Learning the lexicon from raw texts for open-vocabulary Korean word recognition

Sungho Ryu and Jin Hyung Kim
Division of Computer Science, KAIST
Kusung-Dong, Yusung-Gu, Daejon, Korea
E-mail: {shryu, jkim}@ai.kaist.ac.kr

Abstract

In this paper, we propose a novel method of building a language model for open-vocabulary Korean word recognition. Due to the complex morphology of Korean, it is inappropriate to use lexicons based on the linguistic entities such as words and morphemes in open-vocabulary domains. Instead, we build the lexicon by collecting variable length character sequences from the raw texts using a dynamic Bayesian network model of the language.

In simulated word recognition experiments, the proposed language model could find correct words from lattices of character candidates in 94.3% of cases, increasing the word recognition rates by 20.9%.

1. Introduction

In character recognition, the language model provides contextual information that can be used for resolving ambiguities. The likelihoods of recognition results evaluated by the language model are used as cues for choosing more likely one among alternatives.

In open-vocabulary domains, the language model is generally a Markov model that represents phrases and sentences in terms of more basic elements such as words. The set of these basic elements used in a particular language model is called its *lexicon*.

Proper choice of the lexicon is crucial because it is the primary factor that decides the performance and complexity of the language model. In order to get reliable likelihood estimates from the language model, the lexicon should be large enough to get sufficient coverage over the texts of target domain. However, if the lexicon is too large, the language model suffers from the data sparseness problem.

In this paper, we propose a novel language model of Korean for open-vocabulary domains. The proposed language model describes a sentence in terms of variable length character sequences and their joint probability distribution using a dynamic Bayesian network. Both the lexicon and its associated probabilities in the model are trained using raw texts without any linguistic analysis.In the latter part of the paper, we address the shortcomings of linguistic entity-based language models, and explain the structure and the training algorithm of the proposed model.

2. Morphology of Korean

The basic unit that comprises a Korean sentence is called an *eojeol*. The eojeol is roughly equivalent to a word phrase in English, or a *bunsetsu* in Japanese. An eojeol is a sequence of stem morphemes followed by functional morphemes. The functional morphemes define the eojeol's role such as the noun's syntactic role or the verb's tense. The number of morphemes that constitute an eojeol can be arbitrary.

Table 1. Examples of eojeols

Eojeol	Morphemes
□□□ in the sky	□□/ncn + □/jca the sky in
□□□ sound	□□/ncn + □/jcs sound subjective particle

In texts, each eojeol is separated from others by spaces. However, there are no explicit delimiters between morphemes within an eojeol. Therefore, a dedicated morphological analysis is necessary in order to identify the morphemes that constitute an eojeol.

Table 2. Examples of irregular inflections

Eojeol	Morphemes
□□ Blue	□□/paa + □/etm Blue present tense
□□□□ Did	□/pvg + □/ep + □□□/ef do past tense honorific declarative

The morphological analysis in Korean is not a simple task. Inter-morpheme coarticulation effects are common within an eojeol. And the orthography of Korean prefers to transcribe the frequently used eojeols as they are pronounced, rather than preserving the morphemes' own forms. All of these irregularities due to coarticulations should be taken into account when analyzing the morphological structure of eojeols. The examples in table 2 show the discrepancies between the eojeols and the morphemes. In texts containing colloquial style dialogs or dialects, the analysis becomes more complex because the discrepancies are more various and frequent.

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03)
0-7695-1960-1/03 $17.00 © 2003 IEEE

3. Automatic lexicon selection for Korean

3.1. Motivation

In open-vocabulary domains, both eojeols and morphemes have their own shortcomings to be used as the tokens of the lexicon of language models. Because the eojeols have many variations, it is infeasible to build a sufficiently large eojeol-based lexicon and estimate its associated probabilities. Therefore, the eojeol based lexicons suffer from high out-of-vocabulary rates, even if it contains millions of entries.

The morpheme-based language models are less susceptible to the data sparseness problem. These models have been used in large vocabulary continuous speech recognition problems such as the broadcast news recognition[6]. However, the automatic morphological analysis can be applied reliably only to grammatically correct well-formed sentences. Moreover, since the morpheme is not a mathematically defined entity, the implementation of morphological analysis requires subjective judgments on several issues. This makes building the morpheme-based lexicon a manual process, and may lead to incompatible analysis results on the same eojeol in different systems.

In this paper, we try to automatically discover the lexicon from the raw training texts. The variable length character sequences are used as the tokens of the lexicon in order to capture the inherent regularities in the language. The automatic acquisition of linguistic units has been studied in speech recognition and cognitive science community[1,2]. We take the approach that uses a probabilistic framework for the discovery, focusing on building high-order language models for open-vocabulary domains.

3.2. Modeling

Basically, a language model is composed of a lexicon and a set of rules to generate valid sentences using the tokens in the lexicon. A decomposition of a sentence in terms of a lexicon of a particular language model and its associated rules can be regarded as its *interpretation* under the model. If the lexicon happens to be composed of morphemes and the associated rules coincide with the morphology of the language, the interpretation becomes the morphological analysis result.

In our language model for Korean, we enforced following criteria for the lexicon.

- *Each token should have a unique transcription*: Specifically, the tokens should be a sequence of complete characters. This enforces the boundary of tokens strictly lie on the character boundaries, and prevents decomposing a single Hangul character into multiple tokens as happens in morpheme decomposition.
- *All of the individual character is included in the lexicon*: In order to guarantee that at least one interpretation exists for any sentence, all of the individual characters are included in the lexicon. As a consequence, all sentences have at least one interpretation that is a sequence of single character tokens.
- *Each token should have a sufficient frequency of usage in training texts unless it is a single character*: In order to make the lexicon compact, only the character sequences that have been sufficiently observed in the training texts are used in the lexicon.
- *No token can be other token's strict prefix or suffix unless it is a single character*: In order to reduce size of the lexicon and redundant interpretations, a token cannot be a strict prefix or suffix of others unless it is a single character.

In our model, given a sequence of tokens, a sentence is generated deterministically by concatenating the tokens. A token sequence is generated by drawing tokens based on the conditional probability distribution of consecutive tokens. The relationship between tokens *t* and the characters in sentences *c* can be expressed using a dynamic Bayesian network. Figure 1 shows an example of the relationship when first order Markov assumption is applied between tokens.

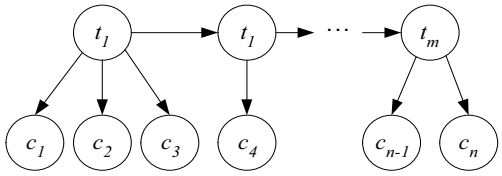

Figure 1. Relationship of tokens and characters

In a sentence, the spaces between eojeols indicate only a subset of token boundaries. There are no explicit token delimiters within an eojeol. As a consequence, a sentence can have multiple interpretations. Therefore, the likelihood of a sentence should be evaluated by summing the expected values of its all possible interpretations. This can be calculated by providing the evidences to the character nodes and treating the token nodes as latent variables. That is, for a sentence *S* and a token sequences T, the likelihood of *S* is calculated as follows:

$$p(\mathbf{S} = c_i \Lambda\ c_n) = \sum_{\mathbf{T}} p(c_1 \Lambda\ c_n \mid \mathbf{T}) p(\mathbf{T})$$

$$= \sum_{\forall \mathbf{T} \to c_1 \Lambda\ c_n} p(\mathbf{T})$$

$$= \sum_{\forall \mathbf{T} \to c_1 \Lambda\ c_n} \prod_{i=1}^{n(\mathbf{T})} p(t_i \mid pa(t_i))$$

where $n(T)$ indicates the number of tokens in T. $T \rightarrow S$ represents T generates S.

The summation over all possible interpretations can be implemented using a dynamic programming algorithm. For example, in a first order Markov model, we use the function $a(i,x)$ that represents the sum of likelihoods of all possible token sequences ending at the character position i in the sentence with the last token x. Then the likelihoods can be evaluated in a recursive form as follows:

$$a(i + len(y), y) = \sum_x a(i, x) + p(t_{cur} = y \mid t_{prev} = x)$$

$$p(c_1 \Lambda\ c_n) = \sum_x a(n, x) + p(t_{cur} =< eos > \mid t_{prev} = x)$$

where <*eos*> indicates the special token for the end of sentence, t_{cur} a token beginning at position i+1, t_{prev} the token preceding t_{cur}, and $len(t)$ represents the length of a token t measured in characters.

Since no token is a prefix or suffix of others, there can be at most 2 tokens for each position within a sentence. Therefore, in a model with m-th order Markov assumption, evaluating the likelihood of a sentence with n characters have the time complexity of $O(2^m n)$. Usually, m does not exceed 3 due to the data sparseness problem, so the time complexity is linearly proportional to the length of the sentence.

If we assume that only one interpretation has much greater likelihood than others, we can approximate the likelihood of sentence with that of the most likely interpretation.

$$p(\mathbf{S} = c_1 \Lambda\ c_n) \approx \max_{\forall \mathbf{T} \rightarrow c_1 \Lambda\ c_n} p(\mathbf{T})$$

$$= \max_{\forall \mathbf{T} \rightarrow c_1 \Lambda\ c_n} \prod_{i=1}^{len(\mathbf{T})} p(t_i \mid pa(t_i))$$

3.3. Training

The lexicon and the probabilities are trained by fixing one and updating the other alternately. The overall flow of the training procedure for the proposed language model is shown in figure 2.

The initial model is built using the positional character n-gram counts acquired from the training texts. Once we have a candidate of the lexicon, the probabilities can be trained by fixing the lexicon and applying EM algorithm that maximizes the log-likelihood of the training text.

$$Q(\boldsymbol{\theta}; \bar{\boldsymbol{\theta}}) = \sum_{\mathbf{S}} \frac{1}{p(\mathbf{S} \mid \boldsymbol{\theta})} \sum_{\mathbf{T}} p(\mathbf{S} \mid \boldsymbol{\theta}) \log p(\mathbf{S} \mid \bar{\boldsymbol{\theta}})$$

$$= \sum_{\mathbf{S}} \frac{1}{p(\mathbf{S} \mid \boldsymbol{\theta})} \sum_{\forall \mathbf{T} \rightarrow \mathbf{S}} p(\mathbf{T} \mid \boldsymbol{\theta}) \log p(\mathbf{T} \mid \bar{\boldsymbol{\theta}})$$

For example, in a first order Markov model, the equation for updated conditional probability for a token y given the previous token x is as follows:

$$\bar{\theta}_{p(y|x)} = \frac{\sum_{\mathbf{S}} \frac{1}{p(\mathbf{S} \mid \boldsymbol{\theta})} \sum_{\forall \mathbf{T} \rightarrow \mathbf{S}} p(\mathbf{T} \mid \boldsymbol{\theta}) count(\mathbf{T}, x, y)}{\sum_{\mathbf{S}} \frac{1}{p(\mathbf{S} \mid \boldsymbol{\theta})} \sum_{\forall \mathbf{T} \rightarrow \mathbf{S}} p(\mathbf{T} \mid \boldsymbol{\theta}) \sum_y count(\mathbf{T}, x, y)}$$

where $count(T,x,y)$ indicates the number of consecutive tokens (x,y) in a token sequence T.

After the probabilities converge, we take expected frequencies in the training texts for each token and remove the ones that were not sufficiently used. Using this updated lexicon, we renormalize the probabilities and train the parameter again until convergence. This procedure is iterated until there are no updates in the lexicon.

Using the final model, the most likely interpretation of the training text is generated. The token counts in this interpretation is used to build a language model that is smoothed with lower-order models using dedicated algorithms such as Katz's back-off[4,7,8].

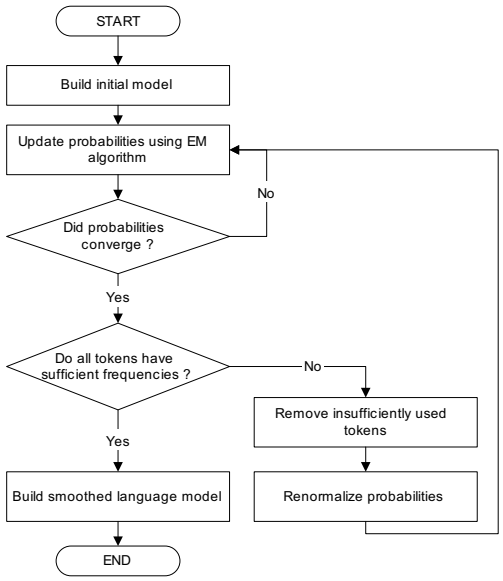

Figure 2. Training procedure

4. Experiment

In order to evaluate the performance of proposed language model, we performed a simulated eojeol recognition experiment. Due to the large size of Hangul character set and the complexity of eojeol's structure, collecting sufficient handwritten eojeol image data is an infeasible task, especially in open-vocabulary domain. Therefore, we synthesized handwritten eojeol images using a raw text corpus and a handwritten Hangul character image database.

The eojeol data were generated using PE92 handwritten Hangul character image database and the '96 KAIST raw text corpus. The PE92 database contains 100

images for each of 2,350 Hangul characters used in KS X 1001 character code set. The texts that were used to generate handwritten eojeol images contain about 850,000 eojeols.

For training the language model, '97 KAIST raw text corpus was used. It is composed of about 15,000,000 eojeols, with about 1,300,000 unique eojeols. The corpus contains various topics such as news, novels, scientific papers and history.

For recognition of handwritten Hangul character images, we used a hierarchical random graph-based recognizer[5]. Currently, it is the only offline handwritten Hangul recognition system that can recognize all Hangul characters. In experiments, we used 5 most likely candidates for each character in the given eojeol and built a fully connected lattice. Therefore, in an eojeol with n characters, up to 5^n eojeol candidates can be generated from the lattice.

For scoring each candidate, we used a likelihood function based on the posterior probabilities of eojeols. That is, for an eojeol W that is consisted of characters $c_1,c_2 \ldots c_n$, its likelihood function is as follows :

$$L(W) = \prod_{i=1}^{n} p(c_i) p(W)^{\lambda}$$

where $p(c_i)$ is the score from the character recognizer and $p(W)$ is the score from the language model. In order to compensate the discrepancies between the language model and the character recognizer, an additional scaling parameter λ was used. The value of λ was determined empirically.

The results of the eojeol recognition experiment are shown in table 3. In 19.5% of cases, at least one of the true label for characters in the eojeol was not included in the candidate lattice due to failures of the character recognizer. Therefore, the upper bound of eojeol recognition rates was 80.5 %.

Table 3. Eojeol recognition rates

Eojeol candidates	Without language model	With language model
1st	55.0% (68.3%)	75.9 % (94.3%)
2nd	66.5% (82.6%)	79.1 % (98.3%)
3rd	70.5% (87.6%)	79.8 % (99.2%)
4th	73.0% (90.7%)	80.1% (99.5%)
5th	74.6% (92.7%)	80.2% (99.6%)

Without using the scores from the language model, only 55% of words were correctly recognized. The word recognition rate was increased to 75.9% by using the proposed language model. Compared to the upper bound, this is equivalent to 94.3% of all possible cases. Moreover, the recognition rate of the first candidate with the language model was higher than that of top 5 candidates without the language model. When considering top 5 candidates with the language model, the recognition rate increased up to 99.6% of the upper bound.

5. Conclusion

The language model is an essential component that can be used for incorporating contextual information in recognition. Due to complex morphology of Korean, both the morphemes and the eojeols have shortcomings to be used as the tokens of the lexicon of the language model, especially in open-vocabulary domains.

In this paper, we proposed a novel language model of Korean for the open-vocabulary word recognition problem. Its lexicon is composed of the variable length character sequences that are acquired from the raw training texts. The probabilistic relationship between the tokens of the lexicon and the characters in sentences are expressed using a dynamic Bayesian network. The tokens in the lexicon are selected by the expected frequency of usage evaluated by the dynamic Bayesian network. After the training, the lexicon shows high relevance to the variable-length regularities that can be found in the Korean language.

In the proposed model, in order to evaluate the likelihood of a sentence, all of its possible interpretations in terms of the tokens in the lexicon are considered. In an extreme case where the sentence is completely novel, it is interpreted as a sequence of single-character tokens. Therefore, the proposed model is capable of evaluating the likelihood of any sentence, without introducing a dedicated token for the out-of-vocabulary entities to the lexicon. As a consequence, the proposed model has 0 out-of-vocabulary rates even in open-vocabulary domains.

In experiments, the eojeol recognition rate increased up to 75.9% by applying the proposed language model. This corresponds to 94% of the possible bound, because the character recognizer failed to include correct label within the candidate lattice in remaining cases. Furthermore, choosing top 5 eojeol candidates showed 99.6% of accuracy. These results support that the proposed language model is suitable for context modeling in the recognition problem.

We believe that the proposed method can be also applied in other languages that have high out-of-vocabulary rate problem in open-vocabulary domain due to complex word structure or lack of explicit word delimiters.

6. Acknowledgement

This work has been supported by Korean ministry of science and technology for National Research Laboratory Program # M1-0104-00-0165.

7. References

[1] Sabine Deligne and Frederic Bimbot, "Language Modeling by Variable Length Sequences: Theoretical

Formulation and Evaluation of Multigrams", *in Proc. of ICASSP*, 1995.

[2] Michael R. Brent, "An Efficient, Probabilistically Sound Algorithm for Segmentation and Word Discovery", *Machine Learning Journal*, vol. 34, 1999, pp.71~106

[3] Stanley F. Chen, "Building Probabilistic Models for Natural Language", Ph.D. Thesis, Havard University, 1996

[4] Stanley F. Chen and Joshua Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling", TR-10-98, Havard University, 1998

[5] H.Y.Kim and J.H.Kim, "Hierarchical random graph representation of handwritten characters and its application to Hangul recognition", *Pattern Recognition*, vol.34,no.2, 2001, pp.187-201

[6] Oh-Wook Kwon, K.Hwang and J.Park, "Korean Large Vocabulary Continuous Speech Recognition using Pseudomorpheme units", *in Proc. of EUROSPEECH*, 1999, pp.483~486

[7] Reinhard Kneser and Hermann Ney, "Improved Backing-off for m-gram Language Modeling", *in Proc. of ICASSP,* 1995

[8] Slava M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer", *IEEE Trans. on Acoustics, Speech and Signal Processing,* vol. 35, no. 3, pp.300~401, 1987

[9] Christopher D. Manning and Hinrich Schütze, *Foundation of Statistical Natural Language Processing*, The MIT Press, Cambridge, Massachusetts, USA, 2000

[10] Roland Hausser, *Foundations of Computational Linguistics*, Springer-Verlag, Berlin, Germany, 1999

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03)
0-7695-1960-1/03 $17.00 © 2003 IEEE