllable	frequency	examples(eojul&frequency)
1	다	1245	가격보다(1),가구다(2)...
2	는	1091	가구수는(1),가는(2)...
3	을	990	가격왜곡을(1),가격을(1)...
...			
26*	시	106	다시(17),당시(14),...

3.2.2. Relative frequencies

Some syllables are simply higher in frequency than others. These syllables will appear at the end of eojuls very often, as well as in the middle. The syllable "시" we saw in the last subsection appeared 317 times in the given text, and among them 106 were at the end of eojuls. Considering that the average length of eojuls is around 3, it is not surprising at all that it appeared 106 times at the end of eojuls. Therefore, a better measure to determine the unusually high frequency at the end of eojuls may be the relative frequency. For instance, if a syllable appeared 100 times in a text but 70 of them were at the end of eojuls, then the syllable has 70% of relative frequency. One serious problem with relative frequency metric is that strings with very low frequency can easily have extremely high relative frequency. For instance we observed many of the suffixes including the ones in Table 2 occurred only once or twice in the text, resulting in 100% of relative frequency. The probability by the maximum likelihood estimation is high, but the sample size is too small for the statistics to be reliable.

Table 2. syllables ranked by relative frequencies.

rank	syllable	ratio	examples(eojul&frequency)
1*	꽃	1	책꽃(2)
2*	씬	1	휠씬(3)
3*	쩔	1	어쩔(1)
...			
40	을	0.96	가격왜곡을(1),가격을(1),...

3.2.3. Usage counts

The usage count means the number of different prefixes to which the given suffix is concatenated, thereby forming eojuls. For instance, if we had three eojuls "내가" [n ae g aa], "학교가" [h aa k y ao g aa], and "내가" [n ae g aa] in the text, the usage count of "가" [g aa] is 2, not 3. We observed that the string "께" [k eh] appeared 35 times in the text. Among them was an instance of "아버님께" [aa b ah n ih m k eh],

and 34 instances of "함께" [h aa m k eh]. In this case, the frequency of "께" [k eh] is 35, while the usage count is only 2. In most cases usage count is more important than the simple number of occurrences of the same eojul. Table 3 shows some highly ranked eojuls in terms of usage counts. The one in rank 28, "장" is not a grammatical morpheme.

Table 3. syllables ranked by usage counts

rank	syllable	usage	examples(eojul&frequency)
1	는	702	가구수는(1),가는(2)...
2	을	668	가격왜곡을(1),가격을(1)...
3	다	596	가격보다(1),가구다(2),...
...			
*28	장	52	가장(18),개장(2),...

3.2.4. T-test values

The construction of a sentence can be regarded as a sequence of probabilistic choices. Suppose the probability of choosing a certain syllable is p_0. Also suppose the probability of the same syllable selected at the end of an eojul is p_0. Assume that syllables other than grammatical morphemes are randomly selected. Then if the syllable were a grammatical morpheme, p will be greater than p_0. Since such selections can be regarded as Bernoulli trials, we can use t-test to test if the occurrence of the syllable at the end of eojuls were purely accidental or not. The t-test statistic Z_0 is defined as follows[8] :

$$Z_0 = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{N}}},$$

where N is the number of trials, that is, the number of syllables.

We can reject the null hypothesis if $Z_0 > 1.96$ with significance level of 0.05, or confidence level of 95%. In other words, if $Z_0 > 1.96$, then we can regard the suffix as a grammatical morpheme with confidence level of 95%.

As we see in Table 4, the highly ranked syllables are frequently used grammatical morphemes. However, the syllable "께" at 25th relies its frequency on a single eojul "함께", resulting in high Z_0 value. Incidentally, "께" in "아버님께" happens to be a grammatical morpheme, but the same syllable in "함께" is not.

Table 4. syllables ranked by t-test

rank	syllable	Z_0	examples(eojul&frequency)
1	는	21.2	가구수는(1),가는(2)...
2	을	20.6	가격왜곡을(1),가격을(1)...
3	다	20.4	가격보다(1),가구다(2),...
...			
*25	께	3.78	아버님께(1),함께(34),...

3.3. Comparison of the statistics

In order to judge the relative "goodness" of the above four statistical measures, we considered recall - the number of identified grammatical morphemes divided by the number of all grammatical morphemes in the text - and the precision - the ratio of correct candidates in the list. These terms are borrowed from information retrieval. Achieving either high recall only or high precision only is trivial. To obtain high recall, we include everything in the list, sacrificing precision. To obtain high precision, we include nothing. The problem is to obtain as high a recall as possible while maintaining reasonable precision. Figure 1 plots the precision versus recall, using the four statistical measures as criteria. With t-test, as the recall goes above 70%, the precision drops rapidly. On the other hand, with relatively low recall (50~60%), using t-test as the criterion for finding grammatical morphemes results in high precision. In case of relative frequencies, extremely low frequency syllables were excluded (less than 4 times). The list of grammatical morphemes was taken from [15].

We can easily see that even though these statistical measures give some hints on whether a given suffix might be a grammatical morpheme, it is not a complete measure at all, and cannot be used for a stemming algorithm.

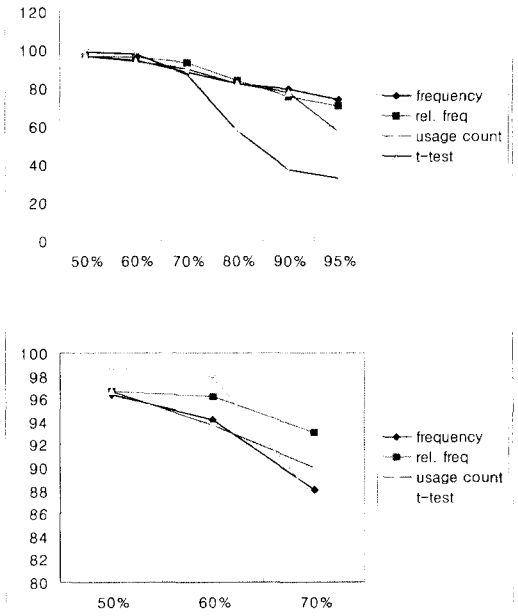

Fig. 1. Precision vs. recall. The bottom one is the exploded view

4. A Multistage method for morphology learning and segmentation

As we alluded to in the previous section, there is no single statistical measure that can be used to identify

grammatical morphemes in eojuls. In this section, we propose a multistage method for morphology learning and stemming based on statistical method. This method makes use of the t-test for the initial construction of the candidate list, while ignoring repeated appearance of the same eojuls. In other words, two of the statistical measures, the usage count and the t-test discussed in the last section were combined. Usage count is adopted because repeated appearance of the same eojul, provided that it indeed consists of the same morphemes, does not give any more statistical information. t-test is adopted because unlike others, it has quantitative measure that lets us know how confident we are as to the decision. For instance, using a significance level of 0.05 implies that we are 95 % sure of the result of the decision. If we needed more accuracy or needed broader coverage, we simply need to adjust the significance level, and therefore the value of Z_0. Other measures, for instance the rank in terms of frequency, may indicate which suffix is more likely to be a grammatical morpheme, but not how much likely it is. Also, unlike previous analysis, which was meant to be simpler, we considered suffixes that do not consist of an integral number of syllables too, simply because there are many such morphemes.

The complete method consists of 5 stages as listed below:

1. Obtain initial list of candidates for grammatical morphemes using t-test.
2. Exclude wrong grammatical morphemes accidentally included in the list from stage 1.
3. Generate list of content morphemes using the list of grammatical morphemes.
4. Exclude more errors from the list of grammatical morphemes using the pattern of morpheme combinations. This is the end of the learning phase.
5. Perform stemming, using the knowledge learned from stages 1 to 4.

Each stage is described in more detail in the following subsections.

4.1. Stage 1: Obtain initial list of candidates for grammatical morphemes using t-test.

In this stage, t-test is performed on all suffixes. Candidate grammatical morphemes are selected based on the Z_0 values. In order to avoid repeated analysis on the same suffixes and to do efficient computation, a backward trie is used. A backward trie is a trie where the root branches according to the rightmost letter of a word, and the next level branches according to the letter second from the right, and so on and so forth. Since in a trie structure identical sequences of letters lead to the same node, the same eojul is not inserted twice. Each Korean character (syllable) consists of three letters, each representing the leading consonant, the vowel, and the trailing consonant. In order to perform t-test, we need to know the prior probability of a given sequence of letters. The prior probabilities are estimated by maximum likelihood estimation as follows:

$$p_0(w) = f(w, C')/Pos(w, C'),$$

where C′ is the (imaginary) corpus derived from C by deleting all duplicate eojuls, $f(w, c)$ the number of occurrences of string w in corpus c, and $Pos(v, c)$ the number of positions that a string of length $length(v)$ might fit in. $Pos(v, c)$ is computed by the following formula:

$$Pos(v, c) = \sum_{i=l_v}^{l_{max}} (i - l_v + 1) f(i, c),$$

where l_{max} is the number of syllables in the longest eojul, and $f(i, c)$ is the number of occurrences of length-i eojuls in corpus c, and $l_v = \lceil length(v)/3 \rceil$. For instance, the number of positions for a single-letter candidate is equal to the total number of syllables in the corpus. A consonant can appear as a leading consonant or a trailing consonant, and they are treated as different letters. A candidate longer than one syllable and shorter than 2 syllables can appear only in eojuls consisting of two syllables or more. The probability of a two-syllable candidate to appear in a three-syllable eojul is somewhat complicated, since appearing in the first and the second syllables and appearing in the second and the third syllables are not independent events: the former keeps the latter from happening. However, since the probability is much smaller than unity, we can pretend that they are independent, and use an approximation.

Suffix probability, that is, the probability that a string comes at the end of eojul is estimated by $p(w) = \frac{f(u\$, C')}{|\{v \in C' | length(v) \geq length(w)\}|}$, where \$ is the imaginary letter that indicates the end of an eojul. Again, this is a maximum likelihood estimation.

Using the above definitions, Z_0 is computed (p replaced by p(w) and p_0 by $p_0(w)$). Using a small corpus of 25,000 eojuls, 517 candidates were selected among 40,000 suffixes using 90% confidence level. The stage 1 algorithm is depicted in Algorithm 1.

$L \leftarrow \{\}$
for each suffix w in corpus C
$$Z_0(w) \leftarrow \frac{p(w) - p_0(w)}{\sqrt{\frac{p_0(w)(1-p_0(w))}{N}}}$$ // p and p_0 as defined above
if $Z_0 > T_1$, $L \leftarrow L \cup \{w\}$
endfor

Algorithm 1. The first stage algorithm using t-test

T_1 represents a threshold constant, and we chose 1.3 for the experiment. This amounts to the significance level of 0.1. N is the number of Bernoulli trials, that is,

the number of eojuls.

4.2. Stage 2: Exclusion of wrong candidates

The candidates selected in stage 1 are those that have higher probability of occurring at the end of eojuls, compared to probability of occurrence in general. However, some of them are chosen not because they are grammatical morphemes, but because their suffixes are. For instance, the copula "이다" [ix d aa] is a grammatical morpheme, and therefore has high suffix probability. This also causes "ㅁ이다", or "ㄱ이다" to occur very frequently, though not as frequent as "이다" itself. However, the probability of "ㅁ이다" occurring in the middle of an eojul is extremely small, causing the Z_0 value to be higher than we expected. On the other hand, the occurrence of "에서는" does not rely solely on the high probability of "는". In other words, the conditional probability of "ㅁ" in front of "이다" is not significantly higher than the unconditional probability of "ㅁ". On the other hand, the conditional probability of "에서" in front of "는" is significantly higher than its unconditional probability.

Let w and δ be strings in C'. In case $Z_0(w) > T_1$ and $Z_0(\delta w) > T_1$, let $p'(\delta, w) = \frac{f(\delta u\$, C')}{f(u\$, C')}$, and $p_0'(\delta) = \frac{f(\delta, C')}{Pos(\delta, C')}$. Using $p'(\delta, w)$ and $p_0'(\delta)$, we can run a t-test to see whether δ's occurrence in front of w is merely a coincidence.

In the actual experiment, the Z_0 value of "ㅁ이다" in stage 1 was 1.4, while in the second stage it dropped to 0.68, indicating that the high frequency is merely by chance. On the other hand, the value for "에서는" soared from 3.46 in the first stage to 20.0 in the second stage. By applying stage 2 algorithm on the same corpus, 308 candidates among 517 chosen in stage 1 have been discarded and only 219 remained. Algorithm 2 depicts the detailed algorithm for stage 2. N_w is the number of occurrence of w.

---

for each suffix $v = \delta w \in L$ // L is from stage 1 algorithm

$$Z_0(\delta, w) \leftarrow \frac{p'(\delta, w) - p_0'(\delta)}{\sqrt{\frac{p_0'(\delta)(1 - p_0'(\delta))}{N_w}}}$$ // p' and p_0' as defined above

if $Z_0(\delta, w) < T_1$, $L \leftarrow L - \{v\}$

endfor

Algorithm 2. The second stage algorithm for removing wrong candidates

4.3. Stage 3: Construction of content morpheme list

If we can identify grammatical morphemes from an eojul, we can identify the content morpheme; if we can identify the content morpheme, we can identify grammatical morphemes. Even though not prefect, we now are able to find many grammatical morphemes. We can now use them to determine which strings are content morphemes. However, it is difficult to separate the right content morpheme from only the list of grammatical morphemes. First of all, there can be two or more suffixes in a given eojul that are valid grammatical morphemes. For instance, in "학교에는" [h aa k yo eh n uh n], "에는" is the grammatical morpheme, but "는" also is a valid grammatical morpheme. Second, there can be errors in the list of grammatical morphemes: the list may include wrong morphemes, and the right ones might not be in the list. Third, a suffix of an eojul being the same string as a valid grammatical morpheme does not necessarily mean it is used as a grammatical morpheme. For instance, "ㄴ" in "먼" (far) is a grammatical morpheme, but not in "관" (coffin). To resolve these problems, we first select strings that can possibly be content morphemes. Here, we take advantage of the observation that "content morphemes and grammatical morphemes concatenate to form eojuls", and take the number of eojuls that can be formed by the candidate concatenated to by a valid grammatical morpheme as a criterion. That is, the numbers $|\{\delta | w\delta \in C', \delta \in L\}|$ for each w. It is tempting to use another t-test on these numbers. Unfortunately, such frequencies in general are too low to analyze statistically. Therefore, we decided to select only the candidates that are coupled with grammatical morphemes for more than a certain number. We chose $T_2 = 3$ as the threshold in the experiment (see Algorithm 3).

$M \leftarrow \{\}$, $\forall x, count(x) = 0$

for all eojul $w = w_1 ... w_k$

for $i = 2$ to $k - 1$

if $w_i ... w_k \in L$ then

$count(w_1 ... w_{i-1})$++

endfor

for all strings w

if $count(w) > T_2$ then $M \leftarrow M \cup \{w\}$

endfor

Algorithm 3. Generation algorithm for content morpheme candidates

4.4. Stage 4 : Exclusion of grammatical morpheme candidates which are coupled to non-content-morphemes

In the previous subsection, we considered a string to be a content morpheme if it is coupled with many grammatical morphemes to form eojuls. Similarly, we can also decide that a certain suffix is not actually a grammatical morpheme, if it is not coupled with many content morphemes. For instance, the syllable "록" [r ao k] has a high Z_0 value, and initially was considered a good candidate for a grammatical morpheme. However,

even though "록" is coupled with many different prefixes, none of the prefixes are content morphemes (e.g., "그토록" [g uh t ao r ao k]). Therefore we can conclude that it is not indeed a grammatical morpheme, but simply a string that appeared at the end of other morphemes. The algorithm is described in Algorithm 4. We used the threshold value 4 for T_3.

```
for each w∈L
  for each v∈M
    if v·w∈C then count(w)++ // C is the corpus
  endfor
endfor
for each w∈L
  if count(w)<T3, then L ← L−{w}
endfor
```

Algorithm 4. Re-consideration of grammatical morpheme list according to the number of coupling with content morphemes

4.5. Stage 5: Stemming phase

In order to find out the right content morpheme from an eojul, the best candidate is determined by the product of the probability that the prefix is a content morpheme and the probability that the suffix is a grammatical morpheme. The probability that a suffix of an eojul is determined as follows. Let the string w appear with probability p in general and with p_s at the end of eojuls. Further let a be the number of occurrences of w at the end of eojuls, b be the total number of occurrences, and a' be the number of times w is actually used as a grammatical morpheme. A is the total number of eojuls, and B is the number of positions where w could appear. Then $p=\frac{a+b}{A+B}$, and $p_s=\frac{a}{A}$. Therefore the probability of w being a grammatical morpheme is $P_{GM}=a'/a=\left(a-b\frac{A}{B}\right)/a=1-\frac{Ab}{Ba}$, by the maximum likelihood estimation. For instance, suppose we have 1,000,000 eojuls where each eojul consists of 3 syllables. If the total number of syllable "는" is 75,000 and among them 70,000 were at the end of eojuls. Then $P_{GM}=1-0.0357=0.9643$. An intuitive interpretation is this: since the syllable is used even if it is not used as a grammatical morpheme, the same will be true for the last syllable. Therefore 2,500 occurrence (5,000/2) out of 70,000 probably was not used as a grammatical morpheme. Therefore P_{GM}= (70,000 − 2,500) / 70000 = 0.9643.

Of course, this applies only to those that are believed to be grammatical morphemes. Other candidates are given a minimal probability based on the concept of smoothing. Content morpheme probability is quantized to have only 3 different values, depending on which group it belongs. Namely, the candidates included in the set M in Algorithm 3, the candidates that have at least one instance of coupling with grammatical morphemes, and all the rest. The reason for the quantization is because in general we don't have enough quantity for the statistics to be significant. Also we have a lot of low-frequency content morphemes, which can easily be imagined from Zipf's law[17]. The three cases were assigned the probability of 0.9, 0.1, and 1/|C|.

When we assume the selection of content morphemes and grammatical morphemes are independent events, stemming becomes the task of deciding the position m that satisfies the following formula:

$$m=\arg\max_k P(head=w_1..w_k, tail=w_{k+1}..w_n \mid w=w_1..w_n)$$
$$=\arg\max_k \frac{P(w=w_1..w_n \mid head=w_1..w_k, tail=w_{k+1}..w_n)P(head=w_1..w_k, tail=w_{k+1}..w_n)}{P(w=w_1..w_n)}$$
$$=\arg\max_k P(head=w_1..w_k, tail=w_{k+1}..w_n) \equiv \arg\max_k P(head=w_1..w_k)P(tail=w_{k+1}..w_n)$$

where argmax represents the index which maximizes the function that follows. "head=xx" means a proposition that xx is the content morpheme part, and "tail=yy" means a proposition that yy is the grammatical morpheme part.

5. Results and Future Work

A small corpus of 25,000 eojuls is used for learning. The text is from various articles from Internet Dong-A Ilbo newspaper, during the fourth week of Jan. 2001. Using the stage 1 list of grammatical morphemes, the success rate (i.e., the correct stemming rate) was 74%. Using the list enhanced by stage 2 algorithm, the success rate went up to 81.5%. After stage 4, success rate was 85%. Success was measured on randomly selected 1000 eojuls. The margin of error is ±1%, with confidence of 95%. When we doubled the size of the corpus, the success rate went from 85% to 87%.

There are many ways open for future enhancement. Some of them are:

1. use of collocation information
2. identification of composite grammatical morphemes
3. identification of composite content morphemes such as composite nouns
4. POS learning by clustering of morphemes
5. Corpus error-resilient techinique.

It is sometimes difficult to stem an eojul in isolation, due to ambiguity problems. For instance, "순은" can be interpreted in two ways: an eojul without any grammatical morpheme, meaning 'pure silver', or a proper noun "순" plus a grammatical morpheme "은" indicating that this is the subject of the sentence. However, if seen in a particular context, there may be hints for resolution. For instance, in "그러자 순은 말했다" (Then Suhn{subject} utter{past}), the only candidate for subject is the second

eojul. However, in "순은 값은 하락중이다" (pure-silver price{subject} fall{present-progressive}), two consecutive eojuls end with "은", and both cannot be subjective josa at the same time. Such intuition can be encoded by bigrams among grammatical morphemes. Take the sentence "순은 값은 하락중이다" (pure-silver price{subject} fall{present-progressive}) for instance. Suppose P(순*은|순은)=0.55 and P(순은*|순은)= 0.45. Without consideration of bigrams, the former will become the most probable candidate. However, the next syllable also has the same josa "은", and we know that two consecutive occurrences of josa "은" is very rare (say, the probability is 1/10 of the probability of a single occurrence). Then we are better off not breaking the eojul at all and conclude that it is a single word.

In general, grammatical morphemes are used in combination, but in this paper composite grammatical morphemes are treated as a single morpheme. This does not bother us in that we do not seek any morphological theory. However, by being able to recognize composite morphemes, we may be able to recognize grammatical morphemes more accurately. Recognition of composite content morphemes, including incorrectly concatenated content morphemes, can also enhance the accuracy. For instance, suppose we know by learning from the text that "예산" (budget) and "위원회" (committee) are content morphemes. Further assume that we do not know if "예산위원회" (budget-committee) is a word. Then when analyzing "예산위원회가", we may end up concluding the whole eojul is a single (unknown) morpheme. On the other hand, if we knew that content morphemes can be concatenated to form a composite morpheme, we can analyze it as "예산" + "위원회" + "가" (budget-committee{subject}).

The ultimate stemming algorithm will be Part-of-speech tagging. Using unsupervised learning to identify word classes is under study[22]. In case of Korean language, the pattern of coupling with other morphemes is a good source of information for grouping morphemes into classes. For instance, nouns are coupled with grammatical morphemes such as "가" ({subject}), "를" ({object}), "에서" ({location}) often, but not with "었다" ({past}). Such knowledge may be used to classify morphemes into different classes. This is the next study item. In unsupervised learning, especially when the corpus is small, errors in the text cause problems, because the learner acquires knowledge solely from the corpus. However, error-free corpus is unlikely to be available, and therefore the learning technique should be error-resilient. Finally, another challenge is to apply the method presented in this paper to another language. For instance, Janapese language is similar to Korean, but differs in that there are no eojul-delimiting spaces. We conjecture that the method might be applicable if combined with some separation algorithm (e.g., [19]).

References

[1] SangHyun Shin, KunBae Lee, JongHyuk Lee, "Two-stage Korean tagger based on Statistics and Rules", *Journal of Korean Information Science Society* (B) vol. 24-2-, pp. 160-169, 1997.2.

[2] YunJin Nam, ChulYung Ok, "Construction of dictionary of noun-derived suffixes based on Corpus analysis," *Journal of Korean Information Science Society* (B), vol. 23- 4, pp. 389-401, 1996.4.

[3] SeungSik Kang, "Morphological Analysis of Korean Irregular verbs and adjectives using syllable characteristics," *Journal of Korean Information Science Society* (B) vol.22-10, pp. 1480-1487, 1995.10

[4] JaeHyung Choi and SangJo Lee, "Method for reduction of lexicon reference in Korean morphological analysis by two-way longest match," *Journal of Korean Information Science Society* vol. 20 no. 10, pp. 1497-1507, 1993.10.

[5] ChulSu Kim et. al, "Korean dictionary using two-way trie structure," *Journal of Korean Information Science Society* (B) vol. 23-1, pp. 85-94, 1996.1

[6] HeeSuk Lim, BoHyun Yoon, HaeChang Lim, Efficient Korean morphology analyzer using exclusion information," Journal of Korean Information Science Society, vol. 22-6, pp. 957-964, 1995.6.

[7] Kwang Sub Shim, "Automatic segmentation using mutual information among syllables," *Journal of Korean Information Science Society*, vol. 23-9, pp. 991-1000, 1996.9

[8] C. Manning and H. Schultze, Foundations of Statistical Natural Language Processing, 1999 MIT Press.

[9] Lovins, J.B., Development of stemming algorithms. Machine Translation and Computational Linguistics, 11, 1968

[10] Patrick Schone and Daniel Jurafsky, "Knowledge-free Induction of Morphology using Latent Semantic Analysis," in proceedings of the ACL99 workshop: Unsupervised learning in Natural Language Processing, University of Maryland.

[11] J. Goldsmith, "Unsupervised learning of the morphology of a natural language," University of Chicago, http://humanities.uchicago.edu/faculty/ goldsmith.

[12] LLuis Marquez, Lluis Padro, and Horacio Rodriguez, "A Machine Learning Approach to POS tagging," Machine Learning, vol. 39, pp. 59-91, 2000

[13] E. Gaussier, "Unsupervised learning of derivational morphology from inflectional lexicons," in proceedings of the ACL99 workshop : Unsupervised learning in Natural Language Processing, University of Maryland.

[14] Dejean, H. 1998. "Morphemes as necessary concepts for structures : Discovery from untagged corpora," University of Caen-Basse Normandie. http://www.info.unicaen.fr/DeJean/travail/article/pg11.htm

[15] HeungGyu Kim and BumMo Kang, "Analysis of usage count of Korean morphemes and words", Korea Univ. Center for National Culture Research, 2000, 7.
[16] M.F.Porter, "An algorithm for suffix stripping," Program, 14(3), pp. 130-137, 1980
[17] Zipf, G.K. Human Behavior and the Principle of Least Effort, Cambridge, MA: Addison-Wesley, 1949
[18] W. Mendenhall and R.J.Beaver, Introduction to Probability and Statistics, Boston, MA, 1995, PWD-Kent publishing co.
[19] R.Ando and L.Lee, "Unsupervised Statistical Segmentation of Japanese Kanji Strings," Technical Report TR99-1756, Computer Science Department, Cornell University, 1999
[20] J.E. Shoup, "Phonological Aspect of speech recognition," in Lea, W.A.(ed.) Trends in Speech Recognition, pp. 125-138, Prentice-Hall, Englewood Cliffs, NJ, 1980
[21] D. Jurafsky and J.H. Martin, Speech and Language Processing, Prentice Hall, NJ, 2000
[22] H. Schutze, "Distributed Syntactic Representations with an Application to Part-of-speech Tagging," in Proc. IEEE International Conference of Neural Networks., pp. 1504-1509

저 자 소 개

조세형(Sehyeong Cho)
1981년 서울대학교 학사(섬유공학).
1983년 서울대학교 석사(계산통계)
1992년 펜실바니아 주립대 박사(전산)

관심분야 : 자연언어 처리, 기계학습
E-mail : shcho@mju.ac.kr