Part-of-Speech Tagging Considering Surface Form for an Agglutinative Language

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

Abstract

The previous probabilistic part-of-speech tagging models for agglutinative languages have considered only lexical forms of morphemes, not surface forms of words. This causes an inaccurate calculation of the probability. The proposed model is based on the observation that when there exist words (surface forms) that share the same lexical forms, the probabilities to appear are different from each other. Also, it is designed to consider lexical form of word. By experiments, we show that the proposed model outperforms the bigram Hidden Markov model (HMM)-based tagging model.

1 Introduction

Part-of-speech (POS) tagging is a job to assign a proper POS tag to each linguistic unit such as word for a given sentence. In English POS tagging, word is used as a linguistic unit. However, the number of possible words in agglutinative languages such as Korean is almost infinite because words can be freely formed by gluing morphemes together. Therefore, morpheme-unit tagging is preferred and more suitable in such languages than word-unit tagging. Figure 1 shows an example of morpheme structure of a sentence, where the bold lines indicate the most likely morpheme-POS sequence. A solid line represents a transition between two morphemes across a word boundary and a dotted line represents a transition between two morphemes in a word.

The previous probabilistic POS models for agglutinative languages have considered only lexical forms of morphemes, not surface forms of words. This causes an inaccurate calculation of the probability. The proposed model is based on the observation that when there exist words (surface forms) that share the same lexical forms, the probabilities to appear are different from each other. Also, it is designed to consider lexical form of word. By experiments, we show that the proposed model outperforms the bigram Hidden Markov model (HMM)-based tagging model.

2 Korean POS tagging model

In this section, we first describe the standard morpheme-unit tagging model and point out a mistake of this model. Then, we describe the proposed model.

2.1 Standard morpheme-unit model

This section describes the HMM-based morpheme-unit model. The morpheme-unit POS tagging model is to find the most likely sequence of morphemes M and corresponding POS tags T for a given sentence W, as follows (Kim et al., 1998; Lee et al., 2000):

$$\begin{aligned}\Gamma(W) &\stackrel{def}{=} \operatorname*{argmax}_{M,T} P(M,T\,|\,W) \\ &= \operatorname*{argmax}_{m_{1,u},t_{1,u}} P(m_{1,u},t_{1,u}\,|\,w_{1,n}) \quad (1)\\ &\approx \operatorname*{argmax}_{m_{1,u},t_{1,u}} P(m_{1,u},t_{1,u}) \quad (2)\end{aligned}$$

In the equation, $u(>= n)$ denotes the number of morphemes in the sentence. A sequence of $W = w_{1,n} = w_1 w_2 \cdots w_n$ is a sentence of n words, and a sequence of $M = m_{1,u} = m_1 m_2 \cdots m_u$ and a sequence of $T = t_{1,u} = t_1 t_2 \cdots t_u$ denote a sequence of u lexical forms of morphemes and a sequence of u morpheme categories (POS tags), respectively.

To simplify Equation 2, a Markov assumption is usually used as follows:

$$\Gamma(W) \approx \operatorname*{argmax}_{m_{1,u},t_{1,u}} \prod_{i=1}^{u} P(t_i\,|\,t_{i-1},p)P(t_i\,|\,m_i) \quad (3)$$

where, t_0 is a pseudo tag which denotes the beginning of word and is also written as BOW. p denotes a type of transition from the previous tag to the current tag. It has a binary value according to the type of the transition (either intra-word or inter-word transition).

As can be seen, the word[1] sequence $w_{1,n}$ is discarded in Equation 2. This leads to an inaccurate

[1] A word is a surface form.

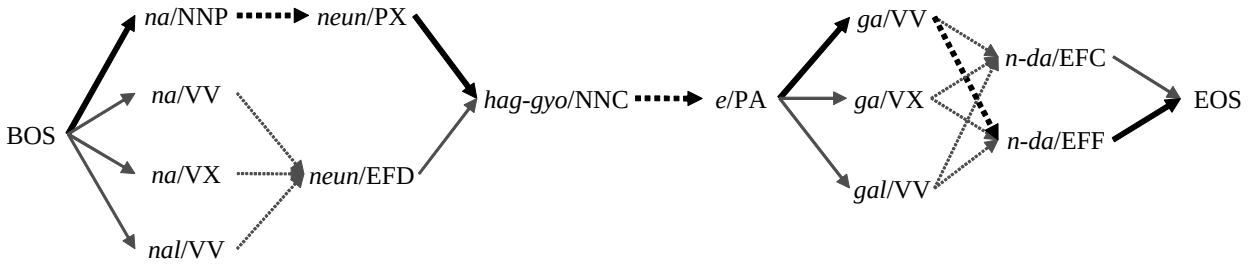

Figure 1: Morpheme structure of the sentence "*na-neun hag-gyo-e gan-da*" (I go to school)

calculation of the probability. A lexical form of a word can be mapped to more than one surface word. In this case, although the different surface forms are given, if they have the same lexical form, then the probabilities will be the same. For example, a lexical form *mong-go/nc+leul/jc*[2], can be mapped from two surface forms *mong-gol* and *mong-go-leul*. By applying Equation 1 and Equation 2 to both words, the following equations can be derived:

$$P(\textit{mong-go}, nc, \textit{leul}, jc \,|\, \textit{mong-gol}) \approx P(\textit{mong-go}, nc, \textit{leul}, jc) \quad (4)$$

$$P(\textit{mong-go}, nc, \textit{leul}, jc \,|\, \textit{mong-go-leul}) \approx P(\textit{mong-go}, nc, \textit{leul}, jc) \quad (5)$$

As a result, we can acquire the following equation from Equation 4 and Equation 5:

$$P(\textit{mong-go}, nc, \textit{leul}, jc \,|\, \textit{mong-gol}) = P(\textit{mong-go}, nc, \textit{leul}, jc \,|\, \textit{mong-go-leul}) \quad (6)$$

That is, they assume that probabilities of the results that have the same lexical form are the same. However, we can easily show that Equation 6 is mistaken: Actually, $P(\textit{mong-go}, nc, \textit{leul}, jc \mid \textit{mong-go-leul}) = 1$ and $P(\textit{mong-gol}, nc \mid \textit{mong-gol}) \neq 0$. Hence, $P(\textit{mong-go}, nc, \textit{leul}, jc \mid \textit{mong-gol}) < P(\textit{mong-go}, nc, \textit{leul}, jc \,|\, \textit{mong-go-leul})$.

To overcome the disadvantage, we propose a new tagging model that can consider the surface form.

2.2 The proposed model

This section describes the proposed model. To simplify the notation, we introduce a variable R, which means a tagging result of a given sentence and consists of M and T.

$$\Gamma(W) \stackrel{def}{=} \underset{M,T}{\operatorname{argmax}} P(M, T \,|\, W) \quad (7)$$

$$= \underset{R}{\operatorname{argmax}} P(R \,|\, W) \quad (8)$$

[2] *mong-go* means Mongolia, *nc* is a common noun, and *jc* is a objective case postposition.

The probability $P(R|W)$ is given as follows:

$$P(R \,|\, W) = P(r_{1,n} \,|\, w_{1,n}) \quad (9)$$

$$= \prod_{i=1}^{n} P(r_i \,|\, w_{1,n}, r_{1,i-1}) \quad (10)$$

$$\approx \prod_{i=1}^{n} P(r_i \,|\, w_i, r_{i-1}) \quad (11)$$

where, r_i denotes the tagging result of ith word (w_i), and r_0 denotes a pseudo variable to indicate the beginning of word. Equation 9 becomes Equation 10 by the chain rule. To be a more tractable form, Equation 10 is simplified by a Markov assumption as Equation 11.

The probability $P(r_i \,|\, w_i, r_{i-1})$ cannot be calculated directly, so it is derived as follows:

$$P(r_i \,|\, w_i, r_{i-1}) = \frac{P(w_i, r_{i-1}, r_i)}{P(w_i, r_{i-1})} \quad (12)$$

$$\approx \frac{P(w_i)P(r_i \,|\, w_i)P(r_{i-1} \,|\, r_i)}{P(w_i)P(r_{i-1})} \quad (13)$$

$$= \frac{P(r_i \,|\, w_i)P(r_{i-1} \,|\, r_i)}{P(r_{i-1})} \quad (14)$$

$$= P(r_i \,|\, w_i)\frac{P(r_{i-1}, r_i)}{P(r_{i-1})P(r_i)} \quad (15)$$

Equation 12 is derived by Bayes rule, Equation 13 by a chain rule and an independence assumption, and Equation 15 by Bayes rule. In Equation 15, we call the left term "morphological analysis model" and right one "transition model".

The morphological analysis model $P(r_i \,|\, w_i)$ can be implemented in a morphological analyzer. If a morphological analyzer can provide the probability, then the tagger can use the values as they are. Actually, we use the probability that a morphological analyzer, ProKOMA (Lee and Rim, 2004) produces. Although it is not necessary to discuss the morphological analysis model in detail, we should note that surface forms are considered here.

The transition model is a form of point-wise mutual information.

$$\frac{P(r_{i-1}, r_i)}{P(r_{i-1})P(r_i)} = \frac{P(M_{i-1}, T_{i-1}, M_i, T_i)}{P(M_{i-1}, T_{i-1})P(M_i, T_i)} \quad (16)$$

$$= \frac{P(m_{1,j}^{i-1}, t_{1,j}^{i-1}, m_{1,k}^{i}, t_{1,k}^{i})}{P(m_{1,j}^{i-1}, t_{1,j}^{i-1})P(m_{1,k}^{i}, t_{1,k}^{i})} \quad (17)$$

where, a superscript i in $m_{1,k}^{i}$ and $t_{1,k}^{i}$ denotes the position of the word in a sentence.

The denominator means a joint probability that the morphemes and the tags in a word appear together, and the numerator means a joint probability that all the morphemes and the tags between two words appear together. Due to the sparse data problem, they cannot also be calculated directly from the test data. By a Markov assumption, the denominator and the numerator can be broken down into Equation 18 and Equation 19, respectively.

$$P(m_{1,k}, t_{1,k}) = \prod_{l=1}^{k} P(t_l \,|\, t_{l-1}) P(m_l \,|\, t_l) \quad (18)$$

$$P(m_{1,j}^{i-1}, t_{1,j}^{i-1}, m_{1,k}^{i}, t_{1,k}^{i}) = \prod_{l=1}^{j} \begin{pmatrix} P(t_l^{i-1} \,|\, t_{l-1}^{i-1}) \\ P(m_l^{i-1} \,|\, t_l^{i-1}) \end{pmatrix} \times P_{inter}(t_1^i \,|\, t_j^{i-1}) \times P(m_1^i \,|\, t_1^i) \times \prod_{m=2}^{k} \begin{pmatrix} P(t_m^i \,|\, t_{m-1}^i) \\ P(m_m^i \,|\, t_m^i) \end{pmatrix} \quad (19)$$

where, $P_{inter}(t_1^i \,|\, t_j^{i-1})$ means a transition probability between the last morpheme of the $(i-1)$th word and the first morpheme of the ith word.

By applying Equation 18 and Equation 19 to Equation 17, we obtain the following equation:

$$\frac{P(r_{i-1}, r_i)}{P(r_{i-1})P(r_i)} = \frac{P_{inter}(t_1^i \,|\, t_j^{i-1})}{P(t_1^i \,|\, BOW)} \quad (20)$$

For a given sentence, Figure 2 shows the bigram HMM-based tagging model, and Figure 3 the proposed model. The main difference between the two models is the proposed model considers surface forms but the HMM does not.

3 Experiments

For evaluation, two data sets are used: ETRI POS tagged corpus and KAIST POS tagged corpus. We divided the test data into ten parts. The performances of the model are measured by averaging over the ten test sets in the 10-fold cross-validation experiment. Table 1 shows the summary of the corpora.

Table 1: Summary of the data

Corpus	ETRI	KAIST
Total # of words	288,291	175,468
Total # of sentences	27,855	16,193
# of tags	27	54

Generally, POS tagging goes through the following steps: First, run a morphological analyzer, where it generates all the possible interpretations for a given input text. Then, a POS tagger takes the results as input and chooses the most likely one among them. Therefore, the performance of the tagger depends on that of the preceding morphological analyzer.

If the morphological analyzer does not generate the exact result, the tagger has no chance to select the correct one, thus an answer inclusion rate of the morphological analyzer becomes the upper bound of the tagger. The previous works preprocessed the dictionary to include all the exact answers in the morphological analyzer's results. However, this evaluation method is inappropriate to the real application in the strict sense. In this experiment, we present the accuracy of the morphological analyzer instead of preprocessing the dictionary. ProKOMA's results with the test data are listed in Table 2.

Table 2: Morphological analyzer's results with the test data

Corpus	ETRI	KAIST
Answer inclusion rate (%)	95.82	95.95
Average # of results per word	2.16	1.81
1-best accuracy (%)	88.31	90.12

In the table, 1-best accuracy is defined as the number of words whose result with the highest probability is matched to the gold standard over the entire words in the test data. This can also be a tagging model that does not consider any outer context.

To compare the proposed model with the standard model, the results of the two models are given in Table 3. As can be seen, our model outperforms the HMM model. Moreover, the HMM model is even worse than the ProKOMA's 1-best accuracy. This tells that the standard HMM by itself is not a good model for agglutinative languages.

4 Conclusion

We have presented a new POS tagging model that can consider the surface form for Korean, which

Figure 2: Lattice of the bigram HMM-based model

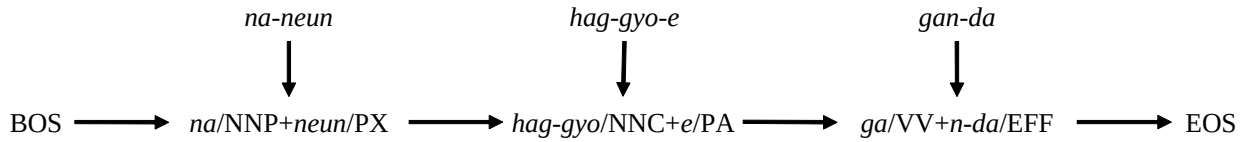

Figure 3: Lattice of the proposed model

Table 3: Tagging accuracies (%) of the standard HMM and the proposed model

Corpus	ETRI	KAIST
The standard HMM	87.47	89.83
The proposed model	90.66	92.01

is an agglutinative language. Although the model leaves much room for improvement, it outperforms the HMM based model according to the experimental results.

Acknowledgement

This work was supported by Korea Research Foundation Grant (KRF-2003-041-D20485)

References

J.-D. Kim, S.-Z. Lee, and H.-C. Rim. 1998. A morpheme-unit POS tagging model considering word-spacing. In *Proceedings of the 1998 Conference on Hangul and Korean Information Processing*, pages 3–8.

D.-G. Lee and H.-C. Rim. 2004. ProKOMA: A probabilistic Korean morphological analyzer. Technical Report KU-NLP-04-01, Department of Computer Science and Engineering, Korea University.

S.-Z. Lee, Jun'ichi Tsujii, and H.-C. Rim. 2000. Hidden markov model-based Korean part-of-speech tagging considering high agglutinativity, word-spacing, and lexical correlativity. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*.