Automatic Annotation for Korean – Approach Based on the Contextual Exploration Method

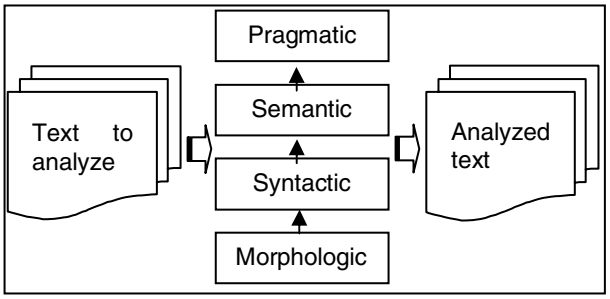
Hyunzoo Chai
LaLLIC, UMR 8139, University of Paris-Sorbonne, Paris, France
hyunzoo.chai@paris4.sorbonne.fr

Abstract

We present an automatic semantic annotation system for Korean based on the Contextual Exploration Method. Creating a morphological analyzer and part-of-speech tagger for the Korean language is difficult as it is a highly agglutinative language. Accordingly, processing Korean in the same order as inflectional languages – morphological analysis, then syntactical and then semantic – has not yielded satisfactory results. Our new method identifies semantic information in Korean text without going through the morphological and syntactical analysis steps. Our initial system properly annotates approximately 88% of standard Korean sentences, and this annotation rate holds across text domains. Previously, the Contextual Exploration Method has been applied successfully to languages as diverse as French and Arabic. Given our success with Korean, we believe that this method can be applied to other agglutinative languages such as Japanese, Turkish and Finnish.

1. Introduction

The ultimate aim of natural language processing is to accurately understand the exact meaning of human language. Most natural language processing methods [Fuchs, 1993][Saint-Dizier, 1995] break the process down into several successive stages using several levels of representation: morphologic, syntactic, semantic then pragmatic analysis. The morphological stage is concerned with the structure of words and the rules of word formation. The syntactic stage deals with the constituent structure of phrases and sentences, while the semantic stage seeks to extract content from text [Desclés et al., 1997].

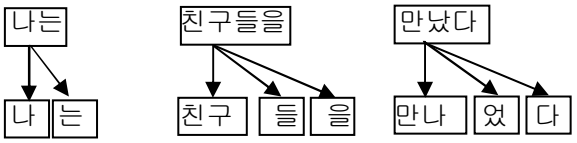


<Figure 1. General natural language processing stages >

Figure 1 shows these stages to decode a natural language sentence into a representation that a computer can understand, and then, ultimately perform a suitable action. Typically, each process in the diagram is progressively harder than the previous one, and less and less about the best method to tackle the problem is known. In this context, early stages such as morphological analysis and part-of-speech tagging have been the primary focus of much research [Ferrari, 2003]. As a result, robust and effective morphological and syntactical processing methods have been developed for many inflectional languages (English, French, Spanish, etc.), and current research concentrates more on the next stage of extracting semantic information. However, compared to that of inflectional languages, the technology in agglutinative language processing such as Korean one is still in difficulty [Sébillot, 2004].

2. Linguistic Characteristics of Korean

Korean is a highly agglutinative language with a very complex affix system, including: postpositions, suffixes and prefixes on nouns; and tense morphemes and conjugational endings on verbs and adjectives. For example, "나는 친구들을 만났다(I met my friends)" consists of 8 morphemes such as:



나(pronoun) 친구(noun) 만나(verb)
는(postposition) 들(suffix) 었(pre-final ending)
을(postposition) 다(ending)

Moreover, flexible sentence patterns make it difficult to determine the part-of-speech in Korean. For example, the above sentence "나는 친구들을 만났다(I met my friends)" could be written in two ways such as;

나는(subject)친구들을(objective)만났다(verb).
친구들을(objective)나는(subject)만났다(verb).

This has led to frequent lexicon lookups, and extensive use of exception rules and tables in typical Korean Natural Language Processing systems [Lee et al., 2003]. Almost all Korean private sector and academic research has been concentrated on finding ways to create a satisfactory morphological analyzer and part-of-speech tagger, and Korean Natural Language Processing research has not yet had a basis for semantic exploration.

3. Contextual Exploration Method

The Contextual Exploration Method provides a method of identifying semantic information in text, without the need for morphological and syntactical analysis stages. The method has been already applied to French and Arabic languages [Motasem et al., 2006]. While the method of identifying syntactical information is limited to a few words around an analyzed sentence, the Contextual Exploration Method takes account of all signs occurring in a given text. For example, the verb "*déraillait*" of the French sentence "*Cinq minutes plus tard, le train déraillait* (Five minutes later, the train ran off the track)" could have different semantic information according to other linguistic signs in the sentence as follows:

(a) ***Malgré*** *toutes les précautions, cinq minutes plus tard, le train déraillait* (In spite of all precautions, five minutes later, the train ran off the track)
(b) ***Sans*** *toutes les précautions, cinq minutes plus tard, le train déraillait* (Without all precautions, five minutes plus tard, the train ran off the track).

In sentence (a), the verb "ran off" really happened while in sentence (b), it has unreal value. Depending on the linguistic clues *Malgré* or *Sans*, we infer quite the opposite information even with the same word "*déraillait*". Thus, the Contextual Exploration Method bases itself on other linguistic clues which must be present in the same context and compensates for difficult ambiguity phenomena in syntactical processing.

The Exploration Contextual Method is based on:
(a) linguistic sign (Indicator) found in a given text,
(b) linguistic clues (Indices) solving ambiguity affecting the indicator in their context,
(c) a set of contextual exploration rules linked with Indicator and Indices.

This method is presented in the following form:

LET Ui BE a linguistic indicator for the A annotation
IF Uk occurs in a sentence S
AND IF linguistic clues Vk occurs in Cik contexts
THEN perform A annotation

In such rules, Ui and Vk are linguistic signs and Cik constitute the contexts which depend on both linguistic Indicators and annotations [Berri et al. 1995]. Using this method, we created an automatic semantic annotation system in order to detect localization relations in a given Korean text.

4. Localization Relation

Our study of the localization relation is situated within the framework of the theory of Cognitive and Applicative Grammar [Desclés, 1990]. *This theory* consists of three levels of representation: a syntactic level, a predicative level and a semantic and cognitive level. The semantic and cognitive level is constructed with primitives defined by a basic semantic and cognitive relaters associated to verbal sense: static, kinematic and dynamic primitives.

Dynamic: FAIRE(make), CONTR(control)…
Kinematic:MOUVT(movement),CHANG(change)…
Static:REP (localization, assignment, differentiation…)

Among them, localization relation is situated within the static primitives that describe essentially the relation of location between a mark and a reference mark. In this model, the localization relation is presented as "X is localized with respect to Y". For example, in the sentence "*Paris est en France (Paris is in France)*", Paris is localized with respect to France. Localization relation is directed from the localized one towards the locator: *X→localisation→Y.*

In order to define the cognitive primitives of localization relation we use concepts inherited from topology. Topological notion is already used in linguistics [Talmy, 1988]. Our work is also based on some quasi-topological operators with specific properties defined on abstract loci. A locus is characterized by topological operators like IN, EX, FR,

FE (respectively take interior, exterior, boundary and closure (interior + frontier)). We add here, la relation VG (*voisinage*, closed-by). Relation VG [Le Priol, 2004] is not described by the traditional topological operators but its behavior approximates them. Similarly, we classify orientation operators such as GAUCHE, DROIT, DESSUS, DESSOUS, DEVANT, DERRIERE (respectively left, right, above, below, front, back).

5. System

The extraction of localization relation is part of the project of automatic annotation, EXCOM (EXploration COntextuelle Multilingue) at the LaLICC laboratory of Paris-Sorbonne University. EXCOM is a XML-based system for an automatic annotation of texts according to semantic categories [Djioua et al., 2006]. The system is based on the theory of the Contextual Exploration Method. The Contextual Exploration Method utilizes principal indices (indicator) and complementary indices together to extract semantic value. The indicator (generally a verb) signals the possible existence of semantic value belonging to a specific semantic category, and complementary indices are used to correctly define this value. Specific semantic information is defined using both linguistic signs (indicator and complementary indices) and contextual exploration rules linke with indicator and complementary indices.

5.1. Linguistic resource

The first step of the extraction of localization relation is to establish lists of indicator and complementary indices. We classified verbs expressing localization for the indicator and chose postpositions as complementary indices for Korean while the indicator for French is the preposition [Le Priol, 2004]. In general, the structure of Korean sentences is subject - object – verb. For example,

"*Seoul-un Hankuk-e issumnida*."
(Seoul in Korea is)

Given this sentence structure, verbs such as *issumnida* are indicators, and postpositions (*josa*) such as -e positioned at the end of nouns are complementary indices. Since Korean verbs almost always appear at the end of sentences, finding an indicator is not difficult. Indicators (verbs) can be classified into the following semantic categories.

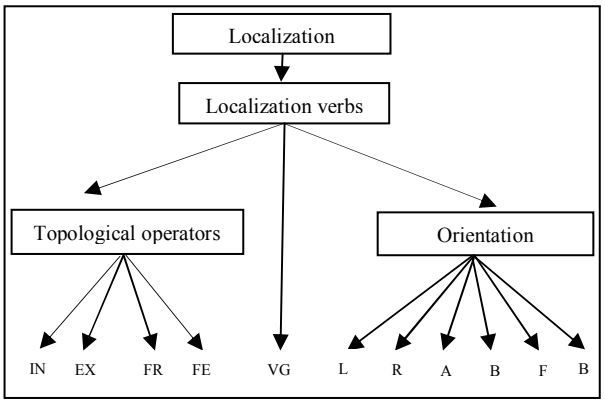
(a)existence ={있다/to be, 존재하다/to exist…},
(b)movement ={움직이다/to move, 이동하다/to transfer…},
(c)arrival ={도착하다/to arrive, 닫다/to reach…},
(d)distance ={멀다/to be distant, 가깝다/to be close…},
(e)adjective ={예쁘다/to be beautiful, 좋다/to be good…},
(f)active ={하다/to do, 놓다/to put…},
(g)passive={연결되다/to be connected, 서게하다/to make stand…}.

Next, following Flaguel's division of spatial prepositions [Flaguel, 1997], we can classify localization relationship complementary indices into five specific semantic categories [Le Priol, 2004].

(a) IntroPlaceIN(interior)= {an&e, naebu&e…}
(b) IntroPlaceEX(exterior) = {bak&e, keol&e…}
(c) IntroPlaceFR(frontier) = {keongkeo&e, kajangjari&e..}
(d) IntroPlaceFE(closure) = {e}
(e) IntroPlaceVG(close-by) = {oelp&e, keot&e…}

Similarly, we can classify complementary indices into six specific semantic categories based on orientation prepositions.

(a) IntroPlaceLeft = {oinzzok&e, joa&e…}
(b) IntroPlaceRight = {olenzzok&e, u&e…}
(c) IntroPlaceAbove = {wi&e, …}
(d) IntroPlaceBelow = {alasszzok&e…}
(e) IntroPlaceFront = {ap&e…}
(f) IntroPlaceBack = {twi&e…}



<Figure 3. Semantic map of localization relation>

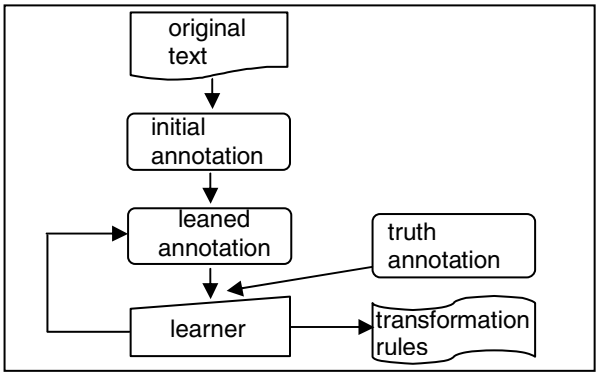
5.2. Contextual Exploration Rule

The contextual exploration rule is the most important part of our system. This module allows us to identify semantic information by taking into account the textual context. The Contextual Exploration rule is not only restricted to the concepts of adjacency and concatenation such as some systems based on finite states automata [Desclés, 2006]. Indeed, the Contextual Exploration rule can apply several linguistic signs located at a very long distance. Finite states automata are a simple example of Exploration

Contextual system. The Contextual Exploration rule is presented with this general form:

IF an Indicator IND classified into a specific semantic category is found,
AND IF one or more complementary indices I1, I2, …,In, classified into the same category as IND are identified,
THEN the specific semantic annotation is applied.

Here, Indicator IND and a string of complementary indices I1 I2,⋯, In are at the same level and are not integrated in an hierarchical dependence. Furthermore, complementary indices I1, …, In, in an EC rule, can be located at a very long distance from an Indicator IND.
To establish contextual exploration rules for Korean, we collected a corpus of Korean texts over several subject domains, including politics, economics, society, culture, sports and information technology. The corpus comes from Naver, the most popular Korean web site to eliminate too much classical language. We then subdivided the corpus into training (2000kB) and test (1000kB) data sets. From this training set, firstly, we segmented sentences by typographical signs, such as, punctuation marks. Unlike Latin languages, there are no capital letters in Korean and each sentence ends with punctuation marks such as periods, question marks, and exclamation marks. We then applied methodology popularized by Eric Brill [Brill, 1995]. for linguistic parsing. In this methodology, illustrated in Figure 4, a system learns a sequence of rules that best labels training data. These rules are then used to annotate previously unseen data.



<Figure 4. Overview of general transformation-based error-driven learning>

In our approach, the initial annotation was very simple. We assigned to each verb localization, Indicator and complementary indices its most likely tag without any regard to context.

(a) salam-i < IntroPlaceVG >keote</ IntroPlaceVG > < verbExistence >issta</ verbExistence > (There is one person next to me).

(b) < IntroPlaceFE >jip-e< IntroPlaceFE > < verbArrival >dochakhata</ verbArrival> (I arrive at home).

Then, the iterative learning algorithm applied transformation rules on the output of a simple first tagging to obtain its final result as follows:

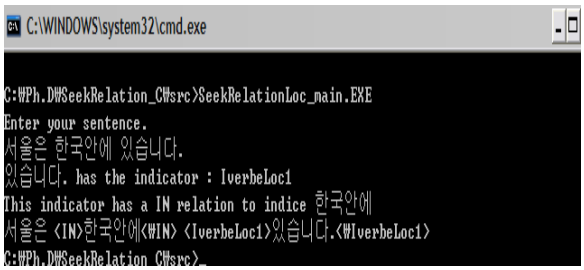
(a) a1 a2-**e** verbe
(1) salam-i keot-e issta.
(2) jip-e dochakhata.
(b) a1-**e** a2 verbe
(3) beol-e mos-ul geollita.
(4) dambae-e bul-ul buthita.

Now, we define rules to perform specific semantic annotation based on the indicators, complementary indices and their classifications defined above. Each rule is linked to an indicator. If one or more complementary indices are found according to a rule, then semantic annotation is applied. For example, we can define:

Rule1,

IF a passive verb is found at the end of sentence,
AND IF a postposition of interior place is found at left of the passive verb,
THEN the specific semantic annotation "be in" is applied

Before implementing into an integrated set of tools, EXCOM, for automatic semantic annotation, we created a simple automatic semantic annotation engine and interface for Korean (Figure 5) and used it to construct 76 such rules for localization relation annotation.



<Figure 5. Automatic semantic annotation interface for Korean>

In Figure 5, when a sentence "서울은 한국안에 있습니다(Seoul is in Korea)" is entered as input, semantic rules processing is applied with indicators and contextual indices for identifying a location relation. As a result, we obtain a structured sentence and semantic annotation metadata:

서울은<IN>한국안에</IN>
<IverbeLoc1>있습니다</IverbeLoc1>

(Seoul<IN>in Korea</IN>
<IverbeLoc1>is</IverbeLoc1>)

5.3. Evaluation

In order to evaluate the annotation results, we measure *precision* and *recall* [Manning and Shutze, 1999]. Precision corresponds to the number of correctly marked annotations divided by the number of annotations produced by the system. The recall rate is the number of annotations assigned a particular classification divided by the number of annotations in the testing set which actually belong to that class.

With our first attempts, we achieved precision of 88% and a recall rate of 86%. To our knowledge, this is the first successful Korean semantic annotation.

6. Conclusions

Without the need for morphological or syntactic analysis, our first-generation automatic semantic annotation system for Korean covers 88% of standard Korean sentences across a wide range of domains. This allows us to sidestep the thorny issues presented by the Korean language's agglutinative nature. Further research and development may lead to significantly higher performance. However, we believe that even the current system can serve as the basis for general cross-domain applications. In addition, in our experience, it is relatively easy to gain high performance by limiting data sets to a single domain.

From a more expansive perspective, the success of the Contextual Exploration Method in Korean, gives us cause to be optimistic about its application to other agglutinative languages, such as Japanese, Turkish and Finnish. Given Contextual Exploration Method's previous successful application to French and Arabic languages, we may even hope for a truly multilingual solution. Indeed, EXCOM is an effort already underway at LaLICC to create an integrated set of tools for automatic semantic annotation for use in many different languages.

10. References

[1] C. Fucha, Danlos, A. Lacheret-Dujour, D. Luzzati, and B. Victorri, *Linguistique et traitement automatique des langues*, Hachette, Paris, 1993.

[2] J. P, Desclés, E. Cartier, A. Jackiewicz and J.L. Minel, "Textual Processing and Contextual Exploration Method", *Context97*, Rio de Janeiro, 1997, PP. 189-197.

[3] G. Ferrari, "A state of the art in Computational Linguistics", *17th International Congress of Linguists-Prague*, Prague, 2003.

[4] P. Sébillot, "Morphological Analysis", WP5 Task 4 State-of-the-Art Natural Language Processing.

[5] D.G. Lee, H.C. Rim, H.S. Lim, "A Syllable Based Word Recognition Model for Korean Noun Extraction". *ACL 2003,* Sapporo, 2003, pp. 471-478.

[6] A. Motasem, H.I. Amr, and J. P. Desclés, "Semantic Annotation of Reported Information in Arabic", *FLAIRS2006*, Florida, 2006.

[7] J. Berri, Le Roux, D., Malrieu D. and Minel, J.L., "SERAPHIN, main sentences automatic extraction system", *Second Language Engineering Convention,* London, 1995**.**

[8] Desclés, J-P., *Langages applicatifs*, Langues naturelles et Cognition, Hermès, Paris, 1990.

[9] L. Talmy, "Force Dynamics in Language and Cognition", *Cognitive Science*, London, 1988, pp. 49-100

[10] F. Le Priol, "La relation de localisation", Lab. – CNRS of Langages, Logiques, Informatique, Cognition et Communication, Paris-Sorbonne University, 2004.

[11] B. Djioua, J. Garcia-Flores, A. Blais, J.P. Desclés, G. Guibert, A. Jackiewicz, F. Le Priol, L. Nait-Baha and B. Sauzay, "EXCOM: an automatic annotation engine for semantic information", *The 19th International FLAIRS Conference*, Florida, 2006, PP. 285-290.

[12] V. Flageul, "Représentation des prépositions spatiales en français". Paris-Sorbonne University, 1997.

[13] J. P. Desclés, "Contextual Exploration Processing for Discourse Automatic Annotation for Texts", *FLAIRS2006*, Florida, 2006.

[14] E. Brill, "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging.", *Computational linguistics*, 1995, pp. 543-566.

[15] C. D. Manning, and H. Shutze, "Foundations of Statistical Natural Language Processing", MIT Press, London, 1999.