http://dx.doi.org/10.3745/KIPSTB.2012.19B.1.043

한국어 의미 표지 부착 말뭉치 구축을 위한 자동 술어-논항 분석기 개발

조 정 현[†] · 정 현 기[†] · 김 유 섭[††]

요 약

의미 역 결정 (Semantic Role Labeling)은 문장의 각 요소들의 의미 관계를 파악하는 연구 분야로써 어휘 중의성 해소와 더불어 자연언어처리에서의 의미 분석에서 매우 중요한 위치를 차지하고 있다. 그러나 한국어의 경우에는 의미 역 결정에 필요한 언어 자원이 구축되지 못하여 연구의 진행이 매우 미진한 상황이다. 본 논문에서는 의미 역 결정에 필요한 언어 자원 중에서 가장 널리 사용되고 있는 PropBank의 한국어 버전의 구축을 위한 시작 단계로써 자동 술어-논항 분석기를 개발하였다. 자동 술어-논항 분석기는 크게 의미 어휘 사전과 자동 술어-논항 추출기로 구성된다. 의미 어휘 사전은 한국어 동사의 격틀 정보를 구축한 사전이며 자동 술어-논항 추출기는 구문 표지 부착된 말뭉치로부터 특정 술어와 관련있는 논항의 의미 부류를 결정하는 모듈이다. 본 논문에서 개발된 자동 술어-논항 분석기는 향후 한국어 PropBank의 구축을 용이하게 할 것이며, 궁극적으로는 한국어 의미 역 결정에 큰 역할을 할 것이다.

키워드 : 의미 분석, 의미 역 결정, PropBank, 자동 술어-논항 분석기, 의미 어휘 사전, 자동 술어-논항 추출기

A Development of the Automatic Predicate-Argument Analyzer for Construction of Semantically Tagged Korean Corpus

Jung-Hyun Cho[†] · Hyun-Ki Jung[†] · Yu-Seop Kim[††]

ABSTRACT

Semantic role labeling is the research area analyzing the semantic relationship between elements in a sentence and it is considered as one of the most important semantic analysis research areas in natural language processing, such as word sense disambiguation. However, due to the lack of the relative linguistic resources, Korean semantic role labeling research has not been sufficiently developed. We, in this paper, propose an automatic predicate-argument analyzer to begin constructing the Korean PropBank which has been widely utilized in the semantic role labeling. The analyzer has mainly two components: the semantic lexical dictionary and the automatic predicate-argument extractor. The dictionary has the case frame information of verbs and the extractor is a module to decide the semantic class of the argument for a specific predicate existing in the syntactically annotated corpus. The analyzer developed in this research will help the construction of Korean PropBank and will finally play a big role in Korean semantic role labeling.

Keywords : Semantic Analysis, Semantic Role Labeling, Automatic Argument-Predicate Analyzer, Semantic Lexical Dictionary, Automatic Predicate-Argument Extractor

1. 서 론

일반적으로 자연언어처리에서는 형태소 분석(Morphological Analysis), 구문 분석(Syntactic Analysis), 의미 분석(Semantic Analysis), 담화 분석(Discourse Analysis), 그리고 대화 분석(Dialogue Analysis)을 기반 기술로 분류한다[1]. 한국어의 경우 형태소 분석 및 구문 분석에서는 매우 많은 연구들이 진행되어 왔으며, 이들 분석과 관련하여 다양한 표지 부착된 한국어 말뭉치가 개발되어 왔다. 대표적인 말뭉치로는 한국전자통신연구원의 품사 부착 말뭉치 및 구문구조 부착 말뭉치, 21세기 세종계획에 의한 형태소 분석 말뭉치, 구문 분석 말뭉치, 의미 어휘 분석 말뭉치, 한국과학기술원의 품사 부착 말뭉치 등이 있다[2].

이러한 말뭉치를 활용한 기계 학습 및 통계 기반 알고리즘의 개발은 한국어 정보처리 기술을 매우 극적으로 발전시켜왔다. 그러나 의미 역 결정(Semantic Role Labeling)과 같

※ 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2009-0074909).
† 준 회 원 : 한림대학교 컴퓨터공학과 박사과정
†† 종신회원 : 한림대학교 유비쿼터스컴퓨팅학과 교수
논문접수 : 2011년 5월 19일
수 정 일 : 1차 2011년 7월 29일
심사완료 : 2011년 8월 18일

은 문장 단위의 의미 분석은 관련 한국어 말뭉치가 부재하여 부트스트래핑(Bootstrapping)이나[3] 비지도 학습(Unsupervised Learning) 방법론[4]에 의지하고 있다. 특히 이러한 말뭉치의 부재는 의미 분석 연구가 한국어가 아닌 영어에 대하여 진행됨으로써[5, 6] 더더욱 한국어 의미 분석 연구를 어렵게 하였다.

Proposition Bank (이하 PropBank) [7]는 동사의 술어-논항 (Predicate-Argument) 구조를 태그해 놓은 말뭉치를 말하고, 현재는 의미 역 결정 관련 연구에 매우 다양하게 사용되고 있다. 기본적으로 PropBank는 동사의 의존 구조(dependency structure)를 우선 분석하고 술어에 의미상 의존하고 있는 여러 문장 성분들에 적합한 논항 번호를 부여함으로써 문장의 술어-논항 관계를 보여주고 있다.

의미 역 결정 관련 연구들은 주로 문장의 술어-논항 구조를 파악하고자 하는데[8], 이 때, 논항은 술어와의 관계를 정의하는 다양한 수준의 역할(role)을 가진다. 우리가 알고 있는 'agent'나 'theme'과 같은 역할은 가장 일반화된 역할이라 할 수 있다. PropBank는 술어에 특화된 논항을 숫자로써 일반화하여 다양한 술어에 대한 다양한 논항을 모두 아우를 수 있도록 하였고, 이것이 문장 단위의 의미 역 결정에 큰 도움이 되었다[8, 9]. PropBank는 단독 또는 다른 언어 자원과 함께 사용되어 의미 분석과 관련한 여러 연구에 활용되어 왔다[10-13]. 또한 PropBank는 그 활용성 때문에 영어뿐만 아니라 중국어[14], 아랍어[15], 바스크어[16] 등 여러 언어에 대해서 구축되고 있다.

본 논문에서는 이처럼 한국어 의미 분석에 매우 중요한 역할을 할 수 있는 한국어 PropBank 구축의 시작 단계로써 자동 술어-논항 분석기를 개발하였다. 자동 술어-논항 분석기는 관련 말뭉치가 부재한 상황에서 자동으로 의미 표지 부착된 말뭉치를 초벌 생성할 수 있고, 이 결과를 향후 수작업에 활용하여 한국어 PropBank의 구축 시간을 크게 단축시킬 수 있다. 실제 영어 PropBank를 구축하는 초기 단계에서는 [17]에서 개발된 자동 분석기를 활용하였다. 물론 중국어, 아랍어, 힌두어의 경우와 같이 병렬 말뭉치를 활용하여 한국어 PropBank를 구축하는 것도 가능하겠으나, 한국어는 병렬 말뭉치가 제대로 구축되어 있지 않아 한국어 PropBank 구축에 이러한 방법론을 활용하는 것에는 제약이 있다.

본 논문에서 제시하는 자동 술어-논항 분석기는 크게 두 개의 부분으로 이루어져 있다. 첫째, 동사의 격틀 사전을 분석 및 가공하여 구축된 의미 어휘 사전이다. 의미 어휘 사전은 기본적으로 기존의 동사 격틀 사전이 가지고 있는 격틀 정보가 부족하기 때문에 새로이 확장, 구축되었다. 본 논문에서는 21세기 세종 계획[1]에서 구축한 동사 격틀 사전을 확장하였다. 세종 격틀 사전의 경우에는 약 1만 5천개의 동사에 대한 격틀이 사전화되어 있으나, 모든 동사가 풍부한 격틀 정보를 가지고 있는 것은 아니다. 이는 말뭉치에 자주 나타나지 않는 동사의 경우 격틀 정보를 구축할 근거가 없었기 때문이다. 본 논문에서는 유사한 격틀 정보를 가지는 동사들을 하나의 클러스터로 묶어 서로 격틀 정보를 공유하게 함으로써 격틀 정보의 부족 현상을 극복하고자 하였다.

둘째는 술어-논항 관계 추출 모듈이다. 기존에 영어에 관하여 구축된 관계 추출기[17]는 구문 표지 말뭉치에 나타나는 구문 패턴에 근거하여 관계를 추출하였다. 그러나 한국어는 어순이 매우 자유로운 반면 조사/어미라는 매우 특수한 형태의 형태소가 술어-논항간의 관계를 표현하는데 적극적으로 활용되기 때문에 패턴보다는 조사의 분석을 통한 의미 역 결정이 더 자연스럽다. 말뭉치의 문장에서 나타난 격틀과 격틀 사전에서 나타난 격틀간의 비교를 위해서는 격틀에 포함되어 있는 논항의 의미를 분석하여야 한다. 이를 위하여 세종 계획에서의 의미 어휘 분류 체계를 단순화하여 논항들의 의미 부류를 결정하였다.

이상을 통합하여 한국전자통신연구원의 구문 표지 부착 말뭉치 문장에서 동사 어휘를 발견하게 되면, 해당 동사의 여러 논항중에서 부사격 논항들을 따로 분리하고 이들 술어-논항 관계를 동사의 의미 어휘 사전에 기술되어 있는 격틀 정보와 비교하여 해당 논항의 의미 역을 결정하였다. 본 논문에서는 분석 대상으로 부사격 논항들만을 삼았는데, 이는 한국어의 경우 주격조사와 목적격조사의 논항들은 AGT나 THM과 같이 그 의미 역이 대부분 정해져 있기 때문이다. 따라서 본 논문에서는 비교적 모호성이 큰 부사격 조사의 논항들만을 분석 대상으로 하였다. 보다 상세한 주격 조사의 논항 및 목적격 조사의 논항은 향후 연구 과제로 잠시 미루고자 한다.

본 논문은 다음과 같은 내용으로 구성된다. 2장에서는 PropBank에 대하여 간략하게 설명하고, 본 과제에서 개발한 전체 시스템의 개요에 대하여 3장에서 설명한다. 3장에서는 의미 어휘 사전과 이를 활용하여 술어-논항 관계를 추출하는 자동 술어-논항 추출 모듈에 관하여 기술한다. 4장에서는 구축된 시스템의 성능에 대하여 기술하고, 마지막 5장에서는 본 논문의 결론을 정리하고, 향후 더욱 발전적인 연구를 위한 제언을 기술할 것이다.

2. Propositional Bank

Propositional Bank (PropBank)는 구문 표지 부착 말뭉치와 의미 표지 부착 정보 두 개의 쌍으로 이루어진다. 한국어 문장 "그는 르노가 3 월말까지 인수제의 시한을 갖고 있다고 덧붙였다."를 구문 분석한 결과 및 이를 PropBank 표지 부착한 결과는 각각 다음과 같다.

```
(S (NP-SBJ 그/NPN+은/PAU)
    (VP (S-COMP (NP-SBJ 르노/NPR+이/PCA)
        (VP (VP (NP-ADV 3/NNU
                월/NNX+말/NNX+까지/PAU)
            (VP (NP-OBJ 인수/NNC+제의/NNC
                시한/NNC+을/PCA)
            갖/VV+고/ECS))
```

1) http://www.sejong.or.kr

있/VX+다/EFN+고/PAD))
덧붙이/VV+었/EPF+다/EFN)
./SFN)

12 2:2-ARG0 4:2-ARGM-TMP 8:2-ARG1 12:0-rel 14:1-AUX
17 0:2-ARG0 2:3-ARG1 17:0-rel

이 예문에서는 2개의 술어('갖다', '덧붙이다')에 대한 논항들에 대한 정보를 기술하고 있다. 첫 번째 술어인 12번째 형태소 '갖다'는 모두 3개의 논항을 가지고 있다. 여기서 2번째(0부터 시작하여) 형태소인 '르노'를 포함하고 있는 최초 2번째 구절 단위인 '(NP-SBJ 르노/NPR+이/PCA)'가 술어의 Arg0 (행위주) 가 된다. 여기서 형태소 '르노/NPR' 이 첫 번째 구절 단위이고 이것이 포함된 가장 작은 구절 단위인 '(NP-SBJ 르노/NPR+이/PCA)'가 두 번째, 그리고 이를 포함하고 가장 작은 구절 단위인 '(S-COMP (NP-SBJ 르노/NPR …)'이 세 번째 구절 단위가 된다. 그리고 4번째 형태소인 '3 '을 포함하고 있는 최초 2번째 구절인 '(NP-ADV 3/NNU .../PAU)'가 ARGM-TMP가 되며 8번째 형태소인 '인수'를 포함하는 최초 2번째 구절인 '(NP-OBJ 인수/NNC ...을/PCA)'가 Arg1 (대상주)가 된다. 마지막 14번째 형태소인 '있'의 경우에는 조동사의 역할을 하는 것으로 분석되었다. 여기서 술어 '갖다'의 arg0가 '행위주'가 되고 arg1이 '대상주'가 된다는 정보는 동사 '갖다'의 프레임 파일에 기술되어 있다.

마찬가지로 두 번째 술어인 17번째 형태소 '덧붙이다'는 2개의 논항을 가지는데, 이 중에서 첫 번째 논항은 0번째 형태소인 '그'를 포함하고 있는 최초 2번째 구절 단위인 '(NP-SBJ 그/NPN+은/PAU)가 Arg0(행위주)가 되고 2번째 형태소인 '르노'를 포함하고 있는 최초 3번째 구절 단위인 '(S-COMP (NP-SBJ 르노/NPR+ 이/PCA) (VP (VP … 있/VX+ 다/EFN+ 고/PAD)) 전체가 ARG1(대상주)가 된다.

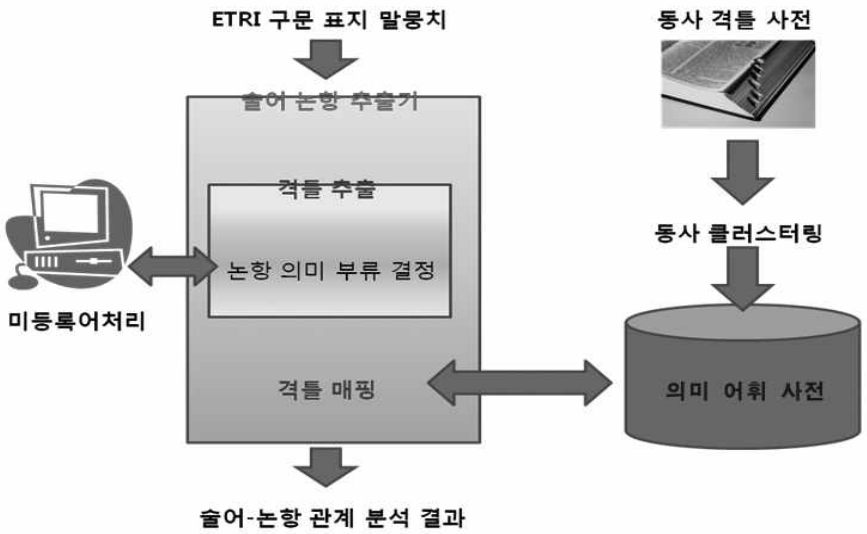
본 논문의 주 목적은 이러한 한국어 PropBank를 구축하는데 있어 초벌 의미 표지 부착 말뭉치 구축을 보다 효율적으로 도와줄 수 있는 자동 술어-논항 분석기의 개발에 있다. 즉, 자동 술어-논항 분석기를 활용하여 구문 표지 부착된 말뭉치에서 술어-논항 쌍을 추출하여 각 논항이 술어에 대해 어떤 의미적인 관계를 가지고 있는가를 1차적으로 찾고자 한다. 이를 위하여 본 논문에서는 세종 계획에서 구축된 동사 격틀 사전으로 한국어 프레임 파일을 대신하고, 한국전자통신연구원에서 구축된 한국어 구문 표지 부착 말뭉치의 문장에 나타난 술어들을 대상으로 술어-논항 쌍을 분석하고자 한다.

3. 자동 술어-논항 분석기

본 논문에서 제안하는 자동 술어-논항 분석기의 전체 시스템 개요는 (그림 1)에서 보여준다.

자동 술어-논항 분석기는 크게 두 개의 모듈로 구성된다. 첫째는 의미 어휘 사전으로 이 사전에는 동사들의 격틀 정보가 저장되어 있는데, 개별 동사의 격틀 정보가 부족하기 때문에 복수의 동사들을 보유하고 있는 격틀의 유사도에 기반하여 클러스터링하여 정보 부족 문제를 해결하고자 하였다. 둘째는 술어-논항 추출기 모듈로써 이 모듈은 구문 표지가 부착된 문장에서 술어-논항의 의미 관계를 파악하는데 필요한 부분을 추출하여 해당 논항이 술어와 어떤 의미 관계에 있는가를 결정해 준다.

ETRI의 구문 표지 부착된 말뭉치에서 처리 대상 문장이 입력되면 분석의 대상이 되는 술어-논항의 쌍이 전처리 과정을 통하여 추출된다. 그리고 각 논항들은 세종 체언 의미 체계를 활용하여 적절한 의미를 부여받는다. 격틀 사전에서의 논항은 이러한 의미 체계로 표시되기 때문에 개별 단어들은 반드시 의미를 부여받아야 한다. 세종 동사 격틀 사전의 각 표제어들은 보유하고 있는 격틀의 유사성에 따라서



(그림 1) 자동 술어-논항 분석기 개요

군집화(Clustering)되는데, 동일한 군집에 분포하는 모든 동사들은 서로 보유하고 있는 격틀 정보를 공유하도록 한다. 이러한 과정을 통하여 의미 어휘 사전이 구축된다. 최종적으로 문장에서 추출된 술어-논항 쌍은 의미 어휘 사전의 격틀 정보를 참조하여 의미 관계가 분석된다.

3.1절에서는 의미 어휘 사전의 구축에 있어서 가장 중요한 군집화 과정과 구축된 사전의 현황에 대하여 설명하고 3.2 절에서는 술어-논항 추출 과정에서 매우 중요한 논항의 의미 부여 과정과 전체 추출 과정에 대하여 설명한다.

3.1 의미 어휘 사전

3.1.1 동사 격틀

본 논문에서는 의미 어휘 사전을 구축하기 위하여 21세기 세종 계획의 용언 사전을 기본 사전으로 활용하였다. 세종 용언 사전에는 총 15,174개의 동사와 1,269 가지의 격틀이 수록되어 있는데, <표 1>은 세종 용언 사전에 나타난 대표적인 격틀과 그 격틀이 나타나는 동사의 빈도를 보여준다. 이 격틀들은 논항의 의미 역과 선택제약을 기준으로 보다 상세히 분류된다.

〈표 1〉 가장 빈번히 나타나는 격틀의 사례

격틀 사례	동사 빈도
X = N0-이 Y = N1-을 V	5339
X = N0-이 V	4425
X = N0-이 Y = N1-에 V	1174
X = N0-이 Y = N1-로 V	663
X = N0-이 Y = N1-을 Z = N2-로 V	629
X = N0-이 Y = N1-에|에게 V	520

3.1.2 동사 유사도 추정

개별 동사들은 자체적으로 가지고 있는 격틀이 매우 제한적이기 때문에 향후 술어-논항 관계를 분석하기에는 많은 제약이 따른다. 따라서 본 논문에서는 [18]에서 제안한 방식으로 격틀을 기반으로 동사간의 유사도를 추정 한 후에, 이 유사도 정보를 기반으로 동사들을 군집화하였다. 본 논문에서는 동일한 군집으로 분류된 동사들은 서로가 가지고 있는 모든 격틀을 공유하게 함으로써 격틀의 부족 현상을 극복하고자 하였다.

각 동사 v_i는 다음과 같이 벡터의 형태로 표현된다.

$$v_i = <sim_{i1}, sim_{i2}, \cdots, sim_{in}>$$

여기서 n은 동사의 전체집합의 크기로써 본 논문에서는 15,174가 된다. 그리고 sim_{ij}는 v_i와 v_j의 격틀 기반 유사도로써 다음과 같이 계산된다. 세종 용언 사전의 격틀 정보는 말뭉치에 출현한 내용을 기반으로 구축되었기 때문에 용언이 자주 나타나는 경우는 격틀 정보가 매우 상세히 구축되었으나, 그렇지 않은 경우에는 많은 격틀 정보가 생략되어 있다. 따라서 말뭉치의 출현 빈도에 따른 격틀 정보의 불균형을 고려하여 다음과 같은 방법으로 sim을 계산하였다.

$$\mathrm{sim}_{ij} = \frac{\text{두 동사가 동시에 보유하고 있는 격틀의 수}}{\text{두 동사 중에서 격틀의 개수가 작은 동사의 격틀 수}} \quad (1)$$

모든 동사마다 벡터로 표현이 가능해지면, 벡터간의 유사도를 아래와 같이 계산한다.

$$\text{유사도}(\vec{x},\vec{y}) = \frac{\vec{x}\bullet\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}x_i^2} \times \sqrt{\sum_{i=1}^{n}y_i^2}} \quad (2)$$

<표 2>는 이와 같은 방식으로 동사 “가꾸다”와 다른 동사간의 유사도를 계산한 결과를 보여준다.

〈표 2〉 “가꾸다 동사와 다른 동사간의 격틀 유사도

동사	가감하다	가공하다	가꾸다	가누다	가다
유사도	0.0272	0.0038	1.0000	0.8731	0.1989
동사	가다듬다	가동하다	가두다	가득차다	가라앉다
유사도	0.7756	0.0783	0.0824	0.0334	0.0667

3.1.3 동사 클러스터 구축

본 논문에서는 클러스터링을 위하여 EM (Expectation-Maximization) 알고리즘[19]과 k-means 알고리즘[20]을 활용하였다. 본 논문에서는 WEKA 패키지[2]를 이용하였고, 하나의 클러스터에 지나치게 많은 동사 어휘가 포함되는 현상을 극복하기 위하여 지나치게 많은 원소를 포함하는 클러스터는 재군집화를 하였다.

본 논문에서는 세종 용언 사전에 등록되어 있는 15,174개의 동사 어휘들을 대상으로 군집화를 시도하였으나, 격틀 정보가 매우 빈약하여 활용도가 전혀 없는 일부 동사를 제외한 14,023개의 동사 어휘들만을 군집화 하였다. 모든 동사들은 3.1.2에서 기술된 방식처럼 벡터로 표현되는데, 하나의 동사 어휘는 14,023개의 원소를 가지는 벡터로 변환된다. 그런데 이러한 차원은 크기가 지나치게 크기 때문에 클러스터링에 직접 활용하기 어렵다. 따라서 본 논문에서는 Singular Value Decomposition(SVD) 알고리즘[21]를 활용하여 차원을 45개로 축소시켰다. SVD를 통하여 차원이 축소된 동사 중에서 원소가 모두 0인 벡터를 제외하면 12,896개의 동사만 남게 되었고 이들이 군집화 되었다.

2) http:// www.cs.waikato.ac.nz/ml/weka/

〈표 3〉 군집 사례

번호	군집에 포함된 동사 리스트	개수
0	가결되다, 가결의되다, 각색되다, 각하되다, 개괄되다, 개신되다, 갱신되다 등	96
1	과대평가하다, 금하다, 내세우다, 두다, 막다, 밝혀내다, 정죄하다	7
2	개별화되다, 객관화되다, 공인받다, 과잉되다, 과잉하다, 관례화되다 등	89
3	객관화시키다, 객관화하다, 논급하다, 대상화하다, 미분화하다, 반감시키다 등	49
4	맞닥뜨리다, 미워하다, 불신하다, 샘하다, 생별하다, 생이별하다, 시큰대다 등	13
5	꿰어차다, 놔두다, 멸균시키다, 산개시키다, 살균하다, 엎어놓다, 처치하다	7
6	간취하다, 감득하다, 감지덕지하다, 관전하다, 되씹다, 맞바라보다, 맞보다 등	10
7	감손되다, 감점되다, 균일화되다, 급등하다, 급락하다, 누진되다, 미분되다 등	18
8	가시화하다, 만족시키다, 수그리다, 일축하다, 주체하다, 진척하다, 질책하다 등	12
9	손대다, 쉬쉬하다	2

본 논문에서는 최초 클러스터링으로 k-means 알고리즘을 활용하였다. 이 때, 최초 클러스터의 개수는 500개로 지정하였다. 이 중 원소가 하나도 없는 3개의 클러스터를 제외한 497개의 클러스터 중에서 원소가 100개 이상 포함하고 있는 대형 클러스터가 13개 발생하였다. 특히 어떤 클러스터는 총 2,185개의 동사 원소를 포함하고 있었는데 이는 전체 동사의 15%를 차지하는 매우 큰 클러스터이다. 하나의 클러스터가 이렇게 큰 비중을 차지하게 되면 해당 클러스터는 거의 모든 종류의 격틀을 포함하게되어 사실상 의미 어휘 사전으로써의 역할을 하기 어렵다. 따라서 본 논문에서는 이렇게 많은 원소들을 포함하고 있는 대형 클러스터들을 다시 클러스터링하여 보다 세밀히 클러스터를 나누었다.

본 논문에서는 13개의 대형 클러스터 중에서 3개의 클러스터를 재클러스터일 하였다. 2,185개의 원소를 포함하는 클러스터는 40개의 초기 클러스터 개수를 초기값으로 하여 k-means 알고리즘을 사용하였다. 그러나 이 중 1,731개의 원소를 포함하는 클러스터가 발생하여 다시 k=20으로 하여 클러스터링을 하였다. 결과적으로 최초 2,185개의 원소를 갖는 클러스터는 59개의 새로운 클러스터로 분할되었다.

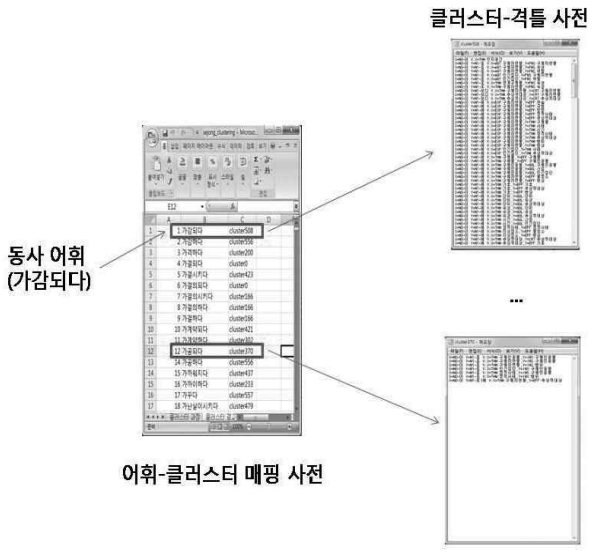
또한 다른 두 개의 재클러스터링 대상 클러스터는 EM 알고리즘으로 재클러스터링을 시도하였다. 여기서 EM 알고리즘을 활용한 이유는, EM의 경우에는 초기 클러스터의 개수를 지정하지 않아도 되며, 동시에 비교적 균등한 숫자로 각 클러스터에 원소들을 할당하는 성격이 더 강하기 때문이다. 그러나 최초에 이를 사용하게 되면 원소의 수가 너무 많아 사실상 실행이 되지 않았다. 결국 이 두 클러스터는 각각 3개의 새로운 클러스터로 재클러스터링되었다.

이러한 과정을 거치면서 군집의 수는 재군집화 과정을 통하여 559개로 결정되었고, 의미 어휘 사전은 총 559개의 표제어로 구성되었다. <표 3>은 이러한 과정을 통하여 구축된 군집 중에서 최초 10개를 보여준다.

3.1.4 의미 어휘 사전 구성

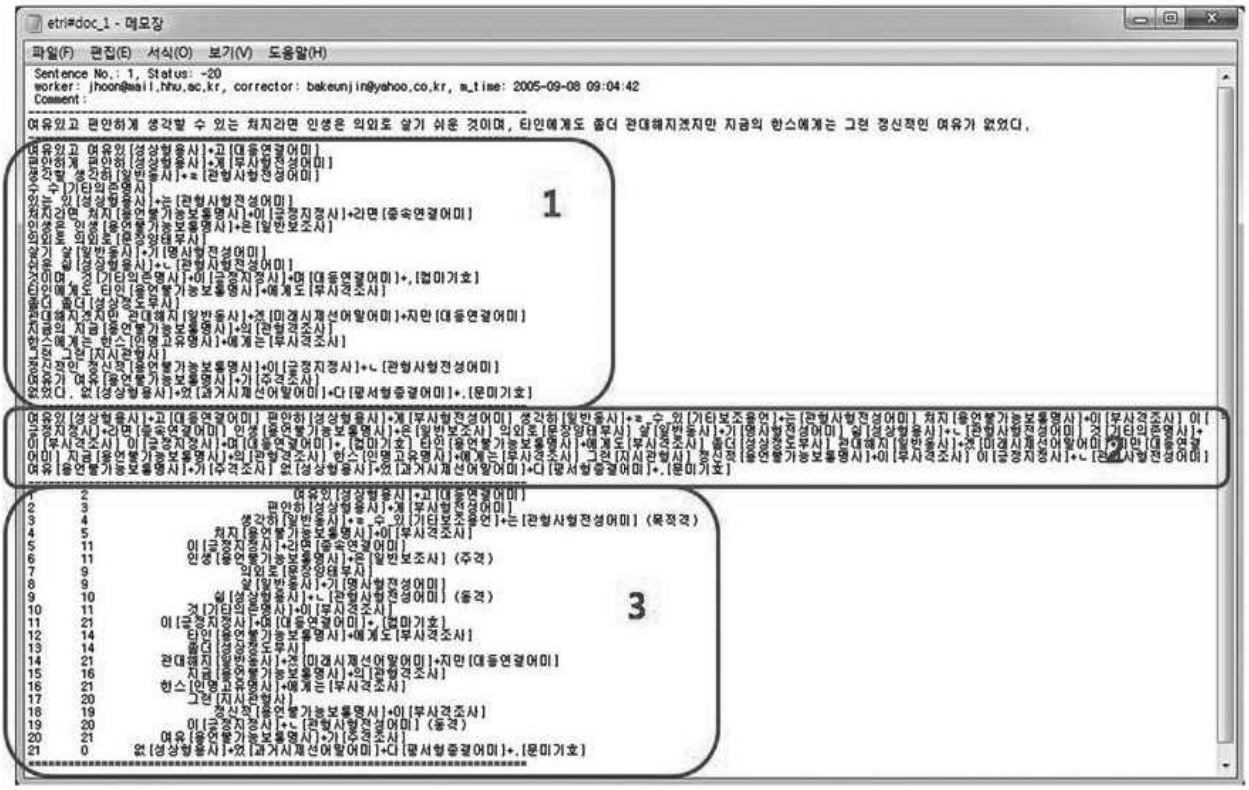
본 논문에서 사용된 의미 어휘 사전은 크게 다음과 같이 두 가지로 구성된다. 첫째는 주어진 단어의 소속 클러스터를 찾는 부분이고, 둘째는 첫 번째 단계에서 검색된 클러스터 번호로 해당 클러스터가 가지고 있는 격틀 정보를 가지고 오는 부분이다. 의미 어휘 사전의 전체 구성은 (그림 2)와 같다.

(그림 2)의 왼쪽에 위치하는 어휘-클러스터 매핑 사전은 말뭉치에서 추출된 동사의 원형을 색인어로하여 해당 동사가 포함되어 있는 클러스터의 번호를 찾는 테이블이다. 이 테이블에는 동사-클러스터 쌍의 정보를 가지고 있다. (그림 2)의 오른편에 위치하는 클러스터-격틀 사전은 각 클러스터별로 보유하고 있는 격틀 정보를 가지고 있다. 클러스터-격



(그림 2) 의미 어휘 사전의 전체 구성

(그림 3) 하나의 클러스터에 포함된 동사 어휘의 격틀 정보



(그림 4) ETRI 구문 표지 부착 말뭉치

틀 사전은 하나의 클러스터에 포함된 모든 동사들이 가지고 있는 격틀 정보들이 모두 합집합 연산되어 하나의 격틀 정보 집합으로 저장된다. 아래 (그림 3)은 '과대평가하다', '금하다', '내세우다', '두다', '막다', '밝혀내다', '정죄하다' 등 총 7개의 동사의 격틀 정보를 모두 합한 격틀 정보를 보여준다.

3.2 술어-논항 관계 추출기

이 절에서는 주어진 구문 표지 부착 문장을 입력으로 받아 술어(동사 술어)와 관련 논항을 추출하고, 이 중 부사격 조사와 결합이 된 논항의 술어와의 의미 관계를 분석하는 모듈에 대하여 논한다. 이 모듈은 첫째, 술어-논항 쌍을 추출하고, 둘째, 논항을 향후 처리를 위하여 변형시키고, 셋째, 의미 어휘 사전을 검색하고 넷째, 적절한 격틀을 찾고 마지막으로 위에서 분석되지 않은 논항을 추가적으로 처리하는 과정으로 이루어진다.

3.2.1 술어-논항 쌍의 추출

본 논문에서는 구문 표지 말뭉치로 ETRI의 구문 표지 부

착 말뭉치[3]를 사용하였다. 다음 (그림 4)는 ETRI의 구문 표지 부착 말뭉치의 사례를 보여준다.

ETRI의 구문 표지 부착 말뭉치는 하나의 문장별로 원문, 개별 어절들의 형태소 분석 결과(영역 1), 문장 형태의 형태소 분석 결과(영역 2), 그리고 의존 관계 분석 결과(영역 3)으로 구성된다. 그림을 보면14번 라인의 어절 "관대해지[일반동사] + 겠[미래시제선어말어미] 지만[대등연결어미]"는 21번 라인의 어절 "없[성상형용사] + 었[과거시제선어말어미] + 다[평서형종결어미] + .[문미기호]"에 의존한다는 것을 보여준다. 마찬가지로 14번 라인의 어절은 12번 라인의 어절 "타인[용언불가능보통명사' + 에게도[부사격조사]"와 13번 라인의 어절 "좀더[성상정도부사]"의 수식을 받는다. 이와 같은 의존 관계를 분석하면 실제 의미 역 결정이 필요한 술어-논항의 쌍을 찾을 수 있는데, 이 경우에는 술어 "관대해지다"에 의존하는 여러 성분 중에서 부사격조사로 연결된 "타인에게도"가 하나의 쌍을 이루어 의미 관계 분석의 대상이 된다.

ETRI 구문 표지 부착 말뭉치에는 총 37,226개의 문장이 있으며, 본 논문에서는 술어-논항(부사격 조사로 연결된)으로 자동 분석의 대상을 제한하였는데, 약 185,538개의 술어-논항 쌍을 추출하였다.

이렇게 추출된 술어-논항 쌍은 3.1에서 구축된 의미 어휘 사전을 검색하여 논항을 분석하여 술어와의 의미 관계를 파악하게 된다.

3.2.2 논항의 변환

말뭉치에 나타난 논항들은 일반적으로 일반명사 또는 고유명사이다. 그러나 구축된 의미 어휘 사전은 논항이 특정 의미 부류로 표기되었기 때문에 이들 논항들은 의미 부류로 변환되어야 한다. 본 논문에서는 세종 용언 격틀 사전을 사용하였기 때문에 논항들은 모두 세종 의미 체계에서 사용되는 부류로 표기되었다. 본 논문에서는 세종 의미 체계의 상위 2번째 단계의 부류들을 의미 부류로 활용하였는데 <표 4>에서 이들 부류들이 나열되어 있다.

예를 들어 "인류", "단체", "민중" 등의 어휘들은 '집단'부류의 하위 부류인 '인간집단' 부류에 속하게 되고 이들 어휘들이 논항에 나타나게 되면 해당 부류로 변환된다. 그러나 말뭉치의 논항에 나타나는 모든 어휘들이 세종 체언에 포함되는 것은 아니기 때문에 이들 미등록 어휘들은 다양한 방식으로 의미 부류를 결정해야 한다. 이 문제는 본 논문의 범위에 벗어나므로 여기서는 논하지 않는다.

3.2.3 의미 어휘 사전 검색

이 부분은 이미 3.1절에서 상세히 설명하였다. 따라서 여기서는 생략하도록 한다.

3.2.4 격틀 매핑 작업

술어-논항 쌍이 해당 클러스터를 의미 어휘 사전에서 찾게 되면, 그 다음에는 보유하고 있는 조사와 논항 정보를 가지고 일치하는 격틀이 있는 지 확인해야 한다. 만일 입력된 술어-논항 쌍이 가지고 있는 모든 조사 및 논항과 일치하는 격틀이 있으면 해당 격틀 정보에 명시되어 있는 의미 정보를 찾고 그 의미 정보를 술어-논항 쌍의 의미 분석 결과로 결정한다.

3.2.5 논항을 제외한 격틀 매핑 작업

이 과정은 클러스터 격틀 사전에 입력된 술어-논항과 일치하는 격틀 정보가 없는 경우에 하는 작업이다. 이 과정에서는 격틀 사전에 수록되어 있는 격틀 정보에서 논항의 부류 정보를 무시하고 오직 조사만을 고려하여 일치하면 주어진 술어-논항 쌍과 일치하는 격틀이 있는 것으로 여기는 것이다.

4. 성능 및 평가

본 논문에서 제안한 자동 술어-논항 분석기의 성능을 분석하기 위하여 추출된 술어-논항 관계의 정확도를 계산하였다. 특히 본 논문에서 제안된 클러스터링에 기반한 의미 어휘 사전의 역할 분석에 초점을 맞추었다. 본 실험에서는 크게 적용력과 정확도의 두 가지 기준으로 성능을 측정하였다. 여기서 적용력과 정확도는 각각 다음과 같이 정의된다.

〈표 4〉 세종 의미 체계의 상위 1,2번째 부류

구체물	구체자연물,구체인공물,속성구체물,관계구체물
집단	인간집단,비인간집단
장소	지상장소,물장소,공중장소,상상적장소,건물,길,다리,굴,부분장소,관계장소,자리좌석,경계,지역,속성기능장소
추상적대상	금전,시간,방법,기술,역할,범주,속성,단위,방향,사실명제,기호,기호체계,자연법칙,규범,관습,권력,권리,의무,학문과목,제도,종교,사조,예술텍스트,작품,방송물,산업,역사,상,별,분야,범위,경로,추상적부분,관계추상대상,개념,기준,상황,개체상황,영상,수학적대상,물리학적대상,언지공간,추상적장애물
사태	정적사태,행위,사건,현상,상태변화

3) http://slrdb.etri.re.kr/db/form.asp?leftNum=3

$$\text{적용력} = \frac{\text{적용된(술어,논항)튜플의 개수}}{\text{전체 테스트(술어,논항)튜플의 개수}}$$

$$\text{정확도} = \frac{\text{올바르게 결정된 의미 역의 개수}}{\text{적용된(술어,논항)튜플의 개수}}$$

적용력이란 실험 대상이 되는 전체 테스트 (술어,논항) 튜플 중에서 실제 의미 어휘 사전에서 격틀을 찾아 의미 관계를 추출하고자 시도한 (술어,논항) 튜플의 비율을 말한다. 즉 이 수치가 높다는 것은 구축된 의미 어휘 사전의 적용 범위가 넓어 의미 관계 추출을 시도한 술어-논항의 쌍이 많다는 것을 말하며, 제안한 클러스터링을 통한 의미 어휘 사전의 구축이 처리 범위를 얼마나 확장시켰는가를 확인할 수 있다. 또한 정확도란 의미 어휘 사전을 활용하여 의미 관계 추출을 시도한 (술어,논항) 튜플 중에서 올바르게 의미 관계가 추출된 쌍의 비율을 말한다. 즉 이 수치가 높으면 그만큼 클러스터링을 통하여 구축된 의미 어휘 사전의 정확도가 높다는 것을 말한다.

본 논문에서는 지금까지 구축한 시스템이 얼마나 높은 적용력과 정확도를 가지고 있는 지 확인하기 위하여 보유하고 있는 술어-논항 쌍 중에서 임의로 1,000개의 쌍을 추출하였다. 이 때, 논항은 의미 부류가 결정된 경우로 제한하였다. 또한 성능의 비교를 위하여 다음과 같은 세 가지 방식의 실험을 하여 그 성능을 비교하였다.

첫째, 개별 동사 어휘의 격틀 사전으로 성능을 측정하였다. 이는 기존에 구축된 격틀 사전만을 가지고 의미 역 결정 성능을 기준 성능으로 하여 본 논문에서 제안된 방법이 성능 향상에 도움이 되는 정도를 확인하기 위함이다(실험1). 둘째, 클러스터 격틀 사전을 이용하되, 논항 및 조사의 엄격한 비교를 요구하는 방식이다(실험 2). 셋째, 클러스터 격틀 사전을 이용하되, 만일 주어진 술어-논항 쌍에 적용될 격틀이 사전에 존재하지 않는다면 논항의 의미 부류 정보는 무시하고 조사만 일치하면 무조건 해당 격틀을 적용하는 방식이다(실험 3).

위의 세 가지 실험의 결과는 아래 <표 5>에서 요약된다.

〈표 5〉 실험 결과 요약

	실험 1	실험 2	실험 3
적 용 력(%)	10.9	24.8	62.5
정 확 도(%)	60.6	55.6	50.7

<표 3>의 결과를 보면 실험 1이 가장 높은 정확도를 보이고 있으나 적용력은 불과 10%이다. 이는 개별 어휘의 격틀 사전이 실제 적용시 그 범위가 얼마나 적은 지를 알 수 있다. 반면에 실험3으로 확장하면서 그 적용력은 60%를 상회하는 것을 알 수 있다. 특히 적용력이 6배 가까이 늘어남에도 불구하고 정확도는 불과 10% 정도밖에 줄어들지 않았다. 이는 본 논문의 기본 방향인 동사 클러스터링에 기반한 의미 어휘 사전의 구축이 올바른 방향성을 가지고 있음을 보여준다. 따라서 앞으로는 보다 적용력을 높이고 반면에 정확도는 떨어지지 않는 방법을 추가하여 전체적으로 거의 대다수의 술어-논항 쌍의 의미 관계 추출이 될 수 있어야 할 것이다. 또한 정확하게 의미 역이 결정되지 못한 데이터를 중점적으로 분석하여 정확도를 향상시킬 수 있는 방법을 찾아야 할 것이다.

실제 본 논문에서는 적용력이 최대 62.5%까지 밖에 나오지 못했다. 이는 클러스터링을 하더라도 실제 문장에 나타난 격틀 구조를 미리 구축한 사전으로 처리하는데 한계를 보인다는 것을 말한다. 따라서 사전의 지속적인 확대 뿐만 아니라 다양한 기계 학습 방법론 및 통계적인 방법론이 추가적으로 고려되어야 할 것이다.

또한 매핑이 이루어진 격틀 역시 오류를 가지고 있을 수 있는데, 이는 실제 문장에서 사용된 격틀은 생각보다 매우 다양하기 때문이다. 이를 극복하기 위해서는 매핑이 이루어진 상황에서도 이를 보다 유연하게 활용할 수 있는 방법을 찾아야 할 것이다.

5. 결 론

본 논문은 향후 한국어 PropBank의 구축을 위한 전 단계로 한국어 자동 술어-논항 분석기의 구현을 그 목적으로 한다. 자동 술어-논항 분석기란 결국 술어에 의존하고 있는 여러 논항들과 술어간의 의미 관계를 분석하는 모듈이라 할 수 있는데, 본 논문에서는 특히 부사격 조사로 의존 관계를 맺고 있는 논항들의 의미 관계 분석에 초점을 맞추었다.

이러한 자동 술어-논항 분석기의 구현을 위하여 본 논문에서는 크게 의미 어휘 사전과 술어-논항 추출기 모듈을 개발하였다. 먼저 의미 어휘 사전은 개별 동사 격틀 사전으로는 그 적용 범위가 너무 작기 때문에 보유하고 있는 격틀이 유사한 동사 어휘들을 k-means 및 EM 알고리즘을 사용하여 클러스터링을 하고 동일 클러스터로 모인 동사들의 격틀을 모두 통합하여 하나의 사전을 만들었다. 따라서 향후 특정 동사 어휘로 이 사전에 접근하게 되면 원래 이 동사가 가지고 있던 격틀 정보뿐만 아니라, 동일 클러스터에 포함되어 있는 여타 동사들의 격틀 정보도 함께 사용할 수 있다.

둘째는 술어-논항 추출기인데, 이 모듈은 실제 구문 표지 부착 말뭉치에서 분석 대상이 되는 술어-논항 쌍을 추출하고, 위에서 구축된 의미 어휘 사전을 활용하여 논항의 의미 관계를 분석하는 모듈이다. 이를 위해서는 문장에서 술어-논항 쌍을 추출하는 과정은 물론이고, 논항 어휘의 의미 부류를 결정해 주는 과정이 필요하다. 의미 부류는 의미 어휘 사전에 기술되어 있는 논항의 의미 부류와 주어진 술어-논항 쌍에서의 논항의 의미 부류를 일치시키기 위하여 사용된다. 끝으로 주어진 술어-논항 쌍은 의미 어휘 사전의 검색을 통하여 논항의 의미 관계를 분석하게 된다.

향후 연구로는 첫째, 거의 프로토타입 수준의 현 시스템의 완성도를 높이는 것이다. 현 시스템은 많은 다양한 모듈과 데이터 사전으로 구성된다. 이들 모듈을 지속적으로 테

스트함으로써 문제가 되는 부분을 고치고, 동시에 사전을 지속적으로 검증하여 문제점을 수정해야 할 것이다. 또한 미등록어의 처리 문제는 매우 중요한 사안이 될 것이다. 미등록어는 결과적으로는 전체 논항의 25%정도 수준이지만 실제로는 형태소 분석 결과를 활용하는 부분도 포함되어야 하기 때문에 그 비중은 매우 높아진다. 따라서 이 부분에 대한 연구는 독립적인 연구로도 진행되어야 한다. 마지막으로 적용력을 높이는 방법을 연구해야 한다. 아무리 정확한 시스템이라 하더라도 적용력이 10% 수준이라면 그 시스템의 실질적인 활용도는 떨어질 수 밖에 없다. 따라서 1차적으로는 적용력을 높이는 방향으로, 그리고 그 다음에 정확도를 높이는 방향으로 시스템 성능 향상과 관련한 연구가 진행될 것이다.

참 고 문 헌

[1] Jurafsky, D. and J.H. Martin, "Speech and Language Processing (2nd Edition)," Prentice Hall, 2008.

[2] 정성원, 권혁철, "자연언어처리를 위한 기계학습," 정보과학회지, 제25권 제3호, pp.57-63, 2007.

[3] 김병수, 이용순, 나승훈, 김병기, 이종혁, "부트스트래핑 알고리즘을 이용한 한국어 격조사의 의미역 결정," 한국정보과학회 2006 한국컴퓨터종합학술대회 논문집(B), pp.4-6, 2006.

[4] 김병수, 이용훈, 이종혁, "비지도 학습을 기반으로 한 한국어 부사격의 의미역 결정," 정보과학회논문지:소프트웨어및응용, 제34권 제2호, pp.112-122, 2007.

[5] Lim, J., Y. Whang, S. Park, and H. Rim, "Semantic Role Labeling using Maximum Entropy Model," Procs. of CoNLL-2004, 2004.

[6] Park, K., Y. Whang, and H. Rim, "Two-Phase Semantic Role Labeling based on Support Vector Machines," Procs. of CoNLL-2004, 2004.

[7] Palmer, M., P. Kingsbury, and D. Gildea, "The Proposition Bank: An Annotated Corpus of Semantic Roles," Computational Linguistics, 31(1), pp.71-106, 2005.

[8] Xue, N., and M. Palmer, "Automatic Semantic Role Labeling for Chinese Verbs," Procs. of International Joint Conference on Artificial Intelligence, 2005.

[9] Kingsbury, P., B. Snyder, N. Xue, and M. Palmer, "PropBank as a Bootstrap for Richer Annotation Schemes," Procs. of sixth Workshop on Interlinguas, Machine Translation Summit IX, 2003.

[10] Johansson, R., and P. Nugues, "Dependency-based Syntactic-Semantic Analysis with PropBank and NomBank," Procs. of CoNLL-2008, 2008.

[11] Giuglea, A., and A. Moschitti, "Knowledge Discovering using FrameNet, VerbNet and PropBank," Workshop on Ontology and Knowledge Discovery at ECML-04, 2004.

[12] Hovy, E., M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel, "OntoNotes: The 90% Solution," Procs of the Human Language Technology Conference of the NAACL, 2006.

[13] Giuglea, A., and A. Moschitti, "Semantic Role Labeling via FrameNet, VerbNet and PropBank," Annual Meeting of Association for Computational Linguistics, 2006.

[14] Xue, N., and M. Palmer, "Annotating the Propositions in the Penn Chinese Treebank," Procs. of the 2nd SIGHAN Workshop on Chinese Language Processing, 2003.

[15] Palmer, M., O. Babko-Malaya, A. Bies, M. Diab, M. Maanouri, A. Mansouri, and W. Zaghouani, "A Pilot Arabic Propbank," Procs. of the 6th International Language Resources and Evaluation (LREC'08), 2008.

[16] Agirre, E., I. Aldezabal, J. Etxeberria, and E. Pociello, "A Preliminary Study for Building the Basque PropBank," Procs. of the 5th International Language Resources and Evaluation (LREC'06), 2006.

[17] Palmer, M., J. Rosenzweig, and S. Cotton, "Automatic Predicate Argument Analysis of the Penn Treebank," Procs. of HLT 2001, First International Conference on Human Language Technology Research, 2001.

[18] 조정현, 정현기, 김유섭, "격틀 구조에 기반한 유사 동사 추출," 제21회 한글 및 한국어 정보처리 학술대회, 2009.

[19] Mustapha, N., M. Jalali, and M. Jalali, "Expectation Maximization Clustering Algorithm for User modeling in Web Usage Mining Systems," European Journal of Scientific Research, Vol.32, No.4, pp.467-476, 2009.

[20] Hartigan, J. A., "Clustering Algorithms," Wiley., 1975.

[21] Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, "Numerical Recipes 3rd Edition: The Art of Scientific Computing," Cambridge University Press, 2007.

조 정 현

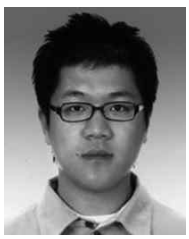
e-mail : showcjh@gmail.com

2008년 한림대학교 컴퓨터공학과(학사)

2010년 한림대학교 컴퓨터공학과(공학석사)

2010년~현 재 한림대학교 컴퓨터공학과 박사과정

관심분야 : 자연언어처리, 정보검색 등



정 현 기

e-mail : oops1212@gmail.com

2007년 한림대학교 컴퓨터공학과(학사)

2009년 한림대학교 컴퓨터공학과(공학석사)

2009년~현 재 한림대학교 컴퓨터공학과 박사과정

2010년~현 재 ㈜GM Creative 연구소팀장

관심분야 : 자연언어처리, 기계번역, mobile web, social web 등

김 유 섭

e-mail : yskim01@hallym.ac.kr

1992년 서강대학교 전자계산 학과(학사)

1994년 서울대학교 컴퓨터공학과 (공학석사)

2000년 서울대학교 컴퓨터공학과 (공학박사)

2002년~현 재 한림대학교 유비쿼터스컴퓨팅학과 교수

관심분야 : 전산금융, 자연언어처리, 기계학습, e-learning 등