Korean Semantic Annotation on the EXCOM Platform

Hyunzoo Chai, Brahim Djioua, Florence Le Priol, Jean-Pierre Desclés*

LaLLIC, UMR 8139, University of Paris-Sorbonne/CNRS
28 rue Serpente, 75006 Paris, France
{hyunzoo.chai, bdjioua, Florence.le-priol, jean-pierre.descles}@paris4.sorbonne.fr

Abstract. We present an automatic semantic annotation system for Korean on the EXCOM (EXploration COntextual for Multilingual) platform. The purpose of natural language processing is enabling computers to understand human language, so that they can perform more sophisticated tasks. Accordingly, current research concentrates more and more on extracting semantic information. The realization of semantic processing requires the widespread annotation of documents. However, compared to that of inflectional languages, the technology in agglutinative language processing such as Korean still has shortcomings. EXCOM identifies semantic information in Korean text using our new method, the Contextual Exploration Method. Our initial system properly annotates approximately 88% of standard Korean sentences, and this annotation rate holds across text domains.

Keywords: semantic annotation, semantic web, agglutinative language, Korean, EXCOM, Contextual Exploration Method.

1. Introduction

The ultimate aim of natural language processing is to enable computers to accurately accomplish tasks, just as humans do. Humans are capable of finding relevant documents on the Web for queries like "Who did Jacques Chirac meet?" Returning such responses as: "Jacques Chirac met with a group of successful young entrepreneurs…", "The French President had lunch with the Dalai Lama at the Elysee Palace…", "French President Jacques Chirac was interviewed by CNN's Jim Bittermann in Paris." and so on. In this context, we are able to understand that words such as have lunch and interviewed could have the same meanings as meet. However, a computer cannot produce the same results without human intervention because languages rely on the concept systems of humans, not the rules of machines. The purpose of semantic annotation is enabling computers to understand human language, so that they can perform more sophisticated tasks.

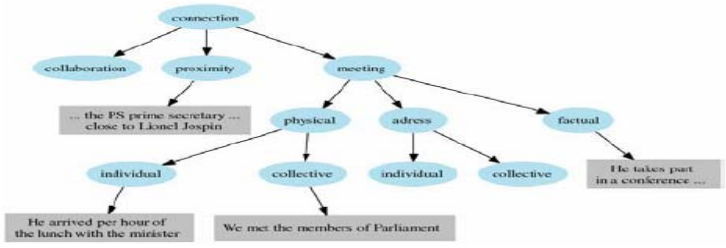


Figure 1: Semantic map for "meet" semantic category

The realization of semantic processing requires the widespread annotation of documents. There are two ways to annotate texts, manual and automatic annotation. Manual annotation is more easily accomplished for a few limited texts but it is an expensive process to tag large documents for semantic processing. Automatic semantic annotation is crucial to semantic processing even though it still has unsolved problems.

In this paper we first give an overview of existing semantic annotation systems. Then, we explain the linguistic characteristics of Korean and present our experiences in creating an engine for automatic semantic annotation of Korean texts. The evaluation results are presented before the conclusion.

2. Semantic Annotation

Semantic annotation is the process of mapping instance data to ontology. Ontology is the conceptualization of a domain that typically is represented using domain vocabulary (Taniar and Rahayu , 2006). For example, the concept “meet” can be expressed using different terms: encounter, get together, have lunch, interview, join, experience… . Semantic annotation can be classified in two categories based on the type of annotation method used: Pattern-based and machine learning-based. Here we briefly present several currently available semantic annotation systems.

Platform	Doc. Type	IE Method	MR	External Input	X	IE Tools	Initial Ontology
AeroDAML	HTML	Rule	Y	Rule	N	AeroText	WordNet; AeroText KB
Armadillo	HTML	PD	Y	Seed	N	Amilcare ANNIE	Address Book; Paper Citation
KIM	HTML	PM	Y	Gazetteer KB population	N	GATE	KIMO
MnM	HTML, Plain Text	WI using ML LP^2	N	Annotated corpus	Y	Amilcare	KMi
MUSE	Plain Text	Rule	Y	Gazetteer Rules	Y	GATE JAPE	User constructed
Ont-O-Mat: Amilcare	HTML	WI using ML LP^2	N	Annotated corpus	Y	Amilcare	User constructed
Ont-O-Mat: PANKOW	HTML	PD	N	Web pages	Y	PANKOW	User constructed
SemTag	HTML	TLM	N	Taxonomy with labels	Y	Seeker platform	TAP with 72K labels

(Doc: Document; IE: Information Extraction; JAPE: Java Annotations Pattern Engine; KB: Knowledge Base; ML: Machine Learning; MR: Manual Rules; N: No; PD: Pattern Discovery; PM: Pattern Matching; TLM: Taxonomy Label Matching; WI: Wrapper Induction; X: Extensible; Y: Yes)

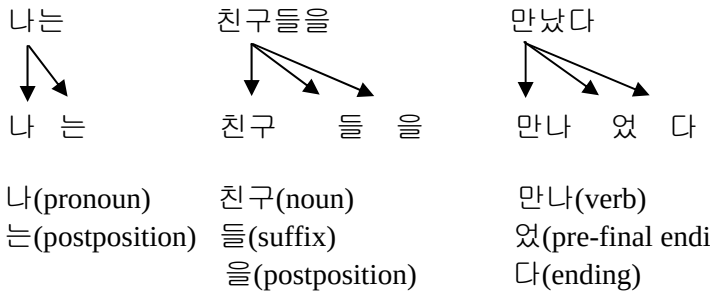
Figure 2: Semantic Annotation Platforms

AeroDAML, Armadillo, KIM, MUSE, PANKOW, SegTag are pattern-based systems. An initial set of entities is defined and the corpus is scanned to find the patterns in which the entities exist. New entities are discovered, along with new patterns. This process continues recursively until no more entities are discovered. Annotations can also be generated by using manual rules to find entities in text (Reeve and Han, 2005). Machine learning-based systems such as MnM and Amilcare utilize probability and induction methods. The locations of entities in the text are predicted by statistical models or wrapper induction.

EXCOM is a pattern-based system for automatic semantic annotation. Most pattern-based annotation systems are based on morphologic and syntactic analysis. 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 extracting semantic information. However, compared to that of inflectional languages, the technology for the processing of agglutinative languages such as Korean still has shortcomings (Sébillot, 2004).

3. 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:



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.

4. EXCOM

EXCOM is an XML-based system for automatic annotation of text according to semantic categories (Djioua, 2006). The system is based on the theory of the Contextual Exploration Method (Desclés et al, 1991). 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 linked with indicator and complementary indices.

4.1.Contextual exploration method

The Contextual Exploration Method (Desclés, 1997) provides a method of identifying semantic information in text, without the need for morphological and syntactical analysis stages. The method has already been applied to French and Arabic languages (Desclés and Motasem, 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 the precautions, five minutes later, the train would have 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 “ran off”. 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:
- linguistic sign (Indicator) found in a given text,
- linguistic clues (Indices) solving ambiguity affecting the indicator in their context,
- 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; Berri 1996).

4.2.Architecture

The process of EXCOM annotation consists of the following steps, illustrated in Figure 3:

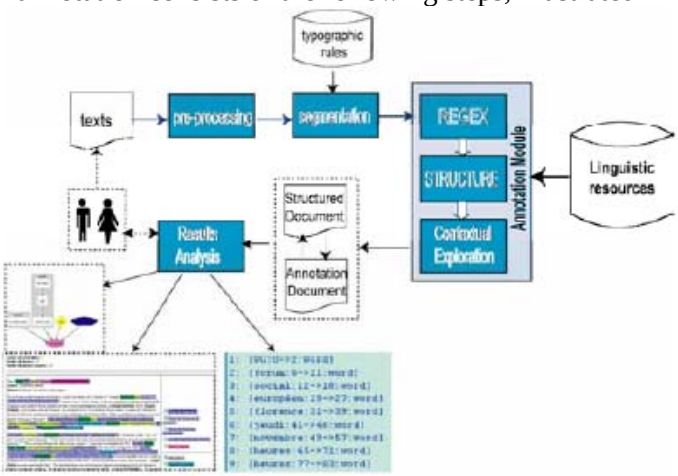


Figure 3: EXCOM Architecture

(a) An input text in HTML-XML-TXT format is encoded on ISO-Latin1 or Unicode.
(b) Encoded documents are converted to plain text format.
(c) Plain text documents are transformed into structured documents with structural annotations such as title, section, paragraph and sentence.
(d) In the annotation module, first, regular expression processing is performed to identify first-level data such as named-entities, locations, dates and temporal expressions. Then, structure rules identify complex structures based on first-level annotation. Semantic rules processing identifies a semantic category with indicators and contextual clues and then negative rules identify negations of semantic categories. Finally, modality rules identify the achieved and possible semantic relations.
(e) As a result, structured document and semantic annotation metadata are obtained.

EXCOM uses an XSLT engine (with XPath parser) to identify nodes in the input XML document and process transformations by adding XML elements and attributes.

4.3.Localization relations for Korean

The extraction of localization relations is part of the automatic annotation project, EXCOM (EXploration COntextuelle Multilingue) at the LaLICC laboratory of Paris-Sorbonne University. The first step of the extraction of localization relations is to establish lists of indicator and complementary indices. For Korean, we classified verbs expressing localization for the indicator and chose postpositions as complementary indices 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 (Flaguel, 1997) division of spatial prepositions, 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…}

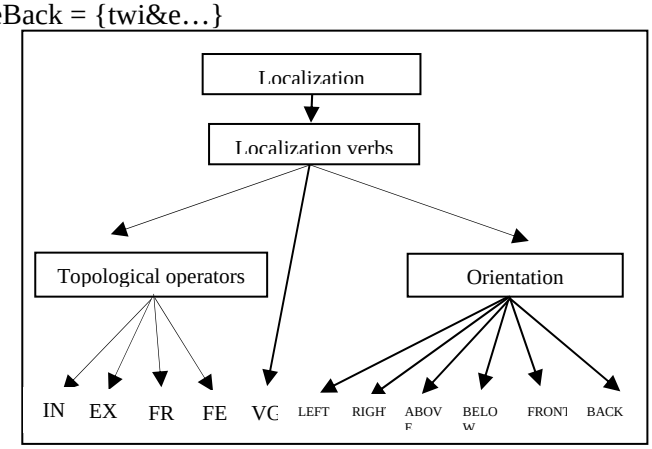


Figure 4: Semantic map of localization relation

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 state automata (Desclés, 2006). Indeed, the Contextual Exploration rule can apply several linguistic signs located at a very long distance. Finite state automata are a simplifed case of a Contextual Exploration 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 a 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 little-used archaic forms. We then subdivided the corpus into training (2000kB) and test (1000kB) data sets. From the training set, we first 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 for linguistic parsing. In this methodology, illustrated in Figure 5, a system learns a sequence of rules that best labels training data. These rules are then used to annotate previously unseen data.

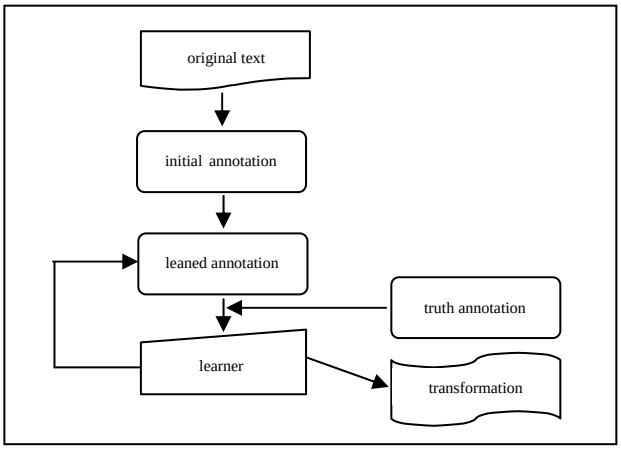


Figure 5: Overview of general Transformation-based Error-Driven learning

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 (Brill, 1995).

(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 a sentence,
AND IF a postposition of interior place is found 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 6) and used it to construct 76 such rules for localization relation annotation.

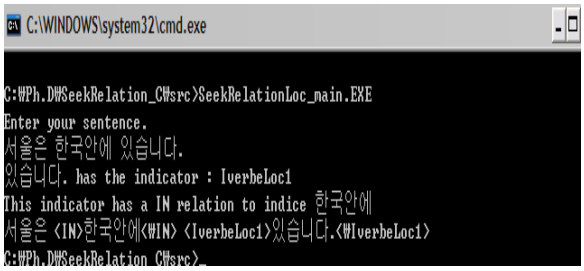


Figure 6: Automatic semantic annotation interface for Korean

In Figure 6, 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>)

4.4.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 practical Korean semantic annotation.

5. Conclusion

EXCOM is a comprehensive framework for creating automatic semantic annotations based on the Contextual Exploration method. It (Djioua et al. 2006) is an effort underway at LaLICC to

create an integrated set of tools for automatic semantic annotation for use in many different languages. This paper shows the application of EXCOM's multilingual automatic annotations for Korean-language semantic categorizations. Our first-generation automatic semantic annotation system for Korean based on the Contextual Exploration Method covers 88% of standard Korean sentences across a wide range of domains. This allows us to sidestep the thorny issues presented by 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.

References

Berri J., Le Roux, D., Malrieu D. and Minel, J.L., SERAPHIN main sentences automatic extraction system, Second Language Engineering Convention, Londres, 1995.

Berri J., 1996, Contribution à la méthode d'exploration contextuelle, applications au résumé automatique et aux représentations temporelles; réalisation informatique du système SERAPHIN, thesis of Univ. Paris Sorbonne.

Brill, E., 1995, Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. Computational Linguistics, December 1995.

Desclés, J-P., 1990, Langages applicatifs, Langues naturelles et Cognition, Hermès, Paris.

Desclés, J. P,; Cartier. E.; Jackiewicz, A.; and Minel, J. L. 1997, Textual Processing and Contextual Exploration Method, 189-197, Context97, Rio de Janeiro.

Desclés, J. P. 2006, Contextual Exploration Processing for Discourse Automatic Annotation for Texts, FLAIRS2006, Melbourne, Florida.

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

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

Flageul, V. 1997, Représentation des prépositions spatiales en français. Ph. D. diss., Lab. – CNRS of LaLLIC, Paris-Sorbonne Univ.

Fucha, C. ; Danlos, ; Lacheret-Dujour, A. ; Luzzati, D. ; and Victorri B. eds. 1993, Linguistique et traitement automatique des langues, Hachette, Paris.

Le Priol, F. 2004, La relation de localisation, Lab. – CNRS of Langages, Logiques, Informatique, Cognition et Communication, Paris-Sorbonne Univ.

Lee, D.G.; Rim, H.C.; Lim, H.S., 2003, A Syllable Based Word Recognition Model for Korean Noun Extraction. ACL 2003, 471-478, Sapporo.

Manning, C. D. and Shutze, H. 1999, Foundations of Statistical Natural Language Processing, London: MIT Press.

Motasem, A.; Amr H.I.; Desclés, J. P. 2006, Semantic Annotation of Reported Information in Arabic, FLAIRS2006, Melbourne, Florida.

Reeve, L and Han, H, 2005, Survey of Semantic Annotation Platforms, 2005 ACM Symposium on Applied Computing, Santa Fe, New Mexico.

Saint-Dizier, P. and Viegas E. 1995, Constraint propagation techniques for lexical semantics descriptions in Computational semantics, 426-440, Cambridge Univ.

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

Talmy, L. 1988, Force Dynamics in Language and Cognition, pp. 49-100, Cognitive Sciences.
Taniar, D and Rahayu, J.W., 2006, Web Semantic Ontology, Hershey, USA.