사전 정보를 이용한 단어 중의성 해소 모형에 관한 실험적 연구*

An Experimental Study on an Effective Word Sense Disambiguation Model Based on Automatic Sense Tagging Using Dictionary Informaiton

이 용 구(Yong-Gu Lee)**, 정 영 미(Young-Mee Chung)***

초 록

이 연구에서는 수작업 태깅없이 기계가독형 사전을 이용하여 자동으로 의미를 태깅한 후 학습데이터로 구축한 분류기에 대해 의미를 분류하는 단어 중의성 해소 모형을 제시하였다. 자동 태깅을 위해 사전 추출 정보 기반 방법과 연어 공기 기반 방법을 적용하였다. 실험 결과, 자동 태깅에서는 복수 자질 축소를 적용한 사전 추출 정보 기반 방법이 70.06%의 태깅 정확도를 보여 연어 공기 기반 방법의 56.33% 보다 24.37% 향상된 성능을 가져왔다. 사전 추출 정보 기반 방법을 이용한 분류기의 분류 정확도는 68.11%로서 연어 공기 기반 방법의 62.09% 보다 9.7% 향상된 성능을 보였다. 또한 두 자동 태깅 방법을 결합한 결과 태깅 정확도는 76.09%, 분류 정확도는 76.16%로 나타났다.

ABSTRACT

This study presents an effective word sense disambiguation model that does not require manual sense tagging process by automatically tagging the right sense using a machine-readable dictionary, and attempts to classify the senses of those words using a classifier built from the training data. The automatic tagging technique was implemnted by the dictionary information-based and the collocation co-occurrence-based methods. The dictionary information-based method that applied multiple feature selection showed the tagging accuracy of 70.06%, and the collocation co-occurrence-based method 56.33%. The sense classifier using the dictionary information-based tagging method showed the classification accuracy of 68.11%, and that using the collocation co-occurrence-based tagging method 62.09%. The combined tagging method applying data fusion technique achieved a greater performance of 76.09% resulting in the classification accuracy of 76.16%.

키워드: 단어 중의성 해소, 자동 태깅, 의미 분류, 사전 추출 정보 기반 태깅, 연어 공기 기반 태깅
word sense disambiguation, automatic tagging, sense classification, dictionary information-based tagging, collocation co-occurrence-based tagging

---

* 이 논문은 연세대학교 대학원 박사학위논문을 요약한 것이며, 2005년도 한국학술진흥재단의 지원에 의해 연구됨(KRF-2005-908-H00004).

** 연세대학교 문헌정보학과 시간강사(yglee@yonsei.ac.kr)
*** 연세대학교 문헌정보학과 교수(ymchung@yonsei.ac.kr)

3. 서론

텍스트에 대한 자연언어 처리 단계는 흔히 형태소 분석, 구문 분석, 의미 분석, 화용 분석의 네 단계로 구분된다. 전문으로 되어 있는 문헌을 자동색인하기 위해 형태소 분석을 수행하여 접두사와 접미사, 어미, 조사 등을 제거한 후, 어근이나 어간을 일정한 기준에 의해 색인어로 선정한다. 특히 의미 분석은 동형이의어나 다의어와 같이 중의성을 갖는 단어의 의미를 정확히 파악하기 위해 필요하다. 이때 필요한 방법 중에 하나가 단어 중의성 해소(word sense disambiguation: WSD) 기법이다. 이 기법은 문맥에서 중의성을 갖는 단어가 어떤 의미로 사용되었는가를 식별하여 단어 중의성을 해소한다.

WSD 기법에서 중의성 단어의 의미를 식별하기 위해서는 다양한 언어 정보원을 이용한다. 영문의 경우, 단어의 중의성을 해소하고 의미를 식별하기 위해 사전의 뜻풀이부터 중의성 해소용 말뭉치(corpus)에 이르기까지 다양한 언어자원을 구축하여 이용하고 있다. 하지만 한글의 경우 일반적인 사전을 전자화한 언어자원 이외에는 이렇다 할 자원이 없다. 중의성을 띠는 대표적인 일부 단어에 대해 의미를 태깅한 말뭉치가 있긴 하지만, 이러한 언어자원의 경우 몇몇 단어만이 의미가 태깅되어 있어 WSD에서 광범위하게 이용하기는 어려운 실정이다. 따라서 한글에서 이러한 언어자원의 제약을 극복하기 위한 중의성 해소 방안의 연구가 절실히 필요하다.

이 연구에서는 문헌의 내용을 대표하는 색인어의 중의성을 해소하기 위해 수작업 태깅 없이 기계가독형 사전을 이용하여 자동으로 의미를 태깅하는 단어 중의성 해소 모형을 제안하고자 한다. 또한 모형의 최적화를 통해 중의성 해소 성능의 향상을 가져올 수 있는지를 평가하고자 하였다. 이 모형에서는 좋은 성능을 가져오기 위해 지도 학습 방법을 적용하되, 언어자원의 제약을 극복하기 위해 학습데이터(training data)를 자동으로 구축하였다.

이 연구에서는 사전의 표제어 아래에 기술된 핵심 단어를 이용한 두 종류의 자동 태깅 기법을 구현하였다. 하나는 사전 추출 정보기반 방법으로, 사전에서 추출한 핵심 단어가 텍스트 문맥의 중의성 단어와 함께 출현하면 이 핵심 단어가 추출된 의미로 중의성 단어를 자동 태깅한다. 다른 모형은 연어 공기(collocation co-occurrences) 기반 방법으로 사전에서 추출한 핵심 단어와 중의성 단어 사이에 공통된 연어가 존재하면 전자와 같이 의미를 자동 태깅한다. 또한 최적의 자동 태깅 성능을 이끌어 내기

위해 두 가지 자동 태깅 기법에 대해 데이터 결합(data fusion)을 적용하였다. 이렇게 구축된 학습 데이터를 이용하여 대표적인 지도학습 모형인 나이브 베이즈 분류기(Naive Bayes classifier)를 구축하고 의미를 분류함으로써 최적의 중의성 해소 성능을 가져올 수 있는 방안을 제시하고자 한다.

4. 중의성 해소에 관한 선행연구

중의성 해소 알고리즘은 중의성 해소에 필요한 정보를 입수하는 방법에 따라 (1) WordNet과 같은 시소러스와 LDOCE(Longman Dictionary of Contemporary English)와 같은 전자적 어휘사전 등의 지식베이스를 이용하는 지식 기반 알고리즘(knowledge-driven WSD), (2) 의미 태깅이 되어 있거나 또는 태깅이 되어 있지 않은 말뭉치를 사용하는 말뭉치 기반 알고리즘(data-driven or corpus-based WSD), (3) 말뭉치와 지식베이스를 함께 사용하는 혼합형 알고리즘(hybrid WSD) 등 세 가지 유형으로 분류한다(Stevenson 2003).

지식베이스를 이용한 단어 중의성 해소 알고리즘은 크게 사전 뜻풀이(dictionary definition), 시소러스, 그리고 2개 언어 사전(bilingual dictionary)을 이용하는 방법으로 구분할 수 있다(Manning and Schütze 1999).

MRD의 뜻풀이를 이용한 대표적인 단어 중의성 해소 연구로는 Lesk(1986)를 들 수 있다. Lesk는 사전의 뜻풀이에 나타난 단어의 중복 정도에 따라 문맥에 출현한 단어의 의미를 식별하였다. 즉 문맥에 출현한 중의성 단어에 대해 의미범주별로 뜻풀이를 MRD로부터 추출하고 단어간의 모든 가능한 의미 조합에 대해 뜻풀이에서 공통으로 출현한 단어의 중복도(definition overlap)를 계산한 후, 중의성 단어의 의미를 가장 높은 중복도를 갖는 의미로 배정하였다. 이 방법은 50~70% 정도의 정확성을 보이기는 하지만, 사전으로부터 얻은 정보를 이용하여 양질의 단어 중의성 해소를 수행하기엔 부족하다. Kilgarriff and Rosensweig(2000)는 이 방법을 변형하여 중의성 단어의 뜻풀이에 포함된 단어와 중의성 단어가 포함된 문맥에 출현한 단어간의 중복 정도를 이용하였다.

MRD를 이용한 단어 중의성 해소에서 표제어의 뜻풀이나 용례 이외에 다른 추가적인 정보를 이용할 수 있다. 추가적인 정보란 문법 태그(syntactic tag), 형태소 정보, 주제 분류코드, 의미론적 문맥(semantic context), 연어, 의미적 단어 연관성, 동사가 요

구하는 보충어의 의미를 기술하는 선택 제약(selectional restriction) 등이 해당된다 (Stevenson 2003). 이러한 부가 정보를 단독으로 이용하거나 여러 개를 조합하여 중의성 해소의 성능을 향상시키고자 하였다. Krovetz and Croft(1989)는 LDOCE를 이용하여 정보검색에서 단어를 의미적인 수준에서 색인하였다. LDOCE에서 중의성 단어가 포함된 문맥차원 내에서 의미를 결정하기 위해 여러 가지 정보원을 수집하고 그것을 이용하였다. 이들 정보원들은 품사, 형태소 분석, 하위범주(subcategorization) 정보, 의미론적 제약, 주제 분류 등 이다.

MRD와 달리 시소러스나 온톨로지의 경우 주제적으로 관련된 단어들을 포함하거나 단어간의 계층적 관계를 포함한다. 단어 중의성 해소에 시소러스나 온톨로지를 정보원으로 이용하는 연구는 이러한 의미적 관계를 이용한다.

시소러스 기반 알고리즘의 기본 개념은 의미 범주화(semantic categorization)이다. 이는 한 문맥 내 단어들의 의미 범주가 문맥 전체의 의미 범주를 결정하며, 다시 이 문맥의 의미 범주가 문맥 내 단어들의 의미를 결정한다는 것이다. 문맥에 포함된 단어 측면에서 보면, 특정 의미범주에 해당하는 문맥에 출현한 중의성 단어는 여러 의미 중에 그 단어가 속한 문맥과 일치하는 하나의 의미범주로만 사용된다. 따라서 이러한 특성을 이용하면, 중의성을 띠는 단어가 특정 문맥에 출현하였을 때 그 단어의 의미범주는 그 단어를 포함하는 문맥의 의미 범주를 이용하여 결정할 수 있다. 이러한 특성을 이용한 대표적인 시소러스 기반 중의성 해소 연구로는 Roget 시소러스의 의미 범주를 이용한 Yarowsky(1992)의 연구가 있다. Resnik(1995)는 WordNet의 IS-A 계층관계에 있는 명사에 대해 의미 유사도(sematic similarity)를 측정하여 단어 중의성 해소에 이용하였으며, Agirre and Rigau(1996)는 개념 밀도(conceptual density)를 적용한 개념간의 거리와 WordNet의 명사 분류(taxonomy)를 이용하여 완전 자동으로 중의성을 해소하였다. 이러한 연구의 문제는 MRD와 같이 시소러스도 마찬가지로 인간을 위해 만들어진 언어자원이므로 단어간의 완벽한 의미 범주 정보를 제공할 수 없다는 한계를 가지고 있다는 것이다.

2개 언어 사전에 기반한 단어 중의성 해소 기법은 2개 언어 사전에서의 대응되는 단어 엔트리를 이용한다. 이 기법에서 제1언어는 중의성 해소의 대상 단어가 속한 언어이며, 제2언어는 중의성 해소에 사용될 단어가 있는 언어를 의미한다. Dagan and Itai(1994)의 알고리즘에서는 중의성 단어의 중의성 해소를 위하여 기본적으로 그 단어가 출현한 구를 확인하고, 그 구의 출현 사례들(instances)을 제2언어의 말뭉치에서 찾

은 후, 이 번역어구들 속에 한 가지 의미의 번역어만이 출현할 경우 이에 대응하는 제1언어의 정의를 중의성 단어의 의미로 부여하였다.

5. 단어 중의성 해소 모형 설계

3.1. 모형 개요

여러 가지 중의성 해소 기법 중에 말뭉치 기반 중의성 해소 기법이 가장 좋은 성능을 보여주고 있다. 특히 지도학습 알고리즘이 비지도학습 알고리즘보다 더 좋은 성능을 보여준다. 말뭉치 기반 중의성 해소 기법 중에 지도학습 알고리즘을 이용하여 중의성을 해소하려면 중의성 단어에 대해 의미 태그가 부착된 말뭉치가 필요하다. 그 이유는 지도학습 알고리즘이 의미 태그가 부착된 말뭉치로부터 학습 데이터를 추출하고 이를 이용하여 분류기를 구축할 수 있기 때문이다. 하지만 한글의 경우 의미 태그가 부착된 중의성 해소용 말뭉치의 부족으로 인해, 분류기를 구축하기 위한 학습 데이터를 획득하는데 어려움이 있다.

중의성을 해소하기 위해서는 단어의 의미 정보가 존재해야 한다. 이 정보는 의미 태그가 부착된 말뭉치에서 추출할 수 있거나 전자적으로 이용 가능한 사전으로부터 추출할 수 있다. 현재 이용 가능한 중의성 해소용 한글 말뭉치는 '눈'과 같은 대표적인 중의성 단어만을 의미 태깅하였으므로, 태깅된 단어 이외에 실험문헌에 나타난 단어의 중의성을 해소하는 것은 불가능하다. 국어사전의 경우 사전의 편집정책에 따라 다르지만, 중의성을 띠는 단어를 포함하여 대부분의 단어에 대해 의미를 구분할 수 있는 정보를 제공한다. 그리고 이들의 몇몇은 컴퓨터 파일이나 인터넷을 통해 전자적으로 이용할 수 있다.

일반적으로 사전은 표제어를 설명하기 위해 여러 가지 종류의 정보원을 제공한다. 표제어의 의미를 기술하는 뜻풀이가 있고, 표제어가 실제로 쓰이는 예를 보여주는 용례가 있다. 또한 파생어, 동의어, 반의어, 관련어 등을 보여주는 부가 정보란이 있으며 활용형이나 발음, 한자 등에 대한 다양한 정보도 기술되어 있다. 이들 정보 중에 대다수는 단어의 중의성을 해소하는 데 이용가능하다.

앞서 살펴보았듯이 학습데이터 획득의 어려움을 극복하고 지도학습 알고리즘을 이용하기 위해서는, 사전에서 추출한 정보로 자동으로 태깅하여 학습데이터를 생성하고, 중의성 해소 분류기를 구축하여 의미 분류를 수행하는 단어 중의성 해소 모형을 적용

하는 것이 좋은 방법 중의 하나이다. 이 연구에서는 중의성 단어의 의미를 자동으로 태깅하기 위한 세 가지 방법을 적용하였다. 먼저 단일 기법으로 사전 추출 정보 기반 방법과 연어 공기 기반 방법을 적용하였으며, 이 두 방법의 자동 태깅 결과를 결합한 방법을 적용하였다.

3.2. 자동 태깅 기법

5.1.1 사전 추출 정보 기반 방법

중의성 단어에 대해 사전으로부터 단어의 의미식별을 가능하게 해 주는 정보를 추출할 수 있으며 이를 이용하여 중의성을 해소할 수 있다. 이와 같은 기법은 주로 지식기반 중의성 해소 기법에서 널리 이용하는 방법이다.

의미적으로 관련된 단어들은 어휘적 응집성(lexical cohesion)을 갖는다. 즉 문맥에서 동의어, 상・하위어, 관련어 등과 같이 관련된 단어들은 함께 사용된다. 이와 같은 특성을 반영하여 두 단어간의 의미적인 관련성을 공기 정보를 이용하여 파악할 수 있다. 따라서 사전 추출 정보 기반 방법은 핵심 단어와 중의성 단어가 특정 크기의 문맥에서 공기하면 핵심 단어가 추출된 의미를 중의성 단어에 부여하였다.

의미를 자동 태깅한 학습데이터를 구성하기 위해서 사전으로부터 단어의 의미를 식별가능하게 해 주는 정보를 추출할 때, <표1>과 같이 정보원과 경험규칙을 이용하였다.

<표 1> 자동 태깅을 위한 경험규칙

유형	규칙	정보원	사례
연어	한 연어 한 의미 (Yarowsky 1993)	용례 동사형 연어	'경기 변동' '눈을 감다'
한자	한자어로 구별	한자정보	實事/實辭/實査
동의어/반의어/관련어	공기 정보	부가정보	實事 : 事實 實辭 : 虛辭
파생어	'-하다/-되다' 접미사	파생정보	實査하다/實査되다
뜻풀이 출현 명사	공기 정보	뜻풀이	實査 : 검사, 조사

연어정보를 이용할 때 형태소 분석과 구문분석 결과를 이용하여 <표2>와 같은 연어의 구문적 패턴을 추출하였다. 사전에서 용례를 대상으로 연어를 추출하였으며, 실

험문헌에서는 중의성 단어와 사전에서 추출한 핵심 단어를 대상으로 연어를 추출하였다.

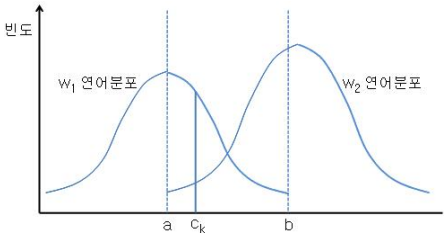
<표 2> 연어의 구문적 패턴 유형

유 형	문법 형태	용 례
유형1	[중의성 단어/핵심 단어]+명사	감자튀김
유형2	명사+[중의성 단어/핵심 단어]	차등감자
유형3	[중의성 단어/핵심 단어](+조사)␣명사	감자의 재배
유형4	명사(+조사)␣[중의성 단어/핵심 단어]	개량한 감자

5.1.2 연어 공기 기반 방법

선행 연구(Yaroskwy 1993; 정영미, 이용구 2005)를 보면 연어가 형성되고 이를 이용하여 중의성 해소가 가능하다면 한글이나 영문, 모두 98%이상의 정확도를 가져온다. 두 번째 자동 태깅 방법인 연어 공기 기반 방법은 이러한 특성을 이용하여 중의성 단어와 사전에서 추출한 핵심 단어 사이에 형성된 연어의 공기 정보를 이용하였다.

특정 단어 w_1이 있고 이 단어와 함께 연어를 구성하는 단어들을 집합으로 표현할 수 있다. 마찬가지로 다른 단어 w_2의 연어를 구성하는 단어들도 집합으로 표현할 수 있다. 만약 다수의 연어를 형성하는 두 단어 w_1과 w_2가 의미적으로 밀접히 관련되어 있다면, 이 연어들 중에 두 단어 w_1과 w_2의 사이에 공기하는 연어(c_3)가 형성될 수 있다. w_1과 w_2의 연어들을 실험문헌 집단 전체에서 추출한다면 이들 연어의 빈도분포는 <그림1>과 같을 것이다. <그림1>에서 x축은 연어로 형성된 단어에 해당하며, y축은 그 단어들의 출현빈도에 해당한다. 두 단어에서 공기한 연어는 a에서 b사이에 해당하는 단어들이며, 그렇지 않은 단어는 나머지 부분에 해당한다. 공기한 연어 c_k의 경우 단어 w_2에서의 출현빈도보다 w_1에서의 출현빈도가 더 많은 것을 알 수 있다. 또한 두 단어의 연어분포가 서로 많이 겹칠수록 의미적으로 밀접하게 관련될 확률이 높다. 즉 두 단어의 공통된 연어를 많이 가질수록 두 단어의 의미는 더 가까워진다고 볼 수 있다.



<그림 1> 두 단어의 연어분포

서로 다른 두 단어에 대해 연어분포를 적용하였지만 한 단어가 여러 의미를 갖는다면 의미별로 다른 연어분포를 갖게 되어 이러한 정보를 이용하여 의미를 식별할 수 있을 것이다. 따라서 다의어에 대해 자동으로 의미를 태깅하기 위하여 연어 공기 기반 방법은 다음과 같은 규칙을 적용하였다. 실험집단에서 사전에서 추출된 여러 정보(뜻풀이 출현 명사, 동의어/반의어/관련어)의 연어와 패턴정보(<표1> 참조)를 추출하였으며, 중의성 대상 단어의 전체 출현문맥에서도 연어와 패턴정보를 추출하였다. 이렇게 추출한 두 집단에서 공통된 연어의 유무에 따라 사전의 정보가 추출된 의미로 태깅하였다.

이 연구에서는 가장 단순하면서 기본이 되는 연어 공기 패턴을 이용하고자 한다. 중의성 단어와 공기 연어의 구문 패턴이 사전으로부터 추출한 핵심 단어와 공기 연어의 구문 패턴과 동일한 것만을 이용하였다. 다만 자동 태깅의 정확도를 높이기 위해 공기 연어의 빈도를 이용하였다. 즉 추출된 연어별로 총 출현빈도를 계산하여 상위 N(10, 30, 50, 100, All)개를 선정하였다. 그리고 이들 단어와 중의성 단어가 연어를 형성한 문맥에 대해 해당 의미로 자동 태깅하였다.

3.3. 의미 분류 기법

지도학습 기법에서는 중의성이 해소된 말뭉치, 즉 중의성을 띄는 단어 w의 모든 출현에 대해 특정 문맥에서 의미 s_k로 의미가 분류된 말뭉치를 학습데이터로 사용한다. 통계적인 자연언어 처리 이론 중 하나를 단어 의미 중의성 해소에 적용한 방법이 나이브 베이즈 분류기이다. 나이브 베이즈 분류기는 Gale(1992), Gale, Church, and Yarowsky(1992a, 1992b, 1993) 등의 연구에서 사용되었다.

나이브 베이즈 분류기는 방대한 문맥에서 중의성을 띄는 단어의 인접 단어들을 이용하여 의미를 식별한다. 즉, 중의성을 띄는 단어의 의미를 파악하는 데 주위의 단어가 유용한 정보를 제공하게 되며, 이러한 주변 단어의 공기빈도(co-occurrence frequency) 정보를 이용하여 통계적인 연상추론을 하게 된다. 나이브 베이즈 분류기는 범주(class)를 결정할 때 오류 확률을 최소화하기 위해 베이즈 결정규칙을 이용한다.

중의성 단어 w가 말뭉치에서 출현한 문맥을 c_j라 하고 중의성 해소를 위해 문맥 자질로 사용된 단어들을 v_j라 할 때, 문맥 자질에 기반하여 단어의 의미 중의성을 해소하는 나이브 베이즈 분류기의 결정 규칙은 다음과 같다(Manning and Schütze 1999). 이 공식에 의해 중의성 단어 w에 의미 s'를 부여하게 된다.

$$Decide\ s'\ \mathrm{if}\ s' = \arg max_{s_k}[\log P(s_k) + \sum_{v_j in\ c} \log P(v_j|s_k)]$$

위 공식에서 $P(v_j|s_k)$와 $P(s_k)$는 중의성이 해소된 학습 말뭉치로부터 최우추정법(maximum-likelihood estimation)에 의해 계산된다. $P(v_j|s_k)$와 $P(s_k)$를 계산하기 위한 공식은 다음과 같다.

$$P(v_j|s_k) = \frac{C(v_j, s_k)}{\sum_{i=1}^{t} C(v_i, s_k)}$$

$$P(s_k) = \frac{C(s_k)}{C(w)}$$

위 공식에서 $C(v_j, s_k)$는 중의성을 띄는 단어 w가 학습 말뭉치에서 의미 s_k를 가질 때 문맥자질 v_j가 출현한 횟수이며, $\sum_{i=1}^{t} C(v_i, s_k)$는 문맥에 나타난 모든 단어들의 총 출현횟수이다. $C(s_k)$는 중의성을 띄는 단어 w가 학습 말뭉치에서 의미 s_k로 출현한 횟수이고, $C(w)$는 중의성을 띄는 단어 w의 총 출현횟수이다.

6. 단어 중의성 해소 실험

6.1 실험 설계

이 연구의 중의성 해소 실험은 크게 사전에서 추출할 수 있는 정보를 이용하여 의미를 자동 태깅하는 단계와, 자동 태깅으로 구축된 학습데이터를 토대로 중의성 해소 분류기를 구축하고 전체 실험문헌 내 나머지 중의성 단어를 의미 분류하는 단계로 나뉜다.

중의성 해소 실험을 하기 전에 중의성 해소 대상 단어 9개를 선정하고, 실험문헌인 신문기사에 대해 형태소 분석과정을 통해 색인어를 추출한 후 해당 색인어의 연어를 구문적 패턴에 맞춰 추출하였다.

전체적인 실험과정은 다음과 같다.

첫째, 기계 가독형 국어사전에서 중의성 해소에 필요한 정보를 추출하였다. 사전에서 표제어를 설명하기 위해 기술된 핵심 단어인 한자, 용례에서 추출한 연어, 파생어, 동의어, 반의어, 관련어 그리고 뜻풀이에서 출현한 명사를 추출하였다. 핵심 단어를 추출할 때 표제어의 의미번호를 함께 추출하였다.

둘째, 사전에서 추출된 핵심 단어를 이용하여 중의성 단어의 의미를 자동 태깅하였으며, 자동 태깅 성능의 최적화를 위해 자질 선정 방법을 적용하였다. 즉 사전으로부터 추출된 핵심 단어가 자동 태깅에 좋은 자질인지 판별하기 위해, 핵심 단어의 정보원에 따라 다양한 문헌빈도 기준을 적용하여 최적화가 가능한지 분석하였다.

셋째, 전 단계에서 자동 태깅된 데이터를 학습데이터로 이용하여 분류기를 구축하고, 자동 태깅이 안된 나머지 중의성 단어에 대해 분류기를 이용하여 의미 분류하여 중의성 단어의 전체를 의미 태깅하였다. 그리고 그 결과를 수작업 태깅과 비교하여 중의성 해소 성능이 어느 정도 향상되는가를 실험을 통해 파악하였다.

이 연구에서 의미 분류를 위해 이용한 분류기는 기존의 중의성 해소 연구에서 성능이 우수하고 자주 이용되는 나이브 베이즈 분류기를 적용하였다. 중의성 해소 분류기에서 문맥 창(context window)의 크기는 기존의 중의성 해소 연구에서 성능이 가장 좋은 좌우 50바이트의 지역 문맥(local context)을 그대로 사용하였다(정영미, 이용구 2005).

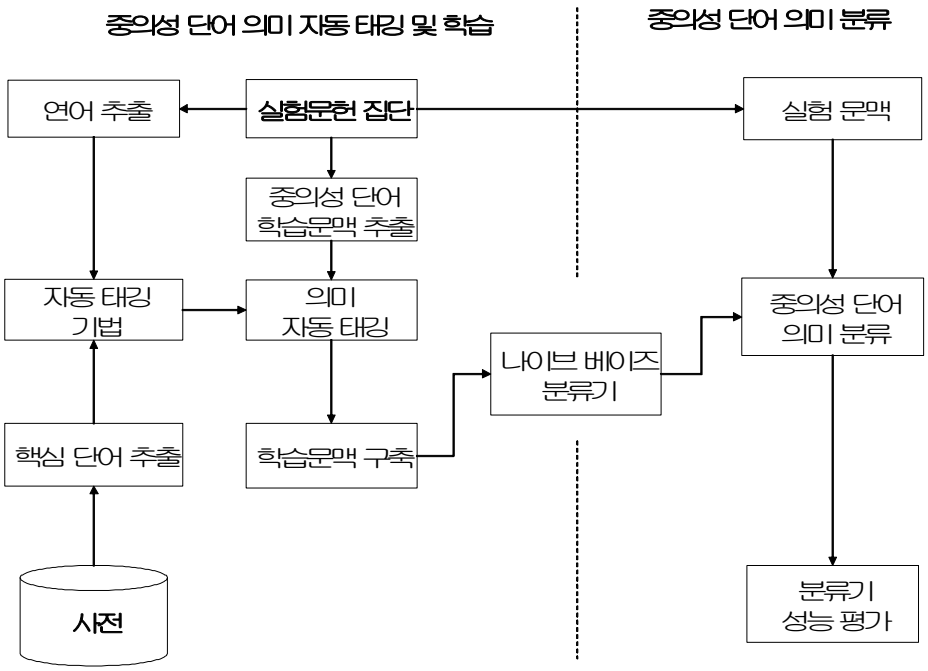
하나의 중의성 단어는 한 문헌에 여러 번 출현 할 수 있으므로 학습문헌은 문헌 단위가 아니라 문맥 단위로 계수하였다. 학습 데이터 또는 학습문맥의 수는 50, 100, 300, 600, 800개로 나누어 사전 실험한 결과, 600개나 800개에서는 어느 정도 성능이

높게 안정된 경향을 보여 더 적은 수인 600개를 학습문맥의 수로 하였다. 이 연구에서는 중의성 해소 대상 단어가 출현한 전체 문맥에서 분류기를 구축하는데 쓰인 학습문맥을 제외한 나머지를 실험문맥으로 하였다.

랜덤으로 학습문맥을 선정할 경우 랜덤 추출에서 오는 오차를 줄이기 위해 중의성 해소 실험을 다른 학습문맥 집단을 대상으로 10회 반복 실험하여 평균 성능을 산출하였다. 또한 학습문맥의 정확성을 높여 의미 분류의 성능을 최적화하기 위해 학습문맥을 구성할 때, 사전에서 핵심 단어가 추출된 정보원의 자동 태깅 성능이 좋은 것부터 순차적으로 학습문맥에 추가하였다.

넷째, 사전 추출 정보 기반 방법과 연어 공기 기반 방법의 자동 태깅 결과를 결합하여 단일 방법보다 더 좋은 성능을 가져오는지를 분석하였다. 사전 추출 정보 기반 방법에서 뜻풀이 출현 명사를 이용한 자동 태깅 결과를 제외한 나머지를 사용하였으며, 연어 공기 기반 방법에서 상위 10개의 출현빈도를 갖는 연어로 자동 태깅한 결과 전체를 이용하여 데이터 결합을 수행하였다. 사전 추출 정보 기반 기법은 뜻풀이 정보원을 제외하면 정확도가 높고 많은 수의 자동 태깅된 학습문맥을 제공하므로 의미 분류의 성능을 최적화할 수 있다. 또한 특정 의미의 출현빈도가 아주 작더라도 의미 분류가 이루어져야 하므로 포괄성이 좋고 성능이 좋은 뜻풀이 정보원이 포함된 연어 공기 기반 방법을 적용하였다.

이 연구의 전체적인 실험 과정의 개요는 <그림2>와 같다.



<그림 2> 단어 중의성 해소 실험 개요

단어의 중의성 해소 실험을 수행하기 위해서는 전문 실험집단을 이용하는 것이 바람직하다. 이 연구는 정영미와 이용구(2005)의 실험집단과 중의성 해소 대상 단어를 사용하였다. 각 중의성 단어가 갖는 의미는 <연세 한국어 사전>(1998)에서 표제어의 어깨번호로 구별되는 가장 큰 의미들로 구분하였다. 만약 중의성 대상 단어의 의미가 실험문헌에는 있지만 <연세 한국어 사전>에는 나와 있지 않다면, <표준 국어 대사전>(1991)에서 그 의미를 차용하였다. 의미를 차용한 단어와 의미번호는 '감자(2)', '경기(4)', '인도(3)', '지원(3)' 등과 같다. 실험문헌에서 출현한 9개의 중의성 단어가 갖는 사전적 의미, 실험문헌에서 출현한 의미별 빈도(의미빈도)와 출현비율 등이 <표3>에 나와 있다.

<표 3> 중의성 해소 대상 단어의 의미 및 출현빈도

단어	의미 번호	사전의 뜻풀이	출현 빈도	출현 비율
감자	1	땅속에서 자라며, 껍질이 얇고 연한 갈색이고, 속이 흰 둥근 덩어리 채소	668	59.9%
	2	기업이 자본금의 액수를 줄이는 일	447	40.1%
경기	1	운동이나 기술 등에서 재주나 능력을 서로 겨루는 것	18,105	48.0%
	2	어린아이가 경련을 일으키고 기절하는 병	33	0.00%
	3	(호황, 불황 따위의) 매매나 거래에 나타난 경제 활동의 상황	11,797	31.3%
	4	서울을 중심으로 한 가까운 주위의 지방, 경기도	7,828	20.7%
기간	1	어느 한 때로부터 다른 때까지의 시간	15,496	98.1%
	2	(어느 분야나 부문에서) 기본이나 중심이 되는 부분	307	1.9%
신병	1	['누구의 ~'꼴로 쓰이어] 법적으로 구속되어 있거나 구속할 사람	283	60.3%
	2	새로 입대한 병사	135	28.8%
	3	(겉으로 잘 드러나지 않는 어른이) 앓는 병	51	10.9%
신장	1	사람의 키	117	12.3%
	2	물체의 크기, 세력, 권리 등을 늘이고 넓게 펴는 것	314	33.0%
	3	(척추 동물의) 몸 안의 불필요한 물질을 오줌으로 배설하게 하는 구실을 하는 기관	490	51.5%
	4	건물 등을 새로 단장하는 것, 새 단장	31	3.3%
연기	1	물건이 탈 때 나는 검든가 희끄무레한 기체	763	14.8%
	2	(어떠한 일의) 정한 시기를 뒤로 미루는 것	1,931	37.5%
	3	(연극이나 영화 따위에서) 배우가 대본의 인물이나 상황에 맞춰 연극, 노래, 곡예 따위의 재주를 보이는 것/사실과 다르게 꾸며서 행동하는 것	2,441	47.4%
	4	(불교에서) 세상의 모든 것이 서로 인연이 있음	12	0.2%
인도	1	사람으로서 마땅히 지켜야 할 도리나 도덕	501	18.2%
	2	사람이 다니는 길	186	6.8%
	3	가르쳐 일깨우는 것/ 길을 안내하는 것/ (종교에서) 종교적 깨달음을 주어 그 종교에 귀의하게 하는 것	133	4.8%
	4	(사람, 물건, 권리 따위를 남에게) 넘겨주거나 넘겨 받는 것	305	11.1%
	5	아시아 남부, 인도 반도 대부분을 차지하는 공화국	1,625	59.1%
지구	1	태양을 중심으로 하여 도는 태양계의 한 행성으로서 인류가 살고 있는 땅덩어리	2,815	30.0%
	2	일정한 기준에 따라 나누고 구별한 지역	6,557	70.0%
지원	1	(정신적, 물질적, 또는 행동적으로) 돕는 것	18,623	87.3%
	2	어떤 일이나 조직에 끼이는 것을 바라고 원하는 것	2,184	10.2%
	3	지방법원이나 가정 법원이 관할 아래 있으면서 일정한 지역에 따로 떨어져 그곳의 법원 사무를 맡아 처리하는 하부 기관	513	2.4%

자동 태깅과 의미 분류의 성능을 평가하기 위한 방법으로 정확도(accuracy)를 이용하였다.

$$\text{정확도} = \frac{\text{중의성 해소에 성공한 건수}}{\text{중의성을 해소한 전체 건수}}$$

정확도는 공식에서 보듯이 중의성 해소 모형이 올바르게 의미 분류한 빈도를 중의성 해소한 전체 빈도로 나누어 산출하였다(Edmonds and Cotton 2001).

6.2 실험 결과 분석 및 평가

6.2.1 학습문맥의 자동 태깅 실험

6.2.1.1 사전 추출 정보 기반 태깅 방법의 성능 평가

(1) 정보원 유형별 자동 태깅 결과

국어사전으로부터 추출한 정보를 이용하여 앞서 기술한 몇 가지 경험규칙에 의해 실험집단의 9개의 중의성 해소 대상 단어에 대해 사전 추출 정보 기반 방법을 적용한 결과, <표4>와 같은 정보원별 자동 태깅 성능을 보였다.

<표 4> 정보원 유형에 따른 자동 태깅 성능

유형	9개 중의성 해소 대상 단어의 성능		
	전체 문맥수	성공 건수	정확도
반의어	237	195	0.8228
연어	3,229	2,931	0.9077
한자	74	74	1.0000
관련어	846	791	0.9350
동의어	2,107	1,598	0.7584
파생어	1,078	1,071	0.9935
뜻풀이어	128,520	60,810	0.5091
합 계	136,091	67,470	0.4958

정보원의 유형에 따라 자동 태깅한 문맥수와 정확도가 달라지는 것을 알 수 있다. 이러한 특성은 추후 중의성 해소 모형의 전체적인 성능을 최적화하는데 필요하다. 그 이유는 중의성 단어에 대해 어떤 유형의 정보원이 자동 태깅과 의미 분류에서 보다 높은 정확도를 가져오는지 알 수 있다면, 이를 활용하여 중의성 해소의 최종 성능을

높일 수 있기 때문이다.

먼저 정보원 유형별 문맥수를 살펴보면, 자동 태깅된 전체 문맥수 136,091건 중 94.44%에 해당하는 128,520건으로 가장 많은 수가 표제어의 뜻풀이 정보를 이용하여 추출되었음을 알 수 있다. 이는 뜻풀이 출현 명사를 이용한 자동 태깅의 성능이 중의성 해소 기법의 최종 성능에 많은 영향을 미칠 수 있음을 의미한다. 하지만 뜻풀이 정보원의 문제점은, 다른 정보원과 달리 추출된 명사가 사전 편찬자의 개념체계에 따라 달라질 수 있다는 것이다. 이 연구에서는 이러한 뜻풀이 정보원에 대한 높은 의존과 문제점을 보완하기 위한 방법으로 부가정보를 보다 적극적으로 활용할 필요가 있다.

정보원의 유형에 따른 자동 태깅 성능 중 가장 좋은 성능을 보이는 것은 한자 정보원이며, 그 다음으로 파생어, 관련어, 연어 순으로 성능이 낮아지고 있는 것을 알 수 있다. 한자 정보원의 자동 태깅 성능을 보면 전체 문맥수와 그에 따른 성공 건수는 가장 작지만 100%의 정확도를 보였다. 파생어의 경우 전체 문맥수 1,078개 중에서 7개 틀린 1,071개로 99.35%의 정확도를 보였다. 반면 뜻풀이 출현 명사의 자동 태깅 성능은 50.91%로 가장 낮았다. 따라서 자동 태깅의 성능을 높이기 위해서는 정확도가 높은 정보원인 한자, 파생어, 관련어, 연어에서 추출된 정보를 이용하여 자동 태깅을 수행하는 것이 요구된다.

(2) 자질 축소에 따른 자동 태깅 결과

이 연구에서는 경험규칙이나 통계에 의해 학습문맥을 추출하므로, 보다 좋은 중의성 해소 성능을 이끌어내기 위해 자질 선정을 수행할 필요가 있다. 자질 축소를 위해 문헌빈도를 사용하였으며, 학습문맥을 형성할 때 이용하는 정보원에 따라 다른 문헌빈도의 기준을 적용하였다.

이 연구에서 문헌빈도는 두 가지 종류를 이용하였다. 하나는 사전에서 표제어의 의미를 기술하기 위해 사용된 뜻풀이 문장을 하나의 문헌으로 간주한 문헌빈도(dicdf)와, 다른 하나는 실험집단에 나타난 문헌빈도(docdf)이다. dicdf를 통해 사전 뜻풀이에서 표제어의 의미를 기술하기 위해 자주 사용되는 고빈도어나, docdf를 통해 실험문헌에 나타난 고빈도어는 문헌의 주제적 특성을 나타내지 못하거나 여러 의미를 가진 다의어이므로 자질 축소의 대상으로 삼았다.

두 종류의 문헌빈도 중에서 먼저 docdf만을 이용하여 자질 축소를 수행한 단일 자

질 축소와, dicdf를 이용한 결과에 docdf를 이용하여 자질 축소를 수행하는 복수 자질 축소 방법을 적용하였다. 다양한 사전 실험을 수행한 결과, 최적의 자동 태깅 성능을 가져오는 자질 축소의 문헌빈도 기준은 다음과 같았다.

정보원의 유형을 고려하지 않는 전체 자동 태깅의 성능에서 docdf는 5000미만이며 dicdf는 300미만일 때 가장 좋았다. docdf는 정보원 유형에 따라 달랐는데 동의어, 반의어, 그리고 관련어의 경우 docdf를 8000미만으로 하고 연어와 뜻풀이 출현 명사에 대해서는 docdf를 5000미만으로 하였을 때 가장 좋은 성능을 가져왔다. 한자와 파생어는 앞서 살펴보았듯이 좋은 자동 태깅 성능을 보여 자질 축소를 하지 않았다. 이 문헌빈도 기준을 단일 자질 축소 방법과 복수 자질 방법 모두에 적용하였다. 다만 복수 자질 축소 방법에서 자동 태깅 성능이 낮은 뜻풀이 출현 명사에 대해서는, 이들 단어가 두 종류의 문헌빈도 기준을 동시에 만족하더라도 전체 실험문헌에서 중의성 단어와 한번도 같은 문장에서 출현하지 않은 경우 자질에서 제외하였다.

두 가지 종류의 자질 축소 방법을 적용하여 자동 태깅을 수행한 결과 <표5>와 같은 결과를 얻었다.

<표 5> 사전 추출 정보 기반 방법의 자동 태깅의 성능

단어	단일 자질 축소			복수 자질 축소		
	전체 건수	성공 건수	정확도	전체 건수	성공 건수	정확도
감자	882	870	0.9864	802	800	0.9975
경기	7,697	4,840	0.6288	6,200	4,833	0.7795
기간	2,128	1,241	0.5973	2,128	1,271	0.5973
신병	364	270	0.7418	299	265	0.8863
신장	741	552	0.7449	653	471	0.7213
연기	4,823	4,192	0.8692	4,732	4,169	0.8810
인도	4,026	2,282	0.5668	3,956	2,274	0.5748
지구	2,207	2,124	0.9624	2,207	2,124	0.9624
지원	6,171	2,171	0.3518	4,826	1,870	0.3875
합계	29,039	18,572	0.6396	25,803	18,077	0.7006

<표5>에서 보면 단일 자질 축소 실험의 경우 전체 문맥 건수인 29,039개에 대해 자동 태깅을 수행하였으며, 그 중 18,572개의 문맥을 올바른 의미로 태깅하여 중의성 해소에 성공하였다. 따라서 자동 태깅을 통한 중의성 해소 성능은 63.39%의 정확도를 보였다. 복수 자질 축소에서는 25,803개의 전체 문맥 건수에 대해 자동 태깅을 수행하여, 18,077개(70.06%)의 문맥에 대해 자동 태깅을 성공하였다. 두 자질 축소 방법을 비

교해보면 70.06%의 정확도를 보이는 복수 자질 축소 방법의 성능이 63.96%의 단일 자질 축소 방법 보다 9.54%의 성능 향상을 보이고 있다.

(3) 복수 자질 축소의 정보원 유형별 성능 평가

복수 자질 축소 방법에서 정보원 유형에 따른 자동 태깅 성능은 <표6>과 같다. <표6>은 9개의 중의성 해소 대상 단어에 대해 사전에서 추출한 정보원 유형에 따른 자동 태깅 성능을 보여주고 있다. 복수 자질 축소 방법의 자동 태깅 정확도가 0.7006로 <표3>의 자질 축소 전의 자동 태깅 성능 0.4958에 비해 41.31% 향상되었다. 정확도를 정보원의 유형별로 보면, 성능이 가장 좋은 한자와 파생어의 경우 각각 100%와 99.38%에 달하며, 연어의 경우 96.57%에 달한다. 동의어, 반의어 그리고 관련어의 경우 각각 94.3%, 82.28%, 93.5%에 달한다. 마지막으로 뜻풀이 출현 명사의 경우 전체 자동 태깅된 24,725건 중에서 82.16%를 차지하는 20,315건을 자동 태깅하였지만, 그 중 성공 건수는 12,842건수로 정확도가 63.21%에 해당한다.

<표 6> 복수 자질 축소의 정보원 유형별 자동 태깅 성능

유형	연어	한자	동의어	반의어	관련어	파생어	뜻풀이어	합계
성공 건수	1,548	74	1,556	195	791	1,071	12,842	18,077
전체 건수	1,603	74	1,650	237	846	1,078	20,315	25,803
정확도	0.9657	1.0000	0.9430	0.8228	0.9350	0.9935	0.6321	0.7006

사전 추출 정보 기반 방법에서 복수 자질 축소 방법을 사용할 때, 중의성 단어의 의미를 자동 태깅하기 위해 사전에서 추출한 정보 중 '경기'에 대해서만 살펴보면 <표7>과 같다.

연어 정보의 경우 용례에서 추출한 '월드컵'과 특정 동사와 잘 어울려 쓰이는 공기 관계 정보에서 추출한 경제활동 상황을 나타내는 '경기(3)'의 '~가 좋다/나쁘다'를 비롯하여 3건을 나타내고 있다. 한자 정보의 경우 '경기'의 네 가지 의미 모두가 다른 한자를 사용하고 실험문헌에 출현하였므로 모두 이용하였다. '경기(1)'의 동의어로 '시합'을 추출하였으며, 파생어 정보로 '경기하다'를 추출하였다.

뜻풀이어 정보의 경우, <표3>의 사전 뜻풀이에서 출현한 명사를 대상으로 자질 축소하여 자동 태깅에 이용하였다. 예를 들면 '경기(3)'의 경우 '매매/불황/호황'만 이용

하여 자동 태깅하였다. 다만 ‘기간(1)’, ‘신병(3)’, ‘인도(2)’, ‘지원(1)’, 그리고 ‘지원(2)’의 경우 뜻풀이어를 이용한 자동 태깅이 불가능하였다. 그 이유는 ‘돕는 것’과 같이 명사 추출이 어려운 짧은 문장을 이용하여 뜻풀이를 하였거나, ‘시간’, ‘병’과 같은 아주 일반적인 단어를 이용하여 뜻풀이를 기술하였기 때문이다.

<표 7> 복수 자질 축소에 사용된 정보원 유형별 사전 추출 정보(‘경기’의 경우)

단어	의미	연어	한자	동의어	반의어	관련어	파생어	뜻풀이어
경기	1	월드컵	競技	시합	–	–	경기하다	재주
	2	~를 일으키다	驚氣	–	–	–	–	경련/어린아이
	3	~가 좋다/나쁘다	景氣	–	–	–	–	매매/불황/호황
	4	–	京畿	–	–	–	–	경기도/주위

6.2.1.2 연어 공기 기반 태깅 방법의 성능 평가

연어 공기 정보를 이용한 자동 태깅 기법은 중의성 대상이 되는 단어와 사전을 통해 추출된 핵심 단어의 공통된 연어를 이용하여 중의성을 해소한다. 이 방법에서도 자질 축소를 수행하였으며, 그 기준은 사전 추출 정보 기반 방법에서 복수 자질 축소 방법을 따랐다.

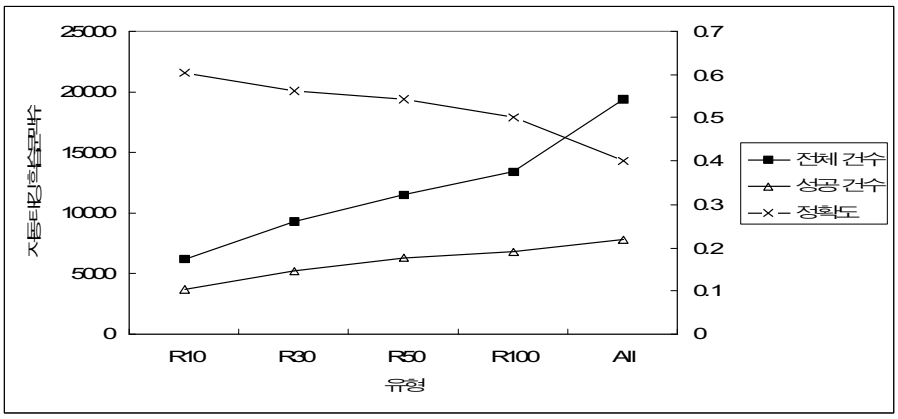
연어 공기 기반 방법을 이용하여 자동 태깅한 성능을 보면 <표8>과 같다. 표에서처럼 보다 높은 정확도를 얻기 위해, 공기한 연어의 출현빈도를 이용하여 상위 10개(R10), 상위 30개(R30), 상위 50개(R50), 상위 100개(R100), 그리고 전체 연어로 세분하여 자동 태깅의 정확도를 계산하였다.

<표 8> 연어 출현빈도에 따른 자동 태깅 성능

유형	전체 건수	성공 건수	정확도
R10	6,155	3,727	0.6055
R30	9,258	5,215	0.5633
R50	11,544	6,264	0.5426
R100	13,432	6,751	0.5026
All	19,436	7,796	0.4009

고빈도 연어만을 이용할수록 정확도가 증가하는 것을 볼 수 있으며, 구체적인 정확

도는 상위 10개의 연어를 이용한 경우(R10)가 60.55%이고, 상위 30개(R30)가 57.86%, 상위 50개(R50)가 54.26% 순으로 점차 낮아지는 것을 볼 수 있다. 자동 태깅된 전체 건수와 성공건수의 경우 정확도에 반비례하여 점차 증가하는 경향을 보였다. <표8>를 그림으로 표현하면 <그림3>과 같다.



<그림 3> 연어 출현빈도 유형에 따른 자동 태깅 성능

가장 좋은 의미 분류 성능을 가져올 수 있는 유형을 알아내기 위해 각각의 유형을 학습문맥으로 이용하여 의미 분류를 수행하였다. 그 결과는 <표12>에 나와 있다. 가장 좋은 분류 성능을 가져온 R30유형의 단어별 자동 태깅 성능을 살펴보면 <표9>와 같다. 이 표에서 주 정보원은 중의성 단어와 공기한 연어가 사전의 어느 필드에서 추출되었는지를 나타낸다. 여러 정보원 중에서 가장 많은 건 수를 차지하는 공기 연어가 추출된 정보원을 대표로 표기하였다. 대부분의 의미들이 뜻풀이어에서 공기된 연어를 통해 자동 태깅된 것을 알 수 있다. 그러나 뜻풀이에서 좋은 자질의 명사를 추출하기 힘든 '기간', '인도', 그리고 '지원'의 경우, 중의성 단어와 사전의 다른 정보원, 즉 용례 연어나 동의어, 관련어에서 공기된 연어가 추출되는 것을 볼 수 있다. 이렇게 사전의 뜻풀이가 부정확하거나 뜻풀이어를 추출하기 어려운 경우, 앞의 사전 추출 정보 기반 방법에서처럼 다른 정보원이 의미 태깅에 좋은 자질을 제공해 주는 것을 볼 수 있다.

<표 9> 연어 공기 기반 방법의 자동 태깅 성능

단어	주 정보원	전체 건수	성공 건수	정확도
감자	뜻풀이어	273	251	0.9194
경기	뜻풀이어	3,540	2,951	0.8336
기간	동의어	1,205	365	0.3029
신병	뜻풀이어	112	67	0.5982
신장	뜻풀이어	101	77	0.7624
연기	뜻풀이어	520	435	0.8365
인도	반의어	277	195	0.7040
지구	뜻풀이어	609	546	0.8966
지원	관련어	2,621	328	0.1251
합계		9,258	5,215	0.5633

연어 공기 기반 방법을 적용하였을 때 사전에서 추출한 정보원 유형에 따른 자동 태깅 성능을 살펴보면 <표10>과 같다. 정확도를 정보원의 유형별로 살펴보면 성능이 가장 좋은 반의어의 경우 92.25%에 달하며, 관련어와 동의어의 경우 각각72.17%, 73.9%에 달한다. 사전 추출 정보 기반 방법에서처럼 뜻풀이 출현 명사의 경우 전체 자동 태깅된 9,258건 중에서 90.25%를 차지하는 8,355건을 이 정보원을 이용하여 자동 태깅하였지만, 그 중 성공 건수는 4,528건수로 정확도가 54.2%에 해당한다.

<표 10> 연어 공기 기반 방법의 정보원 유형별 사전 추출 정보

유형	동의어	반의어	관련어	뜻풀이어	합계
성공 건수	402	119	166	4,528	5,215
전체 건수	544	129	230	8,355	9,258
정확도	0.7390	0.9225	0.7217	0.5420	0.5633

연어 공기 기반 방법에서 중의성 단어와 사전에서 추출된 단어 사이에 공기하는 연어를 정보원의 유형에 따라 추출한 결과 중에 '경기'만 살펴보면 <표11>과 같다. 이 표에서는 공기한 연어를 정보원 별로 나열하였으며, 나열 순서는 출현빈도 순이고, 같은 출현빈도를 가질 경우 자모순으로 상위 3개만을 제시하였다.

<표 11> 연어 공기 기반 방법의 정보원 유형별 주요 연어 정보('경기'의 경우)

단어	의미	동의어	반의어	관련어	뜻풀이어
경기	1	격투기/체스/이종	–	–	타수/매치/승패
	2	–	–	–	복식/천둥/발작
	3	–	–	–	회복세/부양책/하강
	4	–	–	–	광명시/안양시/부천시

두 가지 자동 태깅 기법의 전체적인 성능을 보면, 사전 추출 정보 기반 방법의 경우 복수 자질 축소 방법을 적용한 방법이 70.06%의 태깅 정확도를 보이고, 연어 공기 기반 방법의 경우 56.33%의 태깅 정학도를 보이고 있다. 사전 추출 정보 기반 방법이 연어 공기 정보 기반 방법보다 24.37%의 성능이 향상된 것을 알 수 있다.

6.2.2 실험문맥의 중의성 해소 실험

사전 추출 정보 기반 방법이나 연어 공기 기반 방법에서 살펴보았듯이 이용된 정보원에 따라 자동 태깅의 성능은 많이 차이를 보이고 있다. 뜻풀이 출현 명사를 이용하는 것보다 한자와 파생어를 비롯한 다른 정보원을 이용하여 자동 태깅하는 것이 더 좋은 성능을 가져왔다. 하지만 중의성 단어의 의미를 분류하는데 이러한 정보들이 항상 존재하긴 어렵다.

따라서 의미 분류를 최적화하기 위해서는 자동 태깅 성능이 높은 정보원을 먼저 이용하여 학습문맥을 구성하고, 나머지 부분은 뜻풀이를 이용한 자동 태깅한 학습문맥을 이용하였다.

먼저 사전 추출 정보 기반 방법의 의미 분류 성능을 보면 다음과 같다. 단일 자질 축소 방법을 이용하여 학습문맥을 구축하고 이를 이용한 분류기가 중의성을 해소해야 할 전체 94,851건수 중 53,010건을 올바르게 해소하여 55.89%의 성능을 보였으며, 복수 자질 축소 방법의 분류기는 64,604건을 올바르게 해소하여 68.11%의 성능을 보였다. 후자가 의미 분류에서 21.86% 정도의 향상된 성능을 보이고 있다. 다시 말하면 자동 태깅된 학습문맥의 성능 차이가 추후에 중의성 해소용 분류기 성능에 많은 영향을 미친다고 할 수 있다. 따라서 성능이 좋은 양질의 학습문맥의 구축이 절실히 필요하다.

연어 공기 기반 방법의 유형에 따른 의미 분류 성능은 <표12>와 같다. 표를 보면

출현빈도 상위 30개 연어만 대상으로 한 R30 유형이 가장 좋은 성능을 보여 주고 있다.

<표 12> 연어 공기 기반 방법의 유형에 따른 의미 분류 성능

유형	전체 건수	성공 건수	정확도
R10	94,851	57,201	0.6031
R30	94,851	58,891	0.6209
R50	94,851	56,871	0.5996
R100	94,851	56,916	0.6001
All	94,851	53,218	0.5611

R30 유형의 연어 공기 기반 방법이 태깅한 학습문맥을 이용한 분류기의 단어별 중의성 해소 성능은 총 94,851개의 실험문맥 중에서 58,891개를 올바른 의미로 중의성을 해소하여 정확도 62.09%에 달한다. 두 가지 자동 태깅 기법을 적용하여 생성된 학습문맥으로부터 구축한 분류기의 의미 분류 성능을 보면, 사전 추출 정보 기반 방법이 9.7%의 성능 향상이 되었다.

6.2.3 자동 태깅 기법의 결합 실험

6.2.3.1 결합 기법의 자동 태깅 평가

두 중의성 해소 모형의 자동 태깅 결과를 결합하기 위해 사전 추출 기반 방법에서는 뜻풀이를 제외한 모든 정보원을 사용하였으며, 연어 공기 기반 방법에서는 출현빈도 상위 10개 연어만 대상으로 한 R10 유형의 전체 정보원을 이용하였다.

두 방법의 데이터 결합 후 자동 태깅 결과는 총 11,352건 중 8,637건을 올바르게 중의성을 해소하여 복수 자질 축소를 이용한 사전 추출 정보 기반 방법의 70.06%에 비해 8.59% 향상된 76.09%를 보였다.

결합 기법에서 정보원 유형에 따른 자동 태깅 성능을 비교해 보면 <표13>과 같다. 연어, 한자, 파생어 정보원의 성공 건수, 전체 건수, 그리고 정확도는 사전 추출 모형의 성능과 같다. 뜻풀이 정보원의 경우 연어 공기 기반 방법에서 유형 R10의 것만을 사용하였다. 동의어, 반의어, 관련어 정보원은 사전 추출 정보 기반 방법과 연어 공기 기반 방법의 자동 태깅 결과가 결합되어, 보다 많은 수의 자동 태깅 건수가 결합 기

법에 추가 되었다.

<표 13> 데이터 결합 기법의 정보원 유형에 따른 자동 태깅 성능

유형	연어	한자	동의어	반의어	관련어	파생어	뜻풀이어	총합계
성공 건수	1,548	74	1,856	290	907	1,071	2,891	8,637
전체 건수	1,603	74	2,064	336	978	1,078	5,219	11,352
정확도	0.9657	1.0000	0.8992	0.8631	0.9274	0.9935	0.5539	0.7608

세 가지 자동 태깅 기법간의 전체적인 성능을 비교한 <표14>를 보면, 연어 공기 기반 방법 56.33%, 사전 추출 정보 기반 방법 70.06%, 결합 기법 76.08% 순으로 나타났다. 건수로 보면 사전 추출 정보 기반 방법이 전체 건수 25,803건, 성공 건수 18077건으로 가장 많고 다음으로 결합 기법이 11,352건 중 8,637건으로 약간 낮지만, 가장 높은 정확도로 인해 더 풍부한 자동 태깅 데이터를 제공하는 것을 알 수 있다.

<표 14> 세 기법의 자동 태깅 성능 비교

자동 태깅 방법	전체 건수	성공 건수	정확도
사전 추출 정보 기반 방법	25,803	18,077	0.7006
연어 공기 기반 방법	9,258	5,215	0.5633
결합 기법	11,352	8,637	0.7608

6.2.3.2 결합 모형의 의미 분류 성능 평가

결합 기법을 적용하여 자동 태깅하고 이를 이용한 의미 분류한 결과를 보면 총 94,851건 중 72,237건을 올바르게 중의성을 해소하여 76.16%의 정확도를 보였다. 이 모형에서는 다른 두 모형과 비교하여 '감자', '신장', '지구'에서 그리 큰 차이를 보이지 않지만, '경기', '기간'과 '지원'의 경우에는 훨씬 좋은 성능을 보였다. 이들 단어의 뜻풀이 정보원을 이용한 자동 태깅 건수가 줄어들어 자동 태깅 오류가 적어지고, 다른 정보원에서는 더 풍부한 자동 태깅 데이터를 제공하여 의미 분류에 필요한 자질이 더 좋아졌기 때문인 것으로 보인다.

9개의 중의성 단어에 대한 의미 분류 성능을 모형별로 살펴보면 <표15>와 같다. 결합 모형의 중의성 해소 성능이 76.16%로 사전 추출 기반 방법을 적용한 모형의 68.11%, 연어 공기 기반 방법을 적용한 모형의 62.09%에 비해 각각 11.82%, 22.66% 성능 향상이 되었다.

<표 15> 세 중의성 해소 모형의 의미 분류 성능 비교

중의성 해소 모형	전체 건수	성공 건수	정확도
사전 추출 정보 기반 모형	94,851	64,604	0.6811
연어 공기 기반 모형	94,851	58,891	0.6209
결합 모형	94,851	72,237	0.7616

7. 결론

이 연구에서는 문헌의 내용을 대표하는 색인어의 중의성을 해소하기 위한 방안으로 수작업 태깅없이 기계가독형 사전을 이용하여 자동으로 의미를 태깅하는 단어 중의성 모형에 대해 제시하였다.

단어 중의성 해소 실험은 크게 실험문헌에 나타난 중의성 단어에 대해 자동 태깅 기법을 이용하여 의미를 자동으로 태깅하여 학습문맥을 구축하는 단계와, 구축된 학습문맥을 이용하여 분류기를 구축하고 의미 분류를 수행하는 단계로 나누어 수행하였다. 자동 태깅 기법은 중의성 단어가 사전에서 추출한 핵심 단어와 문맥에서 공기하면 이 핵심 단어가 추출된 의미로 자동 태깅하는 사전 추출 정보 기반 방법과, 중의성 단어와 핵심 단어 사이에 공통된 연어가 존재하면 이 핵심 단어가 추출된 의미로 자동 태깅하는 연어 공기 기반 방법을 적용하였다. 그리고 분류기로는 나이브 베이즈 분류기를 이용하였다.

실험 과정을 통해 나타난 결과는 다음과 같다.

첫째, 사전에서 추출한 정보원의 유형에 따라 자동 태깅의 성능 차이를 보였다. 뜻풀이 출현 단어, 관련어, 연어, 파생어, 한자 순으로 더 많은 수의 학습문맥을 제공하여 중의성 단어에 대한 의미 식별력이 컸으며, 정확도 측면에서는 한자, 파생어, 연어, 관련어, 뜻풀이 출현 단어 순으로 더 좋은 자동 태깅 성능을 보였다.

둘째, 표제어의 뜻풀이를 기술할 때 자주 사용되는 고빈도 단어의 경우 주제적 특성이 없으므로, 이들을 그대로 자동 태깅에 이용할 경우 오류를 가져올 확률이 높다.

따라서 사전에서 추출한 핵심 단어에 대해 자질 축소를 한 결과, 실험문헌의 문헌빈도만 이용한 자질 축소 방법 보다 사전의 뜻풀이를 문헌으로 간주한 문헌빈도를 같이 적용한 축소 방법의 태깅 정확도가 9.54% 향상된 성능을 가져왔다.

셋째, 연어 공기 기반 방법의 경우 공기 연어의 출현빈도 순으로 상위 N(10, 30, 50, 100, All)개로 유형을 나누어 성능의 차이를 분석하였다. 상위 30개의 공기된 연어를 자동 태깅에 사용한 유형의 태깅 정확도가 56.33%로 가장 좋은 성능을 보여 주었다. 따라서 자동 태깅을 위한 두 중의성 해소 모형의 성능을 비교하면, 복수 자질 축소 방법을 적용한 사전 추출 정보 기반 방법의 태깅 정확도가 70.06%로, 연어 공기 기반 방법의 56.33%보다 24.37% 향상된 성능을 가져왔다.

넷째, 사전 추출 정보 기반 방법과 연어 공기 기반 방법을 통해 자동 태깅한 학습문맥을 이용하여 분류기를 구축하고 중의성 단어에 대해 의미 분류를 수행한 결과, 복수 자질 축소에 기반한 사전 추출 정보 기반 방법의 분류기는 68.11%의 분류 정확도를 보였으며, 다른 분류기는 62.09%의 분류 정확도를 보여 사전 추출 정보 기반 방법을 적용한 분류기가 9.7% 향상된 의미 분류 성능을 나타냈다. 이는 사전 추출 정보 기반 방법에서 자동 태깅한 학습문맥의 자질이 더 우수하기 때문인 것으로 보인다.

다섯째, 사전 추출 정보 기반 방법과 연어 공기 기반 방법의 자동 태깅 결과를 결합한 모형이 단일 모형보다 더 좋은 성능을 가져왔다. 결합 모형의 태깅 성능은 복수 자질 축소를 이용한 사전 추출 정보 기반 방법의 70.06%에 비해 8.59% 향상된 76.09% 태깅 정확도를 보였다. 또한 결합된 자동 태깅 결과를 학습문맥으로 이용한 의미 분류 성능은 76.16%의 분류 정확도로 사전 추출 정보 기반 방법의 68.11%, 연어 공기 기반 방법의 62.09%에 비해 각각 11.82%, 22.66%의 성능 향상이 되었다.

이 연구에서는 의미 태깅된 언어자원의 제약으로 사전만을 이용하여 중의성을 해소할 경우 단일 모형에서는 복수의 자질 축소 기준을 적용한 사전 추출 정보 기반 방법을 이용하여 중의성 단어의 의미를 분류하는 방법이 가장 좋은 성능을 보인 것으로 나타났다. 그리고 단일 모형과 이들의 결과를 결합한 모형 중에서는 결합 모형이 더 좋은 성능을 보였다. 중의성 해소 성능은 태깅 정확도가 높아질수록 분류 정확도가 높아지는 경향을 보이지만 두 정확도에 정비례하여 증가하지 않는 것을 알 수 있는데, 이는 자동 태깅 방법이 양질의 자질 집합을 형성하느냐 그렇지 않느냐의 문제로 보인다.

참고문헌

국립국어연구원. 1999.『표준국어대사전』. 서울: 두산동아.

연세대학교 언어정보개발연구원. 1998.『연세한국어사전』. 서울: 두산동아.

정영미, 이용구. 2005.『정보검색 성능 향상을 위한 단어 중의성 해소모형에 관한 연구』. 정보관리학회지, 22(2): 125-145.

Agirre, E. and G. Rigau. 1996. "Word sense disambiguation using conceptual density." *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96),* 16-22.

Dagan, I. and A. Itai. 1994. "Word sense disambiguation using a second language monolingual corpus." *Computational Linguistics*, 20(4): 563-596.

Edmonds, P. and S. Cotton. 2001. "SENSEVAL-2: Overview." *Proceedings of the 2nd International workshop on Evaluating Word Sense Disambiguation Systems*, 1-5.

Gale, W. A. 1992. "A Method for Disambiguating Word Sense in a Large Corpus." *Computers and the Humanities*, 26: 415-439.

Gale, W., K. W. Church, and D. Yarowsky. 1992a. "Estimating upper and lower bounds on the performance of word sense disambiguation Programs." *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, 249-256.

Gale, W., K. W. Church, and D. Yarowsky. 1992b. "One sense per discourse." *Proceedings of the Speech and Natural Language Workshop*, 233-237.

Gale, W., K. W. Church, and D. Yarowsky. 1993. "A method for disambiguating word senses in a large corpus." *Computers and the Humanities*, 26(5-6): 415-439.

Kilgarriff, A., and J. Rosenzweig. 2000. "English Framework and Results." *Computers and the Humanities* 34(1-2): 1-13.

Krovetz, R., and W. B. Croft. 1989. "Word sense disambiguation using machine-readable dictionaries." *Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *SIGIR'89*

127-136.

Lesk, M. 1986. "Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone." *Proceedings of the 1986 SIGDOC Conference*, 24-26.

Manning, C. D. and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Resnik, P. 1995. "Disambiguating Noun Groupings with Respect to WordNet Senses." *Proceedings of the Third Workshop on Very Large Corpora*, 54-68.

Stevenson, M. 2003. *Word Sense Disambiguation: the Case for Combinations for Knowledge Sources*. California: CSLI Publications.

Yarowsky, D. 1992. "Word sense disambiguation using statistical models of Roget's categories trained on large corpora." *Proceedings of the 14th International Conference on Computational Linguistics, COLING'92*, 454-460.

Yarowsky, D. 1993. "One sense per collocation." *Proceeding of ARPA Human Language Technology Workshop*, 266-271.

Yarowsky, D. 1995. "Unsupervised word sense disambiguation rivaling supervised methods." *Annual Meeting of the ACL Archive Proceedings of the 33rd conference on Association for Computational Linguistics*, 189-196.