Probabilistic Models for Korean Morphological Analysis

Do-Gil Lee and **Hae-Chang Rim**
Dept. of Computer Science & Engineering
Korea University
1, 5-ka, Anam-dong, Seongbuk-gu
Seoul 136-701, Korea
{dglee, rim}@nlp.korea.ac.kr

Abstract

This paper discusses Korean morphological analysis and presents three probabilistic models for morphological analysis. Each model exploits a distinct linguistic unit as a processing unit. The three models can compensate for each other's weaknesses. Contrary to the previous systems that depend on manually constructed linguistic knowledge, the proposed system can fully automatically acquire the linguistic knowledge from annotated corpora (e.g. part-of-speech tagged corpora). Besides, without any modification of the system, it can be applied to other corpora having different tagsets and annotation guidelines. We describe the models and present evaluation results on three corpora with a wide range of conditions.

1 Introduction

This paper discusses Korean morphological analysis. Morphological analysis is to break down an Eojeol[1] into morphemes, which is the smallest meaningful unit. The jobs to do in morphological analysis are as follows:

- Separating an Eojeol into morphemes
- Assigning the morpho-syntactic category to each morpheme
- Restoring the morphological changes to the original form

We have to consider some difficult points in Korean morphology: there are two kinds of ambiguities (segmentation ambiguity and part-of-speech ambiguity). Moreover, 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) regal interpretations. Table 1 gives examples of morphological analysis for Eojeols "*na-neun*" and "*gam-gi-neun*".

Previous works on morphological analysis depends on manually constructed linguistic knowledge such as morpheme dictionary, morphosyntactic rules, and morphological rules. There are two major disadvantages in this approach:

- Construction of the knowledge base is time-consuming and labor-intensive. In addition, storing every word in a lexicon is impossible so the previous approch suffers 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.

The proposed morphological analyzer, ProKOMA, tries to overcome these limitations: Firstly, it uses only 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 labor in constructing and maintaining such a knowledge base.

[1] Eojeol is the surface level form of Korean and is the spacing unit delimited by a whitespace.

Table 1: Examples of morphological analysis

na-neun		*gam-gi-neun*	
na/np+*neun*/jx	'I am'	*gam-gi*/pv+*neun*/etm	'be wound'
na/pv+*neun*/etm	'to sprout'	*gam-gi*/nc+*neun*/jx	'a cold is'
nal/pv+*neun*/etm	'to fly'	*gam*/pv+*gi*/etn+*neun*/jx	'to wash is'

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 by the probabilistic models. In Korean, no attempt has been made at probabilistic approach to morphological analysis. Probabilities enable the system to rank the results and to provide the probabilities to the next module such as POS tagger.

2 Related works

Over the past few decades, a considerable number of studies have been made on Korean morphological analysis. The early studies concentrated on the algorithmic research. The following approaches belong to this group: longest matching algorithm, tabular parsing method using CYK algorithm (Kim, 1986), dictionary based approach (Kwon, 1991), two-level morphology(Lee, 1992), and syllable-based approach (Kang and Kim, 1994).

Next, many studies have been made on improving the efficiency of the morphological analyzers. There have been studies to reduce the search space and implausible interpretations by using characteristics of Korean syllables (Kang, 1995; Lim et al., 1995).

There have been no standard tagset and annotation guideline, so researchers have developed methods with their own tagsets and guidelines. The Morphological Analysis and Tagger Evaluation Contest (MATEC) took place in 1999. This is the first trial about the objective and relative evaluation of morphological analysis. Among 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 described so far can be considered as the so called dictionary and rule based approach. In this approach, the quality of the dictionary and the rules govern the system's performance.

The proposed approach is the first attempt to probabilistic morphological analysis. The aim of the paper is to show that this approach can achieve comparable performances with the previous approaches.

3 Probabilistic morphological analysis model

Probabilistic morphological analysis generates all the possible interpretations and their probabilities for a given Eojeol w. The probability that a given Eojeol w is analyzed to a certain interpretation R is represented as $P(R\,|\,w)$. The interpretation R is made up of a morpheme sequence M and its corresponding POS sequence T as given in Equation 1.

$$P(R\,|\,w) = P(M, T\,|\,w) \tag{1}$$

In the following subsections, we describe the three morphological analysis models based on three different linguistic units (Eojeol, morpheme, and syllable[2]).

3.1 Eojeol-unit model

For the Eojeol-unit model, it is sufficient to store the frequencies of each Eojeol (surface level form) and its interpretation acquired from the POS tagged corpus[3].

The probabilities of Equation 1 are estimated by the maximum likelihood estimator (MLE) using relative frequencies in the training data.

[2] In Korean written text, each character has one syllable. We do not distinguish between character and syllable in this paper.

[3] ProKOMA extracts only Eojeols occurred five times or more in training data.

The most prominent advantage of the Eojeol-unit analysis is its simplicity. As mentioned before, morphological analysis of Korean is very complex. The Eojeol-unit analysis can avoid such complex process so that it is very efficient and fast. Besides, it can reduce unnecessary results by only producing the interpretations that really appeared in the corpus. So, we also expect an improvement in accuracy.

Due to the high productivity of Korean Eojeol, the number of possible Eojeols is very large so storing all kinds of Eojeols is impossible. Therefore, using the Eojeol-unit analysis alone is undesirable, but a small number of Eojeols with high frequency can cover a significant portion of the entire ones, thus this model will be helpful.

3.2 Morpheme-unit model

As discussed, not all Eojeols can be covered by the Eojeol-unit analysis. The ultimate goal of morphological analysis is to recognize every morpheme within an Eojeol. For these reasons, most previous systems have used morpheme as a processing unit for morphological analysis.

The morpheme-unit morphological analysis model is derived as follows by introducing lexical form l:

$$P(R|w) \quad = \quad P(l|w)P(R|l,w) \qquad (2)$$

where l should satisfy the following condition:

$$l \in L_w \cap P(l|R) = 1$$

where L_w is a set of lexical forms that can be derived from the surface form w. This condition means that among all possible lexical forms for a given w (L_w), the only lexical form l is deterministically derived from the interpretation R.

$$\begin{aligned} P(l|w)P(R|l,w) &\approx P(l|w)P(R|l) && (3) \\ &\approx P(l|w)P(R) && (4) \\ &= P(l|w)P(M,T) && (5) \end{aligned}$$

Equation 3 assumes the interpretation R and the surface form w are conditionally independent given the lexical form l. Since the lexical form l is underlying in the morpheme sequence M[4], the lexical form l can be omitted as in equation 4. In Equation 5, the left term $P(l\,|\,w)$ denotes "the morphological restoration model", and the right $P(M,T)$ "the morpheme segmentation and POS assignment model".

We describe the morphological restoration model first. The model is the probability of the lexical form given a surface form and is to encode the probability that the k substrings between the surface form and its lexical form correspond to each other. The equation of the model is as follows:

$$P(l\,|\,w) \approx \prod_{j=1}^{k} P(\hat{s}_j\,|\,s_j) \qquad (6)$$

where, s_j and $\hat{s}_j$ denote the jth substrings of the surface form and the lexical form, respectively.

We call such pairs of substrings "morphological information". This information can be acquired by the following steps: If a surface form (Eojeol) and its lexical form are the same, each syllable pair of them is mapped one-to-one and extracted. Otherwise, it means that a morphological change occurs. In this case, the pair of two substrings from the beginning to the end of the mismatch is extracted. The morphological information is also automatically extracted from training data. Table 2 shows some examples of applying the morphological restoration model.

Now we turn to the morpheme segmentation and POS assignment model. It is the joint probability of the morpheme sequence and the tag sequence.

$$\begin{aligned} P(M,T) &= P(m_{1,n}, t_{1,n}) \\ &\approx \prod_{i=1}^{n} P(m_i\,|\,t_i)P(t_i\,|\,t_{i-1}) \\ &\quad \times P(t_{EOW}\,|\,t_n) \qquad (7) \end{aligned}$$

In equation 7, t_0 and t_{EOW} are pseudo tags to indicate the beginning and the end of Eojeol, respectively. We introduce the t_{EOW} symbol to reflect the preference for well-formed structure of a given Eojeol. The model is represented as the well-known bigram hidden Markov model (HMM), which is widely used in POS tagging.

The morpheme dictionary and the morphosyntactic rules that have been used in the previous

[4] A lexical form is just the concatenation of morphemes.

Table 2: Examples of applying the morphological restoration model

Surface form	Lexical form	Probability	Description
na-neun	*na-neun*	$P(na \mid na)P(neun \mid neun)$	No phonological change
na-neun	*nal-neun*	$P(nal \mid na)P(neun \mid neun)$	‘l’ irregular conjugation
go-ma-wo	*go-mab-eo*	$P(go \mid go)P(mab\text{-}eo \mid ma\text{-}wo)$	‘b’ irregular conjugation
beo-lyeo	*beo-li-eo*	$P(beo \mid beo)P(li\text{-}eo \mid lyeo)$	Contraction
ga-seo	*ga-a-seo*	$P(ga \mid ga)P(a\text{-}seo \mid seo)$	Ellipsis

approaches are included in the lexical probability $P(m_i \mid t_i)$ and the transition probability $P(t_i \mid t_{i-1})$.

3.3 Syllable-unit model

One of the most difficult problems in morphological analysis is the unknown word problem, which is caused by the fact that we cannot register every possible morpheme in the dictionary. In English, contextual information and suffix information is helpful to estimate the POS tag of an unknown word. In Korean, the syllable characteristics can be utilized. For instance, a syllable “*eoss*” can only be a pre-final ending.

The syllable-unit model is derived from Equation 4 as follows:

$$P(l \mid w)P(R) = P(l \mid w)P(C,U) \tag{8}$$

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

In the above equation, $P(l \mid w)$ is the same as that of the morpheme-unit model (Equation 6), we use the morpheme-unit model’s result as it is. The right term $P(C,U)$ is referred to as “the POS assignment model”.

The POS assignment model is to assign the m syllables to the m syllable tags:

$$\begin{aligned} P(C,U) &= P(c_{1,m}, u_{1,m}) && (9)\\ &\approx \prod_{i=1}^{m} \begin{pmatrix} P(c_i \mid c_{i-2,i-1}, u_{i-2,i-1}) \\ P(u_i \mid c_{i-1,i}, u_{i-2,i-1}) \end{pmatrix} \\ &\quad \times P(c_{EOW} \mid c_{m-1,m}, u_{m-1,m}) \\ &\quad \times P(u_{EOW} \mid c_m, c_{EOW}, u_{m-1,m}) && (10) \end{aligned}$$

In Equation 10, when i is less than or equal to zero, c_is and u_is denote the pseudo syllables and the pseudo tags, respectively. They indicate the beginning of Eojeol. Analogously, c_{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 the current syllable c_i conditionally depends only on the previous two syllables and two syllable tags. The other is that the probability of the 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 the less strict independent assumption than the HMM.

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-unit tagging with BI symbols are given in Table 3.

4 Experiments

4.1 Experimental environment

For evaluation, three data sets having different tag sets and annotation guidelines are used: ETRI POS tagged corpus, KAIST POS tagged corpus, and Sejong POS tagged corpus. All experiments were performed by the 10-fold cross-validation. Table 4 shows the summary of the corpora.

Table 4: Summary of the data

Corpus	ETRI	KAIST	Sejong
# of Eojeols	288,291	175,468	2,015,860
# of tags	27	54	41

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

Table 3: Examples of syllable tagging with BI symbols

Eojeol	*na-neun* ‘I’		*hag-gyo-e* ‘to school’			*gan-da* ‘go’		
Tagged Eojeol	*na*/np+*neun*/jx		*hag-gyo*/nc+*e*/jc			*ga*/pv+*n-da*/ef		
Morpheme	*na*	*neun*	*hag-gyo*		*e*	*ga*	*n-da*	
Morpheme tag	np	jx	nc		jc	pv	ef	
Syllable	*na*	*neun*	*hag*	*gyo*	*e*	*ga*	*n*	*da*
Syllable tag	B-np	B-jx	B-nc	I-nc	B-jc	B-pv	B-ef	I-ef

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 returned results per Eojeol by the system.

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 of which only one interpretation with highest probability per Eojeol 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 many results, it is likely to include the correct answer in them, but this leads 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. A low AA can reduce the burden of the disambiguation process of the POS tagger. Although the 1A is not used as a common evaluation measure for morphological analysis because previous systems do not rank the results, ProKOMA can be evaluated by this measure because it provides the probabilities for the results. This measure can also be served as a baseline for POS tagging.

4.2 Experimental results

To investigate the performance and the effectiveness of the three models, we conducted several tests according to the combinations of the models. For each test, we also performed the experiments on the three corpora. The results of the experiments are listed in Table 5. In the table, “E”, “M”, and “S” mean the Eojeol-unit analysis, the morpheme-unit analysis, and the syllable-unit analysis, respectively. The columns having more than one symbol mean that each model performs sequentially.

According to the results, when applying a single model, each model shows the significant differences, especially between “E” and “S”. Because of low coverage of the Eojeol-unit analysis, “E” shows the lowest AIR and the highest FR. However, it shows the lowest AA because it produces the small number of results. On the contrary, “S” shows the highest AA but the best performances on AIR and FR, which is caused by producing many results.

Most previous systems use morpheme as a processing unit for morphological analysis. We would like to examine the effectiveness of the proposed models based on Eojeol and syllable. First, compare the models that use the Eojeol-unit analysis with others (“M” vs. “EM”, “S” vs. “ES”, and “MS” vs. “EMS”). When applying the Eojeol-unit analysis, AA is decreased, and AIS and 1A are increased. Then, compare the models that use the syllable-unit analysis with others (“E” vs. “ES”, “M” vs. “MS”, and “EM” vs. “EMS”). When applying the syllable-unit analysis, AIR and 1A are increased, and FR is decreased. Therefore, both models are very useful when compared the morpheme-unit model only.

Compared with the performances of two systems that participated in MATEC 99, we listed the results in Table 6. In this evaluation, the ETRI corpus was used and the number of Eojeols included in the test data is 33,855. The evaluation data used in MATEC 99 and ours are not the same, but are close. As can be

Table 5: Experimental results according to the combination of the processing units

Data	Measure	E	M	S	EM	ES	MS	EMS
ETRI	Answer inclusion rate (%)	54.65	93.87	98.91	94.16	98.81	97.09	97.22
	Average ambiguity	1.23	2.63	6.95	2.10	4.46	2.95	2.41
	Failure rate (%)	45.21	3.81	0.06	3.67	0.06	0.06	0.06
	1-best accuracy (%)	51.66	83.49	89.98	86.41	91.22	86.15	88.92
KAIST	Answer inclusion rate (%)	57.43	94.29	98.36	94.41	98.25	96.97	97.02
	Average ambiguity	1.26	1.84	6.05	1.57	3.80	2.16	1.89
	Failure rate (%)	42.40	3.73	0.06	3.67	0.06	0.06	0.06
	1-best accuracy (%)	54.22	87.51	90.02	89.18	91.02	89.50	91.12
Sejong	Answer inclusion rate (%)	67.79	90.96	99.38	92.17	99.33	96.60	97.09
	Average ambiguity	1.29	2.35	6.60	1.82	3.52	2.72	2.17
	Failure rate (%)	32.13	5.94	0.02	5.21	0.02	0.02	0.02
	1-best accuracy (%)	64.64	83.86	91.56	87.00	92.96	88.72	91.16

Table 6: Performances of two systems participated in MATEC 99

	(Lee et al., 1999)'s system	(Song et al., 1999)'s system
Answer inclusion rate (%)	98	92
Average ambiguity	4.13	1.75

seen, the Lee et al. (1999)'s system is better than ProKOMA in terms of AIS, but it generates too many results (with higher AA).

5 Conclusion

We have presented and described the new probabilistic models used in our Korean morphological analyzer ProKOMA. The previous systems depend on manually constructed linguistic knowledge such as morpheme dictionary, morphosyntactic rules, and morphological rules. The system, however, requires no manual labor 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 three different corpora.

The previous systems take morpheme as a processing unit, but we take three kinds of processing units (e.g. Eojeol, morpheme, and syllable). According to the experiments, we can know that the Eojeol-unit analysis contributes efficiency and accuracy, and the syllable-unit analysis is robust in the unknown word problem and also contributes accuracy. Finally, the system achieved comparable performances with the previous systems.

References

S.-S. Kang and Y.-T. Kim. 1994. Syllable-based model for the Korean morphology. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 221–226.

S.-S. Kang. 1995. Morphological analysis of Korean irregular verbs using syllable characteristics. *Journal of the Korea Information Science Society*, 22(10):1480–1487.

S.-Y. Kim. 1986. A morphological analyzer for Korean language with tabular parsing method and connectivity information. Master's thesis, Dept. of Computer Science, Korea Advanced Institute of Science and Technology.

H.-C. Kwon. 1991. Dictionary-based morphological analysis. In *Proceedings of the Natural Language Processing Pacific Rim Symposium*, pages 87–91.

S.-Z. Lee, B.-R. Park, J.-D. Kim, W.-H. Ryu, D.-G. Lee, and H.-C. Rim. 1999. A predictive morphological analyzer, a part-of-speech tagger based on joint independence model, and a fast noun extractor. In *Proceedings of the MATEC 99*, pages 145–150.

S.-J. Lee. 1992. A two-level morphological analysis of Korean. Master's thesis, Dept. of Computer Science, Korea Advanced Institute of Science and Technology.

H.-S. Lim, S.-Z. Lee, and H.-C. Rim. 1995. An efficient Korean mophological analysis using exclusive information. In *Proceedings of the 1995 International Conference on Computer Processing of Oriental Languages*, pages 255–258.

T.-J. Song, G.-Y. Lee, and Y.-S. Lee. 1999. Morphological analyzer using longest match method for syntactic analysis. In *Proceedings of the MATEC 99*, pages 157–166.