Korean Language Engineering: Current Status of the Information Platform *

Kim, Seongyong and **Choi, Key-Sun**
Department of Computer Science
Korea Advanced Institute of Science and Technology
Taejon, Korea
sykim@csking.kaist.ac.kr, kschoi@world.kaist.ac.kr

Abstract

Language engineering implements functions of a language and information via computers. The need for language engineering platforms has been generally recognized and several researches are being undertaken around the world. Our goal is to establish Korean information platform of linguistic resources and tools for Korean language and information communities. The platform will support researchers and engineers with well-developed and standardized resources and application tools thereby avoiding duplicate activities from scratch and amplifying overall effort on the domain. This paper reports the components and the current status of the project, and the importance of the effort.

1 Korean Language Engineering

1.1 Language Engineering

Language engineering is such an activity that implements various functions related to a language and builds up an information base. It realizes linguistic activities of everyday life and linguistic competence of human beings with the aids of computer science, thereby supporting people's intellectual linguistic productions. The language engineering not only collects and disseminates the information and knowledge of a language among the linguistic society but also serves as a foundation on which linguistic culture and technologies can be based (Oh et al., 1994).

1.2 Korean Language Engineering

Korean language engineering is one for Korean language. It came into birth in early 1980's with the emergence of personal computers (PCs). In the beginning, they focused on Korean alphabets and some scrappy parts of character processing, lacking the global view of the engineering approaches. Technical approaches to Korean began with the formation of the special interest group on Korean information processing under the Korea Information Science Society. And in 1994 *Center for Korean Language Engineering (KLE)* was founded to serve as a central organization for Korean language engineering, which aims to plan and program related projects and works in a consistent, systematic way with long-term goals. It also incorporates academic and research institutes and industries into common goals: the efficient and harmonious drive toward research and development, and establishment of long-range policies and strategies for Korean language engineering.

2 Areas of Korean Language Engineering Researches

According to the level of technologies, KLE partitioned its projects into three classes.

Fundamental technology deals with radical and theoretical researches, collection and manipulation of data, and standardization. In linguistic viewpoint, these include language formalisms, text corpora, and statistical information of a language. On information engineering side, the technology covers information interchange and compression techniques, basic techniques of artificial intelligence such as knowledge representation, searching, and tools for manipulating Korean alphabets. From the cognitive engineering point of view, the research focuses on the structure of Korean alphabets, fonts, command structures, and interdisciplinary works of cognitive science. Another division handles standardization issues for code schemes and vocabularies, keyboard layout, standard text formats, and internationalization.

The second class is called basic technology, which is related to the basic software libraries for Korean language processing. Included in this class are natural language analysis, pattern recognition, multimedia data base, and data conversion tools.

The third class is applications technology. It

*This work is funded by *Ministry of Science and Technology* and *Ministry of Education and Athletics*, as a part of a contract by *Center for Korean Language Engineering*.

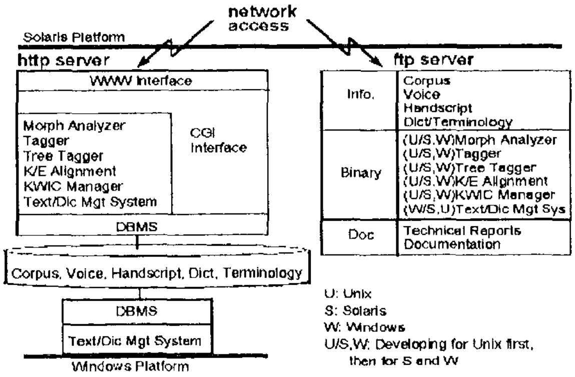

Figure 1: The Conceptual Diagram of the Information Platform

consists of systems for text interchange and compression, hypertext, multimedia, word processing and others. For knowledge processing, it will cover document paraphrasing, indexing and retrieval, computer-based instruction/education, etc.

3 Information Platform

For Korean language engineering, it is necessary to develop systematically all the projects of each area and integrate them into a uniform frame, called an information platform (IP).[1] KLE programs each project according to its priority and state-of-the-art technology. Consequently, IP reflects the status of ongoing projects and is an *as-is* framework on which further researches and development works can be performed.

Figure 1 shows the conceptual diagram of IP. This platform doesn't integrate all the project outcomes but some of the fundamental resources and basic tools, since it reflects the current configuration that is not concrete but open to changes. The whole integration of the project outcomes will be available at the end of the first phase in 1997.

This platform is different from ALEP (Advanced Language Engineering Platform) (Simpkins, 1994) in that ALEP is an environment that can be provided to users as a form of a (customizable) package whereas our platform is a server-client model in pursuit of a web-based service for resources and tools.

Worldwide web is composed of hyperdocuments and hyperlinks to handle multimedia data as well as to provide easy and timely access to electronic information. It uses hypertext markup language (HTML) based on standardized generalized markup language (SGML). Therefore, it guarantees the standardization and straightforward design characteristics, which lead to the ease of system design and flexibility of the system configurrations (Berners-Lee & Connolly, 1993). Its other characteristic lies in the common gateway interface (CGI) which makes it possible to interface with various shell scripts and program codes without difficulties. Yet another point is that the server-client model makes the platform transparent to the users.

IP consists of three parts. First, text corpora, voice and handwritten scripts DBs, dictionaries and a set of terminological DBs constitute the information base. The information base may directly be distributed through ftp server or indirectly accessed by the language tools on the higher layer of the http server configuration.

Secondly, language tools are running on the http server with the aids of CGI as well as being ftp-ed to users as executable codes. Since we aim to provide software versions on Unix, Solaris, and PC Windows altogether, initial hardware requirements for each tool may be different.[2]

Finally, documentation preparation will also be accompanied with the project's progress.

4 Information Base

4.1 Text Corpus

Text corpora are essential to statistical modeling, in developing formal theories of the grammars, investigating prosodic phenomena in speech, and evaluating or comparing the adequacy of parsing models (Marcus et al., 1993). There are four sorts of corpora from contemporary Korean texts.

- Raw corpus
  Two factors are the genre of each source text that is related to the objective(s) in using the corpus, and the category of the text that represents the internal structure of the text. Major sources of the corpus include books, magazines, and newspapers; up to date three million word phrases are gathered.
- Part-of-speech (POS) tagged corpus
  POS tagset for Korean originated from (Kim & Seo, 1994). In version 1 platform we yielded 2.5 million automatically tagged word phrases and 1.5 million post-edited word phrases.
- Tree-tagged corpus
  This can be produced by applying syntactic tagset to the POS tagged corpus. The syntactic tagset is being studied using 100,000 sentences out of POS tagged corpus, and the resultant tree-tagged corpus using a tree tagger will appear at the end of this year.

[1] "http://world.kaist.ac.kr/KLE/KIBS/" is SunOS, version 1 platform and web pages are only in Korean. The 2nd version will be released on Solaris at the address "http://kibs.kaist.ac.kr/KLE/KIBS/."

[2] For example, the text and dictionary management system is currently being built upon PC Windows so that Unix and Solaris executables are not yet available.

- Categorized corpus
  Korean verbs and adjectives are classified into over seventy categories, and a set of sentence styles are investigated for 940 basic verbs of those categories. About thirty five thousand sentences are tangible in version 1 platform.

4.2 Voice Data Base

This resource can be used for speech recognition and synthesis applications. We initially focused on word-level voice data. It includes phonetically balanced words, phonemic sequences pronounced by four different speakers, and narration of sample stories. It also stores the sounds of single syllables, diphones, numerics, high-frequency words, gazetteers, functional words, and consecutive word sequences. The data are stored in server disks and CD-ROMs as a wave form. This effort will be extended to sentence-level collections such as phonetically balanced sentences, speech dialogues, and scenarios.

4.3 Handwritten Scripts Data Base

Since character recognition systems are under the control of applications engineers, the objective of this work is to provide well-formed data and evaluation criteria for those recognition systems. We stepwise our data collection into three phases: to scan, with 300 dpi resolution, one thousand sets of 520 high-frequency syllables in the first year, then of 990 syllables and 2,350 syllables in the following years.[3] At each phase, we develop both the square-hand characters and free-style characters.

4.4 Dictionaries and Terminological Data Base

- Multilingual technical dictionary
  The objective is to set up mappings between technical terms of Korean and other language(s) in both directions. The first work is done for computer science domain, and it has 35,000 entries each for Korean and English. It will be extended to cover Chinese, Japanese, and German as well as more domains including electrical/electronic engineering, medical science, law, etc.
- Monolingual terminology data bank
  Users need definitions and explanations of technical terms during their work on specific domains. This work provides users such terminological details. We assorted 15,000 entries each for culture/art and Korean classical literature.
- Ontology-based lexicon
  Currently available dictionaries are semantically oriented. They don't provide pools of target language expressions but offer basic meanings for entries together with some syntactic and morphological information. Ontology-based lexicon is lexically oriented in that it guides the user to find a pragmatically or contextually equivalent expression corresponding to the source language expression. The work is on the phase of feasibility study with intensive focus on collecting Korean-English bilingual information sources and developing tools for lexicon construction.
- Lexicon for morphological analysis
  The lexicon for Korean morphological analysis is currently being built to have 30,000 entries with off-line management tools, and will grow to 100,000 entries with on-line tools after two more years.[4]

5 Language Engineering Tools

Basically, the tools that we present here are for text corpus and dictionaries, except for voice and character recognizers. The latter two programs are currently under the development and will be integrated later.

5.1 Morphological Analyzer

Morphological analysis is an important but difficult part of the analysis since Korean is an agglutinative language with sophisticated morpheme segmentation rules and morphotactic rules. The morphological analyzer is based on the Korean chart parsing (Lee, 1993). Its' current precision is over 92 percent for the grammatical input sentences. It aims to achieve 98 percent accuracy in two more years. It will be extended to cover special symbols, alien strings, elliptical or abbreviated words, and spell errors to earn higher accuracy.

5.2 Tagger

Because the output of morphological analysis is rather complex due to the characteristics of Korean, the use of a tagger to reduce ambiguities seems important for further processing. (Shin et al., 1995) adopts the hidden Markov model and takes into account the characteristics of Korean word phrase structures for more accurate tagging: a word phrase contains one or more morphemes, syntactic information (grammatical relations by bound morphemes), and semantic information (case roles by postpositions). The experiments revealed 98 % accuracy for the test set of 5,500 word phrases out of 55,000 training data, and 94.7 % for 5,500 untrained test data.

[3] It is possible to compose up to 11,172 syllables out of each Korean alphabet, but Korean Standard Code KSC-5601 prescribes 2,350 complete codes for Korean syllables.

[4] We can conceive much more types of dictionaries: for example, lexicons for syntactic and semantic analyses, and dictionaries that are to be created or extracted from existing ones upon users' or developers' needs. These will be included after the first phase of the project, following future direction of the project.

Another approach is based on the Markov random field (MRF) theory (Jung, 1996), whose Korean version will be added to IP this year.

5.3 Tree Tagger

(Kim, 1995) is a prototype using dependency grammar and adopting statistical methods for ranking the parse trees to get *k-best* parsing results. Its current accuracy is about 80 % for the trained data. While this is a working prototype, we need a tree tagger with better performance so that another tree tagger using partial parsing method (Abney, 1991) is on breadboard.

5.4 Korean/English Alignment System

An alignment system gathers correspondences between surface representations of both languages. (Shin, 1996) experimented expectation-maximization algorithm with 68.7 % accuracy at phrase level, and this will be incorporated into version 2 platform.

5.5 KWIC Manager

Keyword-in-context (KWIC) manager deals with word usage of text corpus. Its functions include indexing and searching word phrases, morphemes or unigrams, applying logic operations (*AND, OR, NOT*) to them, and sorting the results.

5.6 Text/Dictionary Management System

TDMS' goals are twofold: to provide customizable information extraction/indexing/search tools and managerial functions for text data base; and to provide an environment for dictionary development and management as well as converting or merging existing dictionaries to the intended one according to user's specification.

Because of the big size of each text to be stored and lots of keywords to be indexed and searched for each text, it requires special storing and managing mechanisms. This is also the case for the dictionary management. For the extensibility and adaptability, we have devised standard dictionary markup language based on SGML. Templates (dictionary features, text descriptors, and relations among those), specifications for text/dictionary editor and format translator have been also being designed and low-level design is being undertaken. This work is being coded on PC Windows and will output the first draft version this year.

6 Conclusion

To this point we described the motivation and current status of the Korean IP, and took a brief look at resources and tools. We started the project in 1994 to yield version 1 platform in 1995 and are working on version 2 platform. The project will continue till the years of twenty first century. Although the current status is just an opening spot, the long-term goal is to build fully automatic servers for Korean language information. Since IP plays a key role in the effort, we hope that our endeavors would be well geared to the needs of nation-wide language engineering.

References

Abney, Steven. 1991. Parsing by Chunks. Berwick, R., Abney, S., and Tenny, C. (eds.), *Principle-Based Parsing*. Kluwer Academic Publishers.

Berners-Lee, Tim, and Connolly, Daniel. 1993. *Hypertext Markup Language: A Representation of Textual Information and Metainformation for Retrieval and Interchange*. CERN, USA.

Jung, Sung-Young. 1996. it Markov Random Field based English Part-of-Tagging System. M. S. Thesis, Korea Advanced Institute of Science and Technology. (to appear in COLING96.)

Kim, Hiongun. 1995. *Korean Syntactic Analysis with Probabilistic Dependency Grammar*. M. S. Thesis, Korea Advanced Institute of Science and Technology.

Kim, Jae-Hoon, and Seo, Jungyun. 1994. *A Korean Part-of-Speech Tag Set for Natural Language Processing*. Technical report no. CAIR-TR-94-55. KAIST: Center for Artificial Intelligence Research.

Lee, Eun-Chul. 1993. *An Improved Method on Korean Morphological Analysis Based on CYK Algorithm*. M. S. Thesis, Pohang Institute of Science and Technology.

Marcus, Mitchell P., Santorini, Beatrice, and Marcinkiewicz, Mary A. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2): 313-330.

Oh, Gil-Rok, Choi, Key-Sun, and Park, Se-Young. 1994. *Hangul Engineering*. Seoul, Korea: Daeyoungsa.

Shin, Jung-Ho, Han, Young-Seok, Park, Young-Chan, and Choi, Key-Sun. 1995. An HMM Part-of-Speech Tagger for Korean Based on Word-phrase. *Recent Advances in Natural Language Processing*, Bulgaria.

Shin, Jung-Ho. 1996. *Aligning a Parallel Korean-English Corpus at Word and Phrase Level*. M. S. Thesis, Korea Advance Institute of Science and Technology. (to appear in COLING96.)

Simpkins, N. K. 1994. *ALEP (Advanced Language Engineering Platform): An Open Architecture for Language Engineering*. CEC and Cray Systems, Luxembourg.