DOI: 10.3745/KIPSTB.2010.17B.1.085

부분 구문 분석 결과에 기반한 두 단계 부분 의미 분석 시스템

박 경 미[†] · 문 영 성[††]

요 약

부분 의미 분석 시스템은 문장의 구성 요소들이 술어와 갖는 관계를 분석하는 것으로 문장에서 술어의 주체, 객체, 도구 등을 나타내는 의미 논항을 확인하게 된다. 본 논문에서 개발한 부분 의미 분석 시스템은 두 단계로 구성되어 있는데, 먼저 부분 구문 분석 결과로부터 의미 논항의 경계를 찾는 의미 논항 확인 단계를 수행하고 다음으로 확인된 의미 논항에 적절한 의미역을 부착하는 의미역 할당 단계를 수행한다. 순차적인 두 단계 방법을 적용하는 것에 의해서, 학습 성능 저하의 주요한 원인인 클래스 분포의 불균형 문제를 완화할 수 있고, 각 단계에 적합한 자질을 선별하여 사용할 수 있다. 본 논문에서는 PropBank 말뭉치에 기반한 CoNLL-2004 shared task의 데이터 집합 및 평가 프로그램을 사용하여 각 단계가 시스템의 전체 성능에 기여하는 정도를 보인다.

키워드 : 부분 구문 분석, 부분 의미 분석, 지지 벡터 기계, 의미 논항 확인, 의미역 할당

Two-Phase Shallow Semantic Parsing based on Partial Syntactic Parsing

Kyung-Mi Park[†] · Youngsong Mun[††]

ABSTRACT

A shallow semantic parsing system analyzes the relationship that a syntactic constituent of the sentence has with a predicate. It identifies semantic arguments representing agent, patient, instrument, etc. of the predicate. In this study, we propose a two-phase shallow semantic parsing model which consists of the identification phase and the classification phase. We first find the boundary of semantic arguments from partial syntactic parsing results, and then assign appropriate semantic roles to the identified semantic arguments. By taking the sequential two-phase approach, we can alleviate the unbalanced class distribution problem, and select the features appropriate for each task. Experiments show the relative contribution of each phase on the test data.

Keywords : Partial Syntactic Parsing, Shallow Semantic Parsing, Support Vector Machine, Semantic Argument Identification, Semantic Role Assignment

1. 서 론

일반적으로 자연 언어 처리는 형태소 분석(morphological analysis), 구문 분석(syntactic analysis), 의미 분석(semantic analysis) 등의 순차적인 단계들을 거치게 된다. 이 중 구문 분석 단계에서는 부분 구문 분석(partial syntactic parsing) 또는 완전 구문 분석(full syntactic parsing)이 수행될 수 있다. 부분 구문 분석은 문장을 구성하는 요소들을 분석하기 위하여 문장에서 명사구, 동사구와 같은 기본구를 인식하는 것으로, 문장을 구성하는 요소들 간의 의존 관계를 확인하는 완전 구문 분석에 비해 분석 속도가 빠르다. 또한 문제의 복잡도가 완전 구문 분석에 비해 낮아 더 높은 정확도를 보인다. 예를 들어, 문장 [it has already delivered 793 of the shipsets]에 부분 구문 분석을 적용하면, [it_명사구] [has already delivered_동사구] [793_명사구] [of_전치사구] [the shipsets_명사구]를 확인할 수 있다.

부분 의미 분석(shallow semantic parsing)은 문장에서 특정 동사에 의해서 표현된 명제를 확인하는 것으로 이 과정을 통해 문장 구성 요소들 중에서 해당 동사의 주체(agent), 객체(theme), 도구(instrument) 등의 의미역을 표현

※ 이 논문 또는 저서는 2006년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(KRF-2006-005-J03802).
† 정 회 원 : 숭실대학교 정보미디어기술연구소 전임연구원
†† 정 회 원 : 숭실대학교 컴퓨터학부 교수
논문접수 : 2009년 8월 25일
수 정 일 : 1차 2009년 10월 29일, 2차 2009년 11월 12일
심사완료 : 2009년 11월 26일

하는 구성 요소들이 무엇인지 확인할 수 있다. 예를 들어, 문장 'it has already delivered 793 of the shipsets'에 부분 의미 분석을 적용하면, 'delivered'란 행위의 주체는 'it'이고 객체는 '793 of the shipsets'임을 알 수 있다. 부분 의미 분석은 자연 언어 처리의 다양한 분야에서 활용될 수 있는데 대표적으로 텍스트로부터 개체 간 상호 작용 관계를 인식하는 정보 추출 과제에 적용될 수 있다. 개체명 및 이들 간의 관계에 대한 최신의 정보는 연구 문헌과 같은 텍스트 형태로만 존재하기 때문에, 대량의 텍스트로부터 유의미한 새로운 정보를 자동으로 추출하기 위해서는 부분 의미 분석과 같은 자연 언어를 처리하는 기술이 필요하다. 예를 들어, 생의학 텍스트로부터 문장 'distinct cytokines can activate the same Stat protein'이 주어지고 'cytokines'와 'Stat protein'이 단백질 이름이라는 개체명 인식 결과가 존재할 때, 부분 의미 분석을 통해 'activate'이란 행위의 주체가 'distinct cytokines_명사구'이고 객체가 'the same Stat protein_명사구'임을 확인하면, 두 생의학 개체 'cytokines'와 'Stat protein' 사이에 'activate' 관계가 존재한다는 것을 확인할 수 있다.

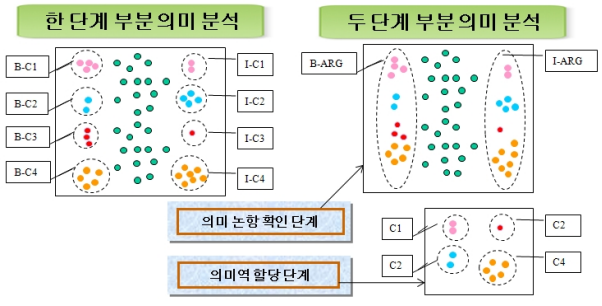
본 논문에서 개발한 부분 의미 분석 시스템은 부분 구문 분석 결과에 기반한다. 기본구 인식 결과를 사용하여 문장 구성 단위를 단어 대신에 명사구, 동사구와 같은 기본구로써 표현하고, 단어가 아닌 기본구 단위로 각 기본구가 의미 논항 구성 요소인지 아닌지를 판별하고자 한다. 또한 부분 의미 분석을 위한 자질로써 주변 문맥을 추출할 때 단어 대신에 각 기본구의 중심어를 사용함으로써, 문장에서 중요한 단어만을 선별하여 사용하고자 한다.

본 논문에서는 지지 벡터 기계(SVM : Support Vector Machine) 기반의 교사 학습(supervised learning) 방법을 사용한다. 지지 벡터 기계는 고차원의 자질 공간에서 높은 일반화 성능을 보이는 것으로 알려진 기계 학습 알고리즘이다. 또한 지지 벡터 기계의 polynomial kernel 함수들은 다양한 자질 조합들을 학습할 수 있게 한다[13]. 그러나 지지 벡터 기계는 이진 분류기이기 때문에 멀티 클래스 분류 과제(multiclass classification task)에 적용하는 경우, 클래스의 수가 증가할수록 더 심화되는 클래스 분포의 불균형 문제(unbalanced class distribution problem)에 직면하게 된다. 부분 의미 분석은 주로 동사에 의해 표현되는 행위의 주체, 객체, 도구 등의 다양한 의미역 클래스가 존재하는 멀티 클래스 분류 과제이기 때문에, 지지 벡터 기계를 적용하기 위해서는 클래스 분포의 불균형 문제를 완화할 수 있는 방법을 찾는 것이 필요하다.

개념적으로 부분 의미 분석은 두 가지 하위 과제로 나누어 질 수 있다. 주어진 문장에서 의미 논항의 경계를 찾는 확인 단계와 확인된 논항의 의미역을 결정하는 할당 단계이다. 부분 의미 분석이 의미 논항 확인(semantic argument identification)과 의미역 할당(semantic role assignment)의 두 단계를 거치면, 첫 단계에서 클래스의 수를 줄일 수 있기 때문에, 지지 벡터 기계를 사용하면서도 클래스 분포의 불균형 문제를 완화할 수 있다. 즉 의미역의 종류가 증가할수록 부정 학습 예제의 수에 비해 각 의미역의 긍정 학습 예제의 수가 현저히 적어지기 때문에 본 논문에서는 부분 구문 분석을 두 단계로 구분하여 의미 논항 확인 단계에서는 의미역의 개수와 상관없이 3개(B-ARG, I-ARG, O)의 SVM 분류기만을 학습하고자 한다. 여기서 B-ARG 클래스는 의미 논항을 구성하는 첫 번째 요소를 의미하고, I-ARG 클래스는 의미 논항을 구성하는 요소이면서 시작 위치에 있지 않다는 것을 나타낸다. O 클래스는 의미 논항을 구성하지 않는 요소를 의미한다.

(그림 1)은 4개의 의미역(C1, C2, C3, C4)이 존재할 때 클래스 분포의 불균형 문제의 예를 보인다. 점선에 포함된 요소들은 각 클래스에 속하는 긍정 학습 예제를 나타내고, 점선에 포함되지 않은 요소들은 부정 학습 예제를 나타낸다. 두 단계 부분 의미 분석의 의미 논항 확인 단계에서는 긍정 학습 예제들이 논항을 구성하는 첫 번째 요소 B-ARG 클래스와 논항을 구성하면서 시작 위치에 있지 않은 요소 I-ARG 클래스로 구분되기 때문에 한 단계 방법에 비해 클래스 분포의 불균형이 완화됨을 알 수 있다. 다음으로 의미역 할당 단계에서는 논항으로 확인된 요소들만을 대상으로 하기 때문에 부정 학습 예제들은 제외되고 각 의미역에 속하는 긍정 학습 예제들을 대상으로 분류기를 생성하게 된다. 첫 번째 단계를 수행한 후 논항이 아닌 문장 구성 요소들은 고려되지 않기 때문에 첫 단계에 비해 두 번째 단계는 학습 비용을 현저히 줄일 수 있다.

문장에서 의미 논항의 경계를 확인하는데 중요한 자질들과 적절한 의미역으로 분류하는데 중요한 자질들이 다르기 때문에, 두 단계로 구분하면 각 단계에 적절한 자질 집합을 구성할 수 있다. 본 연구는 두 단계 방법을 적용하여 각 단계에 적절한 다른 자질 집합을 구성한다. 의미 논항 확인 단계에서는 대상 기본구를 논항 구성 기본구와 논항 비구성 기본구로 구별하기 위해 문장 구성 요소와 술어 사이의 구문적 의존성을 발견하는데 초점을 맞춘다. 해당 술어와 의존 관계를 갖는 대부분의 논항 구성 기본구들은 술어로부터 근거리 내에서 발생하고 술어로부터 기본구까지의 구문 경로의 종류가 제한적이라는 특징이 있기 때문이다. 의미역 할당 단계에서는 인식된 의미 논항들을 행위의 주체, 객체, 도구 등의 클래스 중 하나로 분류하기 위해 논항 구성 기본구로부터 추출한 어휘 자질에 초점을 맞춘다. 클래스 부착의 단서는 술어와 논항을 구성하는 기본구의 중심어이기 때



(그림 1) 클래스 분포의 불균형의 예

문이다.

본 논문의 구성은 다음과 같다. 2장에서는 부분 의미 분석과 관련된 기존의 연구들을 살펴본다. 3장에서는 두 단계 부분 의미 분석 시스템에 대해 기술한다. 부분 구문 분석과 같은 자연 언어를 처리하는 기술을 적용하여 추출한 어휘 정보와 같은 자질들에 대해 설명한다. 4장에서는 여러 실험을 통해 제안한 방법의 도입이 부분 의미 분석에 유용한지를 검증한다. 끝으로 5장에서 본 연구를 통해 얻은 결론과 연구의 기여에 대해 기술한다.

2. 관련 연구

이 장에서는 부분 의미 분석 과제에 대한 대표적인 연구들을 소개하고자 한다. 각 연구들은 다양한 학습 방법 및 자료 부족 문제를 완화할 수 있는 새로운 자질들을 제안하였다. 대부분의 연구들이 자연 언어 처리의 순차적인 단계에 따라 구문 분석 결과에 기반하여 수행되었다. 부분 의미 분석 자체에 대한 연구뿐만 아니라 자연 언어 처리의 응용 분야에 부분 의미 분석 결과를 사용하는 것이 효과적임을 보인 연구들도 있다. 예를 들어, 부분 의미 분석을 일반 도메인에 대한 정보 추출에 적용하여 술어-논항 구조에 기반한 효과적인 정보 추출이 가능함을 보인 연구도 있었다 [11].

(D. Gildea and D. Jurafsky, 2002)는 확률 모형을 사용하여 부분 의미 분석을 수행하였다 [4]. 실험 데이터로 FrameNet[1]을 사용하였고, 자료 부족 문제로 인해 복잡한 평탄화(smoothing) 과정을 수행하였다. 이 연구에서는 각 문장 구성 요소의 중심어와 술어의 공기 빈도가 다른 자질들에 비해 높은 식별력을 가짐을 보였다. 그러나 자료 부족 문제로 인해 각 구성 요소의 중심어와 술어의 공기 빈도는 0인 경우가 많았다. 이 문제에 초점을 맞추어, 각 문장 구성 요소의 중심어 정보를 일반화하는 3가지 방법을 제안하였다. 첫 번째 방법은 같은 동사와 자주 발생하는 명사들은 의미적으로 유사하다는 가정을 사용하여 대량의 말뭉치로부터 명사의 자동 클러스터링을 수행하였다. 두 번째 방법은 의미적 계층을 표현하기 위해 WordNet을 사용하였다. 세 번째 방법은 태그 정보가 부착되지 않은 원시 데이터로부터 bootstrapping을 수행하였다.

(S. Pradhan et. al., 2003)은 지지 벡터 기계를 사용하여 부분 의미 분석을 수행하였다 [7]. (D. Gildea and D. Jurafsky, 2002)가 각 문장 구성 요소의 중심어 정보를 일반화하기 위해 명사 클러스터링을 제안하였는데 이 연구에서는 유사한 의미를 갖는 동사들은 유사한 직접 목적어를 갖는 경향이 있다는 직관을 이용하여 동사 클러스터링을 수행하였다. 또한 동일한 구 레이블이 반복될 때는 하나로 압축한다는 규칙을 이용하여 트리 경로 자질을 일반화하였다.

현재 가장 좋은 성능을 보이는 (S. Pradhan et. al., 2005)는 부분 의미 분석의 첫 단계인 논항 확인 단계의 성능을 높여 오류 전파(error propagation)를 줄이고자 하였다 [8]. 논항 확인 단계에서는 구문 분석기의 오류로 인해 구문 분석 결과로부터 추출한 구문 구성 요소가 논항 구성 단어와 논항 비구성 단어를 동시에 포함하는 경계의 오류가 많이 발생하였다. 기존의 시스템들은 하나의 구문 분석기로부터 구문 구성 요소를 추출하여 해당 술어의 논항인지 여부를 판별하는데, 구문 분석기가 기본구의 경계를 잘못 인식하면 부분 의미 분석 단계로 오류가 전파된다. 이 연구에서는 Charniak의 구문 분석기, 의존 분석기 Minipar, 기본구 인식기와 같은 3가지 시스템에 기반하여 다양한 구문 구성 요소들을 추출함으로써 구문 분석 단계에서 이미 정답 논항이 배제되는 논항 인식의 오류를 줄이고자 하였다.

(N. Xue and M. Palmer, 2004)는 최대 엔트로피 (ME : Maximum Entropy) 모형에 기반하여 부분 의미 분석을 수행하였고 식별력이 있는 새로운 자질을 제안하는데 연구의 초점을 맞추었다 [12]. 지지 벡터 기계에 비해 학습 속도가 빠른 최대 엔트로피 분류기와 상대적으로 적은 수의 자질만으로도 (S. Pradhan et. al., 2003) 시스템에 견줄만한 성능을 보였다. 말뭉치의 정답 구문 트리를 이용해 성능을 측정하였는데, 자동 구문 분석기를 사용했을 때도 같은 효과를 얻는지는 실험하지 않아 구문 분석기로부터의 오류 전파에 견고한지 여부는 보이지 않았다.

전처리 단계에서 논항 후보의 수를 크게 줄이어 학습 비용을 감소시킬 수 있는 효과적인 방법을 제안한 (N. Kwon et. al, 2004)는 부분 의미 분석을 위해 문장 segment의 개념을 제시하였다 [6]. 이것은 구문 트리에서 술어를 포함하지 않은 가장 상위의 구문 구성 요소를 의미하고, 평면 표현(flat representation) 상태에서 부분 의미 분석을 수행하게 한다. 문장 segment 확인, 논항 확인, 논항 분류의 각 단계에서 Viterbi 탐색을 통해 n-best를 유지하고 최종적으로 재순위화(re-ranking) 방법을 통해 하나의 최종 결과를 선택하였다.

3. 기본구 기반 두 단계 부분 의미 분석

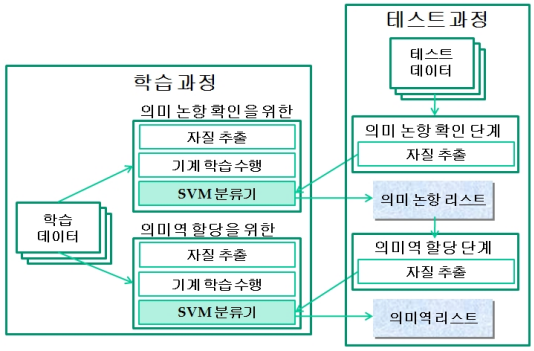
본 논문에서 개발한 부분 의미 분석 시스템은 문장 구성 요소들 중에서 술어와 의존 관계를 갖는 의미적 논항을 찾고, 그 논항의 의미적 역할이 주체인지 객체인지 혹은 다른 의미적 역할을 하는지 분석하는 것으로, 의미 논항 확인 단계와 의미역 할당 단계로 구성되며 전체적인 구조는 (그림 2)와 같다.

<표 1>은 부분 의미 분석 과정의 예로써 각 행이 의미하는 바는 다음과 같다. 첫 번째 행은 주어진 문장을 나타내고 두 번째 행은 술어(P : Predicate)에 대해 문장 구성 요소들이 명사구, 동사구와 같은 기본구 단위로 논항 후보(C : Candidate)를 구성하는 것을 나타낸다. 단어 단위로 논항 후보를 생성할 때보다 후보의 수를 줄여 학습 비용을 감소시킬 수 있다. 반면에 기본구 인식 결과의 오류가 전파된다는 문제점을 가지고 있다. [Rockwell said]처럼 술어를 포함하

1) http://framenet.icsi.berkeley.edu/

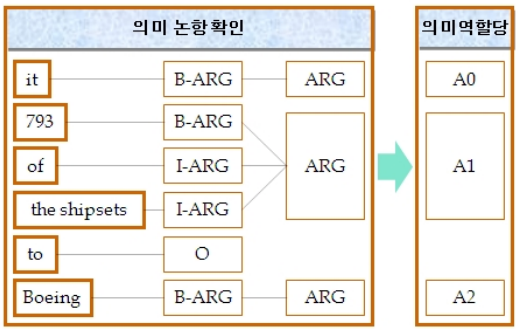
〈표 1〉 부분 의미 분석 과정

Under	the existing contract	,	Rockwell said	,	it	has already delivered	793	of	the shipsets	to	Boeing	.
C	C	C	C	C	C	P	C	C	C	C	C	C
B-ARG	I-ARG	O	O	O	B-ARG	P	B-ARG	I-ARG	I-ARG	O	B-ARG	O
ARG					ARG	P	ARG				ARG	
AM-LOC					A0	P	A1				A2	



(그림 2) 시스템 구성도

지 않는 하위절(subclause)이 존재하는 경우, 하위절을 구성하는 각 기본구를 논항 후보로 하지 않고 하위절 전체를 술어에 대한 논항 후보로 간주한다. 세 번째 행은 의미 논항 확인 단계의 결과이다. SVM 분류기를 통해 각 논항 후보에 B-ARG, I-ARG 또는 O 클래스 중 하나를 할당한다. 이 과정을 통해 네 번째 행처럼 각 의미 논항(ARG : ARGument)들의 경계를 확인할 수 있다. 예문에서는 [Under the existing contract] [it] [793 of the shipsets] [Boeing]와 같은 의미 논항이 존재한다. 마지막 다섯 번째 행은 SVM 분류기를 통한 의미역 할당 단계의 결과이다. AM-LOC, A0, A1, A2는 본 연구에서 실험 말뭉치로 사용한 PropBank[2]에 정의되어 있는 의미역 이름으로써 각각 [Under the existing contract]이 위치 정보이고, [it]이 배달하는(deliver) 행위의 주체이고, [793 of the shipsets]이 배달되는 대상이고, [Boeing]이 배달을 받는 대상임을 나타낸다. 이 과정에 대한 이해를 돕기 위해 (그림 3)을 추가하였다.



(그림 3) 부분 의미 분석 과정의 예

2) http://verbs.colorado.edu/~mpalmer/projects/ace.html

3.1 의미 논항 확인 단계

이 단계는 논항이 기본구의 열로 구성된다고 가정하여, 부분 구문 분석 결과로부터 추출한 명사구, 동사구 등의 기본구 단위로 각각이 논항을 구성하는 요소인지 아닌지 확인한다. 주어진 술어의 논항을 확인하기 위해서는 각 기본구와 술어 사이의 의존 관계를 확인하는 것이 필요하다. 본 연구에서는 술어와 대상 기본구의 거리, 경로와 같은 부분 구문 분석 결과로부터 추출한 구문적 자질들을 활용해 술어와 각 기본구 간의 의존 관계를 확인하고자 한다.

의미 논항 확인 단계에서는 술어와 기본구 사이의 의존 관계를 파악하기 위해 <표 2>와 같은 자질들을 사용한다. 첫 번째 열에 기술된 것처럼 자질들은 발생 위치에 따라 각각 [동사구와 기본구 사이의 정보], [동사구], [동사구-1], [기본구], [기본구-2], [기본구-1], [기본구+1]로 구분된다. 여기서 -와 +는 동사구 또는 기본구를 기준으로 위치 정보를 나타내는 것으로 -는 왼쪽 문맥을 +는 오른쪽 문맥을 상징한다. 또한 -2, -1, +1 등의 숫자가 의미하는 것은 동사구 또는 기본구를 기준으로 위치 및 거리를 나타내는 것으로 예를 들어 [동사구-1]은 동사구를 기준으로 왼쪽 첫 번째 기본구를 나타내고 [기본구-2]는 기본구를 기준으로 왼쪽 두 번째 기본구를 표시한다. <표 1>의 예에서 기본구 [the shipsets]가 술어 [deliver]의 논항 구성 요소인지 여부를 판별하기 위해 자질을 추출한다고 할 때 [동사구-1]은 기본구 [it]을 가리키고, [기본구-2]는 기본구 [793]을 가리킨다.

<표 2>의 자질 집합에서 먼저 [동사구와 기본구 사이의 정보]에서는 술어와 대상 기본구 사이에 구문적 의존성이 존재하는지 확인하기 위해 위치, 거리 정보 및 둘 사이에서 특정 품사, 기본구, 문장 부호 등이 발생하는지 확인한다. 각

〈표 2〉 자질 집합

자질		자질값의 예
동사구와 기본구 사이의 정보	위치	-2, -1, +1
	거리	0, 1, 2,
	#VP, #NP, #SBAR	0, 1, 2,
	#CC, #쉼표, #콜론	0, 1, 2,
	큰따옴표	-1, 0, 1
	경로	VP-PP-NP,
동사구	중심어, 중심어 품사, 구 유형	
	첫 단어의 품사	MD, TO, VBZ,
동사구-1	중심어, 중심어 품사, 구 유형	
기본구	중심어, 중심어 품사, 구 유형	
기본구-2	중심어, 중심어 품사, 구 유형	
기본구-1	중심어, 중심어 품사, 구 유형	
기본구+1	중심어, 중심어 품사, 구 유형	

자질이 의미하는 바는 다음과 같다.

• **[위치]** 자질은 대상 기본구가 술어의 왼쪽에서 발생하는지 오른쪽에서 발생하는지를 나타낸다. 이 자질은 -2, -1, +1 등의 3가지 값을 갖는데 -1은 왼쪽 발생을 +1은 오른쪽 발생을 나타낸다. 대상 기본구가 술어를 포함하는 가장 짧은 길이의 절의 바깥쪽에 존재하면 -2의 값을 갖는다.

• **[거리]** 자질은 대상 기본구와 술어 사이에 존재하는 기본구의 수를 나타낸다. 이 자질은 0, 1, 2 등의 값을 갖는데 예를 들어 <표 1>에서 대상 기본구가 [the shipsets]이고 술어가 [deliver]일 때 둘 사이의 거리는 기본구 [793], [of]가 존재하기 때문에 2이다. 자질값 0의 의미는 대상 기본구가 술어와 인접한 것이다. 의미없는 자질의 값을 제거하기 위해 거리 자질의 값이 특정값 이상이면 하나의 값으로 대체하는 방법을 적용하였다.

• **[#VP, #NP, #SBAR]** 자질은 대상 기본구와 술어 사이에 존재하는 동사구(VP : Verb Phrase), 명사구(NP : Noun Phrase), 종속 접속사(SBAR)의 수를 나타낸다. 두 기본구 사이에서 발생하는 동사구 등의 수가 증가할수록 의존 관계가 성립할 가능성은 낮아진다.

• **[#CC, #쉼표, #콜론]** 자질은 대상 기본구와 술어 사이에 존재하는 품사인 등위 접속사(CC : Coordinating Conjunction), 쉼표, 콜론의 수를 나타낸다. 이전 자질과 마찬가지로 쉼표 등의 수가 증가할수록 두 기본구 사이에 의존 관계가 존재할 확률이 낮아진다.

• **[큰따옴표]** 자질은 대상 기본구와 술어 사이에 존재하는 앞큰따옴표와 뒤큰따옴표의 수의 차이를 나타낸다. 이 자질은 -1, 0, 1의 3가지 값을 갖는데 -1은 뒤큰따옴표의 수가 앞큰따옴표의 수보다 많다는 것을, 1은 반대로 앞큰따옴표의 수가 뒤큰따옴표의 수보다 많다는 것을 의미한다. 자질값 0이 의미하는 것은 앞큰따옴표의 수와 뒤큰따옴표의 수가 같다는 것이다. 문장은 여러 개의 하위절로 구성되는데 주로 동사에 의해서 구분된다. 이 연구에서는 인용절을 구분하는 큰따옴표 자질을 통해 대상 기본구가 술어의 절 안에 포함되는지 자질로써 표현하고자 한다. 자질값이 -1 또는 1을 갖는다면 대상 기본구가 술어의 절 안에 포함되지 않아 의존 관계가 존재할 가능성이 낮아진다.

• **[경로]** 자질은 술어와 대상 기본구 사이의 구문 경로를 나타낸다. 술어와 기본구 사이에서 하위절이 발생하면 하위절을 구성하는 기본구들의 유형을 표시하지 않고 하나의 절로 나타낸다.

[동사구], [동사구-1], [기본구], [기본구-2], [기본구-1], [기본구+1]에서는 공통적으로 [중심어, 중심어 품사, 구 유형] 정보를 추출하여 사용하였다. 구 안에서 중심어를 찾기 위해 Buchholz가 개발한 규칙을 사용하였다 [1]. 이 규칙은 각 구 안에서 중심어를 찾는 우선 순위를 정해놓은 것이다. 예를 들어 규칙의 예 'ADJP=JJ|RB|VB|IN|UH|FW|RP'이 의미하는 것은 형용사구 ADJP를 구성하는 단어 중 형용사 JJ가 있으면 그것을 중심어로 하고, 형용사 JJ가 없으면 다음의 부사 RB를 중심어로 하고, 부사 RB도 없으면 동사 VB를 중심어로 하는 방식으로 적용된다. 하위절로부터 자질을 추출하는 경우 절 안의 여러 개의 기본구 중에서 첫 번째로 발생한 것을 [구 유형] 자질로 선택하였다. [동사구]에서는 [첫 단어의 품사] 자질을 사용하였는데 이것은 동사구의 구문 중심어의 품사를 나타낸다. 자질값은 [MD TO VB VBD VBG VBN VBP VBZ]이고 동사구의 특성을 표현할 수 있다. 예를 들어 자질값이 TO인 경우 동사가 to 부정사 형태로 발생함을 나타낸다.

<표 3>과 <표 4>의 예제를 통해 자질 추출의 결과를 확인할 수 있다. <표 3>은 본 연구에서 사용한 학습 말뭉치의 예로써 2개의 동사 exist와 deliver를 포함하는 하나의 문장을 나타낸다. 첫 번째 열은 단어 정보이고 두 번째 열은 수작업으로 부착한 품사 부착 결과이다. 세 번째 열과 네 번째 열은 자동 분석기를 적용해 얻은 기본구 및 절 인식 결과이다. 다섯 번째 열은 어떤 동사에 대한 명제를 찾을 것인지 그 기준이 되는 동사를 표시하고 있다. 마지막의 두 열은 수작업으로 부착한 각 동사에 대한 부분 의미 분석의 결과로써, 동사 exist는 의미적 역할이 A1인 [contract]만을 논항으로 갖는다. deliver는 5개의 논항을 갖는데 예를 들어 의미적 역할이 A1인 [793 of the shipsets]이 그 중 하나이다. <표 4>는 <표 3>의 학습 말뭉치로부터 추출한 자질 벡터의 예이다. [자질값1] 열은 동사가 deliver이고 대상 기본구가 [the existing contract]일 때의 결과이고 [자질값2] 열은 동사가 deliver일 때 기본구 [the shipsets]에 대한 자질 추출 결과이다. 이 과정에 대한 이해를 돕기 위해, 동사구가 'has already deliverd'이고 기본구가 'the shipsets'일 때 각 주변 문맥의 위치를 (그림 4)에 표시하였다.

〈표 3〉 학습 말뭉치의 예

단어	품사	기본구	절	동사	exist	deliver
Under	IN	B-PP	(S*	-	*	(AM-LOC*
the	DT	B-NP	*	-	*	*
existing	VBG	I-NP	*	exist	(V*V)	*
contract	NN	I-NP	*	-	(A1*A1)	*AM-LOC)
,	,	O	*	-	*	*
Rockwell	NNP	B-NP	(S*	-	*	*
said	VBD	B-VP	*S)	-	*	*
,	,	O	*	-	*	*
it	PRP	B-NP	*	-	*	(A0*A0)
has	VBZ	B-VP	*	-	*	*
already	RB	I-VP	*	-	*	(AM-TMP*AM-TMP)
delivered	VBN	I-VP	*	deliver	*	(V*V)
793	CD	B-NP	*	-	*	(A1*
of	IN	B-PP	*	-	*	*
the	DT	B-NP	*	-	*	*
shipsets	NNS	I-NP	*	-	*	*A1)
to	TO	B-PP	*	-	*	*
Boeing	NNP	B-NP	*	-	*	(A2*A2)
.	.	O	*S)	-	*	*

〈표 4〉 자질 벡터의 예

자질		자질값1	자질값2
동사구와 기본구 사이의 정보	위치	-1	1
	거리	3	2
	#VP	1	0
	#NP	2	1
	#SBAR	0	0
	#CC	0	0
	#쉼표	2	0
	#콜론	0	0
	큰따옴표	0	0
	경로	NP-O-S-O-NP-VP	VP-NP-PP-NP
동사구	중심어	deliver	deliver
	중심어 품사	VBN	VBN
	구 유형	VP	VP
	첫 단어의 품사	VBZ	VBZ
동사구-1	중심어	it	It
	중심어 품사	PRP	PRP
	구 유형	NP	NP
기본구	중심어	contract	shipsets
	중심어 품사	NN	NNS
	구 유형	NP	NP
기본구-2	중심어	NONE	793
	중심어 품사	NONE	CD
	구 유형	NONE	NP
기본구-1	중심어	under	Of
	중심어 품사	IN	IN
	구 유형	PP	PP
기본구+1	중심어	,	To
	중심어 품사	,	TO
	구 유형	O	PP

(그림 4) 논항 후보가 'the shipsets'일 때 주변 문맥의 위치

3.2 의미역 할당 단계

이 단계는 첫 번째 단계에서 확인된 의미 논항에 적절한 의미역을 부착한다. 학습 데이터에는 26개의 의미역이 존재하는데 각 의미역이 무엇을 의미하는지는 PropBank Annotation Guidelines[3]를 참조하기 바란다. 본 연구는 SVM 분류기를 학습하는 비용을 줄이기 위해 모든 의미역을 고려하지 않고 학습 데이터에서 발생한 빈도를 고려해 18개의 고빈도 의미역만을 대상으로 학습을 진행한다.

의미역 할당 단계의 자질 집합 구성은 이전 단계에서 사용한 [큰따옴표]를 제외한 모든 자질에 덧붙여 [동사의 태] 자질을 사용한다. [동사의 태]는 동사가 수동태인지 능동태인지를 나타내는 것으로 동사의 앞 단어의 품사 및 동사의 품사에 따라 결정된다. 동사가 수동태이면 동사의 앞에는 행위의 객체에 해당하는 논항이 동사의 뒤에는 행위의 주체에 해당하는 논항이 발생한다. 능동태는 반대의 경우가 성립한다.

논항이 기본구들로 구성된다고 가정하여 문장을 기본구 단위로 구분하기 때문에, 술어를 포함한 동사구는 논항 후보에서 배제된다. 그런데 학습 데이터를 관찰한 결과 술어를 포함한 동사구 안에서 의미 논항들이 많이 발생하였다. 그래서 동사구 안에서 발생한 논항들을 찾기 위해 술어 앞과 술어 뒤로 나누어 분석하여 후처리를 수행하였다.

먼저 <표 1>의 already처럼 동사구에서 술어 앞에 위치한 단어들은 학습 데이터에서의 빈도수에 기반해 4가지 수동 규칙 및 211개의 자동 규칙에 의해서 후처리하였다. 수동 규칙의 예로써 동사구에서의 단어가 n't not 또는 Not이고 그 단어의 품사가 RB이고 단어와 술어 사이의 거리가 4 미만이면 그 단어에 의미역 AM-NEG을 할당하였다. 말뭉치로부터 얻은 자동 규칙의 예로써 동사구에서 단어 already가 발생하고 그 단어의 품사가 RB이면 그 단어에 의미역 AM-TMP을 할당하였다.

학습 데이터에서 의미 논항이 술어 바로 다음의 단어부터 시작하는 경우가 자주 관찰되었다. 논항 경계의 일치를 위해서 술어가 동사구의 끝 단어가 아니면 술어 바로 다음에 나오는 단어를 새로운 기본구의 시작 단어로 간주하였다. 즉 기본구 인식 결과에서 I-VP 태그를 B-VP 태크로 변경하는 후처리를 수행하였다.

4. 실 험

본 연구에서 사용한 실험 말뭉치는 PropBank에 기반한 CoNLL-2004 shared task 데이터 집합으로 <표 3>의 예와 같다 [2-3]. 이 말뭉치는 완전 구문 분석기, 사전 또는 의미 지식 베이스를 사용하지 않고, 단지 부분 구문 분석 결과만을 포함하고 있다. PropBank는 Penn Treebank II에 의미역 정보를 추가한 것으로 논항의 구문적 위치(syntactic position)에 상관없이 의미역을 할당하였다. 즉 'break'에 대해 깨트린 대상을 나타내는 'the window'가 문장에서 주어, 목적어, 또는 보어로 발생하였는지와 관계없이 같은 의미역을 할당받는다. 이 말뭉치는 동사의 의미마다 별도의 'frameset'이라는 하위범주화(subcategorization) 프레임을 가지고 있다. 예를 들어, 'pass'라는 동사가 '법안을 가결하다'란 의미로 쓰일 때는 3개의 논항('legislative body', 'bill', 'law')을 갖는 것으로 '추월하다'라는 사용될 때는 2개의 논항('entity moving ahead', 'entity falling behind')을 갖는 것으로 정의되어 있다.

PropBank의 의미역은 'core argument(A로 표기)'와 'adjunct argument(AM으로 표기)'로 구성되어 있다. 먼저 'core argument'는 각 동사에 대해 정의된 고정된 수의 의미적 역

3) http://verbs.colorado.edu/~mpalmer/projects/ace/PBguidelines.pdf

할이 있고, 그것은 0부터 5까지의 숫자로 표기한다. 예를 들어 'break'란 동사의 깨트리다란 의미에 대해 'A0:the breaker', 'A1:thing broken', 'A2:instrument', 'A3:pieces'로 정의되어 있다. 논항의 의미적 역할이 같은 A0로 표기되었다 할지라도 동사에 따라 다른 의미로 사용될 수 있다. 다음으로 'adjunct argument'는 동사와 밀접하게 관련되어 있지 않다. 예를 들어 'yesterday'는 'break'뿐만 아니라 다양한 동사에서 시간을 나타내는 같은 의미적 역할 'AM-TMP'로 사용될 수 있다. 'adjunct argument'는 언어학의 'argument/adjunct' 구분과 차이가 있는데, 'adjunct'에 속하지 않는 의미적 역할 'MOD'와 'NEG'가 추가되었다. 'MOD'는 직설법/명령법/가정법을 'NEG'는 긍정/부정을 나타낸다. 'adjunct'의 유형을 나타내는 레이블은 총 12개로 다음과 같다: 'DIR, LOC, MNR, TMP, EXT, REC, PRD, PRP, DIS, ADV, MOD, NEG'

학습을 위해서 SVM-Light [10]를 활용했는데, 의미 논항 확인 단계와 의미역 할당 단계 모두에서 'one-vs-rest' 분류 방법의 degree 2의 polynomial kernel을 사용하였다. 다중 분류를 수행하기 위해 'one-vs-rest' 방법을 사용하였는데, 이 방법은 클래스의 수만큼 분류기를 학습한다. 해당 클래스에 속하지 않는 학습 예제들은 모두 부정 학습 예제로 간주한다. 최종적으로 각 클래스의 분류기가 생성한 값들을 비교해 가장 높은 값을 가진 클래스로 분류합니다.

4.1 실험 결과

<표 5>는 테스트 집합(test set)에서 전체 성능 및 각 의미역별 부분 의미 분석의 성능을 나타낸다. 시스템은 62.43%의 재현율, 65.63%의 정확률, 63.99%의 F-score를 보였다. 학습 비용을 줄이기 위해 저빈도 의미역들은 배제했기 때문에 0의 성능을 보이는 의미역들이 많지만 큰 비율을 차지하지 않는 의미역들이기 때문에 전체 성능에 미치는 영향은 미미하였다. 학습 말뭉치에서 고빈도로 발생하는 행위의 주체를 나타내는 A0와 행위의 객체를 나타내는 A1은 시스템의 전체 성능보다 높은 성능을 보였다. 학습 말뭉치에서 고빈도로 발생한 의미역들이기 때문에 학습이 잘 이루어져 이러한 결과를 얻었다.

<표 6>은 개발 집합(development set)에서의 실험 결과로 부분 의미 분석의 전체 성능 및 각 단계의 독립적인 성능을 나타낸다. 논항 확인 단계에서는 논항 후보들 중에서 논항만을 찾는 일이기 때문에 논항과 논항이 아닌 것을 올바르게 인식했는지 나타내기 위해, 시스템이 할당한 결과를 재현율과 정확률이란 값으로 말뭉치상의 정답 논항 집합과 비교하였다. 반면에 의미역 할당 단계에서는 각 논항에 적절한 의미역이 할당됐는지만을 정답과 비교하기 때문에 정확도로 평가하였다. <표 6>의 부분 의미 분석 항목은 논항 확인 단계와 의미역 할당 단계를 순차적으로 수행한 결과로, 오류를 포함한 본 시스템의 논항 확인 결과가 의미역 할당 단계의 입력으로 사용되었다. 반면에 의미역 할당 항목의 성능은 말뭉치상의 정답 논항을 입력으로 사용해 의미역 할당 단계만의 성능을 측정한 것이다. 본 시스템은 개발 집합

〈표 5〉 각 의미역별 부분 의미 분석의 성능

의미역	재현율	정확률	F-score
전체	62.43	65.63	63.99
A0	74.60	78.24	76.38
A1	66.46	65.83	66.14
A2	43.70	49.84	46.57
A3	34.00	56.04	42.32
A4	44.00	62.86	51.76
A5	0	0	0
AM-ADV	44.30	45.18	44.74
AM-CAU	22.45	36.67	27.85
AM-DIR	20.00	20.00	20.00
AM-DIS	58.22	56.62	57.41
AM-EXT	57.14	61.54	59.26
AM-LOC	31.14	26.01	28.34
AM-MNR	35.69	43.54	39.22
AM-MOD	91.10	97.46	94.17
AM-NEG	88.19	94.92	91.43
AM-PNC	28.24	40.00	33.10
AM-PRD	0	0	0
AM-TMP	45.38	51.83	48.39
R-A0	83.02	80.49	81.73
R-A1	51.43	75.00	61.02
R-A2	33.33	100.00	50.00
R-A3	0	0	0
R-AM-LOC	0	0	0
R-AM-MNR	0	0	0
R-AM-PNC	0	0	0
R-AM-TMP	0	0	0
V	96.66	96.66	96.66

〈표 6〉 개발 집합에서 각 단계의 성능

단계	재현율	정확률	F-score	정확도
부분 의미 분석	64.36	67.27	65.78	-
논항 확인 의미역 할당	72.30 -	75.96 -	74.08 -	- 85.45

에서 64.36%의 재현율, 67.27%의 정확률, 65.78%의 F-score를 보였다. 논항 확인 단계는 74.08%의 F-score를 보였고 의미역 할당 단계는 85.45%의 높은 정확도를 나타내었다. 의미역 할당 단계에서 보인 성능은 오류 전파없이 의미역 할당만을 수행했을 때의 결과이다. 전체 성능을 좌우하는 논항 확인 단계에서는 상대적으로 낮은 성능을 얻었는데 제안한 자질 외에 더 좋은 식별력을 지닌 자질을 개발하는 것이 필요하다.

CoNLL-2004 shared task 테스트 집합에서 가장 좋은 성능을 보인 시스템은 지지 벡터 기계를 분류기로 사용한 (K. Hacioglu et. al., 2004)로 F-score 69.49%를 보였다[5]. 두 번째로 좋은 성능을 보인 시스템은 SNoW를 분류기로 사용한 (V. Punyakanok et. al, 2004)로 F-score 66.39%를 보였다[9]. 이러한 시스템들이 본 시스템보다 더 나은 성능을 보인 이유는 좋은 자질들을 선별해 사용하였기 때문으로 향후 자질 개발에 대한 연구를 진행할 필요가 있다.

5. 결 론

본 논문은 부분 구문 분석 결과에 기반한 두 단계 부분 의미 분석 시스템을 제안하였다. 부분 구문 분석은 완전 구문 분석보다 높은 정확도 및 낮은 복잡도에 따른 빠른 속도를 얻을 있으며, 문장 구성 요소를 단어 단위가 아닌 기본 구 단위로 표현하는 것을 가능하게 하였다. 본 연구는 기본 구 단위의 부분 의미 분석을 사용하였는데, 이를 통해 논항 후보 수의 감소에 따른 낮은 학습 비용이라는 효과를 얻을 수 있었다. 분류기 학습을 위해서는 지지 벡터 기계를 사용하였는데 이진 분류기를 멀티 클래스 분류 과제에 효과적으로 적용하기 위해 두 단계 방법을 제안하였다. 먼저 의미 논항의 경계를 찾는 의미 논항 확인 단계를 수행하고 그 후에 확인된 논항들을 적절한 의미역으로 분류하는 의미역 할당 단계를 수행하였다. 부분 의미 분석의 단계를 구분함으로써 부정 학습 예제들로부터 야기되는 클래스 분포의 불균형 문제를 완화할 수 있었고 다른 일을 수행하는 각 단계에 적합한 다른 자질 집합을 구성할 수 있었다. 실험 결과 본 시스템은 테스트 집합에서 63.99%, 개발 집합에서 65.78%의 F-score를 얻었다. 또한 개발 집합을 사용해 F-score 74.08%의 논항 확인 성능과 정확도 85.45%의 의미역 할당의 성능을 독립적으로 제시하였다.

참 고 문 헌

[1] Buchholz, S., Memory-Based Grammatical Relation Finding, *PhD. thesis, Tilburg University*, 2002.

[2] Carreras, X. and Marquez, L., Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling, In *Proceedings of the Eighth Conference on Natural Language Learning*, 2004.

[3] CoNLL shared task, http://www.lsi.upc.edu/~srlconll/home.*html*

[4] Gildea, D. and Jurafsky, D., Automatic Labeling of Semantic Roles, *Computational Linguistics*, Vol. 28, No. 3, pp. 1-45, 2002.

[5] Hacioglu, K., Pradhan, S., Ward, W., Martin, J. and Jurafsky, D., Semantic Role Labeling by Tagging Syntactic Chunks, In *Proceedings of the Eighth Conference on Natural Language Learning*, 2004.

[6] Kwon, N., Fleischman, M. and Hovy, E., FrameNet-based semantic parsing using maximum entropy models, In *Proceedings of the International Conference on Computational Linguistics*, 2004.

[7] Pradhan, S., Hacioglu, K., Krugler, V., Ward, W., Martin, J. and Jurafsky, D., Shallow Semantic Parsing Using Support Vector Machines, *Technical Report, TR-CSLR-2003-03*, 2003.

[8] Pradhan, S., Ward, W., Hacioglu, K., Martin, J. and Jurafsky. D., Semantic Role Labeling using Different Syntactic Views, In *Proceedings of the Association for Computational Linguistics*, 2005.

[9] Punyakanok, V., Roth, D., Yih, W., Zimak, D. and Tu, Y., Semantic Role Labeling Via Generalized Inference Over Classifiers, In *Proceedings of the Eighth Conference on Natural Language Learning*, 2004.

[10] SupportVector Machine(SVM)-Light, *http://svmlight.joachims.or*

[11] Surdeanu, M., Harabagiu, S., Williams, J. and Aarseth, P., Using Predicate Arguments Structures for Information Extraction, In *Proceedings of the Association for Computational Linguistics*, 2003.

[12] Xue, N. and Palmer, M., Calibrating Features for Semantic Role Labeling, In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2004.

[13] Yamada, H. and Matsumoto, Y., Statistical Dependency Analysis with Support Vector Machines, In *Proceedings of the 8th International Workshop of Parsing Technologies*, 2003.

박 경 미

e-mail : parkkyungmi75@gmail.com

1998년 연세대학교 식품영양학과(학사)

2000년 연세대학교 정보산업공학(학사)

2002년 연세대학교 컴퓨터산업시스템공학과 (공학석사)

2008년 고려대학교 컴퓨터학과(이학박사)

2009년~현 재 숭실대학교 정보미디어기술연구소 전임연구원

관심분야 : 자연언어 처리, 바이오 텍스트 마이닝, 정보 추출

문 영 성

e-mail : mun@computing.ssu.ac.kr

1983년 연세대학교 전자공학과(학사)

1986년 알버타대학교 전자공학과(공학석사)

1987년~1994년 한국통신 연구원

1993년 텍사스대학교 컴퓨터공학과(공학박사)

1994년~현 재 숭실대학교 컴퓨터학부 교수

관심분야 : 모바일 아이피, 네트워크 보안, 그리드 네트워킹