

A Visualization Tool for Korean Morphological Analyzer

Key-Sun Choi, Yongseok Choi, Jonghoon Oh, Choongwon Seo,
Kilyeon Kim, Jaeho Kim, Sunbae Kim, Juho Lee

Korterm, Computer Science Department, KAIST
373-1 Yosong-gu kusong-dong Taejon 305-701 Republic of Korea
{kschoi,angelove,rovellia,cwseo,gykim,jjaeh,yuntan,mywork}@world.kaist.ac.kr

Abstract

This demonstration is to show a visualization tool for morphological analysis. This tool is to show all of the steps in-between the internal flow of morphological analysis and furthermore the user can modify the graph structure and the value of its node in the middle of processing. The steps are briefly introduced as follows (see figure 1):

The First step is a pre-processing one to utilize the already well-analyzed results for frequent occurrences. A well-formed substring table is maintained for the purpose of fetching up the well-analyzed word phrases. We call this well-formed substring table as “eojeol dictionary” where “eojeol” stands for a word phrase or a spacing unit in Korean. For example, this eojeol dictionary contains the word phrases

when they occur more than 1000 times in the KAIST POS-tagged corpus.

In the second step, each word phrase is segmented into all possible sequences of words (or morphemes). This module consists of five sub-modules: dictionary searching, numeric expression identification, phonetic change recovery, unknown word guessing, and POS connection rule checking.

The third step is a rule-based post-processing to cut off the ridiculous tag sequences when its extended contexts can be considered.

Finally, a Korean tagger is applied with considering the word-phrase structure. A Korean sentence is a sequence of word phrases and one word phrase equals to one spacing unit that consists of a concatenated sequence of words (or morphemes). A hidden Markov model is applied with reflecting the word phrase structure of

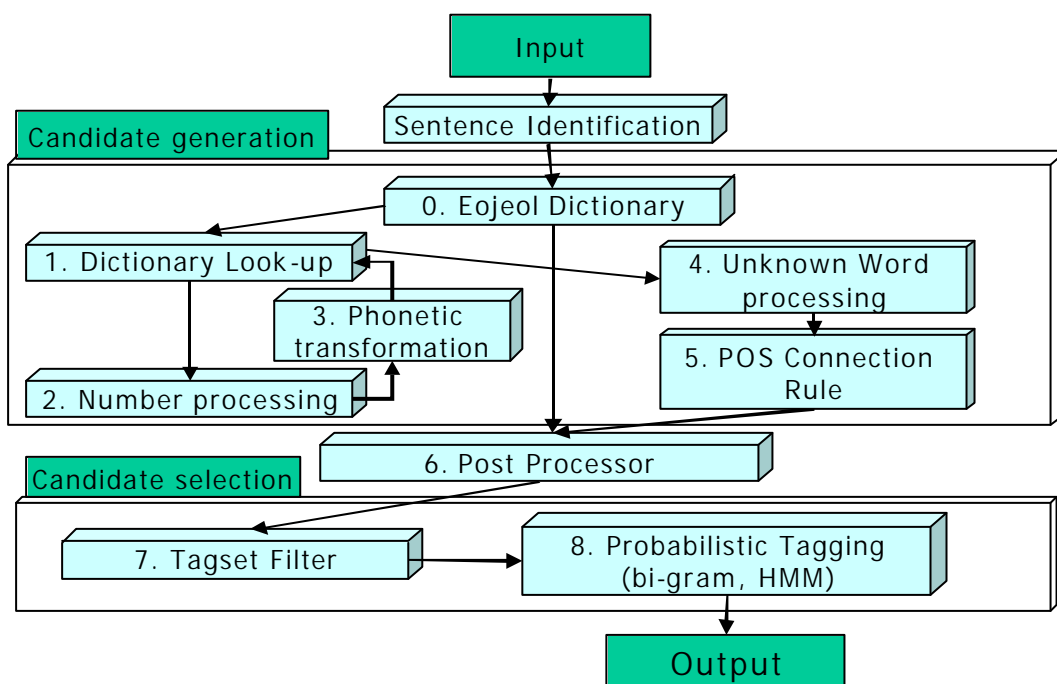


Figure 1 System Configuration of Korean Morphological Analyzer

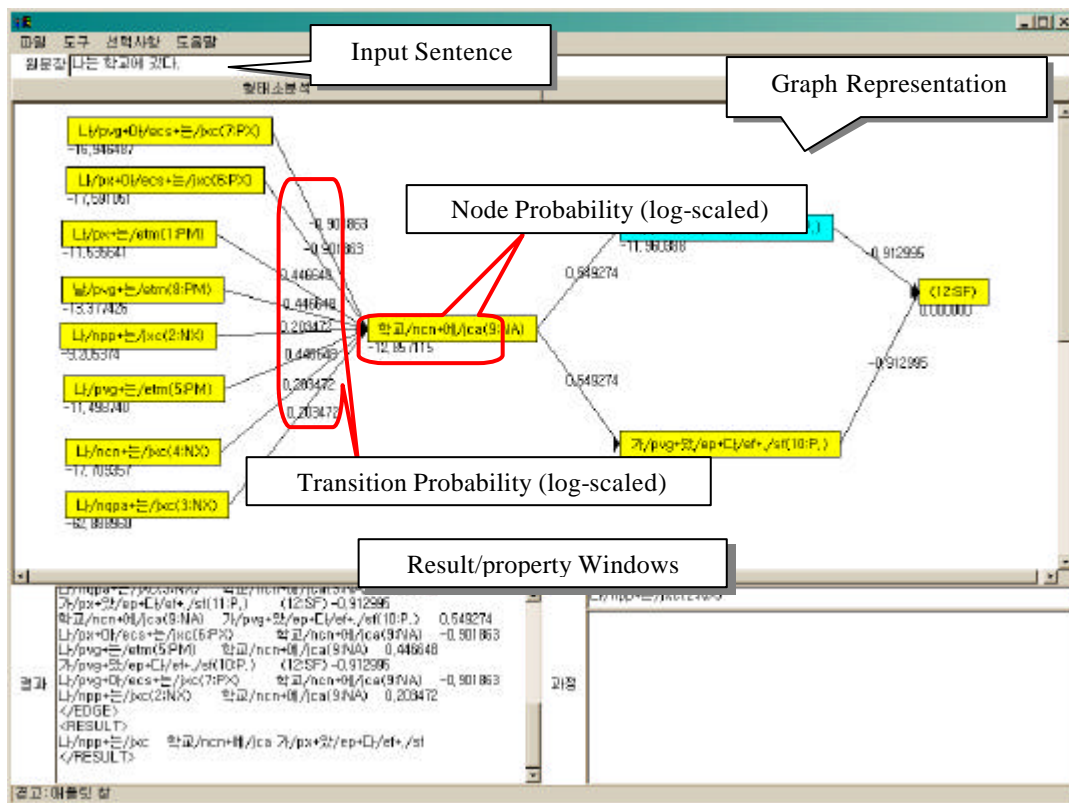


Figure 0 Snapshot of the visualization tool in phase of candidate selection (POS Tagging by HMM model).

Korean.

Our visualization tool shows the numeric value of transition and node probability with graph representation. The probability values can be adjusted and the graph representations can be saved into the files. The tool provide an understandable view of morphological analyzer to researchers and developers.

Acknowledgements

This work was supported by the R&D fund of Ministry of Science and Technology under the project "On Development of Deep-Level Processing and Quality Management Technology for Very Large Korean Information Base", a project of plan STEP2000. And this work was partly supported by KOSEF through the "Multilingual Information Retrieval" project at the AITrc.

References

Nam, K.S. (1986). *Grammar for Standard Korean*. Seoul, Korea: Top Press.[Korean version only]

T. Yamashita, M. Fujio and Y.Matsumoto (1999) *Language Independent Tools for Natural Language Processing* Proceedings of the Eighteenth International Conference on Computer Processing of Oriental Languages, pp.237-240.

Park, J., J. Kang, W. Hur and K.-S. Choi (1998) *Machine Aided Error-Correction Environment for Korean Morphological Analysis and Part-of-Speech Tagging*, Proceedings of the 17th International Conference on Computational Linguistics (COLING'98), pp. 1015-1019, Montreal, Canada.