정보검색 성능 향상을 위한 단어 중의성 해소 모형에 관한 연구*

Improving the Retrieval Effectiveness by Incorporating Word Sense Disambiguation Process

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

초 록

이 연구에서는 문헌 및 질의의 내용을 대표하는 주제어의 중의성 해소를 위해 대표적인 지도학습 모형인 나이브 베이즈 분류기와 비지도학습 모형인 EM 알고리즘을 각각 적용하여 검색 실험을 수행한 다음, 주제어의 중의성 해소를 통해 검색 성능의 향상을 가져올 수 있는지를 평가하였다. 실험문헌 집단은 약 12만 건에 달하는 한국어 신문기사로 구성하였으며, 중의성 해소 대상 단어로는 한국어 동형이의어 9개를 선정하였다. 검색 실험에는 각 중의성 단어를 포함하는 18개의 질의를 사용하였다. 중의성 해소 실험 결과 나이브 베이즈 분류기는 최적의 조건에서 평균 92%의 정확률을 보였으며, EM 알고리즘은 최적의 조건에서 평균 67% 수준의 클러스터링 성능을 보였다. 중의성 해소 알고리즘을 통합한 의미기반 검색에서는 나이브 베이즈 분류기 통합 검색이 약 39.6%의 정확률을 보였고, EM 알고리즘 통합 검색이 약 36%의 정확률을 보였다. 중의성 해소 모형을 적용하지 않은 베이스라인 검색의 정확률 37%와 비교하면 나이브 베이즈 통합 검색은 약 7.4%의 성능 향상률을 보인 반면 EM 알고리즘 통합 검색은 약 3%의 성능 저하율을 보였다.

ABSTRACT

This paper presents a semantic vector space retrieval model incorporating a word sense disambiguation algorithm in an attempt to improve retrieval effectiveness. Nine Korean homonyms are selected for the sense disambiguation and retrieval experiments. The total of approximately 120,000 news articles comprise the raw test collection and 18 queries including homonyms as query words are used for the retrieval experiments. A Naive Bayes classifier and EM algorithm representing supervised and unsupervised learning algorithms respectively are used for the disambiguation process. The Naive Bayes classifier achieved 92% disambiguation accuracy, while the clustering performance of the EM algorithm is 67% on the average. The retrieval effectiveness of the semantic vector space model incorporating the Naive Bayes classifier showed 39.6% precision achieving about 7.4% improvement. However, the retrieval effectiveness of the EM algorithm-based semantic retrieval is 3% lower than the baseline retrieval without disambiguation. It is worth noting that the performances of disambiguation and retrieval depend on the distribution patterns of homonyms to be disambiguated as well as the characteristics of queries.

키워드: 정보검색, 중의성 해소, 나이브 베이즈 분류기, EM 알고리즘, 클러스터링
information retrieval, word sense disambiguation, Naive Bayes classifier, EM algorithm, clustering, retrieval effectiveness

---

* 이 논문은 2003년도 한국학술진흥재단의 지원에 의하여 연구되었음.(KRF-2003-041-H00024)
** 연세대학교 문헌정보학과 교수(ymchung@yonsei.ac.kr)
*** 연세대학교 시간강사(yglee@yonsei.ac.kr)
■ 논문접수일자 : 2005년 5월 20일
■ 게재확정일자 : 2005년 6월 20일

1. 서 론

1950년대 말부터 시작된 정보검색 관련 연구는 1990년대 이후 온라인 접근이 용이한 전문(full-text) 데이터베이스의 증가와 인터넷의 활성화로 인하여 새로운 국면에 접어들었다. 특히 전문을 대상으로 한 자연어 키워드 검색 방법이 보편적으로 사용되고 있는 현재의 환경에서는 텍스트로부터 정확한 키워드를 색인어로 추출하기 위해서 의미분석 수준의 자연언어 처리 기법이 요구되고 있다.

전문을 대상으로 한 정보검색은 자연언어의 모호성 때문에 여러 가지 문제를 야기할 수 있다. 특히 부적절한 색인어 및 탐색어의 선정은 검색 성능을 저하시키는 원인이 되고 있으며, 동형이의어(homonym/homograph)나 다의어(polysemy)와 같은 중의성을 지닌 단어를 탐색어로 사용할 경우 이용자가 원하지 않는 의미로 이 단어가 사용된 문헌을 잘못 검색하는 결과를 가져올 수 있다. 예컨대 "배"라는 동형이의어가 어떤 문맥에서 사용되었는가에 따라 과일, 교통수단, 신체일부 등 여러 의미를 갖기 때문에 문맥을 고려하지 않고 단순히 이 질의어를 탐색어로 하여 문헌을 검색할 경우 이 단어를 포함하는 모든 문헌이 검색되는 결과를 가져온다. 또한 "국회가 오랫동안 문을 닫고 있다"라는 문장에서 "문"은 동형이의어는 아니지만 문맥에 따라 단순히 물리적인 형태의 문을 의미하는 것이 아니라 다른 의미를 가진다는 것을 알 수 있다. 영어의 경우에는 다의어를 많이 찾아볼 수 있는데 예컨대 WordNet에서 "country"라는 단어가 명사로 사용된 경우를 보면 "territory occupied by a nation", "the people who live in a nation or country" 등 모두 다섯 가지 다른 의미를 갖는 것으로 나타나 있다.

따라서 여러 가지 의미를 갖는 단어가 특정 문맥 속에서 어떤 의미로 사용되었는가를 판단하는 단어의 중의성 해소(WSD : word sense disambiguation)를 통해 검색 성능의 향상을 도모할 수 있을 것이다. 단어의 중의성 해소는 중의성을 띄는 단어가 출현하였을 때 그 단어가 어떤 의미로 사용되었는가를 식별하는 일(sense discrimination)을 의미한다. 이러한 의미 식별은 특정 단어가 문헌집단에서 두 번 이상 출현했을 때 이들이 같은 의미로 쓰였는지 여부를 결정함으로써 단어의 의미를 범주화하는 것과 같다(Schütze 1998).

이 연구에서는 문헌 및 질의의 내용을 대표하는 주제어의 중의성 해소를 위해 대표적인 지도학습 모형인 나이브 베이즈 분류기(Naive Bayes classifier)와 비지도학습 모형인 EM 알고리즘을 각각 적용하여 검색 실험을 수행한 다음, 주제어의 중의성 해소를 통해 검색 성능의 향상을 가져올 수 있는지를 평가하였다.

2. 단어 중의성 해소와 정보검색

단어 중의성 해소는 일찌기 1950년대 이래로 기계번역과 같은 자연언어 처리 분야에서 관심의 대상이 되어 왔다. 단어 중의성 해소는 다양한 응용 시스템에서 요구되는 자연언어 처리 과정에서 대개 중간 단계의 작업(intermediate task)으로 수행되고 있다(Ide and Veronis 1998). 즉 중의성 해소 작업은 그 자체가 최종 목표가 되는 것이 아니라 그 이상의 단계를 수

행하기 위해 필요한 과정이라고 볼 수 있다.

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

말뭉치 기반 기법은 다시 의미 태그가 부착된 말뭉치를 이용하는 지도학습 알고리즘(supervised learning algorithm)과 의미 태그가 부착되지 않은 말뭉치를 이용하는 비지도 학습 알고리즘(unsupervised learning algorithm)으로 구분된다. 지도학습 알고리즘을 사용할 경우의 문제점은 학습용 말뭉치를 수작업으로 의미 태깅하는 일이 쉽지 않으며, 미리 의미가 태깅되어 있는 말뭉치도 극소수라는 점이다(Stevenson 2003; Ide and Veronis 1998). 실제 중의성 해소 실험에서 지도학습 알고리즘의 성능이 비지도학습 알고리즘의 성능보다 우수하게 나타나고 있다(Gale, Church, and Yarowsky 1993; Schütze 1998; Levinson 1999). 그러나 부트스트래핑(bootstrapping) 기법을 사용한 Yarowsky(1995)의 연구에서는 비지도학습 알고리즘이 지도학습 알고리즘에 버금가는 우수한 중의성 해소 성능을 보인 바 있다.

1990년대 이후 단어 중의성 해소 기법을 정보검색과 결합한 연구들이 수행되었으나 몇몇 연구들을 제외하고는 대부분의 연구에서 주목할 만한 성능 향상을 가져오지 못한 것으로 나타나 있다(Krovetz and Croft 1992; Voorhees 1993; Sanderson 1994; Schütze and Pedersen 1995; Schütze 1998; Stokoe, Oakes, and Tait 2003).

단어의 중의성 해소 기법을 정보검색 환경에 적용할 경우 다음의 몇 가지 사항을 고려할 필요가 있다. 첫째, 단어 중의성 해소 알고리즘의 정확성이다. 이것은 중의성 해소 모형의 평가를 의미하며, 중의성 단어가 특정 문헌에서 어떤 의미로 사용되었는지를 얼마나 정확하게 예측하는가를 측정하는 것이다. 선행연구(Sanderson 1994, 2000)에서도 중의성 해소 알고리즘의 정확성이 낮을 때에는 중의성 해소가 정보검색 성능 향상에 크게 도움이 되지 못한다는 것을 지적하고 있다. 단어 중의성 해소 기법을 적용하여 단어 의미를 잘못 해석할 경우 그 결과가 검색에 직접적으로 반영되기 때문에 오히려 검색성능을 저하시키는 요인으로 작용할 것이기 때문이다. 따라서 중의성 해소 알고리즘의 오분석 결과를 어떻게 정보검색에 반영해야 할 것인지는 해결해야 할 과제이다.

둘째, 웹과 같이 전문적이기 보다는 일반적인 용어가 질의어로 많이 사용되며, 질의의 길이가 짧은 웹과 같은 검색 환경에서의 중의성 해소 효과가 크다는 점이다. 웹 검색엔진의 탐색 관련 특성에 관한 연구에서 AlltheWeb은 질의어의 수가 1-3개인 질의가 전체 질의의 84%에 달하는 것으로 분석되었으며(Jansen and Spink 2005), Excite는 하나의 질의가 평균 2.2개의 질의어로 구성된 것으로 나타났다

(Jansen, Spink, and Saracevic 2000).

셋째, 정보검색 시스템에서 중의성 단어의 의미를 식별하기 위해 어떤 정보원을 이용할 것인가 하는 점이다. 이것은 검색 대상 문헌들이 갖는 주제적 특성에 따라 중의성 해소에 이용할 수 있는 정보원이 제한될 수 있음을 의미한다. 즉 검색 대상 문헌이 특수한 주제 분야이거나 전문화된 분야라면 외부 지식 정보원의 이용이 용이하지 않다. 왜냐하면 특수 주제 분야의 경우 이미 발행된 사전이나 백과사전 등의 자원을 활용하기 어렵기 때문이다. 이러한 경우에는 지식기반 기법 보다는 말뭉치 기반 중의성 해소 기법을 적용하는 것이 바람직하다.

3. 실험 설계

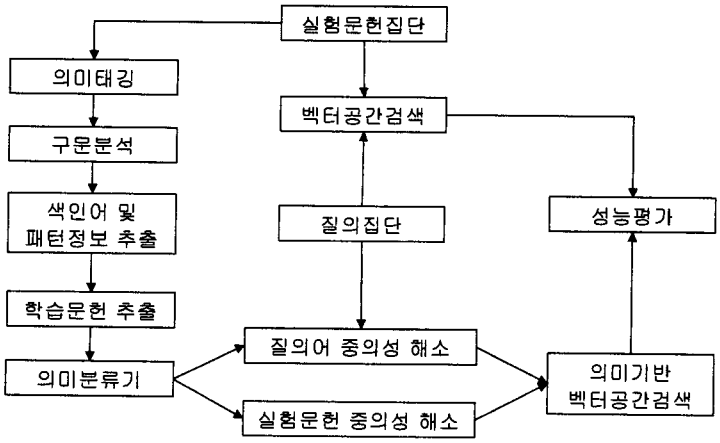
3. 1 실험 개요

이 연구에서는 단어의 중의성이 정보검색의 성능을 저하시킬 수 있다고 보고, 질의 및 문헌에 출현하는 중의성 단어의 올바른 의미를 식별한 후 검색할 경우 검색 성능이 어느 정도 향상되는가를 실험을 통해 파악하였다. 먼저 중의성 해소 실험을 통해 질의 및 검색 대상 문헌에 출현한 중의성 단어의 의미를 식별한 후, 의미기반 검색 모형을 사용하여 검색 실험을 수행하였다.

중의성 해소 실험은 지도학습 기법과 비지도학습 기법을 각각 적용하여 수행하였다. 지도학습 기법으로는 연어(collocation)를 이용한 알고리즘과 나이브 베이즈 분류기를 사용하였으며, 비지도학습 기법으로는 EM 알고리즘을 사용하였다.

중의성 해소 실험에서 최적의 성능을 보이는 조건을 파악하기 위해 중의성 해소 대상 단어가 출연한 문맥창(context window)의 크기와 학습문맥의 수를 달리하여 실험하였다. 문맥창의 크기는 좌우 3단어, 한 문장, 좌우 50바이트 등 세 가지 지역 문맥(local context)과 한 문헌의 텍스트 전체를 대상으로 하는 전역 문맥(global context)을 사용하였다.

학습문맥의 수는 대부분의 범주화 실험에서 대체로 학습문헌와 실험문헌의 비율을 3대 1로 한다는 점을 고려하여 결정하였다. 문헌을 특정 범주로 할당하는 범주화 실험에서는 문헌 단위로 학습을 하게 되지만 중의성 해소 실험에서는 중의성 단어가 대상이므로 학습 단위가 하나의 문헌이 아니라 단어의 출현 문맥이 된다. 다시 말해 하나의 중의성 단어는 한 문헌에 여러 번 출현할 수 있으므로 학습문헌은 문헌 단위가 아니라 문맥 단위로 계수하였다. 따라서 중의성 단어의 전체 출현빈도와 중의성 해소 모형의 성능을 고려하였을 때 학습문맥과 실험문맥의 비율을 3대 1로 하였다. 대부분의 중의성 단어에 대해 600개의 학습문맥과 200개의 실험문맥을 사용하였으며, 전체 출현 문맥이 800개가 넘지 않는 '신병'에 대해서는 300개의 학습문맥과 100개의 실험문맥을 적용하였다. 학습문맥과 실험문맥은 전체 실험문헌 집단에서 랜덤하게 추출하였기 때문에 오차를 줄이기 위해 각 중의성 해소 알고리즘을 다른 학습문맥 집단과 실험문맥 집단을 대상으로 30회 반복해서 실험하였다. 각 반복 실험에 대해 성능을 체크한 결과 대략 15회에 도달했을 때 성능에 큰 변화를 보이지 않았으므로 각 중의성 해소 알고리즘을 15번 돌린 후 평균 성능을 산출하였다.



〈그림 1〉 중의성 해소 및 의미기반 검색 실험 개요

검색 실험은 먼저 의미기반 검색 모형에서 사용할 의미 분류기를 학습을 통해 구축한 다음 수행하였다. 비교 대상이 될 베이스라인 검색 실험에서는 전통적인 벡터공간 모형(vector space model: VSM)을 사용하였으며, 중의성이 해소된 후 적용한 의미기반 벡터공간 검색 모형(semantic vector space model: SVSM)은 전통적인 벡터공간 검색 기법이 검색한 상위 1,000건의 문헌을 대상으로 적용하였다.

이 연구의 실험 개요는 〈그림 1〉과 같다.

3. 2 실험문헌 집단

질의어의 중의성 해소를 통해 검색 성능을 높이고자 한 대부분의 선행연구들은 CACM, Cranfield, TREC 등의 검색용 실험집단을 사용하여 실험을 수행하였다. 이러한 검색용 실험집단을 살펴보면 문헌에 출현한 단어들 가운데 중의성을 띠는 단어가 많지 않으며, 질의어의 경우도 마찬가지이다. 또한 이러한 연구들은 대개 질의의 길이가 길기 때문에 이미 질의 안에서 중의성 단어의 중의성이 어느 정도 해소되는 효과를 갖는다. 따라서 이러한 실험집단을 대상으로 한 검색 실험에서는 단어 중의성 해소 기법을 도입한다고 해도 단어의 중의성 해소가 실제로 검색 성능에 어떠한 영향을 미치는지를 파악하기가 어렵다.

이 연구에서는 기존의 연구들과 달리 단어 중의성 해소 기법이 실제 검색 성능에 어느 정도의 영향을 미치는지 파악하고자 새로운 실험집단을 구축하였다. 실험문헌 집단은 중의성을 띠는 단어를 다수 포함하도록 하였고, 질의도 충분한 중의성을 갖도록 구성하였으며, 이들을 대상으로 중의성 해소 실험과 검색 실험을 수행하였다. 실험문헌 집단은 국내 주요 일간지인 조선일보, 동아일보, 한겨레신문의 2004년 기사 전체를 포함하도록 구축하였다. 실험문헌 집단은 〈한국언론재단〉에서 운영 중인 뉴스검색 데이터베이스인 카인즈(KINDS)로부터 확보하였다.

실험문헌 집단을 구성하는 기사 수는 총 127,641건에 달하며, 기사당 평균 어절 수는 약

209개, 기사당 평균 색인어 수는 154개인 것으로 나타났다. 실험문헌 집단에 부여된 고유 색인어의 수는 82,133개이다. 실험문헌으로부터 색인어와 연어를 추출하기 위하여 "21세기 세종계획"에서 제공되는 "지능형 형태소 분석기 2.0"을 이용하여 구문분석을 수행하였다.

3. 3 중의성 해소 대상 단어 및 실험질의 집단

중의성 해소 연구에서는 의미 태깅한 학습 말뭉치 구축의 어려움 때문에 대개 10개 미만의 중의성 단어를 실험 대상으로 삼고 있다. 이 연구에서는 질의어로 사용할 수 있는 한국어 동형이의어 9개를 선정하여 중의성 해소 대상으로 삼고, 각 동형이의어를 포함하는 질의를 2개씩 생성하여 모두 18개의 질의를 검색 실험에 사용하였다.

질의어 및 색인어로 사용될 9개의 동형이의어에 대해 실험문헌 집단에 포함된 모든 기사를 대상으로 의미 태깅 작업을 수행하였다. 〈표 1〉은 각 동형이의어가 출현한 기사의 수와 실제 출현빈도를 보여 준다.

〈표 1〉 동형이의어 출현기사 수 및 총 출현빈도

단어	출현기사 수	총 출현빈도
감자	622	1,115
인도	2,022	2,750
경기	18,484	37,730
지구	4,017	9,372
기간	11,255	15,803
신장	703	952
신병	360	469
연기	3,227	5,147
지원	12,577	21,320

각 동형이의어가 갖는 의미는 〈금성판 국어대사전〉에서 표제어의 어깨번호로 구별되는 가장 큰 의미들로 구분하였다. 〈표 2〉에는 각 동형이의어가 갖는 사전적 의미와 실험문헌 안에서 각 의미로 출현한 빈도(의미빈도)와 출현비율이 나와 있다. 동형이의어에 따라 의미의 분포 패턴이 다른 것을 볼 수 있다. 예컨대 '감자'나 '경기'는 여러 의미가 고르게 사용되고 있는 반면에 '기간'이나 '지원'은 한 가지 의미로 사용이 집중되어 있는 것을 볼 수 있다. 5개의 의미를 갖는 '인도'의 경우에는 한 가지 의미가 50%를 넘게 차지하고 있다.

질의를 구성하는 질의어의 수는 신문기사나 웹 페이지를 검색하는 일반 이용자들의 탐색 행태를 고려하여 결정하였다. 동형이의어 하나만을 질의어로 사용할 경우에는 중의성을 해소할 수 있는 방법이 없으므로 질의어가 하나인 질의는 제외하였다. 실제로는 더 많은 수의 후보 질의를 가지고 일차적으로 기사를 검색하여, 검색된 기사들에 각 동형이의어가 다양한 의미로 출현하였는지를 카인즈 검색 결과 확인한 다음 최종 질의를 선정하였다. 즉, 실험문헌 집단의 주제어 가운데 중의성을 갖는 단어들을 먼저 핵심 질의어로 추출한 후에 각 질의어가 다른 질의어와 결합되어도 중의성을 갖도록 질의를 구성하였다. 예를 들어 '감자'의 경우, 다른 질의어로 '가격'과 '동향'의 두 단어가 결합되어 '감자 가격'과 '감자 가격 동향'의 질의가 만들어졌다. 이들 질의는 두세 단어가 결합되어도 검색 결과 여전히 모호성을 갖고 있음을 알 수 있다.

〈표 3〉에는 검색 실험에 사용된 18개의 질의와 각 질의의 의미가 나와 있다. 검색 실험 평가에서는 질의에서 정의한 의미로 동형이의어가 사용되었을 경우 적합한 것으로 판정하였다.

〈표 2〉 중의성 해소 대상 단어의 의미 및 출현빈도

단어	의미번호	사전상의 의미	출현빈도	출현비율
감자	1	식용 식물	668	59.9%
	2	자본금의 액수를 줄이는 것	447	40.1%
인도	1	사람이 다니는 길	186	6.8%
	2	인간으로서 마땅히 지켜야 할 도리	501	18.2%
	3	국가명(India)	1,625	59.1%
	4	남에게 넘겨줌	305	11.1%
	5	길을 안내함	133	4.8%
경기	1	매매나 거래 따위에 나타난 경제활동 상황	11,797	31.3%
	2	서로 겨룸	18,105	48.0%
	3	행정구역명 : 京畿	7,828	20.7%
지구	1	인류가 살고 있는 천체	2,815	30.0%
	2	일정한 구역	6,557	70.0%
기간	1	어느 일정한 시간 동안	15,496	98.1%
	2	어떤 조직이나 이룬 것 가운데 중심이 되는 것	307	1.9%
신장	1	내장의 하나	490	51.5%
	2	키를 일컬음	117	12.3%
	3	길게 늘임	314	33.0%
	4	새로 단장함 (예:신장개업)	31	3.3%
신병	1	당사자의 몸	283	60.3%
	2	새로 입대한 병사	135	28.8%
	3	몸의 병(身病)	51	10.9%
연기	1	물건이 탈 때 생기는 빛깔 있는 기체	763	14.8%
	2	관객 앞에서 연극 따위의 재주를 나타내 보임	2,441	47.4%
	3	기한을 물림	1,931	37.5%
	4	불교에서의 '연기'	12	0.2%
지원	1	뜻하여 바람 (예: ~ 대학에 지원하다)	2,184	10.2%
	2	뒷받침하거나 편들어 도움	18,623	87.3%
	3	지방법원에 따로 분설된 하부기관	513	2.4%

〈표 3〉 검색 실험 질의

단어	질 의	의 미
감자	감자 가격	식물 감자의 가격에 관한 기사
	감자 가격 동향	감자의 가격의 변화에 관한 기사
지구	지구 환경	지구 전체의 환경에 관한 기사
	지구 환경 회의	지구 환경을 논의하는 전 세계적인 회의
경기	경기 전망	경기 전망에 관한 기사
	내년 경기 전망	내년도 경기 전망에 관한 기사
인도	인도 외교	인도와의 외교에 관한 기사
	인도 외교 정책	인도의 외교 정책 혹은 인도와의 외교정책
기간	투자 기간	투자 기간에 관한 기사
	기간 산업 투자	기간산업에 투자하는 내용의 기사
신장	신장 이상	체내의 신장 기능의 이상에 관한 기사
	신장 기능 이상	신장 이상을 구체화
신병	신병 교육	신병 교육에 관한 기사
	군대 신병 교육	신병 교육을 좀 더 구체화
연기	연기 대상	매년 말에 방송국에서 주최하는 연기대상
	올해 연기 대상	올해의 연기 대상에 관한 기사
지원	대학 지원	대학 지원에 관한 기사
	대학 지원 방법	대학 지원방법에 관한 기사

3. 4 중의성 해소 모형

본 연구에서는 선행연구에서 비교적 좋은 성능을 보인 나이브 베이즈 분류기와 EM 알고리즘을 각각 지도학습 알고리즘과 비지도학습 알고리즘으로 선정하여 중의성 해소 실험을 수행하였다(Gale 1992; Gale, Church, and Yarowsky 1992a, 1992b, 1993; Schütze 1998).

3. 4. 1 나이브 베이즈 분류기

지도학습 기법에서는 중의성이 해소된 말뭉치, 즉 중의성을 띄는 단어 w의 모든 출현에 대해 특정 문맥에서 의미 s_k로 의미가 분류된 말뭉치를 학습 데이터로 사용한다. 나이브 베이즈 분류기는 방대한 문맥에서 중의성을 띠는 단어의 인접 단어들을 이용하여 의미를 식별한다. 즉, 중의성을 띠는 단어의 의미를 파악하는 데 주위의 단어가 유용한 정보를 제공하게 되며, 이러한 주변 단어의 공기빈도(co-occurrence frequency) 정보를 이용하여 통계적인 연상추론을 하게 된다.

중의성 단어 w가 말뭉치에서 출현한 문맥을 c_i라 하고 중의성 해소를 위해 문맥 자질로 사용된 단어들을 v_j라 할 때, 나이브 베이즈 분류기는 다음과 같은 결정 규칙을 사용하여 중의성 단어 w에 의미 s'을 부여하게 된다(Manning and Schütze 1999).

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

〈공식 1〉

〈공식 1〉에서 $P(v_j|s_k)$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.

$$P(v_j \mid s_k) = \frac{C(v_j, s_k)}{C(s_k)} \qquad \langle 공식\ 2 \rangle$$

$$P(s_k) = \frac{C(s_k)}{C(w)} \qquad \langle 공식\ 3 \rangle$$

위 공식에서 $C(v_j, s_k)$는 중의성을 띠는 단어 w가 학습 말뭉치에서 의미 s_k를 가질 때 문맥 자질 v_j가 출현한 횟수이며, $C(s_k)$는 중의성을 띠는 단어 w가 학습 말뭉치에서 의미 s_k로 출현한 횟수이고, $C(w)$는 중의성을 띠는 단어 w의 총 출현횟수이다.

3. 4. 2 EM 알고리즘

EM 알고리즘을 사용하는 비지도학습 모형은 나이브 베이즈 분류기와 같은 지도학습 모형과는 달리 중의성이 해소된 학습 말뭉치에 기반한 파라미터 추정(parameter estimation)을 하지 않는다. 대신 파라미터 $P(v_j|s_k)$와 $P(s_k)$에 임의의 초기값을 부여하고 두 파라미터 값을 EM 알고리즘에 의해 재추정한다. 임의의 초기값을 부여한 후에 의미 s_k로 분류된 단어 w가 문맥 c_i에 출현할 확률 $P(c_i|s_k)$를 계산한다. EM 알고리즘은 각 단계에서 주어진 데이터를 이용하여 모형의 우도(likelihood)를 증가시키고, 우도가 더 이상 유의미하게 증가하지 않을 때 알고리즘이 멈추게 된다. EM 알고리즘에 의해 두 파라미터 값이 추정되면 중

의성 단어 w의 문맥을 식별함으로써 중의성을 해소하게 된다. 즉, 문맥에서 중의성 단어와 함께 출현한 단어 v_j에 기반하여 이 중의성 단어가 각각의 의미를 갖게 될 확률을 계산하게 되며, 이때 나이브 베이즈 결정규칙(공식 1)을 이용하여 가장 확률 값이 큰 의미를 부여하게 된다.

이 연구에서 단어 중의성 해소에 사용한 EM 알고리즘은 다음과 같다(Manning and Schütze 1999). 아래의 알고리즘에서 K는 의미 클러스터링 과정에서 원하는 의미의 수, $c_1, c_2, \ldots c_I$는 중의성 단어의 문맥, $v_1, v_2, \ldots v_J$는 중의성 해소 자질로 사용되는 단어를 나타낸다.

1. 모형 μ의 파라미터 $P(v_j|s_k)$, $P(s_k)$를 초기화한다. ($1 \le j \le J$, $1 \le k \le K$)
2. 모형 μ가 주어졌을 때 말뭉치 C의 우도(likelihood)를 계산한다.

$$I(C \mid \mu) = \log \prod_{i=1}^{I} \sum_{k=1}^{K} P(c_i|s_k) P(s_k) = \sum_{i=1}^{I} \log \sum_{k=1}^{K} P(c_i|s_k) P(s_k)$$

$$P(c_i) = \sum_{k=1}^{K} P(c_i\ s_k) P(s_k)$$

3. $I(C \mid \mu)$가 향상되는 동안 추정(estimation)과 최대화(maximization) 단계를 반복한다.

(a) 추정 단계: s_k가 c_i를 생성할 사후확률 h_{ik}를 추정한다.

$$h_{ik} = \frac{P(c_i\ s_k)}{\sum_{k=1}^{K} P(c_i\ s_k)}$$

$$P(c_i\ s_k) = \prod_{v_j \in c_i} P(v_j\ s_k)$$

(b) 최대화 단계: 최우추정법에 의해 $P(v_j|s_k)$, $P(s_k)$를 재추정한다.

$$P(v_j\ s_k) = \frac{\sum_{i=1}^{I} \sum_{\{c_i : v_j \in c_i\}} h_{ik}}{Z_j}$$

(Z_j는 정규화 상수)

$$P(s_k) = \frac{\sum_{i=1}^{I} h_{ik}}{\sum_{k=1}^{K} \sum_{i=1}^{I} h_{ik}} = \frac{\sum_{i=1}^{I} h_{ik}}{I}$$

3. 5 검색 모형

이 연구에서는 기본적인 검색 모형으로 벡터공간 모형(VSM)을 적용하였다. 벡터공간 모형은 각 문헌과 질의를 벡터로 표현하며, 코사인 유사계수나 내적 계수 공식을 사용하여 두 벡터 간의 유사도를 산출함으로써 유사도가 높은 문헌들을 유사도 순으로 순위화하여 제공하는 검색 모형이다.

중의성 해소 이전에 적용된 벡터공간 모형에서는 문헌의 용어 가중치로 로그 TF (log tf +1), 질의어의 가중치로는 이진 TF (1, 0)를

사용하였고, 문헌벡터와 질의벡터간의 유사도 척도로는 코사인 유사계수를 사용하였다. 이 벡터공간 검색 시스템은 의미기반 검색 성능 평가를 위해 베이스라인 시스템으로 사용하였다. 베이스라인 시스템에서는 각 질의에 대해 상위 1,000건의 기사를 검색한 다음 질의별로 30위내 기사를 대상으로 정확률을 산출하였다.

중의성 해소 후에 적용한 검색 모형은 의미벡터공간 모형(SVSM)이다. 이 검색 모형은 다음과 같이 구현하였다.

① 각 질의별로 중의성 해소 알고리즘을 구현한다. 지도학습 기법의 경우 의미별 통계 정보를 추출하여 나이브 베이즈 분류기를 구축하고, 비지도학습 기법의 경우 k개의 의미벡터를 형성하여 EM 알고리즘을 적용한다.

② 구축된 분류기를 통해 각 질의의 의미 모호성을 해소한다. 즉, 각 질의에 대해 질의어들이 동시에 출현한 문헌 안에서 가중치(TF*IDF) 값이 임계치 이상인 용어를 이용하여 문맥을 형성한 다음 동형이의어의 중의성을 해소한다.

③ 베이스라인 검색시스템의 검색 결과 상위 N(=1000)개 문헌에 출현한 중의성 해소 대상 단어에 대해 각각의 분류기를 사용하여 중의성을 해소한다.

④ 의미 분류된 질의어를 이용하여 의미 벡터공간 검색 모형을 구현한다. 이때 의미 분류된 질의어의 용어 가중치로는 의미빈도(sense frequency: SF)를 TF 대신 적용한 로그 SF를 사용한다. SF는 검색 결과 상위 N(=1000)개의 문헌에서 중의성 질의어가 특정한 의미로 사용된 빈도를 의미한다.

⑤ 각 질의에 대해 순위가 새로 부여된 검색 결과를 출력한 다음 상위 n개 문헌의 정확률 P(n)에 의해 검색 성능을 평가한다.

검색 실험 결과 성능 평가 척도로는 TREC 실험에서 사용되고 있는 척도 가운데 하나인 검색된 n개 문헌에 대한 정확률을 사용하였다 (TREC-7 1999). 이 연구에서는 신문기사를 검색하는 실제 이용자의 탐색 성향을 반영하여 컷오프 값을 30으로 정하고, 30위까지의 검색 문헌을 대상으로 각 순위 n에서의 정확률 P(n)을 계산한 다음 그 평균을 최종 정확률로 선택하였다. 이 척도는 보다 상위 순위에서 적합문헌이 검색되는 경우 더 높은 정확률을 갖도록 하기 위한 것이다.

4. 단어 중의성 해소 모형의 성능 평가

일반적으로 정보검색 환경에서 질의어 및 문헌에 나타난 주제어의 의미가 모호할 경우, 이들 단어의 중의성 해소를 통해 검색 성능의 향상을 가져오기 위해서는 상당히 높은 수준의 중의성 해소 성능이 요구된다. 단어 중의성 해소 알고리즘의 성능이 낮을 경우 오히려 중의성 해소 이전보다 검색성능이 저하될 수가 있다. 이 연구에서는 단어의 중의성 해소 목적이 검색 성능의 향상에 있기 때문에 최적의 성능을 보이는 중의성 해소 알고리즘을 찾아낼 필요가 있다. 따라서 파라미터 값들을 변화시켜 가면서 최적의 알고리즘을 찾아내기 위해 문맥창과 학습문헌 집단의 크기를 달리 하여 실험

을 수행하였다.

4. 1 연어를 이용한 중의성 해소 실험

단어 중의성 해소 알고리즘들은 흔히 단어들의 동시출현(co-occurrences) 정보와 연어(collocation) 정보를 이용한다. 특히 연어 정보를 이용할 경우, 중의성을 갖는 단어는 한 문헌에서 한 의미를 가질 뿐만 아니라 한 연어에서도 하나의 의미만을 갖는 경향을 이용할 수 있다(Yarowsky 1995). 따라서 중의성을 띠는 단어가 연어를 갖는다면 우선적으로 연어 정보를 이용하여 중의성 해소 과정을 보다 효율적으로 처리할 수 있다.

이 연구에서는 중의성 해소 대상 질의어를 포함하는 연어를 추출하기 위해 형태소 분석과 구문분석 결과를 이용하여 〈표 4〉와 같은 연어의 구문적 패턴을 추출하였다.

연어의 유형을 보면 유형1과 유형2는 복합명사 형태로서 중의성 해소 대상 단어가 복합명사를 구성하는 다른 명사의 앞이나 뒤에 위치한다. 유형3과 유형4는 중의성 해소 대상 단어가 명사구에 포함된 형태로 두 명사 사이에 빈칸(␣)이 존재하면 앞의 단어에 명사가 붙을 수 있다. 유형5, 유형6, 유형7은 중심어의 중의성 해소를 위해 명사가 아닌 형용사, 명사를 서술하는 동사, 또는 명사를 수식하는 동사 등을 이용하는 경우이다.

중의성 해소 알고리즘을 적용하기에 앞서 실험문헌 텍스트를 살펴본 결과 대체로 하나의 동사는 여러 다른 명사와 연결되어 있는 경우가 많기 때문에 동사를 이용한 중의성 해소 성능은 명사만을 사용할 경우에 비해 낮아질 수 있을 것으로 예상하였다. 다시 말해 동사를 이용할 경우 학습률은 높일 수 있으나 학습의 오류율 또한 높아질 가능성이 크다. 허정/옥철영(2001)의 연구에서도 명사만을 이용한 경우의 중의성 해소 성능이 더 높았고, 용언에는 명사보다 훨씬 낮은 가중치를 부여하여 명사와 함께 사용하였을 경우가 성능이 가장 높은 것으로 나타났다. 본 연구의 중의성 해소 실험에서는 유형6과 유형7을 배제한 실험과 두 유형을 포함한 실험의 두 가지 실험을 수행하여 그 결과를 분석하였다.

각 동형이의어에 대해 학습문맥은 600개, 실

〈표 4〉 연어의 구문적 패턴 유형

유 형	문법형태	용 례
유형1	WSD단어+명사	감자튀김
유형2	명사+WSD단어	차등감자
유형3	명사(+조사)␣WSD단어	개량한 감자
유형4	WSD단어(+조사)␣명사	감자의 재배
유형5	형용사␣WSD단어	구수한 감자
유형6	WSD단어(+조사)␣동사	감자를 먹다
유형7	동사␣WSD단어	으깬 감자

〈표 5〉 연어를 이용한 단어 중의성 해소 실험 결과(유형6, 유형7 제외)

단어	연어부재 건수	학습불능 건수	중의성해소 성공 건수	중의성해소 오류 건수	정확률	중의성 해소 성공률
감자	16.60	69.20	108.67	5.53	0.9516	0.5433
지구	17.80	80.20	96.40	5.60	0.9450	0.4820
경기	13.87	71.47	105.87	8.80	0.9233	0.5293
인도	38.60	85.87	57.67	17.87	0.7644	0.2883
기간	9.60	67.33	121.87	1.20	0.9902	0.6094
신장	21.33	71.47	102.80	4.40	0.9590	0.5140
신병	15.20	52.00	122.53	10.27	0.9228	0.6127
연기	26.53	92.07	73.13	8.27	0.8984	0.3657
지원	12.40	82.80	97.67	7.13	0.9320	0.4884
평균	**19.10**	**74.71**	**98.51**	**7.67**	**0.9206**	**0.4924**

험문맥은 200개를 사용하여 중의성 해소 알고리즘을 15회 반복하여 처리하였다. 〈표 5〉에서 유형6과 유형7을 제외한 실험 결과를 살펴보면 각 동형이의어에 대한 실험문맥 200개 중에 평균 19.10개가 앞에서 정의한 연어 패턴을 포함하고 있지 않았으며, 평균 74.71개의 문맥이 연어 패턴을 포함하고 있지만 해당 연어가 학습문맥에는 포함되어 있지 않아서 학습이 불가능한 경우였다. 나머지 실험문맥 가운데 학습 결과 평균 95.51개는 정확한 의미로 분류되었으며, 평균 7.35개는 부정확한 의미로 분류되었다. 따라서 학습이 가능한 경우 연어를 이용한 중의성 해소 성능은 92.1%의 정확률을 보였다. 그러나 연어가 존재하지 않거나 학습이 불가능한 경우까지 포함한다면 연어기반 중의성 해소 알고리즘의 실제 정확률은 47.8%에 지나지 않는 것으로 나타났다.

연어를 이용한 중의성 해소 모형의 경우 단어에 따라 편차는 있지만, 연어를 형성할 수 없는 사례가 발생하며, 또한 추출된 연어가 학습 데이터에 존재하지 않아 학습을 할 수 없는 사례가 많아 이 모형을 단독으로 사용하는 것은 바람직하지 못하다. 하지만 학습집단의 크기를 크게 할수록 학습이 안 되는 연어의 비율이 줄어들면서 중의성 해소 성능이 향상되는 것을 볼 수 있었다.

유형 6과 유형7의 연어 패턴을 포함한 중의성 해소 실험에서는 중의성 해소 정확률(0.9276)과 중의성 해소 성공률(0.5038) 모두 두 패턴을 제외하였을 경우에 비해 약간 높게 나타났다. 실험 결과 예상했던 대로 연어부재 건수와 학습불능 건수가 감소하였으며, 오류율도 오히려 감소함으로써 전반적으로 중의성 해소 성능이 좋게 나타난 것으로 보인다.

4. 2 나이브 베이즈 분류기를 이용한 중의성 해소 실험

나이브 베이즈 분류기를 사용한 중의성 해소 모형에서 최적의 파라미터 값을 찾기 위하여 문맥창과 학습문맥의 수를 달리하여 실험을 수행하였다. 문맥창의 크기는 좌우 3단어, 한 문장, 좌우 50바이트, 전체 텍스트(전역)의 네 가지로 구분하고, 학습문맥의 수는 600개와 400개

〈표 6〉 나이브 베이즈 분류기를 이용한 중의성 해소 실험 결과(학습문맥 수=600)

문맥창	감자	지구	경기	인도	기간	신장	신병	연기	지원	평균
좌우3단어	0.9678	0.9233	0.8911	0.7447	0.9648	0.8931	0.9017	0.8204	0.9417	0.8943
한 문장	0.9782	0.9384	0.9183	0.7772	0.9782	0.9106	0.8768	0.8677	0.9377	0.9092
좌우50바이트	0.9933	0.9457	0.9203	0.7967	**0.9796**	**0.9109**	**0.8997**	**0.9092**	**0.9427**	**0.9220**
전역	**0.9936**	**0.9588**	**0.9223**	**0.8070**	0.9782	0.9032	0.8909	0.9015	0.9370	0.9214

로 나누었다.

먼저 학습문맥의 수를 600으로 하였을 때의 실험 결과는 〈표 6〉과 같다. 모든 단어들이 좌우 50바이트나 전역일 경우 최소 0.8070에서 최대 0.9936의 정확률을 보이면서 가장 좋은 성능을 나타내고 있다. 반면 문맥의 크기가 좌우 3단어인 경우의 성능이 가장 낮았고, 문맥이 한 문장일 경우는 좌우 3단어와 큰 차이가 없었다. 좌우 50바이트의 문맥과 전역의 경우는 평균 성능이 92% 정도로 나타나 별 차이를 보이지 않고 있다. 중의성 해소 성능을 단어별로 분석한 결과 모두 5개의 의미를 갖는 '인도'가 80% 수준의 가장 낮은 성능을 보이고 있다. 학습문맥의 수를 400개로 하였을 경우의 중의성 해소 성능은 모든 문맥창에서 학습문맥이 600개일 경우에 비해 낮게 나타났다. 따라서 나이브 베이즈 분류기를 사용한 알고리즘에서 최적의 파라미터로 학습문맥의 수는 600개, 문맥창 크기는 좌우 50바이트로 설정하였다.

4. 3 연어 및 나이브 베이즈 혼합 기법 실험

연어를 이용한 중의성 해소 기법은 충분한 수의 연어를 추출할 수 있다면 처리 속도나 성능 면에서 우수한 기법이라고 할 수 있다. 그러나 중의성 해소에 사용할 만한 연어가 실험문헌에 출현하지 않거나 학습이 불가능할 경우에는 사용하기가 어렵다. 따라서 효율성과 성능의 두 가지 측면을 고려하여, 먼저 연어를 이용하여 중의성을 해소한 다음, 연어를 이용할 수 없는 단어에 대해서는 나이브 베이즈 분류기를 사용하여 중의성을 해소하는 혼합 기법을 실험해 보았다. 실험 결과 혼합 기법이 처리 속도는 나이브 베이즈 분류기만을 사용한 경우보다 빠르지만 중의성 해소 정확률 값은 0.9138로서 나이브 베이즈 분류기만을 사용하였을 경우의 0.9220보다 다소 낮게 나타났다.

4. 4 EM 알고리즘을 이용한 중의성 해소 실험

비지도학습 모형인 EM 알고리즘은 학습집단을 이용한 변수 추정(parameter estimation)을 하지 않는 대신 클러스터링 과정을 통해 k개의 의미 클러스터를 생성한다. 따라서 중의성 해소 결과를 지도학습 모형과 동일한 척도를 사용하여 평가하기가 어렵다. 이 연구에서는 EM 알고리즘을 사용한 중의성 해소 실험의 성능 평가를 위해 클러스터링 성능 평가 척도를 사용하였다. 클러스터링 성능 평가를 통해 최적의 클러스터링 성능을 가져오는 클러스터 개수를 결정하여 이를 검색 실험에 반영하였다.

클러스터링 성능 평가는 생성된 클러스터 자체의 품질을 평가하는 것으로서 이 연구에서는 문헌쌍을 하나의 데이터 단위로 하여 두 클러스

터 집합간의 유사도를 측정하는 *CSIM* 척도를 성능 평가 척도로 사용하였다(Chung and Lee 2001).

EM 알고리즘을 사용한 중의성 해소 실험에서는 생성될 클러스터의 개수와 문맥창의 크기 등 2개의 파라미터 값을 최적화시키고자 하였다. k 값의 최적화가 필요한 이유는 중의성 해소 대상 단어가 갖는 의미의 수가 각각 다를 뿐더러 클러스터링의 경우 최적의 성능을 보여주는 k의 수는 의미 수와 직접적인 대응관계에 있을 필요가 없기 때문이다. 예컨대 한 단어의 의미가 5개라고 하더라도 검색 대상 문헌 속에서는 질의어가 갖는 의미로 사용된 문헌 클러스터와 나머지 4개의 다른 의미로 사용된 문헌 클러스터의 두 클러스터로 분류가 된다고 해도 별 문제가 없을 것이기 때문이다. 또한 학습문맥을 랜덤하게 추출하였으므로 중의성 해소 대상 단어의 의미별 빈도를 볼 때 극소수의 의미로 사용된 문맥은 추출되지 않을 수도 있다.

먼저 각 중의성 해소 단어에 대한 최적의 클러스터 개수를 찾기 위해 문맥창은 문장으로 하고 학습문맥의 수를 600개로 설정한 다음 k를 2에서 5까지 변화시켜가며 클러스터링 성능을 *CSIM*으로 측정하였다. 〈표 7〉에 나타난 실험결과를 보면 '감자'와 '경기'를 제외한 모든 단어에 대해 k가 2일 때 가장 좋은 성능을 보이고 있으며, *CSIM* 평균은 63% 수준으로 전체적으로 비교적 좋은 성능을 보이고 있다.

클러스터의 개수를 k=2로 고정시키고 문맥창의 크기와 학습문맥 수에 따라 EM 알고리즘의 성능을 평가한 결과 〈표 8〉에서와 같이 학습문맥의 수가 600개일 때의 성능이 높았으며, 문맥창의 크기는 전역일 경우가 가장 높게 나타났다. 전역과 좌우 50바이트를 비교하였을 때 두 학습집단에서 모두 큰 성능 차이는 없는 것으로 나타났다. 따라서 이 연구에서는 중의성 해소 알고리즘의 처리속도를 감안하여 최적의 문맥창 크기를 좌우 50바이트로 설정하였다.

〈표 7〉 k 값에 따른 EM 알고리즘의 중의성 해소 성능 평가(문맥창=문장, 학습문맥 수=600)

k	감자	지구	경기	인도	기간	신장	신병	연기	지원	평균
2	0.7027	**0.6921**	0.5506	**0.6288**	**0.9001**	**0.4910**	**0.5152**	**0.5489**	**0.6876**	**0.6352**
3	**0.7546**	0.5971	**0.6246**	0.5216	0.8000	0.4683	0.4955	0.4774	0.6214	0.5956
4	0.7091	0.5094	0.5131	0.5025	0.7401	0.4248	0.4584	0.4569	0.5895	0.5449
5	0.7149	0.4679	0.5182	0.4655	0.6807	0.4211	0.4181	0.4344	0.6153	0.5262

〈표 8〉 문맥창 크기에 따른 EM알고리즘의 성능 평가(k=2, 학습문맥 수=600)

문맥창	감자	지구	경기	인도	기간	신장	신병	연기	지원	평균
좌우3단어	**0.8337**	0.5994	0.5800	0.5612	0.6839	0.6103	0.5651	0.5055	0.6455	0.6205
한 문장	0.7023	**0.6918**	0.5346	**0.6143**	**0.8927**	0.4893	0.5147	0.5315	0.7307	0.6335
좌우50바이트	0.8076	0.5919	0.6214	0.4909	0.7034	**0.6716**	**0.7718**	0.6183	0.7289	0.6673
전역	0.6946	0.6479	**0.7286**	0.5719	0.7042	0.4940	0.6740	**0.6559**	**0.8663**	**0.6708**

5. 의미기반 검색 실험

단어 중의성 해소 모형을 벡터공간 검색 모형(VSM)에 통합한 의미기반 벡터공간 모형(SVSM)에서는 앞의 중의성 해소 실험 결과 파악한 최적의 알고리즘과 파라미터 값을 사용하였다. 지도학습 중의성 해소 알고리즘으로는 나이브 베이즈 분류기를 사용하였으며, 학습문맥의 수는 600개, 문맥창의 크기는 좌우 50바이트로 설정하였다. 비지도학습 알고리즘으로 채택한 EM 알고리즘에서는 학습문맥의 수는 600개, 클러스터 개수는 2, 문맥창의 크기는 좌우 50바이트로 설정하였다.

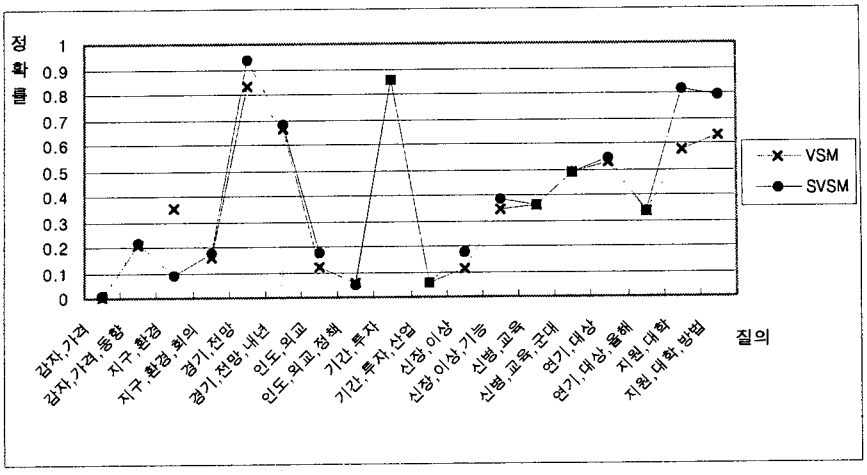
5. 1 나이브 베이즈 분류기를 통합한 의미기반 검색 실험

나이브 베이즈 분류기를 사용하여 중의성 해소 이전과 중의성 해소 이후의 검색 성능을 비교한 결과가 〈표 9〉에 나와 있다. VSM의 경우 정확률은 37%이며, 특히 1, 4, 7, 8, 10, 11번 질의의 검색 성능은 매우 낮게 나타났다. 이 질의들에 대해서는 검색 결과 30위내에 드는 적합문헌 수가 매우 적었기 때문이다. 중의성 해소 후 적용한 SVSM의 검색 성능은 약 39.6%로서, 성능 향상률 측정이 불가능한 1번 질의를 제외했을 때 중의성 해소 전에 비해 약 7.4% 향상된 것이다.

단어의 중의성 해소 이전과 이후의 검색 성능을 질의별로 분석해 보면 다음과 같다. 먼저

〈표 9〉 나이브 베이즈 분류기를 통합한 검색 실험 결과

질의번호	질의	VSM	SVSM	차이	향상률(%)
1	감자,가격	0.0000	0.0073	0.0073	-
2	감자,가격,동향	0.2052	0.2163	0.0111	5.41
3	지구,환경	0.3543	0.0852	-0.2691	-75.95
4	지구,환경,회의	0.1592	0.1776	0.0184	11.56
5	경기,전망	0.8289	0.9365	0.1076	12.98
6	경기,전망,내년	0.6611	0.6778	0.0167	2.53
7	인도,외교	0.1166	0.1765	0.0599	51.37
8	인도,외교,정책	0.0525	0.0443	-0.0082	-15.62
9	기간,투자	0.8543	0.8543	0.0000	0.00
10	기간,투자,산업	0.0582	0.0582	0.0000	0.00
11	신장,이상	0.1159	0.1734	0.0575	49.61
12	신장,이상,기능	0.3413	0.3865	0.0452	13.24
13	신병,교육	0.3566	0.3566	0.0000	0.00
14	신병,교육,군대	0.4903	0.4903	0.0000	0.00
15	연기,대상	0.5265	0.5458	0.0193	3.67
16	연기,대상,올해	0.3363	0.3397	0.0034	1.01
17	지원,대학	0.5763	0.8122	0.2359	40.93
18	지원,대학,방법	0.6323	0.7953	0.1630	25.78
평균		**0.3703**	**0.3963**	**0.0260**	**7.44**



〈그림 2〉 나이브 베이즈 분류기를 통합한 검색 실험 결과

1번 질의의 경우 VSM은 30위 안에 적합문헌을 하나도 검색해내지 못함으로써 정확률이 0의 값을 갖게 되었으며, 중의성 해소 후에도 성능은 거의 향상되지 못하였다. 이것은 '감자'의 두 가지 의미가 실험문헌 집단에서 거의 비슷한 비율로 사용되고 있으나 검색 결과 상위에 출력된 문헌들이 모두 질의와는 다른 의미로 사용되었기 때문이며, 중의성 해소 성능이 낮은 것도 '감자'가 두 가지 의미에서 모두 '가격'이란 단어와 함께 사용될 수 있기 때문인 것으로 보인다.

중의성 해소 결과 오히려 성능이 저하된 질의는 3번 질의와 8번 질의이며, 특히 3번 질의는 약 76% 의 성능 저하율을 보였다. '지구'라는 단어는 두 가지 의미 중에 두 번째 의미의 출현비율이 2배 이상인데 질의에서는 첫 번째 의미를 갖기 때문에 중의성 해소가 오히려 부적합문헌의 검색을 초래한 것으로 보인다. 3번 질의를 제외하고 성능을 비교하면 SVSM이 12.65%의 성능 향상률을 보이고 있다.

중의성 해소 전에 비해 성능 향상률이 현저히 높은 질의는 7, 11, 17, 18번 질의이며, 9, 10, 13, 14번 질의는 성능의 변화가 없었다. 9, 10번 질의를 구성하는 '기간'의 경우 첫 번째 의미의 출현비율이 약 98%로서 중의성 해소의 효과가 거의 없었고, 따라서 검색 결과 성능의 변화가 없는 것으로 보인다. 또한 질의 13, 14번의 '신병'의 경우는 중의성 해소 대상 단어가 '교육', '군대' 등의 단어와 조합되면서, 질의 자체에서 이미 '신병'의 중의성이 이미 해소되어 중의성 해소 효과가 없었던 것으로 분석된다. 의미기반 검색 결과 대략 전체의 절반 정도의 질의에서는 중의성 해소가 전혀 도움이 되지 않았거나 오히려 검색 성능을 저하시킨 것으로 나타났다.

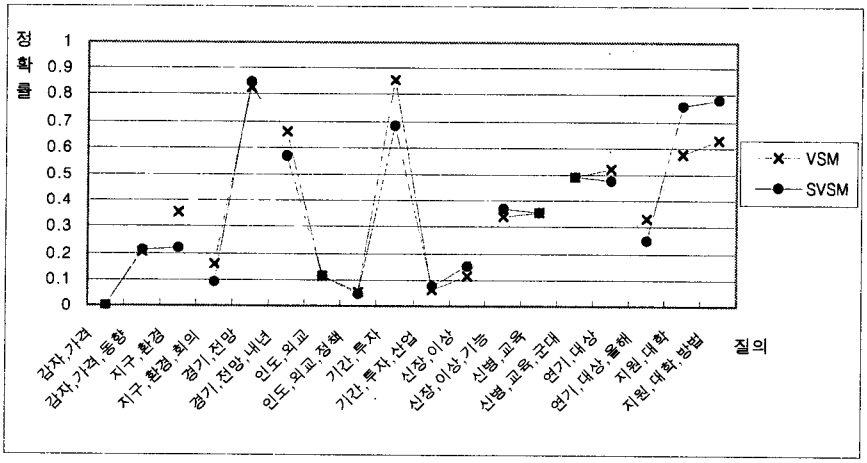
〈그림 2〉는 VSM과 나이브 베이즈 분류기를 통합한 SVSM의 질의별 검색 성능을 그래프로 표현한 것이다.

5.2 EM 알고리즘을 통합한 의미기반 검색 실험

EM 알고리즘을 통합한 의미기반 검색에서는 문맥창의 크기를 좌우 50바이트로 하여 중

〈표 10〉 EM 알고리즘을 통합한 검색 실험 결과

질의번호	질의	VSM	SVSM	차이	향상률(%)
1	감자,가격	0.0000	0.0024	0.0024	-
2	감자,가격,동향	0.2052	0.2097	0.0045	2.19
3	지구,환경	0.3543	0.2193	-0.1350	-38.10
4	지구,환경,회의	0.1592	0.0913	-0.0679	-42.65
5	경기,전망	0.8289	0.8452	0.0163	1.97
6	경기,전망,내년	0.6611	0.5691	-0.0920	-13.92
7	인도,외교	0.1166	0.1151	-0.0015	-1.29
8	인도,외교,정책	0.0525	0.0420	-0.0105	-20.00
9	기간,투자	0.8543	0.6830	-0.1713	-20.05
10	기간,투자,산업	0.0582	0.0722	0.0140	24.05
11	신장,이상	0.1159	0.1482	0.0323	27.87
12	신장,이상,기능	0.3413	0.3735	0.0322	9.43
13	신병,교육	0.3566	0.3566	0.0000	0.00
14	신병,교육,군대	0.4903	0.4903	0.0000	0.00
15	연기,대상	0.5265	0.4747	-0.0518	-9.84
16	연기,대상,올해	0.3363	0.2467	-0.0896	-26.64
17	지원,대학	0.5763	0.7557	0.1794	31.13
18	지원,대학,방법	0.6323	0.7784	0.1461	23.11
평균		**0.3703**	**0.3596**	**-0.0107**	**-3.10**



〈그림 3〉 EM 알고리즘을 통합한 검색실험 결과

의성을 해소한 후 검색 실험을 수행하였다. 〈표 10〉의 검색 실험 결과를 보면 중의성 해소 결과 검색 성능이 오히려 약 3% 저하된 것으로 나타나 있다.

질의별로 살펴보면 나이브 베이즈 분류기를 통합한 검색에서와 마찬가지로 3번 질의와 8번

질의는 검색 성능이 저하되었고, 나이브 베이즈 통합 검색에서 성능의 변화가 없었던 9, 13, 14번 질의는 질의가 저하되거나 변화가 없는 것으로 나타났다. 15, 16번 질의는 나이브 베이즈 통합 검색에서는 미미한 성능 향상률을 보였으나 EM 알고리즘 통합 검색에서는 성능이 저하되었다. 두 가지 의미기반 검색 실험 결과 질의별로 중의성 해소가 검색 성능에 미친 영향을 분석해 보면 대체적으로 비슷한 경향을 보이고 있음을 알 수 있다. 〈그림 3〉은 SVM과 EM 알고리즘을 적용한 SVSM의 질의별 검색 성능을 보여 준다.

비지도학습 모형인 EM 알고리즘의 중의성 해소 성능은 클러스터링 성능 평가 척도를 사용하였기 때문에 중의성 해소 실험 결과 최적의 파라미터로 선택한 문맥창 크기(좌우 50바이트)가 가장 좋은 검색 성능을 가져올 것이라는 가정에는 의문을 가질 수 있다. 따라서 EM 알고리즘을 적용한 SVSM 검색에서 네 가지 문맥창을 모두 사용하여 실험을 수행하였는데 검색 성능은 문장과 좌우 50바이트를 사용한 경우가 비슷하였고, 중의성 해소 실험에서 좋은 성능을 보였던 전역은 오히려 6% 정도 낮은 정확률을 보였다.

6. 결 론

이 연구에서는 단어의 중의성을 해소함으로써 검색 성능을 향상시킬 수 있으리라는 가설하에 최적의 검색 성능을 가져올 수 있는 중의성 해소 모형을 실험을 통해 발견하고자 하였다. 지도학습 모형인 나이브 베이즈 분류기와 비지도학습 모형인 EM 알고리즘을 사용하여 중의성 해소 실험을 수행한 결과 나이브 베이즈 분류기는 최적의 조건에서 평균 92%의 정확률을 보였으며, EM 알고리즘은 최적의 조건에서 평균 67% 수준의 클러스터링 성능을 보였다.

중의성 해소 알고리즘을 적용한 의미기반 검색에서는 나이브 베이즈 분류기 통합 검색이 약 39.6%의 정확률을 보였고, EM 알고리즘 통합 검색은 약 36%의 정확률을 보였다. 중의성 해소 모형을 적용하지 않은 베이스라인 검색의 평균 정확률 37%와 비교하면 나이브 베이즈 통합 검색은 약 7.4%의 성능 향상률을 보인 반면 EM 알고리즘 통합 검색은 약 3%의 성능 저하율을 보였다.

이 연구에서 단어의 중의성 해소 실험을 통해 발견한 사실은 다음과 같다:

(1) 단어의 중의성 해소 성능은 중의성 해소 대상 단어의 특성에 따라 큰 차이를 보인다. 단어의 특성으로는 의미별 출현비율, 중의성 해소에 결정적인 역할을 하는 단서어의 출현 위치 등이 있다.

(2) 연어를 이용한 중의성 해소 모형은 대상 단어가 학습집단과 실험집단 안에서 충분한 수의 연어를 갖는다면 정확률 90% 수준의 높은 성능을 가져올 수 있으므로 효과적이며 동시에 효율적인 모형이 될 수 있다. 그러나 실험 결과 거의 절반 가까운 경우에 연어를 통한 중의성 해소가 불가능한 것으로 나타났기 때문에 단독으로 사용하는 것은 바람직하지 않다.

(3) 나이브 베이즈 분류기는 92%의 중의성 해소 정확률을 보임으로써 검색에 효과적으로 적용할 수 있는 수준의 성능을 갖는 것으로 평가된다.

중의성 해소 모형을 통합한 의미기반 검색 실험을 통해 발견한 사실은 다음과 같다:

(1) 베이스라인 검색과 의미기반 검색 모두 질의에 따라 큰 성능 차이를 보이고 있다. 특히 중의성 해소 이후 검색 성능의 향상률은 질의의 특성에 따라 큰 차이를 보이고 있으며, 질의를 구성하는 중의성 해소 대상 단어의 실험집단내 의미 분포에 의해 중의성 해소가 검색에 끼치는 영향이 달라진다.

(2) 나이브 베이즈 분류기는 질의에 따라 현저한 성능 향상을 가져올 수 있다. 특히 2개 이상의 질의어가 조합되어도 여전히 중의성이 크거나 중의성이 어느 정도 해소된 후에도 검색 성능 자체가 낮은 질의에 대해서는 의미기반 검색을 통해 검색 성능을 향상시킬 수 있을 것이다.

(3) EM 알고리즘을 검색 모형에 통합하기 위해서는 클러스터링 개수, 문맥창의 크기 등 파라미터의 최적화 과정과 중의성 해소 성능 평가 방법에 대한 보다 다각적인 연구가 필요하다.

결론적으로 중의성을 갖는 단어와 이 단어를 포함하는 질의의 특성을 감안하여 중의성 해소 과정을 선택적으로 적용하는 것이 검색 성능을 최적화할 수 있을 것으로 보인다. 지도학습 모형인 나이브 베이즈 분류기는 우수한 성능을 보였지만 학습집단을 구축하기 위한 부담이 너무 크다. 따라서 수작업 의미 태깅을 최소화하면서도 성능을 높일 수 있도록 지식 기반 기법과 말뭉치 기반 기법을 혼합한 모형의 도입이 필요할 것으로 보인다.

참 고 문 헌

허정, 옥철영. 2001. 사전의 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템. 『정보과학회논문지: 소프트웨어 및 응용』, 28(9): 688-698.

Chung, Y. M., and Lee, J. Y. 2001. "A Corpus-based Approach to Comparative Evaluation of Statistical Term Association Measures." *Journal of the American Society for Information Science and Technology*, 52(4): 283-296.

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

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

Gale, W., Church, K. W., and Yarowsky, D. 1993. "A method for disambiguating

word senses in a large corpus." Computers and the Humanities, 415-439.

Ide, N., and Veronis, J. 1998. "Word sense disambiguation: the state of the art." *Computational Linguistics*, 24(1): 1-40.

Jackson, P., and Moulinier, I. 2002. *Natural Language Processing for Online Applications : Text Retrieval, Extraction, and Categorization.* Amsterdam: John Benjamins Publishing Company.

Jansen, B. J., and Spink, A. 2005. "An analysis of web searching by European AlltheWeb.com users." *Information Processing and Management*, 41: 361-381.

Jansen, B. J., Spink, A., and Saracevic, T. 2000. "Real life, real users, and real needs: a study and analysis of user queries on the web." *Information Processing and Management*, 36: 207-227.

Krovets, R., and Croft, W. B. 1992. "Lexical ambiguity and information retrieval." *ACM Transactions on Information Retrieval Systems*, 10(2): 115-141.

Levinson, D. 1999. "Corpus-based method for unsupervised word sense disambiguation." Proceedings of the Workshop on Machine Learning in Human Language Technology, Advanced Course on Artificial Intelligence (ACAI '99), Chania, Greece, 267-273.

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

Sanderson, M. 1994. "Word sense disambiguation and information retrieval" *Proceedings of the 17th international ACM SIGIR*, 49-57.

Sanderson, M. 2000. "Retrieving with good. sense." *Information Retrieval*, 2(1): 49-69.

Schütze, H. 1998. "Automatic word sense discrimination." *Computational Linguistics Archive*, 24(1): 97-123.

Schütze, H., and Pederson, J. 1995. "Information retrieval based on word sense." *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval*, 161-175.

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

Stokoe, C., Oakes, M. J., and Tait, J. 2003. "Word sense disambiguation in information retrieval revisited." *Proceedings of the 26th ACM SIGIR*, 159-166.

TREC-7. 1999. Proceedings of the Seventh Text Retrieval Conference. Appendix A. Evaluation Techniques and Results.

NIST Publication 500-242.

Voorhees, E. M. 1993. "Using WordNet to disambiguate word senses for text retrieval." *Proceedings of SIGIR '93*, 171-180.

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.