Introduction of KIBS (Korean Information Base System) Project[1]

Young-Soog Chae, Key-Sun Choi

KORTERM (Korea Terminology Research Center for Language and Knowledge Engineering)
Korea Advanced Institute of Science and Technology
373-1, Kusung-dong, Yusong-gu,Taejon, 305-701, KOREA
{yschae, kschoi}@korterm.kaist.ac.kr

Abstract

This project has been carried out on the basis of resources and tools for Korean NLP. The main research is the construction of raw corpus of 64 million tokens and Part-of-Speech tagged corpus of about 11 million tokens. And we develop some analytic tools to construct and some supporting tools to navigate them. This paper represents the present state of the work carried out by the KIBS project. We introduce a KAIST tag set of POS and syntax for standard corpus and annotation principles. And we explain several error types represented in tagged corpus.

1. Introduction

In 1994 the KAIST decided to promote current research activities supported by Ministry of Culture and Tourism, Ministry of Science and Technology, and Korea Institute of Science & Technology Evaluation. One of the results of this initiative is the construction of a Korean corpus of written texts, a name of "KAIST (Korea Advanced Institute of Science & Technology) corpus". The purpose of this project is to provide machine-readable corpora and analytic tools that are to serve as the empirical basis for a number of language specific analyses of Korean texts.

The KIBS (Korean Information Base System) project was initiated with the aim of (1) the development of an integrated environment and support management system, (2) the standardization and the specification for Korean Information Base, (3) the construction of Korean Information Base, and (4) the development and management system of electronic dictionary for sentence analysis and generation. In the following sections, we explain the characteristics of corpus and tag set as a Korean information base.

2. Characteristics of KAIST Corpus

Digitalized corpora have proved to be excellent resources for a wide range of research tasks. In the first place, they have provided a more realistic foundation for the study of language than earlier studies. Secondly, they have become a particularly fruitful basis for comparing different varieties of language and for exploring the quantitative and probabilistic aspects of language [Svartvik J. 1991].

For the purpose of language engineering as well as linguistic study, we determined to collect materials of various fields such as literature, newspaper, academic thesis, and so on. They say the printed materials have an effect on the linguistic contents of about 100 million word units of the corpus. A word unit is marked by a space in Korean and is usually a combination of world and functional morphemes. KORTERM/KAIST constructs 64 million of raw corpus, 11.7 million tokens of POS-tagged corpus, and 20 thousand sentences of tree-tagged corpus.

- Raw Corpus

The basic factor of category classification is related to the purpose to use the corpus and the domain of text that represented in the internal content of text. Major sources of the corpus include books, magazines, and newspapers and 64 million word phrases are gathered. The corpus is stored in plain text files with head information (title, author, edited year, and so on) and Korean Standard Codes (KSC 5601) are used. The files are coded using SGML mark-up language that includes the information of document structure including KDC (Korean Dewey decimal Code) classification code of domain.

Category		Percent (%)
Literary Style	Novel	36.8
	Scientific Prose	12.2
	Educational Prose	46.1
	Artistic Prose	3.4
	Memoirs	0.4
Conversational Style		1.1
Total		100.0

[Table 1] The Category State of KAIST Raw Corpus

- Part-of-Speech Tagged Corpus

We construct POS tagged corpus with the standard POS tag set for Korean morphological analysis that is represented morphological and syntactic information in a sentence. We annotate with a morphological analyzer included tagger automatically and then linguistic experts correct some errors turned out contrary to the determined annotation principles.

Genres	Number of Units (thousand)	Percent (%)
Novel	739	62.8
Scientific Prose	92	7.9
Educational Prose	228	19.4
Artistic Prose	14	1.2
Separate Volume	103	8.8
Total	1,176	100.0

[Table 2] The Genre State of KAIST POS-Tagged Corpus

[1] This paper has been supported by AITrc and The Korean Ministry of Science and Technology.

- Tree-Tagged Corpus

We produce Tree-tagged corpus with applying syntactic tag set. We are adapting two grammar rules. The basic syntactic tag set is unique, but the bracketing is different from parsing rules. Therefore we construct 100 thousands sentences.

3. Tag set

We make use of the standardized tag set to build POS-tagged corpus and tree-tagged corpus that have a wide usage as an international common Korean corpus. We establish our standard tag set for analyzing Korean language bases with samples and annotations. We use 54 morphological standard tag set to annotate for part-of-speech and 8 syntactic tag set to annotate for syntactic tree structure.

3.1. Part-of-Speech Tag set

The part-of-speech tag set has the morphological and syntactic information. Postposition and ending have the important function to analysis the surface structure in Korean sentence. The constructed corpus uses in various parts related to language research. They are used (1) to make use of learning data to develop a tagger (2) to analysis for variety of linguistic state. Therefore the unit of tagging is the minimal one to have a meaning as a morpheme. We have great principles to determine them. First, we make the utmost use of rules of composition in our grammar textbook. Second, the view of morphological analysis is more important than the syntactic analysis. Third, we classify the tag set into 3 levels with hierarchical classes to apply it to various application field and we tag with 3rd level.

3.2. Syntactic Tag set

We explain a Korean syntactic tag set for building a large tagged corpus of Korean syntactic trees. The syntactic tree annotated corpora are beginning to serve as an important knowledge source for solving problems in natural language processing as well as in theoretical linguistics. From these corpora, we can extract various linguistic data for Korean such as subcategorisation, the types of sentence, and the patterns of phrase. We describe the principle needed in building the syntactic tag set and show the various examples of usage of the tag set. Establishment of the syntactic tag set is based on the observation of real sentences. The theoretic basis of the syntactic parsing is the standard theory of transformational generative grammar, and the parsing is based on surface structures rather than deep structures.

Name of Tagset	Meanings
S	Sentence
NP	Noun Phrase
VP	Verbal Phrase
AP	Adnominal Phrase
PP	Postpositional Phrase
ADVP	Adverbial Phrase
IP	Independent Phrase
AUXP	Auxiliary Verbal Phrase

[Table3] The Syntactic Tag set

4. Principles of Annotation

4.1. Principles of POS-tagged Annotation

To annotate corpus needs the interaction of an expert in annotation and the tagging support programs. The linguistic annotation or analysis of corpus demonstrates a need for a tightly coupled relationship between analyst processing, computer processing software, and corpus data. There is a truly interactive relation. They have three steps and iteratively carry out those processes. They improve the software and corpus data litter by litter. First process is machine processes of new corpus data. Second process is human expert evaluates and corrects output, and Third process programmer enhances the tagging support programs.

We annotate each word as the unit of spacing words. The tagging unit is a word consisted of a morpheme or one more morphemes. The objects of consideration have to include linguistic state in a real world. So we must analyze a word with our standard tag set that cannot explain with linguistic theory and the rules of current Korean spelling. We write horizontally the tagging results like "Word $Morpheme_1/Tag_1$ + $Morpheme_2/Tag_2$+…" per line and insert a blank line per sentence. Our principles of POS-tagged annotation are as follows.

- The word that has alternative use of the different forms of a functional morpheme

We acknowledge the different forms with same meaning and function in functional morphemes and annotate different tag set each other. Pre-final ending (ep) '-Ass', '-Euss' and '-Ss' is a representative example.

(e.g) Muk-Ess-Ta (ate) :
Muk/pvg+Ess/ep+Ta/ef
(eat) (past) (final ending)
Kass-Ta (went) :
Ka/pvg+Ss/ep+Ta/ef
(go) (past) (final ending)

- The morpheme used abbreviation or contracted forms

This morpheme doesn't re-generate and take a tag with a plain form. For example, auxiliary particle '-i' is usually omitted if an initial sound of next morpheme is a vowel.

(e.g) I-Kus-Un Hak-Kyo-Ta. ('This is a school.') :
Hak-Kyo/ncn+Ta/ef
(school) (final ending)

- The word used compound and indecomposable forms

If we can't analyze the structure of a word to several constituent units any more in the phonology level and the word makes a compound form, we tag only a unit. For example, "Mwe" is a composed form of "Mwu-Es".

(e.g) Mwe-Ka Cho-Ul-Kka? ('Which do you prefer?') :
Mwe/npd+Ka/jcs
(what) (final ending)

- The restoration of irregular verb and adjective

When the irregular verb and adjective are derived or conjugated, the stem of them changes because of the initial sound of ending or suffix, such as vowel and consonant. In this case we can analyze the original form and give a tag to the restored form.

(e.g) Kkun-Kwa Kkun-Ul I-Ess-Ta.
('They link strings together'):
Is/pvg+Ess/ep+Ta/ef
(link) (past) (final ending)

- The compound noun

The types of compound noun are a fusion compound noun, a parallel compound noun, and a subordinate compound noun. If the fusion compound noun and parallel compound noun are analyzed into a unit, they are lost themselves meaning. Accordingly they are annotated a unit. However the subordinate compound noun consists of two or more words with each meaning in itself. In this case it has two or more tags in a word.

(e.g) a. Fusion Compound Noun : Nun_Mul/ncn (tear)
b. Parallel Compound Noun :
Kak-Kyo-Saeng-Hwal (school life) :
Hak-Kyo/ncn+Saeng-Hwal/ncn

- The word has two more function

When a word has two more function, it has different tags according to function of a sentence.

(e.g) Noun and Adverb
O-Nul-I To-Yo-Il-I-Ta. (Today is Saturday)
: O-Nul/ncn+I/jcs
Na-Nun O-Nul Hak-Kyo-Ei Kass-Ta.
(I went to school today.) :
O-Nul/mag

(e.g) Pronoun and Adnoun
Ku-Nun Hak-Kyo Sen-Saing-Nim-I-Ta.
(He is a teacher.) :
Ku/npp
Ku Chaik Pyo-Ji-Nun Bulk-ta.
(The book has a red cover.) :
Ku/mmd

4.2 Principles of Tree-tagged Annotation

The basic principle to construct of tree tagged corpus is based on the surface structure. It is very efficient that we can transport a morphologic construction to syntactic analysis entire. The parser analysis a sentence to a flat tree structure using syntactic tag set.

- Noun Phrase

Noun Phrase consists of noun and case particle (a postpositional word functioning as an auxiliary to a main word).

NP Rule:
NP -> {N|NP|VP|PP} +X/{etn, ef} +
(I/Ka/Ul/Lul/Wa/Kwa/Un/Nun/To/Man)
(e.g) a. (NP Cham-Sae /ncn+Ka/jcs)
b. (NP
(AP Sae-Lop/paa+ n/etm)
(AP Jak-Pum/ncn+Ui/jcm)
(NP Ka-Chi/ncn+Lul/jco))

- VP(Verbal Phrase)

Verbal Phrase includes verb and adverb. VP is consisted sentence with termination sentence sign and combines noun phrase, postpositional phrase, adnominal phrase, adverbial phrase, independent phrase, and so on.

VP Rule:
VP -> XP V*
VP -> {NP|ADVP|PP|AP|IP} V
V -> …+ X/{ecx, ecc, ecs, ef}
(e.g) (VP (NP Tae-Sang/ncn + I/jcs)
(V (V Toi/pvg+Ko/ecx)
(AUXP Iss/px+Ta/ef))

- AP(Adnominal Phrase)

Adnominal Phrase can't use independently and is positively necessary noun phrase, postpositional phrase, or verbal phrase including noun.

AP Rule: AP -> {NP|PP|VP|ADVP}
[X/{pvg|paa|px|xsv}]+ X/etm*
(e.g) (NP (AP Wang/ncn+Ui/jcm)
(NP Pyeng-Hwan/ncn+Ul/jco))

- PP(Postpositional Phrase)

Postpositional phrase consists of several independent components, except noun phrase.

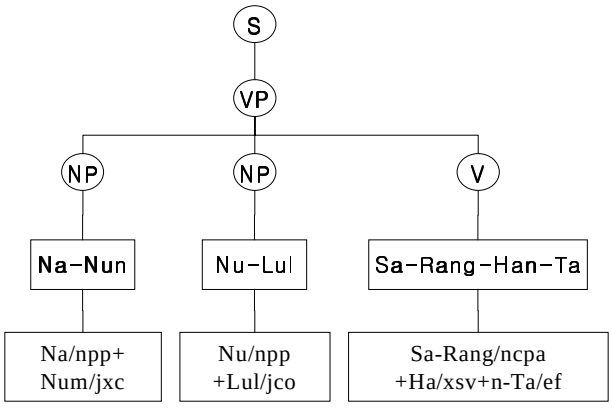
PP Rule: PP -> {NP|VP|PP}
+{Ei|Ei-Kei|Ei-Se|Pu-Te|…}
(e.g) (PP
(AP Yel/ncn+Ui/jcm)
(PP Hyeng-Tae/ncn+Ro/jca))

- Sentence

A sentence consists of phrase categories and the sign of sentence termination (., !, ?, and so on).

S Rule:
S-> XP + { ., !, ?}/sf
(Xp: Vp, NP, ADVP, …)
(e.g) Na-Nun Nu-Lul Sa-Rang-Han-Ta. (I love you.)

(S (VP
(NP Na/npp+Nun/jxc)
(NP Nu/npp+Lul/jco)
(VP Sa-Rang/ncpa+Ha/xsv+n-Ta/ef)
)
)



[Fig1] The Tree structure of a Sentence

5. Tools for Browse

KCP (Korean Concordance Program) and KAIST Raw Corpus Browser are tools for end user who wants to get the linguistic information from raw corpus and POS-tagged corpus. Specially, KCP give some examples including a word to search and probability of word frequency, POS frequency and so on.

[Fig 1] The KCP Execution Screen Capture

6. Conclusion and Future Plans

This project has been carried out only on the basis of tools and resources for Korean. We automatically tagged and manually corrected corpus of about 1 million tokens. This paper represents the present state of the work carried out by KIBS project. We collect the error types of tagging and difficult tagging examples. It requires consideration from various viewpoints.

The annotation of corpora will continue. Semantic analysis and discourse analysis of corpora are likely to be the next stage in this development of corpus annotation. This project has new opportunities for collaboration to the collection and creation of language resources that can be shared.

7. References

Svartvik J. 1991 English Corpus Linguistics, Longman.

EAGLES (1996a) Recommendations on corpus encoding. EAG-TCWG-CES/R-F. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.

EAGLES (1996c) Recommendations for the morphosyntactic annotation of corpora. EAG-TCWG-MAC/R. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.

Kong Joo Lee, Jae Hoon Kim, Key-Sun Choi, Gil Chang Kim, "Korean Syntactic Tagset for Building a Tree Annotated Corpus", Korean Journal of Cognitive Science, Vol. 7, No. 4, pp. 7-24, 1996.

Gi-Chul Yang, Key-Sun Choi, A Way of Constructing a Knowledge Base by Analyzing Korean Text, Korean Journal of Cognitive Science, Vol. 7, No. 4, pp. 203-216, 1996.

The Part-Of-Speech Tag set for Korean

Level I	Level II	Level III					
Symbol (s)		1.	sp	Comma	2.	sf	Sentence closer
		3.	sl	Left quotation and parenthesis mark	4.	sr	Right quotation and parenthesis mark
		5.	sd	Connection mark	6.	se	ellipsis
		7.	su	Unit	8.	sy	other symbols
Foreign word (f)		9.	f	Foreign word			
Nouns (n)	Common Noun (nc)						
	Predicative Noun (ncp)	10.	ncpa	Active common noun	11.	ncps	Stative common noun
	Non-PredicativeNoun (ncn)	12.	ncn	Non-predicative noun			
	Proper Noun(nq)	13.	nq	Proper noun			
	Bound Noun(nb)	14.	nbu	Unit bound noun	15.	nbn	Noun-unit bound noun
	Pronoun(np)	16.	npp	Personal Pronoun	17.	npd	Demonstrative pronoun
	Numeral(nn)	18.	nnc	Numeral	19.	nno	Ordinal Numeral
Predicates(p)	Verb(pv)	20.	pvd	Demonstrative Verb	21.	pvg	General Verb
	Adjective(pa)	22.	pad	Demonstrative Adjective	23.	paa	Aspect Adjective
	Auxiliary Verb (px)	24.	px	Auxiliary Verb			
Modifiers (m)	Adnoun(mm)	25.	mmd	Demonstrative Adnoun	26.	mma	Aspect Adnoun
	Adverb(ma)	27.	mad	Demonstrative Adverb	28.	maj	Conjunctive Adverb
		29.	mag	General Adverb			
Independents (i)	Interjection (ii)	30.	ii	Interjection			
Particles (j)	Case Particle (jc)	31.	jcs	Nominative	32.	jco	Objective
		33.	jcc	Complementary	34.	jcm	Adnominal
		35.	jcv	Vocative	36.	jca	Predicative
		37.	jcj	Conjunctive	38.	jct	Cooperative
		39.	jcr	Quotative			
	Auxiliary Particle (jx)	40.	jxc	General Auxiliary Particle	41.	Jxf	Final Auxiliary Particle
	Auxiliary Particle(jcp)	42.	jcp	Auxiliary Particle			
Endings (e)	Prefinal Ending (ep)	43.	ep	Prefinal Ending			
	Conjunctive Ending (ec)	44.	ecc	Equal	45.	ecs	Subordinative
		46.	ecx	Auxiliary			
	Tansform Ending (et)	47.	etn	Norminal ending	48.	etm	Adnominal ending
	Final Ending (ef)	49.	ef	Final Ending			
Suffixes (x)	Prefix(xp)	50.	xp	Prefix			
	Suffix(xs)	51.	xsn	Noun-derived suffix	52.	xsv	Verb-derived suffix
		53.	xsm	Adjective-derived suffix	54.	xsa	Adverb-derived suffix