Syllable-Based Probabilistic Morphological Analysis Model of Korean

Do-Gil Lee and Hae-Chang Rim

Dept. of Computer Science & Engineering
Korea University, Korea
dglee@nlp.korea.ac.kr, rim@nlp.korea.ac.kr

Abstract
In this paper, we present a syllable-based probabilistic morphological analysis model of Korean. While the previous morphological analyzers that regard morpheme as a processing unit, the model exploits syllable as a processing unit in order to endure the unknown word problem. Actually, it does not use any morpheme dictionary. In contract to the previous systems that depend on manually constructed linguistic knowledge, the proposed system can fully automatically acquire the linguistic knowledge from annotated corpora. Besides, without any modification, the system can be applied to other corpus having a different tagset and annotation guidelines. We describe the model and present experimental results on two corpora.

1. Introduction
This paper discusses Korean morphological analysis. Morphological analysis is to break down an Eojeol, which is the surface level form of Korean, into morphemes, which is the smallest meaningful unit. Table 1 gives an example of morphological analysis for an Eojeol na-neun.

Korean is a highly agglutinative language; an Eojeol, which is a spacing unit delimited by whitespace, is composed of one or more combined morphemes. Moreover, Korean is a morphologically complex language. Korean words are formed through compounding and derivation. Unlike inflectional languages such as English, Korean is very productive so the number of Eojeols may appear in real texts are almost infinite. Therefore, all the possible morphological variants cannot be registered in a lexicon. For these reasons, morphological analysis is necessary in Korean language processing, but is difficult than other languages.

There are some difficult points we have to be considered in Korean morphology: there are two kinds of ambiguities (segmentation ambiguity and Part-of-Speech ambiguity). Moreover, due to compounding and derivation, morphological changes to be restored are very frequent. In contrast to part-of-speech (POS) tagging, morphological analysis is characterized by producing all the grammatically possible interpretations.

There have been no standard tagset and annotation guidelines, so researchers have developed methods with their own tagsets and guidelines. The Morphological Analysis and Tagger Evaluation Contest (MATEC) was held in 1999 [1]. This is the first trial about the objective and relative evaluation of morphological analysis. Among all the participants, some newly implemented the systems and others converted the existing systems’ results through postprocessing steps. In both cases, they reported that they spent much effort and argued the necessity of tuning the linguistic knowledge.

All the systems that participated in MATEC can be considered as the so called dictionary and rule based approach. This approach depends on manually constructed linguistic knowledge such as morpheme dictionary, morphosyntactic rules, and morphological rules. There are two major disadvantages as follows:

- Construction of the knowledge base is time-consuming and labor-intensive. In addition, storing every word in a lexicon is impossible so they suffer from the unknown word problem.
- There is a lack of portability. Because the results produced by a morphological analyzer are limited to the given tagset and the annotation guidelines, it is very difficult to apply the system to other tagsets and guidelines. For example, according to the annotation guidelines, the Eojeol sa-lang-haess-da can be analyzed as sa-lang+ha+ass+da, sa-lang+ha+eoss+da, or sa-lang-ha+ass+da.

Our model tries to overcome these limitations: Firstly, it only uses POS tagged corpora as an information source and can automatically acquire a knowledge base from these corpora. Hence, there is no necessity for the manual labour in constructing and maintaining such a knowledge base. Although constructing such corpora also requires a lot of efforts, the amount of annotated corpora is increasing every year. Secondly, regardless of tagsets and annotation guidelines, it can be applied to any training data without modification. Finally, it can provide not only analyzed results but also their probabilities to be analyzed by the probabilistic models. In Korean, no attempt has been made at probabilistic approach to mor-
phological analysis. Probabilities enable the system to rank the results and to provide the probabilities with the
next module such as a POS tagger.

One of the most difficult problems in morphological analysis is the unknown word problem, which is caused
by the fact that we cannot register all morphemes in the
dictionary. In English, contextual information and suf-
fix information help to estimate POS tags of unknown
words. In Korean, the characteristics of Korean sylla-
cable can be used. For instance, a syllable eoss can only
be a pre-final ending. [2] uses only syllable information
for noun extraction without any dictionary. This allows
the possibility of morphological analysis based only on
syllable information. We will describe the details of the
syllable-based model in the next section.

2. Syllable-based probabilistic model

Probabilistic morphological analysis generates all the
possible interpretations and their probability for a given
Eojeol w. The probability that a given Eojeol w is an-
alyzed to a certain interpretation R is represented as
P(R | w). The interpretation R is made up of a mor-
pheme sequence M and its corresponding POS sequence
T as given in Equation 1.

\[ P(R | w) = P(M, T | w) \] (1)

From Equation 1, the model is derived as follows:

\[ P(R | w) = P(l | w) P(R | l, w) \] (2)
\[ \approx P(l | w) P(R | l) \] (3)
\[ = \frac{P(l | P(w | l) P(R | l) P(l | R)}{P(l)} \] (4)
\[ = \frac{P(w | l)}{P(w)} \times P(R) \] (5)
\[ \approx P(w | l) \times P(R) \] (6)

In Equation 2, a surface level form w corresponds to
the lexical level form l. Equation 3 assumes the inter-
pretation R and the surface level form w are condition-
ally independent given the lexical level form l. By Bayes
rule, Equation 3 becomes Equation 4. Equation 5 regards
P(l | R) of Equation 4 as 1. It is because the lexical level
form l is deterministically derived from the interpretation
R. We can write Equation 5 as Equation 6 because P(w)
is a constant and removing it will not change the rankings
among other interpretations. In Equation 6, the left term
denotes “the morphological restoration model”, and the
right one denotes “the morpheme segmentation and POS
assignment model”.

We describe the morphological restoration model first.
As mentioned before, there are various morpho-
logical phenomena in Korean (e.g. conjugation, con-
traction, elision, etc.). In general, to deal with these
phenomena properly, many morphological rules should
be applied and frequent dictionary look-up is required.
However, our system uses the morphological restoration
model with morphological information instead of mor-
phological rules.

The morphological restoration model is to encode the
probability that the k substrings between the surface form
and its lexical form correspond to each other. The equa-
tion of the model is as follows:

\[ P(w | l) \approx \prod_{j=1}^{k} P(s_j | \tilde{s}_j) \] (7)

where, \( s_j \) and \( \tilde{s}_j \) denote the jth substrings of the surface
form and the lexical form, respectively.

If a surface form (Eojeol) and its lexical form are the
same, each syllable pair of them is mapped one-to-one.
Otherwise, it means that a morphological change occurs.
In this case, the mapping between the two forms, which
is a pair of two longest parts from beginning to end of the
mismatch, is used to calculate the probability. We call a
list of such pairs “morphological information”, which is
also automatically extracted from training data. Table 2
shows examples of applying the morphological restaura-
tion model.

Then, the morpheme segmentation and POS assign-
ment model from Equation 6 is to assign the m syllables
to the m syllable tags, as follows:

\[ P(R) = P(C, U) \] (8)
\[ = P(c_{1,m}, u_{1,m}) \] (9)
\[ \approx \prod_{i=1}^{m} \left( P(c_i | c_{i-2,i-1, i-1,i-1}) P(u_i | c_{i-1,i, i-2,i-1}) \right) \]
\[ \times P(c_{EOW} | c_{m-1,m, u_{m-1,m}}) \]
\[ \times P(u_{EOW} | c_{m}, c_{EOW}, u_{m-1,m}) \] (10)

where, \( C = c_{1,m} \) is the syllable sequence of the lexical
form, and \( U = u_{1,m} \) is its corresponding syllable tag
sequence.

<table>
<thead>
<tr>
<th>na-neun</th>
<th>gam-gi-neun</th>
</tr>
</thead>
<tbody>
<tr>
<td>na/hp+neun/jx</td>
<td>gam-gi/pv+neun/etm</td>
</tr>
<tr>
<td>na/pv+neun/etm</td>
<td>gam-gi/nc+neun/jx</td>
</tr>
<tr>
<td>na/pv+neun/etm</td>
<td>gam/pv+gi/etn+neuni/jx</td>
</tr>
</tbody>
</table>

Table 1: Examples of morphological analysis
In Equation 10, \( c_i \)'s and \( u_i \)'s denote the pseudo syllables and the pseudo tags, respectively, to indicate the beginning of Eojeol when \( i \) is less or equal to zero. Analogously, \( e_{EOW} \) and \( u_{EOW} \) denote the pseudo syllables and the pseudo tags to indicate the end of Eojeol, respectively.

Two Markov assumptions are applied in Equation 10. One is that the probability of a current syllable \( c_i \) conditionally depends only on the previous two syllables and two syllable tags. The other is that the probability of a current syllable tag \( u_i \) conditionally depends only on the previous syllable, the current syllable, and the previous two syllable tags. This model can consider broader context by introducing a less strict independent assumption than HMMs.

In order to convert the syllable sequence \( C \) and the syllable tag sequence \( U \) to the morpheme sequence \( M \) and the morpheme tag sequence \( T \), we can use two additional symbols (‘B’ and ‘I’) to indicate the boundary of morphemes: a ‘B’ denotes the first syllable of a morpheme and an ‘I’ any non-initial syllable. Examples of syllable tagging with BI symbols are given in Table 3.

### 2.1. Implementation

POS tagging models can use the Viterbi algorithm, which is a dynamic programming, to efficiently find the most probable path, while morphological analyzers cannot use the algorithm because they should produce all the results. To avoid the explosion of the search space when using breadth-first search, we adopted a beam search algorithm by exploring only some of the most probable nodes at each state. In this case, it is not guaranteed to find all the (grammatically) possible interpretations, but this enables the system to speed up the process considerably.

Since Maximum likelihood estimation suffers from sparse data; probability estimates of low frequency events led to inaccurate estimates, we use a linear interpolation from Equation 10 as follows:

\[
P(c_i \mid c_{i-2}, \ldots, i-1, u_{i-2}, \ldots, i-1)
\]

\[
= \alpha_1 \cdot P(c_i \mid c_{i-2}, \ldots, i-1, u_{i-2}, \ldots, i-1) + \alpha_2 \cdot P(c_i \mid c_{i-1}, u_{i-2}, \ldots, i-1) + \alpha_3 \cdot P(u_{i-1} \mid c_{i-1}, u_{i-1}) + \alpha_4 \cdot P(c_i \mid u_{i-1})
\]

\[
P(u_i \mid c_{i-1}, \ldots, i, u_{i-2}, \ldots, i-1)
\]

\[
= \beta_1 \cdot P(u_i \mid c_{i-1}, \ldots, i, u_{i-2}, \ldots, i-1) + \beta_2 \cdot P(u_i \mid c_{i-1}, u_{i-1}) + \beta_3 \cdot P(u_i \mid c_{i-1}) + \beta_4 \cdot P(u_i \mid c_i)
\]

We use [3]'s algorithm for calculating the \( \alpha \)s and \( \beta \)s directly.

### 3. Experiments

#### 3.1. Experimental environment

For evaluation, two data sets with different tag sets and annotation guidelines are used: one is ETRI POS tagged corpus and the other is KAIST POS tagged corpus. Randomly selected 33,000 Eojeols and 20,000 Eojeols are used as test data and the rests are used as training data for the ETRI POS corpus and the KAIST POS corpus, respectively. Table 4 shows the summary of the corpora.

In this paper, we use the following measures in order to evaluate the system:

**Answer inclusion rate (AIR)** is defined as the number of Eojeols among whose results contain the gold standard over the entire Eojeols in the test data.

**Average ambiguity (AA)** is defined as the average number of the returned results per Eojeol.
Table 4: Summary of the data

<table>
<thead>
<tr>
<th>Corpus</th>
<th>ETRI</th>
<th>KAIST</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Eojeols in training data</td>
<td>255,291</td>
<td>155,468</td>
</tr>
<tr>
<td># of Eojeols in test data</td>
<td>33,000</td>
<td>20,000</td>
</tr>
<tr>
<td>Total # of Eojeols</td>
<td>288,291</td>
<td>175,468</td>
</tr>
<tr>
<td># of tags</td>
<td>27</td>
<td>54</td>
</tr>
</tbody>
</table>

Failure rate (FR) is defined as the number of Eojeols whose outputs are not produced over the number of Eojeols in the test data.

1-best tagging accuracy (1A) is defined as the number of Eojeols whose result with the highest score is matched to the gold standard over the entire Eojeols in the test data.

There is a trade-off between AIR and AA. If a system outputs the results as many as possible, it is more likely to include the correct answer among them, but this lead to an increase of the ambiguity, and vice versa. The higher AIR is, the better the system. The AIR can be an upper bound on the accuracy of POS taggers. On the contrary to AIR, the lower AA is, the better the system. Low AA value can reduce the burden of the disambiguation process of POS tagger. Although the 1A cannot be used as a common evaluation measure for morphological analysis because previous systems cannot rank the results by a certain criterion, ProKOMA can be evaluated by the measure because it can provide the probabilities of the results. This measure can also be served as a baseline for the POS tagging.

3.2. Experimental results

To investigate the performance and the effectiveness of the model, we conduct several tests and also perform the experiments on the two corpora. The results of the experiments are shown in Table 5.

Table 5: Experimental results with ETRI corpus and KAIST corpus

<table>
<thead>
<tr>
<th>Corpus</th>
<th>ETRI</th>
<th>KAIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer inclusion rate (%)</td>
<td>98.70</td>
<td>98.17</td>
</tr>
<tr>
<td>Average ambiguity</td>
<td>3.08</td>
<td>2.56</td>
</tr>
<tr>
<td>Failure rate (%)</td>
<td>0.08</td>
<td>0.06</td>
</tr>
<tr>
<td>1-best accuracy (%)</td>
<td>89.90</td>
<td>89.95</td>
</tr>
</tbody>
</table>

For comparison with other systems, we list the performances of two systems participated in MATEC 99 in Table 6. The ETRI corpus was used in MATEC 99. The number of Eojeols used in the test data is 33,855. The evaluation data used in MATEC 99 and ours are not exactly same, but are close. As can be seen, the [4]’s system is better than ours in terms of AIS but generates too many results. Our system outperforms the [5]’s system.

Table 6: Performances of two systems participated in MATEC 99

<table>
<thead>
<tr>
<th></th>
<th>[4]’s system</th>
<th>[5]’s system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer inclusion rate (%)</td>
<td>98</td>
<td>92</td>
</tr>
<tr>
<td>Average ambiguity</td>
<td>4.13</td>
<td>1.75</td>
</tr>
</tbody>
</table>

4. Conclusion

We have presented and described the new probabilistic models used in our Korean morphological analyzer. The previous systems depend on manually constructed linguistic knowledge such as morpheme dictionary, morphosyntactic rules, and morphological rules. The system, however, requires no manual labour because all the information can be automatically acquired by the POS tagged corpora. We also showed that the system is portable and flexible by the experiments on two different corpora.

The previous systems take morpheme as a processing unit, but we take syllable. According to the experiments, the system achieved comparable performances with the previous systems.

5. References


