단어 중의성 해소를 위한 지도학습 방법의 통계적 자질선정에 관한 연구*

A Study on Statistical Feature Selection with Supervised Learning for Word Sense Disambiguation

이 용 구(Yong-Gu Lee)**

초 록

이 연구는 지도학습 방법을 이용한 단어 중의성 해소가 최적의 성능을 가져오는 통계적 자질선정 방법과 다양한 문맥의 크기를 파악하고자 하였다. 실험집단인 한글 신문기사에 자질선정 기준으로 정보획득량, 카이제곱 통계량, 문헌빈도, 적합성 함수 등을 적용하였다. 실험 결과, 텍스트 범주화 기법과 같이 단어 중의성 해소에서도 자질선정 방법이 매우 유용한 수단이 됨을 알 수 있었다. 실험에 적용한 자질선중 기준 중에 정보획득량이 가장 좋은 성능을 보였다. SVM 분류기는 자질집합 크기와 문맥 크기가 클수록 더 좋은 성능을 보여 자질선정에 영향을 받지 않았다. 나이브 베이즈 분류기는 10% 정도의 자질집합 크기에서 가장 좋은 성능을 보였다. kNN의 경우 10% 이하의 자질에서 가장 좋은 성능을 보였다. 단어 중의성 해소를 위한 자질선정을 적용할 때 작은 자질집합 크기와 큰 문맥 크기를 조합하거나, 반대로 큰 자질집합 크기와 작은 문맥 크기를 조합하면 성능을 극대화 할 수 있다.

ABSTRACT

This study aims to identify the most effective statistical feature selecting method and context window size for word sense disambiguation using supervised methods. In this study, features were selected by four different methods: information gain, document frequency, chi-square, and relevancy. The result of weight comparison showed that identifying the most appropriate features could improve word sense disambiguation performance. Information gain was the highest. SVM classifier was not affected by feature selection and showed better performance in a larger feature set and context size. Naive Bayes classifier was the best performance on 10 percent of feature set size. kNN classifier on under 10 percent of feature set size. When feature selection methods are applied to word sense disambiguation, combinations of a small set of features and larger context window size, or a large set of features and small context windows size can make best performance improvements.

키워드: 단어 중의성 해소, 통계적 자질선정, 문맥 크기, SVM, 나이브 베이즈 분류기, kNN 분류기
Word Sense Disambiguation, Statistical Feature Selection, Context Size, SVM, Naive Bayes Classifier, kNN Classifier

---

* 이 논문은 2011년 한국비블리아학회 춘계 학술대회에서 발표한 내용을 수정・보완한 것임.
** 계명대학교 문헌정보학과 전임강사(yonggulee@kmu.ac.kr)
논문접수일자 : 2011년 5월 13일 논문심사일자 : 2011년 5월 20일 게재확정일자 : 2011년 6월 10일

1. 서 론

인간의 언어에서 기본 단위인 단어는 중의성을 가진다. 즉 하나의 단어가 여러 의미를 가지며 사용된다. 이러한 인간의 언어를 컴퓨터가 이해하기 위해서는 다양한 방법들이 존재하지만, 그 중에서 가장 기본이 되는 것은 의미 분석을 바탕으로 한다.

의미 분석을 통해 인간의 언어, 좀 더 세분하여 문장 내지 단어에 대한 의미 분석 작업을 수행하는 것은 지금과 같이 디지털화된 문헌 또는 전문(full-text)이 대량으로 생산되는 시대에 가장 기초적이면서 필수적으로 필요한 기술이다. 이러한 기술 내지 방법을 자연 언어 처리(natural language processing)라고 하며, 크게 형태소 분석, 구문 분석, 의미 분석, 화용 분석의 네 단계로 나누어진다.

자연 언어 처리 단계에서 의미 분석은 문장 내 각각의 형태소에 대해 의미해석 규칙이나 의미 자질 사전 등을 이용하여 의미를 부여(tagging)하는데, 인간의 언어에 존재하는 동형이의어(homograph)나 동음이의어(homonym), 다의어(polysemy)와 같이 중의성을 갖는 단어의 의미를 정확히 파악하기 위한 방법이 필요하다. 이렇게 중의성을 갖는 단어가 어떤 의미로 쓰였는지 식별하는 작업을 단어 의미 중의성 해소 또는 단어 중의성 해소(word sense disambiguation: wsd)라 한다.

단어 의미 중의성 해소 방법에서 지도학습(supervised learning)에 기반한 알고리즘은 비교적 성능이 좋아 자주 사용되고 있다. 다만 이러한 알고리즘을 적용하기 위해서는 중의성 단어에 대해 의미가 태깅된 언어자원(corpus)이 필요하다. 또한 일반적으로 지도학습 기반의 의미 분류기를 구축할 때 수천 또는 수만 개의 자질집합(feature set)을 사용하게 되며, 이는 분류기의 성능이나 비효율적 처리를 가져온다.

이 연구에서는 단어 중의성을 해소하기 위해 지도학습 기법을 적용할 때, 효율적인 중의성 해소 환경을 갖추기 위한 방법으로 다양한 자질 선정 실험을 수행하였다. 즉 지도학습 기법에 기반을 둔 의미 분류기를 구축하고자 할 때, 분류기의 성능을 일정정도 유지 내지 향상시킬 수 있는 통계적 자질선정 방법(statistical feature selection)과 분류기별 특성을 가져오는 요인들을 파악하고자 하였다. 또한 이들이 기존의 텍스트 범주화와 어떤 차이가 있는지 살펴보고자 하였다.

2. 단어 중의성 해소와 자질 선정

2.1 단어 중의성 해소

단어 중의성 해소는 특정 문맥(context)에서 사용된 단어의 의미를 평가하여 결정하는 능력 또는 기능을 말한다. 달리 말하면 특정 문맥에 출현한 하나의 단어에 대해 사전에 정의된 여러 의미 중에 가장 적당한 의미범주(class)를 부여하는 작업이다. 이러한 측면에서 단어 중의성 해소는 분류 작업(classification task)으로 볼 수 있다. 따라서 단어의 의미가 범주가 되며 단어의 문맥이나 외부 지식 자원(external knowledge sources)으로부터 관련정보를 추출하여, 이를 기반으로 하나 또는 그 이상의 의미범주를 부여하는 자동 분류 방법을 사용한다(Navigli 2009).

넓은 의미에서 단어 중의성 해소는 중의성 해소를 위해 필요한 정보나 자원, 그리고 중의성 해소 대상 단어의 의미부여 여부에 따라 크게 두 가지로 구별할 수 있다. 하나는 이미 존재하는 의미범주와 무관하게 특정 단어에 대해 그 단어의 용법(usage)에 따라 서로 다른 그룹을 부여하는 과업인 단어 의미 식별(word sense discrimination) 방법이다. 이 방법은 관련 정보나 자원을 필요로 하지 않으며 의미 범주도 필요로 하지 않는다. 다른 하나는 사전이나 시소러스 등과 같은 언어 자원에서 추출한 의미 목록으로부터 실제 쓰인 의미를 선택하는 과업이다. 일반적으로 이를 좁은 의미에서 단어 중의성 해소라 한다.

또한 중의성을 해소해야 할 대상 단어의 범위에 따라, 단어 중의성 해소는 텍스트 내 모든 단어에 대해 각 단어가 갖는 여러 의미들을 구분하는 종류와, 특정 출현 단어에 대해 적절한 의미를 부여하는 단계로 구성된다(Ide and Veronis 1998). 전자에서는 명사, 동사, 형용사, 부사 등과 같이 텍스트에 나타난 모든 단어에 대해 미리 정의된 의미범주를 사용하는데, 이러한 의미범주는 사전에 등록된 의미목록이나 시소러스의 동의어처럼 상호 연관된 단어나 자질들로부터 확보한다. 후자에서는 중의성 해소 대상 단어가 출현한 문맥, 즉 특정한 단어가 출현한 텍스트에 포함된 정보를 사용하며 대개 사람이 손으로 의미를 부여한 사례를 이용하여 그 사례에 정확한 의미를 부여한다(Navigli 2009).

단어 중의성 해소는 일찍이 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 수행하기 위해 필요한 과정이라고 볼 수 있다. 따라서 단어 중의성 해소는 기계번역, 정보검색 및 질의응답, 그리고 음성 처리(speech processing) 및 텍스트 처리(text processing) 분야에서 유용하게 이용되거나 직접적으로 요구되는 핵심적인 알고리즘이다.

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).

두 번째 유형인 말뭉치 기반 알고리즘에서 특정한 의미 범주를 갖는 중의성 단어의 문맥으로부터 자질을 추출한 후, 이들 자질로부터 학습 집합(training set)을 형성하고 의미 분류기를 학습시키는 지도 학습 기법을 가장 일반적으로 이용한다.

문헌의 자동분류에 사용되는 지도학습 기법인 텍스트 범주화(text categorization)는 사전에 분류된 학습문헌 집합에 근거하여 이미 정의 되어 있는 주제범주들을 새로운 문헌에 배정하는 작업이다(정영미 2005). 일반적으로 이

기법은 학습문헌 집합이 필요 없는 비지도학습(unsupervised learning) 방법보다 더 좋은 성능을 보인다. 단점으로는 분류기를 구축하기 위해 의미가 태깅된 말뭉치가 필요하며, 이 부분이 수작업 내지 반 수작업 과정이 필요하여 비용이 많을 들어간다.

지도학습 방법에 의한 중의성 해소 과정은, 먼저 의미 분류기를 구축하기 위한 학습 과정이 필요하며, 구축된 분류기를 이용하여 중의성 해소 대상 단어에 대한 의미 범주의 할당하는 분류과정인 중의성 해소 과정이 필요하다. 이 각각의 과정을 세분화하여 설명하면 다음과 같다.

Ⅰ 학습과정
① 적절한 의미로 태깅된 각각의 중의성 단어에 대해 출현(학습) 문맥 추출
② 문맥에서 출현한 단어로부터 분류기에 필요한 자질 추출 또는 벡터 생성
③ 분류기 구축
Ⅱ 중의성 해소 과정
① 새로운 검증 문맥에 출현한 대상 단어에 대한 자질 추출 또는 벡터 생성
② 분류기를 통하여 의미(범주) 부여하여 중의성 해소

흔히 사용되는 지도학습 알고리즘으로는 나이브 베이즈 분류기(naive Bayes classifiers: NB), kNN(k-Nearest Neighbor classifiers), 결정트리(decision trees), 결정규칙(decision rules), 신경망(neural networks) 분류기, 지지벡터기(support vector machines: SVM) 등을 들 수 있다(Sebastiani 2002).

나이브 베이즈 분류기는 초기 단어 중의성 해소 연구부터 많이 사용되어 왔다. 이 분류기는 특정 범주에 출현한 자질의 확률 통계를 사용하는 비교적 간단한 분류기지만 좋은 성능을 보인다(Jackson and Moulinier 2002; Gale et al. 1992; Ng 1997; Escudero et al. 2000). 사례 기반 학습 방법의 대표적인 분류기인 kNN 분류기는 자질을 어떻게 표현하느냐와 부적합한 자질의 출현에 따라 매우 민감하지만, 이러한 문제만 해결된다면 나이브 베이즈 분류기보다 우수한 것으로 보인다(Ng and Lee 1996; Stevenson and Wilks 2001). 일반적으로 가장 좋은 성능을 보이는 커널 기반 SVM 분류기도 wsd에서 매우 강력하고 매우 좋은 결과를 가져오는데, 심지어 잡음 자질이나 도움이 안 되는 자질들이 존재하거나 자질집합의 크기가 큰 경우도 좋은 성능을 보인다(Joachims 2001; Lee and Ng 2002; Strapparava et al. 2004).

2.2 자질선정

분류기를 구축하기 위해 학습 문맥으로부터 자질을 추출하는데, 분류 성능에 모든 자질이 긍정적인 영향을 미치는 것은 아니다. 어떤 자질은 분류 성능에 도움이 되지만, 어떤 자질은 방해가 되기도 한다. 따라서 불필요한 자질을 제거하거나 적절한 자질의 선정을 통해 자질집합 또는 자질공간의 축소를 가져올 수 있다.

자질선정 또는 자질축소의 혜택은 데이터 시각화와 데이터 이해의 촉진, 측정과 저장공간의 감소, 학습과 활용 시간의 감소, 예측 성능의 향상시키기 위한 차원의 저주(curse of dimensionality)에 도전 등을 들 수 있다(Guyon and Elisseeff

2003). 즉 다시 말하면, 자질선정은 분류 성능의 잡음 내지 도움이 되지 않는 자질들을 제거함으로써 분류 성능이 향상이 가능할 수 있으며, 자질공간의 축소를 가져오게 할 수 있다.

자질선정을 하기 위한 방법은 자질 하부집합의 선정과 새로운 자질의 구축으로 나누어 볼 수 있다(Sebastiani 2002). 자질 하부집합 선정은 기존의 자질집합으로부터 좋은 분류성능을 가져오는 유용한 자질들을 선택하고, 반대로 중복되거나 부가적인 자질들의 제외시키는 것을 말한다. 새로운 자질의 구축은 원래의 자질로부터 조합과 변형을 통한 새로운 자질을 생성하는 것을 말한다. 이 방법에는 어간이 동일한 단어들의 접사나 어미를 제거하는 스테밍(stemming), 용어 클러스터링, 잠재의미색인(LSI) 방법 등을 통해 여러 자질을 하나의 자질로 변화시킨다.

자질의 하부집합을 선정하기 위한 방법으로 문헌빈도와 같은 단어의 출현빈도를 이용하는 방법이 있다. 또한 상호정보량이나 정보획득량과 같은 정보이론적 기준, 카이제곱 통계량, 상관계수, 적합성 점수 승산비와 같은 기준을 적용할 수 있다. 이 중에서 정보획득량과 카이제곱이 가장 효과적이었으며, 문헌빈도도 비슷한 성능을 보인 것으로 나타났다(Yang and Pedersen 1997; Sebastiani 2002). 이러한 유형의 자질선정을 통계적 자질선정이라고 한다.

Yang과 Pedersen(1997)의 연구를 보면, 로이터 말뭉치에서 kNN 분류기를 정보획득량을 이용하여 자질의 98%를 축소하였음에도 약간의 분류성능이 향상되는 것을 보여주었다. 또한 문헌빈도는 정보획득량이나 카이제곱 기준보다 계산측면에서 가장 낮은 비용을 가지면서 유사한 성능을 보였다.

기존의 기계학습에서 사용되는 자질선정 방법 이외에 단어 중의성 해소에서 사용되는 자질선정 방법으로 다양한 문맥 크기나 품사, 문맥 내의 단어 또는 바이그램, 다양한 문법적 속성(예로 어근, 품사, 격) 등을 적용하는 연구들이 있어 왔다(Mihalcea 2002; Orhan and Altan 2006; Suarez and Palomar 2002; Fragos 2008). 다만 이러한 방법의 자질선정을 수행하려면, 앞서 설명한 여러 유형의 자질들에 대한 정보를 말뭉치 내에 추가해야 가능하다. 현재 이러한 정보를 담고 있는 한글 말뭉치는 구하기 어려우므로 이 연구에서는 기존의 자질 선정 방법인 통계적 자질 선정을 대상으로 연구를 수행하였다.

3. 실험 설계

3.1 실험 개요 및 데이터

이 연구는 지도학습 기반의 단어 중의성 해소 방법을 보다 효율적으로 활용하고자 지도 학습 분류기의 통계적 자질선정에 대한 실험을 수행하였다. 즉, 단어 중의성 해소에서 통계적 자질선정이 어떤 의미를 갖는지를 다양한 분류기와 분류 환경을 실험적으로 분석하고자 하였다.

텍스트 범주화는 문헌을 대상으로 자동 분류할 때, 문헌에 특정 주제범주(category)를 부여한다. 반면 단어 중의성 해소는 중의성 단어에 의미 분류, 즉 단어에 특정 의미를 부여하는 유사한 방법을 취한다. 따라서 전자가 분류대상으로 문헌을 취하므로, 분류기를 학습시키기

위해 자질을 추출할 때 학습 문맥의 단위가 문헌이 된다. 후자는 분류 대상이 중의성 대상 단어가 되므로 학습 문맥이 중의성 단어가 출현한 부분이 되고, 범주는 그 단어의 의미가 된다. 이때 중의성 단어를 중심으로 어느 정도의 학습 문맥의 크기(window)를 설정하느냐가 분류 성능의 변수가 될 수 있다. 이러한 문맥 창의 크기로 중의성 해소 대상 단어를 중심으로 좌우 3단어, 그 단어가 출현한 한 문장, 좌우 50바이트, 한 문헌 전체로 달리 적용할 수 있다. 이 연구에서도 문맥창의 크기로 지역문맥 3개(좌우 3단어, 한 문장, 그리고 좌우 50바이트)와 전역문맥 1개(신문기사 전체)를 적용하였다.

이 실험에서 사용한 자질 가중치 방법은 이진빈도(binary)를 부여하였다. 이진빈도의 경우 출현 여부만을 나타내는 가장 기초적인 가중치 부여 방식으로, 다양한 분류기에 적용할 수 있는 장점이 있다. 다만 이 방법은 자질의 빈도에 따라 가중치를 부여하는 방식에서 더 좋은 성능을 보이는 kNN과 같이 분류기에는 적합하지 않을 수 있어 이 연구의 제한점으로 두고자 한다.

일반 텍스트 문서화 기법에서처럼 학습 문헌에 해당하는 학습 문맥의 수는 600개를 설정하고 의미 분류 대상인 검증 문맥은 200개(학습 문헌과 검증 문헌의 비율 = 3 : 1)로 하였다. 다만 중의성 대상 단어 중 '신병'의 경우 총 출현빈도가 800개가 되지 않아 300개의 학습 문맥과 100개의 검증 문맥으로 적용하였다.

각각의 중의성 해소 단어에 대한 학습문맥과 검증문맥은 전체 실험문헌 집단에서 랜덤하게 추출하였기 때문에 이로 인한 오차를 줄이기 위해 학습문맥 집단과 검증문맥 집단을 10회 반복해서 추출하고 이를 대상으로 각각의 분류기를 구축하고 중의성 해소 성능을 산출하여 평균화 하였다. 또한 10번 랜덤으로 추출된 집단(학습집단+검증집단) 하나하나에 대해 서로 다른 자질선정 방법과 다수의 분류기를 구축하여 동일한 환경이 되도록 하였다.

이 연구에서 사용된 중의성 해소 대상 단어 9개에 대한 태깅된 의미, 이들 단어의 출현빈도와 출현비율은 〈표 1〉과 같다. 또한 이들의 출현기사 수 및 총 출현빈도는 〈표 2〉와 같다(정영미, 이용구 2005).

〈표 1〉 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%

단어	의미 번호	사전의 뜻풀이	출현 빈도	출현 비율
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 넘겨받는 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%

〈표 2〉 중의성 해소 대상 단어의 출현기사 수 및 총 출현빈도

중의성 단어	출현기사 수	총 출현빈도
감자	497	1,115
경기	17,294	37,763
기간	11,239	15,803
신병	339	469
신장	629	953
연기	3,089	5,147
인도	1,639	2,750
지구	3,866	9,372
지원	11,047	21,321

3.2 의미 분류기 및 자질 선정 기준

이 연구에서 중의성 해소를 위해 사용한 분류기는 NB 분류기, kNN 분류기, SVM로 단어 중의성 해소 연구에서 주로 사용되는 분류기를 선정하였다. 이 분류기들은 Orange 마이닝 툴 패키지(http://orange.biolab.si/)에 파이선(python)을 연동하여 사용하였다.

이 연구에서 사용된 분류기 중 kNN 분류기와 SVM 분류기는 분류기를 구축할 때 파라미터 값을 필요로 한다. 따라서 kNN 분류기는 k 값을 설정하기 위해 사전 실험을 수행하였으며 SVM 분류기의 다양한 파라미터는 Orange 패키지의 기본값(default)을 그대로 사용하였다.

지도학습 기법을 이용하여 단어의 중의성을 해소하고자 할 때, 성능을 평가하기 위해서는 일반적으로 문헌 자동분류의 실험결과를 평가하기 위해 각 범주별로 정확도, 재현율, 정확률, 그리고 F1 척도를 사용할 수 있다. 더 나아가 전체 범주의 성능을 평가하기 위해 평균 정확도, 평균 재현율, 평균 정확률, 마이크로평균 F1 척도와 매크로평균 F1 척도를 사용할 수 있다.

하지만 하나 또는 그 이상의 주제 범주가 부여되는 일반 문헌 자동분류와 달리, 단어 중의성 해소는 중의성 단어가 특정 문맥에서 한 의미로 사용되며 이를 찾아내는 방법이다. 즉 여러 의미를 지니는 중의성 단어라도 특정 출현에서는 한 의미로 사용된다. 이러한 특성을 반영하여 이 연구에서는 얼마나 정확하게 중의성 단어의 의미를 분류하였는지를 체크할 수 있는 백분율에 의한 정확도를 적용하였다. 이는 전체 중의성 단어에 대해 올바르게 의미를 분류한 수로 나눈 비율을 뜻한다.

이 연구에서는 자질 선정 기준으로 카이제곱 통계량(ch), 문헌빈도(df), 적합성 함수(rv), 정보획득량(ig)을 적용하였다. 이 기준들은 학습 문헌 집단에서 특정 범주에 하나의 자질이 출현하느냐 또는 하지 않느냐에 대한 정보를 산출한다(Sebastiani 2002). 예를 들어 카이제곱 통계량은 특정 자질이 특정 범주에 자주 출현할수록 큰 값을 가지며(의존적임), 범주에 대해 독립적일수록 작은 값을 가진다.

특정 자질선정 기준을 적용하여 자질을 선정하거나 축소하는 방법은 그 기준의 공식에 따라 전체 자질들을 계산한 후, 가장 큰 상위 n개의 단어나 미리 설정한 기준치를 넘는 단어를 분류자질로 선정한다. 만약 카이제곱 통계량을 사용한다면 해당 공식에 의해 모든 자질에 대해 통계량을 계산하고 이를 최댓값부터 선택하여 사용한다.

4. 자질선정 실험결과

4.1 사전실험

kNN 분류기는 입력문헌과 유사도가 가장 높은 k개의 최근접 이웃문헌을 학습문헌 집합으로부터 찾아낸 다음 이 이웃문헌들에 배정된 범주들에 근거하여 입력문헌을 분류할 하나 이상의 범주를 선정한다. 따라서 적절한 k 값의 선정이 선행되어야 한다(정영미 2005).

이 연구에서 사용된 중의성 해소 대상 단어 9개에 대하여 kNN 분류기의 성능을 파악하기 위해 다양하게 k 값을 변화시킨 사전 실험을 수행하였으며, 〈표 3〉에 실험 결과를 제시하였다.

〈표 3〉 k 값에 따른 kNN 분류기 성능

k 값	감자	경기	기간	신병	신장	연기	인도	지구	지원	평균
1	0.8463	0.6663	0.9763	0.7150	0.7188	0.5938	0.6063	0.8575	0.7950	0.7528
3	0.8463	0.6663	0.9763	0.7150	0.7188	0.5938	0.6063	0.8575	0.7950	0.7528
5	0.8463	0.6663	0.9763	0.7150	0.7188	0.5938	0.6063	0.8575	0.7950	0.7528
7	0.8463	0.6663	0.9763	0.7150	0.7188	0.5938	0.6063	0.8575	0.7950	0.7528
10	0.8125	0.6675	0.9763	0.6600	0.7075	0.5850	0.6238	0.8500	0.9313	**0.7571**
15	0.8225	0.6625	0.9763	0.6675	0.7138	0.5700	0.6250	0.8438	0.9300	0.7568
20	0.8038	0.6550	0.9763	0.6700	0.7100	0.5525	0.6200	0.8088	0.9288	0.7472
25	0.7675	0.6425	0.9763	0.6700	0.7150	0.5388	0.6088	0.7700	0.9225	0.7346
30	0.7425	0.6375	0.9763	0.6825	0.7188	0.5200	0.6025	0.7475	0.9188	0.7274
40	0.7113	0.6438	0.9763	0.6800	0.7113	0.5050	0.5900	0.7200	0.9163	0.7171
50	0.7100	0.6475	0.9763	0.6975	0.6838	0.5050	0.5875	0.7038	0.9113	0.7136
70	0.7013	0.6513	0.9763	0.6825	0.6313	0.5138	0.5775	0.7138	0.8900	0.7042
100	0.7075	0.6588	0.9763	0.6675	0.5913	0.5138	0.5613	0.7125	0.8838	0.6969
130	0.7038	0.6575	0.9763	0.6575	0.5613	0.5138	0.5575	0.7088	0.8813	0.6908
150	0.7088	0.6500	0.9763	0.6700	0.5500	0.5175	0.5575	0.7075	0.8800	0.6908

이 표를 보면, k 값이 1에서 150까지 변화시켰을 때, k = 10이 가장 좋은 분류 성능을 보였다. 다만, k = 15일 때와 큰 차이를 보이지는 않았다. 전체적으로 보면, k 값이 10 미만일 경우 0.7528로 모두 동일한 값을 가졌으며, 10 이상일 때는 점차 낮아지는 것으로 나타났다. 단어별로 보면, '감자', '신병', '연기', '지구' 등은 k값이 10 미만일 때 제일 좋은 성능을 보였다. '경기', '인도', '지원'은 k 값이 10 내지 15에서 가장 좋은 성능을 보였다. '신장'은 k 값이 작을 때, '지원'은 k 값이 중간값(10~50)에서 좋은 성능을 보였다. 다만, '기간'은 k 값과 무관하게 동일한 성능을 보였는데, 이는 '기간'이 시간을 나타내는 의미로 98.1%의 출현비율을 갖기 때문인 것으로 보인다.

4.2 자질 통계 정보

실험에 쓰인 중의성 단어들의 자질집합 크기를 알아보기 위해 자질의 통계 정보를 실험 문헌으로부터 추출하였다. 우선 문맥의 크기에 따라 실험 대상 단어별 자질집합의 크기를 살펴보면 〈표 4〉와 같았다. 전역 문맥크기에 추출된 학습문헌내의 자질 크기의 평균은 12,827개로 다른 문맥 크기에 비해 크기가 매우 컸다. 특히 좌우 3단어인 1,590개에 비해 8배 정도의 크기를 보였다. 한 문장과 좌우 50바이트의 경우 좌우 50바이트가 크긴 하지만 큰 차이를 보이진 않았다.

단어별로 보면, 대체로 '연기'와 '인도'가 전역, 좌우 50바이트, 좌우 3단어에서 가장 많은 자질을 가진 것으로 나타났다. 다시 말하면, 이

들 단어들이 출현한 문헌들은 다른 단어들의 문헌에 비해 상대적으로 다양한 자질들이 출현하는 것을 의미한다.

특히 이들 단어가 좌우 3단어에서 '경기', '기간', '지원'과 같이 총 출현빈도가 매우 많은 단어들(〈표 2〉 참조)에 비해 더 많은 자질을 갖는다는 것은 더더욱 이러한 사실을 뒷받침한다고 할 수 있다. 반대로 중의성 대상 단어들의 총 출현빈도를 고려해보면, '경기', '기간', '지원'이 상대적으로 적은 수의 자질을 갖는다는 것은 이들 단어가 많은 수의 공통된 단어를 자질로 갖는 것을 의미하며 이러한 공통된 단어들은 연어일 가능성이 크다고 할 수 있다.

다만 한 문장 내에서 중의성 대상 단어의 자질집합 크기를 보면, '경기'가 가장 많은 자질을 갖는 것으로 나타났으나 이는 문장 길이에 의한 결과인지 단어 자체의 특성으로 인한 결과인지 더 분석이 필요한 부분이다. 이 분석에서는 학습문헌 집합이 300개인 '신병'은 비교할 수 없으므로 제외하였다.

이 연구에서는 자질선정 기준에 의해 순위화된 결과를 이용하여 자질집합의 크기를 제한하였다. 전체 자질집합의 크기에 따른 성능을 비교하기 전에 자질선정 기준에 대해 상위 10개 단어를 제시하면 〈표 5〉, 〈표 6〉과 같다. 〈표 5〉는 카이제곱 통계량 기준으로 전역 문맥에서 선정된 상위 자질 10개이며, 〈표 6〉은 문헌빈도 기준으로 좌우 3단어에서 선정된 상위 자질 10개이다. 이들 표에서 선정된 자질의 순위는 좌에서 우로 순서화 되어 있다. 즉 〈표 5〉의 '감자'의 상위 10개 자질을 살펴보면, 주주, 지분, 채권단 순으로 카이제곱 통계량이 작아진다. 카이제곱 통계량과 문헌빈도 기준에 의해 상위 10위에 위치한 단어 대부분은 특정 의미에서 자주 출현하는 단어들로 채워져 있는 것으로 보인다. 예를 들어 '감자'에 대한 자질들의 카이제곱 통계량 상위 10개를 보면 주주, 지분, 채권단, 매각, 소액, 증권, 채권, 인수, 금융 등으로 '기업이 자본금의 액수를 줄이는 일'이란 두 번째 의미로 사용된 핵심적인 자질들이 제시되었다. 이 자질들의 카이제곱 통계량이 특정 의미 범주에 의존적일수록 높은 값을 가지기 되는데, 이들 단어들이 두 번째 의미 범주에서만 나타나기 때문에 높은 값을 가지는 것이다.

〈표 4〉 문맥크기에 따른 학습문헌내의 자질집합 크기

단어	좌우3단어	한 문장	좌우 50B	전역
감자	1,449	2,333	3,273	11,465
경기	1,674	4,199	3,973	12,281
기간	1,667	3,055	4,006	14,242
신병	825	1,773	2,106	6,804
신장	1,634	3,187	3,932	13,729
연기	1,962	3,245	4,663	15,753
인도	1,851	3,317	4,245	14,098
지구	1,574	3,251	3,696	13,369
지원	1,677	3,571	3,919	13,706
평균	1,590	3,103	3,757	12,827

〈표 5〉 카이제곱 통계량 자질선정 결과(상위 10개 단어, 전역)

대상	선정 자질
감자	주주, 지분, 채권단, 매각, 소액, 은행, 증권, 채권, 인수, 금융
경기	회복, 침체, 팔다리, 시, 서울, 열, 지역, 아기, 체온, 용인
기간	통신망, 사업자, 망, 데이콤, 시청료, 통신, 국가, 사업법, 민영, 공영
신병	훈련, 교육, 치료, 입대, 군대, 군, 부대, 훈련소, 육군, 주간
신장	개업, 환자, 병원, 치료, 수술, 신당, 이식, 농구, 선수, 키, 민주당
연기	영화, 배우, 화재, 드라마, 작품, 출연, 인드라, 참선, 불교, 불
인도	지원, 신병, 범죄인, 차도, 협의, 단속, 대북, 북한, 주차, 주차장
지구	아파트, 세계, 분양, 주택, 미국, 우주, 인간, 건설, 단지, 과학
지원	지법, 서울, 모집, 대학, 서부, 판사, 남부, 북부, 부, 형사

〈표 6〉 문헌빈도 자질선정 결과(상위 10개 단어, 좌우 3단어)

대상	선정 자질
감자	주주, 소액, 튀김, 주식, 완전, 안, 카드, 비율, 지분, 뒤
경기	회복, 팀, 침체, 전망, 지역, 서울, 기업, 리그, 건설, 말
기간	지난해, 작년, 이상, 동안, 연장, 일정, 말, 유예, 계약, 증가
신병	확보, 처리, 인도, 교육, 훈련, 치료, 검찰, 김, 전, 미군
신장	이식, 기능, 인권, 질환, 매출, 간, 투석, 수술, 사람, 체중
연기	배우, 말, 때, 영화, 역, 드라마, 안, 요청, 뒤, 사람
인도	중국, 지원, 미국, 일본, 한국, 세계, 대한, 국가, 파키스탄, 말
지구	개발, 택지, 아파트, 지정, 투기, 과열, 분양, 지역, 계획, 예정
지원	정부, 대한, 대학, 자금, 미국, 경제, 사업, 북한, 이라크, 경우

문헌빈도 기준으로 좌우 3단어에서 추출한 자질을 살펴보면, 정보검색에서 문헌빈도가 높은 경우 주제적 특성이 약한 자질들의 상당수가 의미 분류에 도움이 되는 것을 알 수 있다. 예를 들어 '경기'의 경우 자질 '지역'이나, '연기'나 '감자'의 자질 '뒤'를 들 수 있다.

4.3 의미분류 성능분석

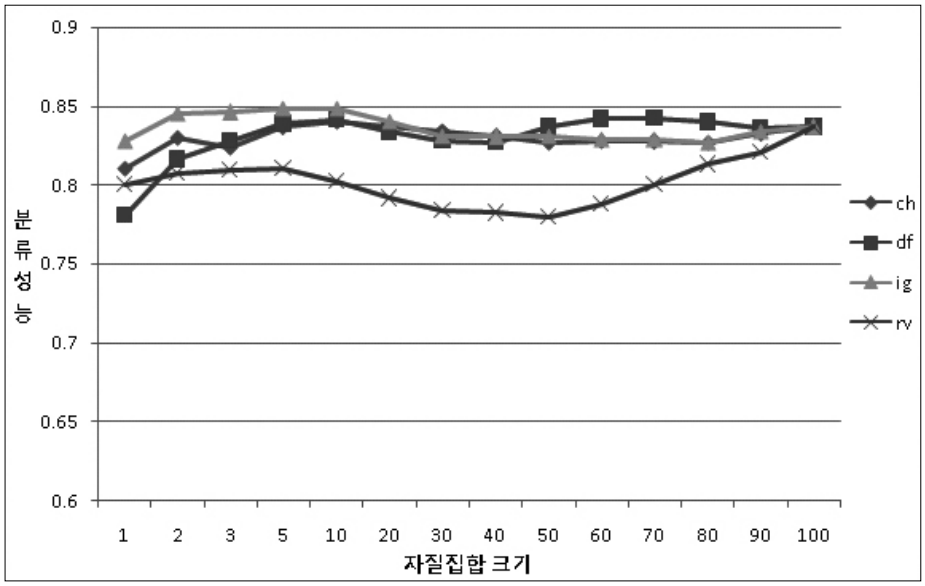
이 연구에서 구축한 의미 분류기의 성능을 분석하기 위해 자질선정 방법, 분류기, 문맥 크기, 단어에 따라 분류기 성능 실험 결과를 분석하였다.

먼저 자질선정 방법에 따른 의미 분류 성능을 보면, 이 연구에서 사용된 세 분류기(NB 분류기, kNN 분류기, SVM 분류기)의 분류 성능의 평균값을 이용하여 자질집합의 크기에 따른 성능을 조사하여 〈표 7〉과 〈그림 1〉을 얻었다.

이 표에서 자질선정 방법에 따른 성능을 살펴보면, 정보획득량(ig)이 0.8361로 가장 좋은 분류 성능을 보였으며, 문헌빈도(df) 0.8305, 카이제곱 통계량(ch) 0.8300, 적합성 함수(rv) 0.8018 순으로 나타났다. 이와 같은 결과는 기존의 연구와 같았다. 자질 선정 방법 각각에 대해 〈표 7〉과 〈그림 1〉을 통해 분류 성능을 좀 더 살펴보면, 정보획득량은 각각의 자질집합 크기에서

〈표 7〉 자질선정 방법과 자질집합 크기에 따른 분류 성능(분류기 평균)

자질집합	ch	df	ig	rv	평균
1%	0.8107	0.7811	0.8281	0.8001	0.8050
2%	0.8301	0.8163	0.8453	0.8075	0.8248
3%	0.8238	0.8278	0.8459	0.8093	0.8267
5%	0.8368	0.8389	0.8484	0.8107	**0.8337**
10%	0.8400	0.8414	0.8485	0.8021	0.8330
20%	0.8365	0.8338	0.8406	0.7916	0.8256
30%	0.8338	0.8277	0.8311	0.7838	0.8191
40%	0.8304	0.8272	0.8305	0.7822	0.8176
50%	0.8267	0.8367	0.8309	0.7800	0.8186
60%	0.8279	0.8419	0.8287	0.7880	0.8216
70%	0.8275	0.8424	0.8293	0.8002	0.8248
80%	0.8272	0.8399	0.8269	0.8130	0.8268
90%	0.8325	0.8359	0.8344	0.8208	0.8309
100%	0.8365	0.8365	0.8365	0.8365	0.8365
평균	0.8300	0.8305	**0.8361**	0.8018	0.8246



〈그림 1〉 자질집합 크기에 따른 자질선정 기준의 분류성능(분류기 평균)

전체적으로 좋은 성능을 가져왔으며 문헌빈도는 1%에서 낮은 성능을 보였고 60%, 70%에서 가장 좋은 성능을 보였다. 카이제곱 통계량은 정보획득량과 유사한 모습을 보였다. 다만 최적화되지 않은 kNN을 제외하면 정보획득량의 10% 자질집합 크기에서 가장 좋은 분류 성

능을 보였다.

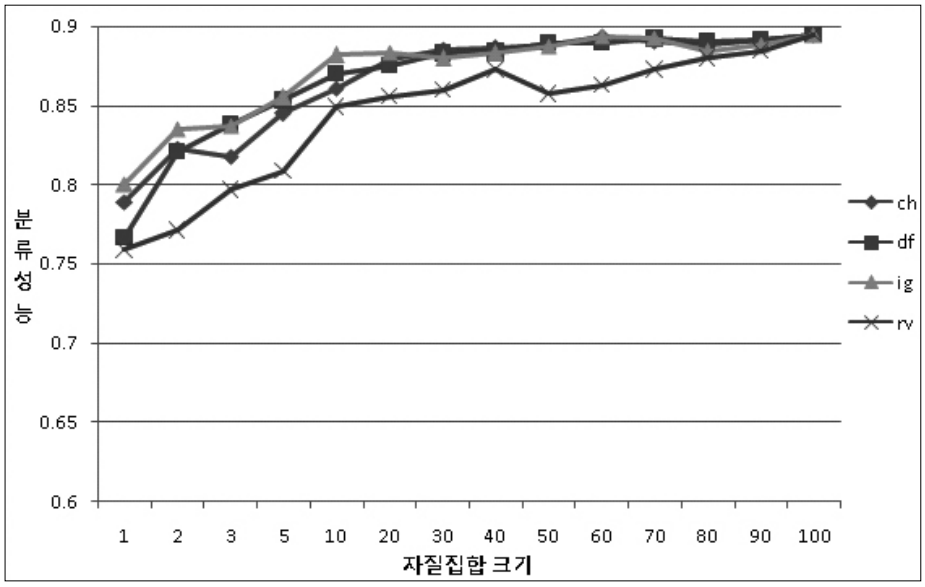
자질집합 크기의 경우, 전체 자질을 100%로 보고 각각의 자질선정 기준으로부터 상위 1%, 2%, 3%, 5%, … 등 일부분을 사용하여 분류 성능을 조사하였다. 그 결과 자질집합 크기 상위 5%와 10%에서 가장 좋은 성능을 보였다. 주목할 점은 자질집합의 크기에 큰 영향을 받지 않고 작은 자질집합 크기에서도 비교적 좋은 성능을 유지하였다는 점이다. 즉 텍스트 범주화 기법처럼 단어 중의성 해소에서도 자질선정 방법이 매우 유용한 수단이 될 것으로 생각된다.

자질선정 방법과 자질집합 크기에 따른 각각의 분류기의 성능을 표현하면, SVM 분류기의 성능은 〈그림 2〉, NB 분류기는 〈그림 3〉, kNN 분류기는 〈그림 4〉와 같았다.

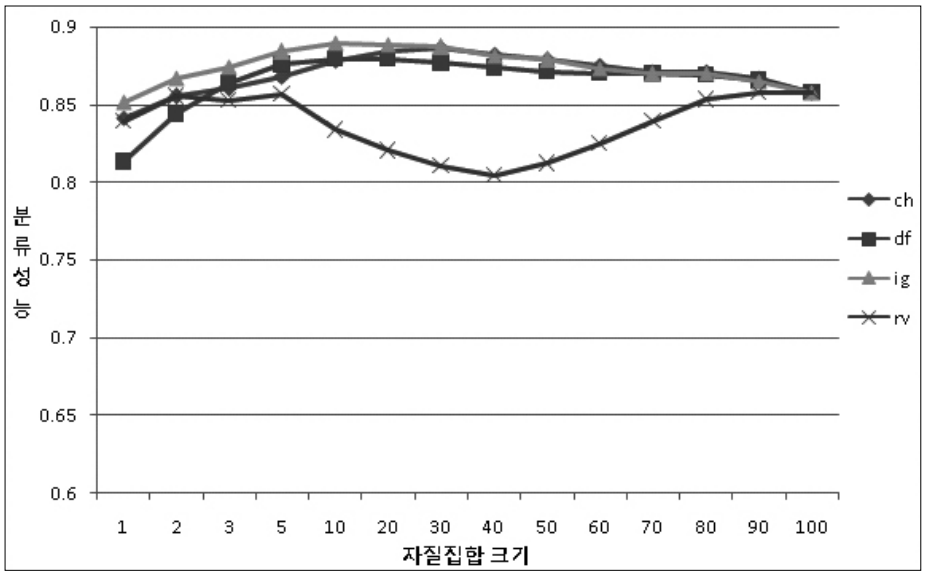
〈그림 2〉에서 SVM 분류기는 자질선정 방법 보다 자질집합의 크기가 클수록 더 좋은 성능을 보였다. 다만 자질집합의 크기 10% 이후부터는 큰 차이를 보이지 않았다. 또한 전체 자질집합 크기에서 다른 분류기와 성능을 비교해 보면, SVM 분류기의 성능이 0.8946으로 NB (0.8579), kNN(0.7570)에 비해 상당히 좋은 성능을 보였다. 이는 SVM 분류기가 자질 선정에 그리 영향을 받지 않는다는 것을 의미한다.

〈그림 3〉을 보면, NB 분류기는 전반적으로 10% 전후에 자질집합 크기에서 가장 좋은 성능을 보였다(적합성 함수 기준 제외). 특히 정보획득량을 이용한 NB 분류기의 성능은 자질집합 크기 10%에서 0.8899로, SVM 분류기의 최고 성능인 0.8946에 근접할 정도로 좋은 성능을 보였다.

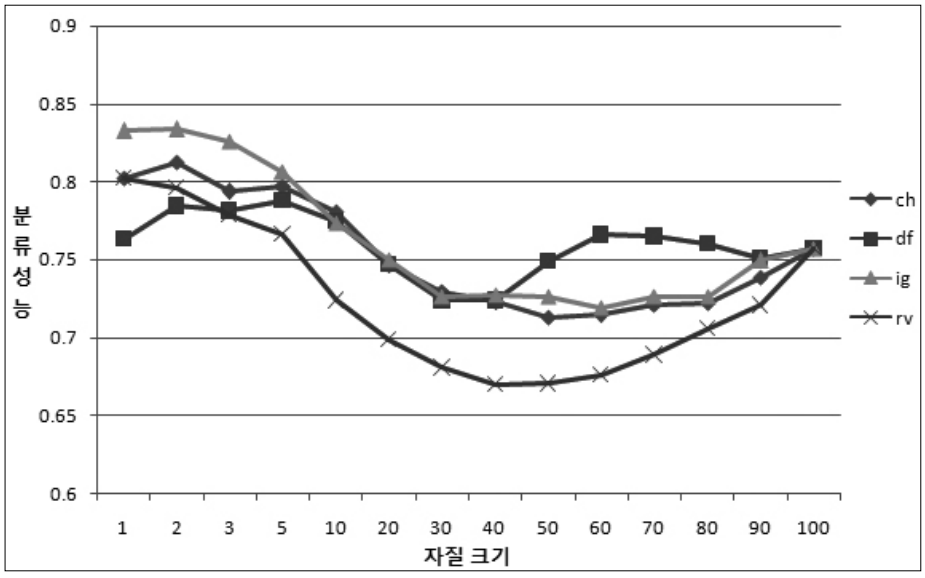
〈그림 4〉의 kNN 분류기는 자질집합 크기가 10% 이하에서 좋은 성능을 보였다. 다만 이 연구에서는 이원 가중치를 사용하였으므로 좀 더 다양한 실험을 통한 보완이 필요하다.



〈그림 2〉 자질집합 크기에 따른 자질선정 기준의 SVM 분류기 성능



〈그림 3〉 자질집합 크기에 따른 자질선정 기준의 NB 분류기 성능

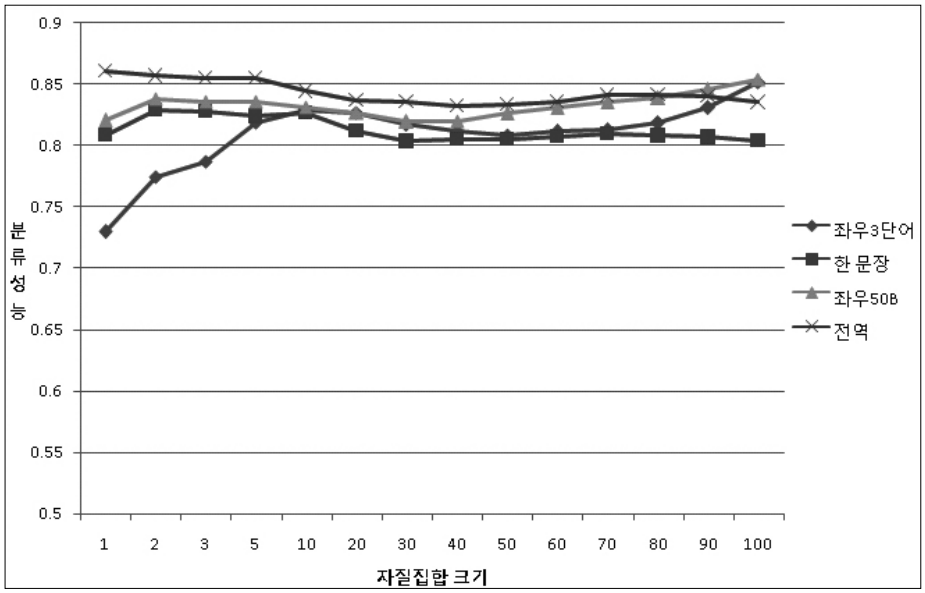


〈그림 4〉 자질집합 크기에 따른 자질 선정 기준의 kNN 분류기 성능

다음으로 자질집합 크기와 문맥 크기에 따른 의미 분류 성능을 조사하여 〈표 8〉과 〈그림 5〉를 얻었다. 자질집합 크기를 100%, 즉 전체 자질을 적용했을 경우 좌우 50바이트의 문맥에서 0.8541로 가장 좋은 성능을 보였으며, 좌우 3단어(0.8522), 전역(0.8354), 한 문장(0.8044) 순

〈표 8〉 자질집합 크기에 따른 문맥 크기의 분류 성능(분류기 평균)

자질집합	좌우3단어	한 문장	좌우50B	전역	평균
1%	0.7297	0.8088	0.8208	**0.8607**	0.8050
2%	0.7746	0.8295	0.8381	0.8570	0.8248
3%	0.7874	0.8284	0.8358	0.8552	0.8267
5%	0.8194	0.8241	0.8363	0.8550	0.8337
10%	0.8285	0.8269	0.8318	0.8447	0.8330
20%	0.8266	0.8120	0.8270	0.8369	0.8256
30%	0.8175	0.8038	0.8195	0.8357	0.8191
40%	0.8121	0.8054	0.8199	0.8329	0.8176
50%	0.8086	0.8057	0.8263	0.8337	0.8186
60%	0.8115	0.8079	0.8309	0.8361	0.8216
70%	0.8126	0.8102	0.8354	0.8412	0.8248
80%	0.8184	0.8083	0.8388	0.8416	0.8268
90%	0.8308	0.8071	0.8455	0.8403	0.8309
100%	0.8522	0.8044	**0.8541**	0.8354	0.8365
평균	0.8093	0.8130	0.8329	0.8433	0.8246



〈그림 5〉 자질집합 크기와 문맥 크기에 따른 분류 성능(분류기 평균)

으로 나타났다. 좌우 50바이트와 좌우 3단어 사이의 성능 차이가 그리 크지 않았다.

자질집합 크기와 문맥 크기를 동시에 고려하면 전역 문맥에서 자질집합 크기 1%가 0.8607로 가장 좋은 성능을 보였으며, 전역 문맥의 자질집합 크기 80%까지 가장 좋은 성능을 보이

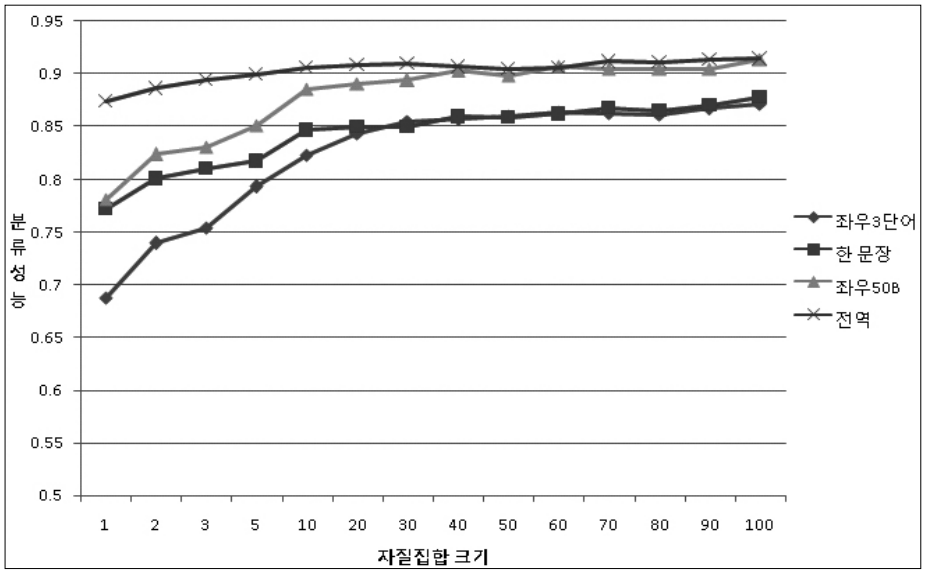
다가 90% 이상에서는 좌우 50바이트가 더 좋은 성능을 보였다. 뿐 만 아니라 전체 자질집합(100%)을 적용했을 때는 좌우 3단어도 비슷한 성능을 보였다. 이는 〈그림 5〉에서 보면 더 쉽게 알 수 있다. 이러한 결과로 볼 때, 전역 문맥 크기에서 추출한 모든 자질을 사용하는 것이 분류 성능에 저하를 가져오는 것을 알 수 있다.

최적화되지 않은 kNN을 제외하면 자질집합 크기 10% 미만에서는 전역 문맥이 가장 좋은 성능을 가져왔으며, 10%에서 40%까지는 전역 문맥과 좌우 50바이트가 거의 유사한 성능을 보였으며, 그 이후에는 좌우 50바이트가 가장 좋은 분류 성능을 보였다.

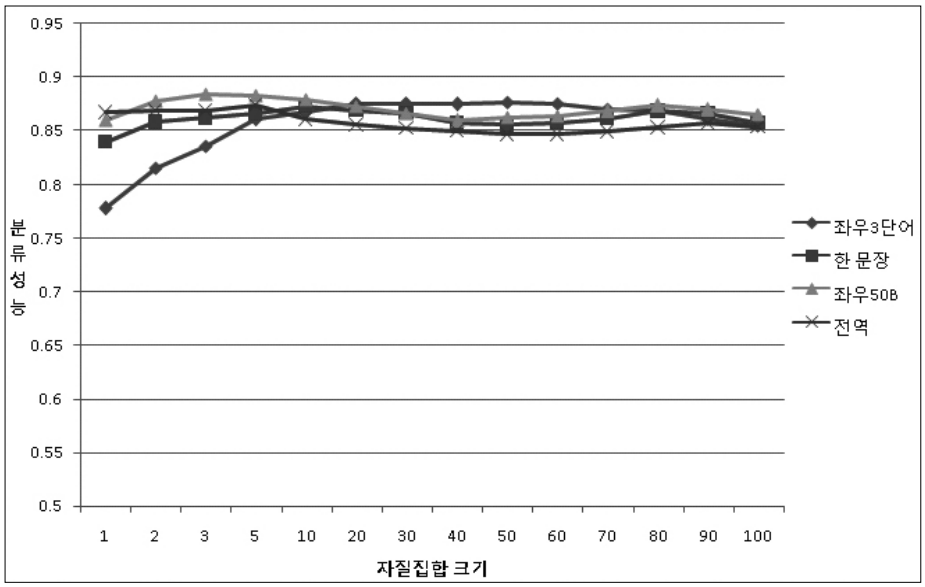
좌우 3단어 또는 좌우 50바이트처럼 문맥 크기가 작은 경우 자질집합을 되도록 크게 가져가야 하며, 반대로 전역과 같이 문맥 크기가 큰 경우 자질집합의 크기를 작게 하면 더 좋은 성능을 보이는 것으로 생각된다. 다만, 이 연구에서는 이러한 결과가 kNN 분류기의 성능에 의해 많은 영향을 받는 것으로 생각되어 좀더 깊이 있는 연구가 이루어져야 할 것으로 보인다.

분류기별로 자질집합 크기와 문맥 크기의 분류 성능을 살펴보면, 먼저 SVM 분류기는 세 분류기 중 가장 좋은 성능(자질집합 크기 100%, 전역 문맥: 0.9144)을 가져왔으며 대체로 큰 문맥 크기와 자질집합 크기로부터 좋은 성능을 가져왔다. 〈그림 6〉을 보면, 문맥 크기가 성능에 많은 영향을 미치는 것을 알 수 있다.

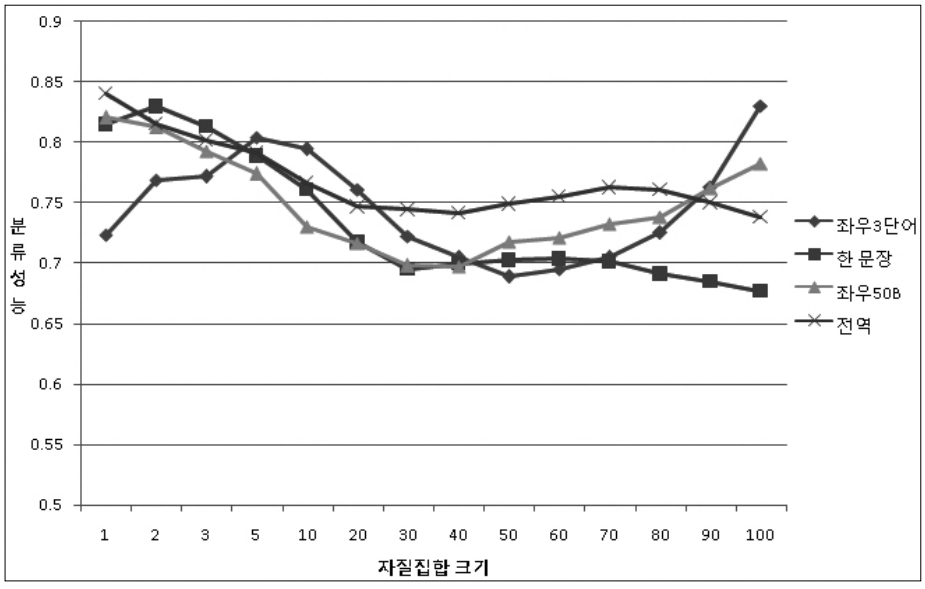
NB 분류기의 경우 큰 자질집합으로부터 약간의 성능 저하를 가져왔다. 〈그림 7〉을 보면 좌우 50 바이트와 자질집합의 크기 3%에서 가장 좋은 성능(0.8841)을 보였다. kNN 분류기는 극명하게 작은 자질집합에서는 큰 문맥이, 큰 자질집합에서는 작은 문맥크기가 가장 좋은 성능을 보였다(〈그림 8〉 참조). NB 분류기와 kNN 분류기의 경우 잡음 자질이나 의미 없는



〈그림 6〉 자질집합 크기와 문맥 크기에 따른 SVM 분류기 성능



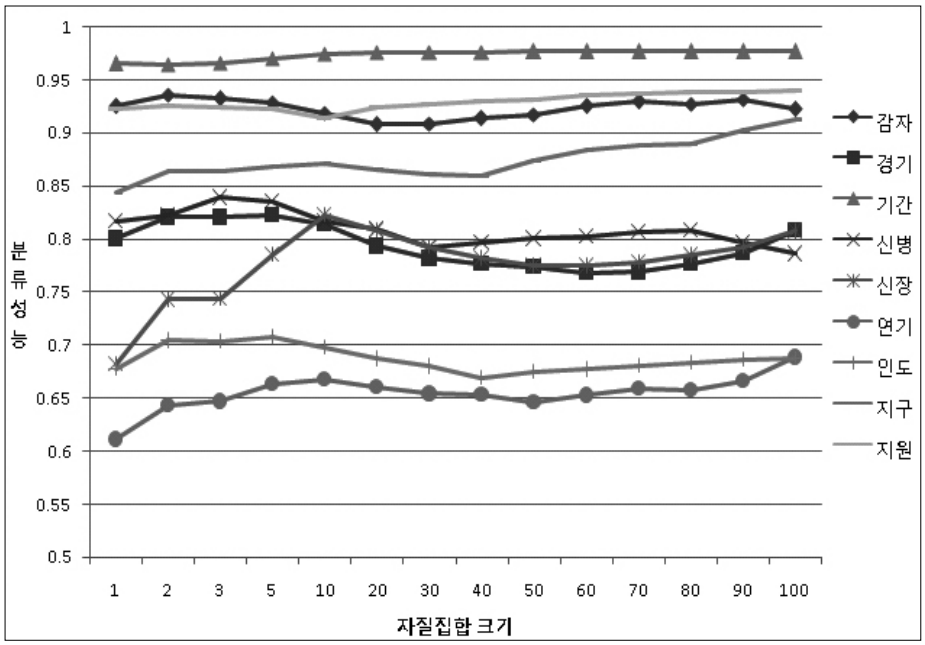
〈그림 7〉 자질집합 크기와 문맥 크기에 따른 NB 분류기 성능



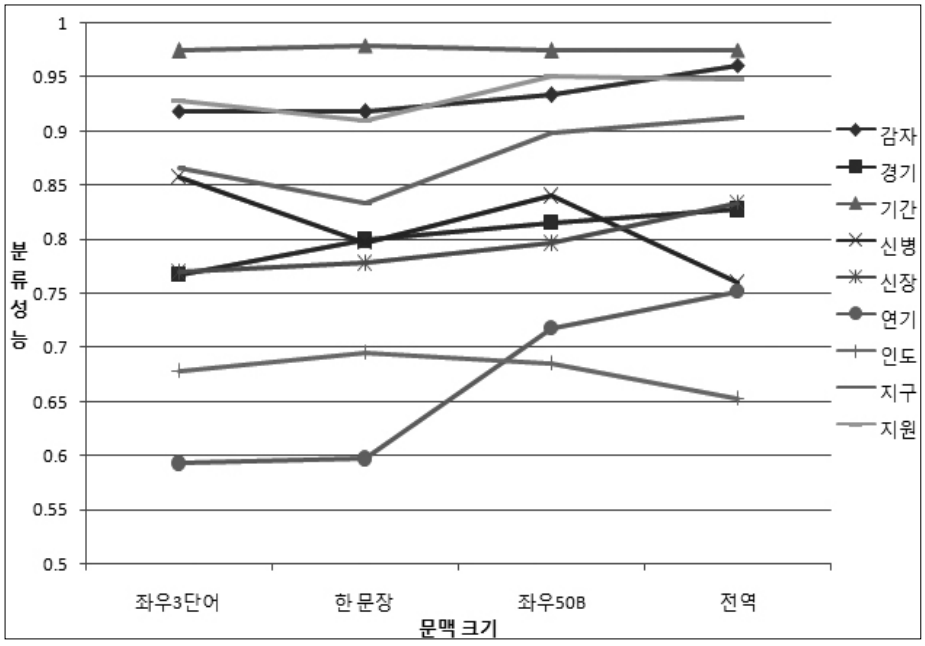
〈그림 8〉 자질집합 크기와 문맥 크기에 따른 kNN 분류기 성능

자질이 영향을 미치는 것으로 나타났다.

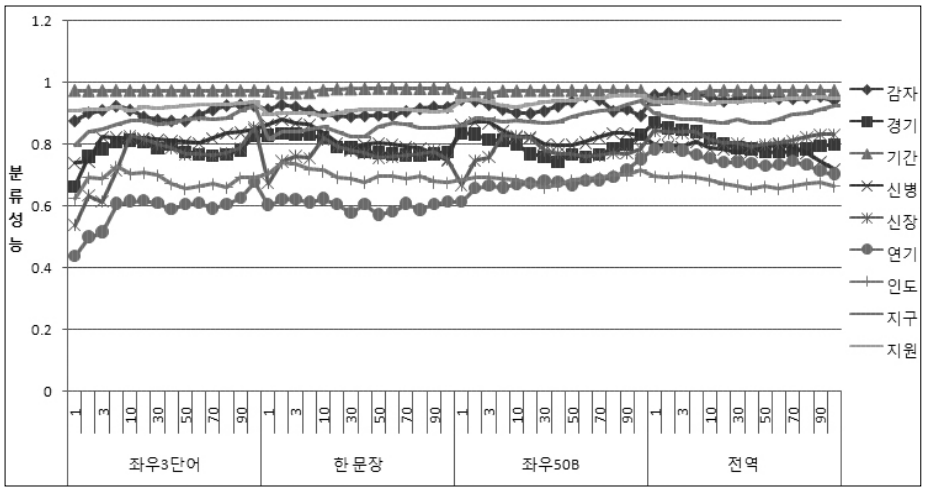
다음으로 자질집합 크기와 문맥 크기에 따른 단어별 의미 분류 성능을 파악하였다. 먼저 자질집합 크기에 따른 단어별 성능을 살펴보면, 〈그림 9〉와 같았다. 또한 문맥 크기에 따른 단어별 성능은 〈그림 10〉과 같았다. 문맥 크기



〈그림 9〉 자질집합 크기에 따른 단어별 분류 성능



〈그림 10〉 문맥 크기에 따른 단어별 분류 성능



〈그림 11〉 문맥 크기내 자질집합 크기에 따른 단어별 성능

에서 자질집합 크기에 따른 성능은 〈그림 11〉과 같았다.

이들을 보면, '기간'은 자질집합의 크기나 문맥의 크기에 상관없이 동일한 성능을 보였다. 이 단어의 출현빈도 분포와 의미 수에 따른 영향으로 보인다. '감자', '지구', '지원'과 '연기'는 자질집합의 크기와 문맥의 크기가 커질수록 좋은 성능을 보였다. '경기'와 '신장'은 자질집합 크기보다 문맥의 크기에 영향을 받는 것으로 보였다. 특히 작은 문맥에서 좋은 성능을 보여 연어효과의 가능성을 보였다. '신병'은 자질집합의 크기와 문맥의 크기가 커질수록 성능이 낮아졌다. 마지막으로 '인도'는 자질집합의 크기와 문맥의 크기에 그다지 영향을 받지 않았다.

5. 결 론

이 연구에서는 단어의 중의성을 해소하기 위한 지도학습 기반의 방법에서 분류 성능을 효율적으로 높이기 위한 방안으로 자질선정에 대하여 실험하였다. 자질선정을 위해 정보획득량, 카이제곱 통계량, 문헌빈도, 적합성 함수 등을 사용하여 다양한 크기의 학습문맥 창에 적용하였다. 이를 통해 자질선정 방법, 문맥의 크기, 분류기, 단어별 의무 분류 성능을 분석하여 각각의 요인에 대한 특징을 제시하였다.

이 연구를 통해 단어 중의성 해소 실험에서 발견한 사실은 다음과 같다.

첫째, 일반 텍스트 범주화 기법과 같이 단어 중의성 해소에서도 자질선정 방법이 의미 분류에 매우 유용한 수단이 됨을 알 수 있었다. 자질선정 기준은 다른 연구와 유사하게 정보획득량, 문헌빈도, 카이제곱 통계량이 좋은 성능을 보였다.

둘째, 자질집합의 크기와 문맥의 크기에 따라 분류기별 성능을 보면, SVM 분류기는 자질집합 크기와 문맥 크기가 클수록 더 좋은 성능을 보여 자질선정에 영향을 받지 않았다. 반면 나이브 베이즈 분류기는 10% 정도의 자질집합

크기에서 가장 좋은 성능을 보였다. kNN의 경우 10% 이하의 자질집합에서 가장 좋은 성능을 보였다.

셋째, 단어 중의성 해소를 위한 자질선정을 적용할 때 작은 자질집합 크기과 큰 문맥 크기를 조합하거나, 반대로 큰 자질집합 크기와 작은 문맥 크기를 조합하면 성능을 극대화 할 수 있다.

넷째, 자질집합과 문맥의 크기에 따라 더 좋은 성능을 보이는 단어('감자', '지구', '지원', '연기')도 있고, 그렇지 않은 단어('인도')도 있다. '경기'와 '신장'은 자질집합 크기보다 문맥 크기에 영향을 받았다. 특히 연어를 갖는 단어('신병')는 매우 작은 문맥 크기에서 좋은 성능을 보였다.

다만, 이 연구의 결과를 좀 더 일반화하기 위해 다른 문헌집단, 좀 더 다양한 분류기, 가중치 방법 사용할 필요성이 있으며, 추후 단어 중의성 해소만을 위한 고유한 자질, 예를 들어 다양한 문법적 속성이나 연어 등을 이용한 자질 선정 기법을 적용할 필요성이 있다.

참 고 문 헌

정영미. 2005. 『정보검색연구』. 서울: 구미무역(주) 출판부.

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

Escudero, G., L. Marquez, and G. Rigau. 2000. "Naive Bayes and Exemplar-based Approaches to Word Sense Disambiguation Revisited." *Proceedings of the 14th European Conference on Artificial Intelligence*, 421-425.

Fragos, K. 2008. "Disambiguation of Greek Polysemous Words Using Hierachical Probabilistic Networks and a Chi-square Feature Selection Strategy." *International Journal on Artificial Intelligence Tools*, 17(4): 687-701.

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

Guyon, I. and A. Elisseeff. 2003. "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research*, 3: 1157-1182.

Hoste, V., W. Daelemans, I. Hendrickx, and A. Bosch. 2002. "Evaluating the Results of a Memory-based Word-expert Approach to Unrestricted Word Sense Disambiguation." *Proceedings of the Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, 95-101.

Jackson, P. and I. Moulinier. 2002. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and*

Categorization. Amsterdam: Benjamins Publishing Co.

Joachims, T. 2001. *Learning to Classify Text Using Support Vector Machines*. Boston: Kluwer Academic Publishers.

Mihalcea, R. 2002. "Word Sense Disambiguation with Pattern Learning and Automatic Feature Selection." *Natural Language Engineering*, 8(4): 343-358.

Navigli, R. 2009. "Word Sense Disambiguation: A Survey." *ACM Computing Surveys*, 41(2): 1-69.

Ng, T. and H. B. Lee. 1996. "Integrating Multiple Knowledge Sources to Disambiguate Word Senses: An Exemplar-based Approach." *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, 40-47.

Orhan, Z. and Z. Altan. 2006. "Impact of Feature Selection for Corpus-based WSD in Turkish." *Proceedings of the MICAI 2006: Advances in Artificial Intelligence (LNCS 4293)*, 868-878.

Sebastiani, F. 2002. "Machine Learning in Automated Text Categorization." *ACM Computing Surveys*, 34(1): 1-47.

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

Stevenson, M. and Y. Wilks. 2001. "The Interaction of Knowledge Sources in Word Sense Disambiguation." *Computational Linguistics*, 27(3): 321-349.

Strapparava, C., A. Gliozzo, and C. Giuliano. 2004. "Pattern Abstraction and Term Similarity for Word Sense Disambiguation: IRST at Senseval-3." *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 229-234.

Suarez, A. and M. Palomar. 2002. "Improving Feature Selection for Maximum Entropy-based Word Sense Disambiguation." *Proceedings of the PorTAL 2002(LNAI 2389)*, 15-23.

Yang, Y. and J. O. Pedersen. 1997. "A Comparative Study on Feature Selection in Text Categorization." *Proceedings of the 14th International Conference on Machine Learning*, 412-420.