저빈도어를 고려한 개념학습 기반 의미 중의성 해소

김동성·최재웅*†
고려대학교

Dong-Sung Kim and Jae-Woong Choe. 2006. Word Sense Disambiguation based on Concept Learning with a focus on the Lowest Frequency Words. *Language and Information 10.1*, 21–46. This study proposes a Word Sense Disambiguation (WSD) algorithm, based on concept learning with special emphasis on statistically meaningful lowest frequency words. Previous works on WSD typically make use of frequency of collocation and its probability. Such probability based WSD approaches tend to ignore the lowest frequency words which could be meaningful in the context. In this paper, we show an algorithm to extract and make use of the meaningful lowest frequency words in WSD. Learning method is adopted from the Find-Specific algorithm of Mitchell (1997), according to which the search proceeds from the specific predefined hypothetical spaces to the general ones. In our model, this algorithm is used to find contexts with the most specific classifiers and then moves to the more general ones. We build up small seed data and apply those data to the relatively large test data. Following the algorithm in Yarowsky (1995), the classified test data are exhaustively included in the seed data, thus expanding the seed data. However, this might result in lots of noise in the seed data. Thus we introduce the "maximum a posterior hypothesis" based on the Bayes' assumption to validate the noise status of the new seed data. We use the Naive Bayes Classifier and prove that the application of Find-Specific algorithm enhances the correctness of WSD. (**Korea University**)

Key words: 의미중의성 해소 (Word Sense Disambiguation), 기계학습 (machine learning), 저빈도어 (lowest frequency words), 개념학습 (concept learning), 동음이의어 (homonym), 나이브 베이즈 구분자(Naive Bayes Classifier)

* 본 논문에 대하여 세밀하고 건설적인 조언을 해 주신 세 분의 익명 심사자께 깊은 감사를 드린다. 그러한 조언과 지적 덕분에 본 논문의 구조나 스타일, 그리고 논증 등이 눈에 띄게 좋아졌다고 생각한다. 수정 과정을 통해서 그러한 지적을 최대한 반영하도록 노력하였다. 아직 남아있을지 모르는 본 논문의 약점이나 흠결은 물론 전적으로 필자들의 책임이다. 본 연구의 주요 내용은 The 2006 KALS-KASELL International Conference on English and Linguistics (Pusan National University, Korea, June 29-30, 2006)에서 "The Supervised Boostrapping Word Sense Disambiguation Methods, Using the Statistical Cues of the Lowest-Frequency Words"라는 제목으로 발표된 바 있다.

† 서울시 성북구 안암동 5가 고려대학교 언어학과. Email: dsk202@korea.ac.kr, jchoe@korea.ac.kr

1. 서론

본 연구에서는 동음이의어의 주변 맥락에 나타난 저빈도어를 고려하는 개념학습 기반의 의미 중의성 해소 알고리즘을 개발하고 이를 평가하고자 한다. 동음이의어에 대한 의미 중의성 해소는 비단 자연언어 의미 처리뿐만이 아니라 정보검색, 기계번역, 질의 검색 시스템과 같은 여러 인공지능 분야의 기초가 된다. 그런데 의미 중의성은 기본적으로는 화자의 직관에 의해 해소될 수 있으나, 모든 자료를 직관에 의존해 분류하기에는 너무 많은 시간과 노력이 필요하므로 실질적으로 불가능한 일이다. 따라서 중의적 표현을 기계적으로 자동 분류하는 방식에 대한 연구가 필요하다.

중의적 표현에 대한 자동 의미 분류 방식 중 대표적인 것으로 기존의 확률 및 통계 기반 알고리즘을 들 수 있다. 이는 어휘의 절대 빈도에 기반을 두어 중의적 표현의 의미를 자동 분류하는 방식으로, 이러한 알고리즘에서는 저빈도어는 실질적 고려의 대상이 되지 않았다 (Weeber, Vos, and Baayen, 2000).

그러나 실제 코퍼스를 조사해 보면 저빈도 어휘들이 전체 어휘의 다수를 차지한다. 또한 중의성 해소의 단서가 되는 단어들이 낮은 빈도로 출현하는 경우들도 있으므로 저빈도어가 고려되지 않는다면 의미 중의성 해소에 필요한 중요 정보가 무시되게 된다. 따라서 의미 중의성 해소에 기여하는 저빈도어를 활용하는 알고리즘이 필요하다고 할 수 있다. 본 연구에서는 화자의 언어직관을 반영하여 저빈도어를 포함한 의미 있는 주변어를 추출하는 방식을 제안하고, 아울러 저빈도어가 의미 중의성 해소에 어떻게 기여하는지를 논한다.

본 논문에서는 연어 관계를 기초로 한 통계적 기법을 도입하여 3개의 동음이의어 '배, 신부, 거리'를 대상으로 의미 중의성 해소를 연구하였다. 연구대상이 된 자료는 세종계획 2단계 연구 결과 중 한국어 학습용 어휘 선정을 위한 기초 조사에 활용된 '현대국어 사용 빈도 조사' 말뭉치 100만 어절(1998년에서 2002년, 이하 [세종100만어절말뭉치])과 21세기 세종계획 연구 교육용 '현대국어 균형 말뭉치' 1,000만 어절(1998년에서 2002년, 이하 [세종1000만어절말뭉치])이다. 전부 형태소 분석이 된 코퍼스를 연구 자료로 선정하였다.

전산언어학의 주요한 가설 중에 하나로 제시된 "분포가설(Distributional hypothesis: Harris (1964))"에 따르면, 단어의 의미적 특성을 주변 맥락에 근거하여 파악할 수 있다. 이러한 가설에 따르면, 의미 중의성 해소는 주변 맥락에 나오는 단어들의 분포적 특성과 관련이 있다. 즉, 주어진 핵심어(keyword)의 주변어를 분석함으로써 그 어휘의 의미적 중의성을 해소하기 위한 단서를 찾아 낼 수 있다는 것이다. 본 연구에서는 통계처리 방식인 t-스코어를 이용하여 주변어의 통계적 유의미성을 검증하게 될 것이다.

자연어처리에서 미리 화자 직관을 이용하여 분류된 데이터를 활용하는 방식이 많이 쓰이고 있으나, 실제 데이터를 분류하는 데 많은 노력과 시간이 걸리는 이른바 "지

식습득의 병목현상(knowledge acquisition bottleneck)"의 문제가 대두 된다 (Ide and Vernois, 1998). 이런 문제를 해결하기 위해 본 연구에서는 상대적으로 적은 수의 학습 집단에서 추출된 정보를 활용하여 많은 수의 테스트 집단을 분류하는 부트스트래핑(bootstrapping) 방식을 활용하였다. 이 방식에 따르면 화자 직관을 활용하여 얻은 적은 양의 학습 데이터를 바탕으로 많은 양의 실험 데이터를 분류할 수 있다. 아울러 Mitchell (1997)에서 제안된, 개념학습을 활용하는 Find-Specific 알고리즘 방식의 기계학습 방식을 이용하였다.

본 연구에서는 Yarowsky (1995)의 제안에 따라 실험 데이터[1]를 처리하여 얻은 분류된 데이터에서 유의미한 공기어휘 집합을 추출한 뒤에 이를 이미 획득된 학습 공기어휘 집합에 추가하는 방식을 활용하였다. 이렇게 하면 적은 데이터로 많은 데이터를 처리할 수 있는 반면, 추가되는 데이터가 많은 노이즈(noise)를 포함할 경우 실험의 정확도가 떨어지게 된다는 문제점을 들 수 있다. 따라서 추가되는 데이터가 실제로 노이즈가 없고 정확성 있는 데이터인지 검증할 필요가 있다. 이러한 검증을 위해 베이즈 확률의 '최대 이후 가설 (Maximum a posteriori hypothesis)'을 이용하였다. 이러한 검증을 통해서 결과적으로 Find-Specific 알고리즘을 적용한 방식에서 정확도가 증가하는 것으로 나타났으며, 추가되는 데이터가 노이즈가 적고 신뢰될 수 있다는 것이 입증되었다.

본 논문의 구성은 다음과 같다. 2절에서는 우선 본 논문에서 주안점인 저빈도어에 대한 논의와 더불어 기존 연구를 소개하고 본 연구의 차이점을 간략하게 설명할 것이다. 3절에서는 Find-Specific 알고리즘과 관련된 개념학습을 소개하고, 본 논문의 알고리즘을 제시하며, 유의미한 연어를 추출하기 위한 통계적 장치인 t-스코어 검증을 논의할 것이다. 4절에서는 베이즈 확률의 '최대 이후 가설'을 적용한 Find-Specific 알고리즘의 정확도 검증의 방편으로 나이브 베이즈 구분자를 활용할 것이다. 5절은 실험과 관련된 논의를 진행하고, 6절은 의미 중의성 해소 실험 결과에 대한 논의를 할 것이다.

2. 저빈도어와 기존 연구

2.1 저빈도어

본 연구는 핵심어의 중의성 해소를 위해 그 핵심어 좌우 맥락에 나오는 주변어들의 통계값을 활용한다. 주변어는 핵심어의 중의성 해소와 연관이 있는 어휘들과 그렇지 않은 어휘들로 나뉠 수 있다. 또 연관이 있는 어휘들 중에는 출현 빈도가 높은 어휘도 있는 반면 낮은 어휘들도 있다. 본 연구에서는 특히 전체 코퍼스에서 출현 빈도가 낮으면서도 해당 핵심어의 의미 중의성 해소에 유의미한 어휘들까지 어휘 중의성 알고리즘에 활용하는 것을 목표로 하고 있다. 이러한 어휘군을 일단 "저빈도어"라 칭하기로 한다.

[1] "실험데이터"는 실험코퍼스에서 추출된 핵심어를 포함하는 문장을 뜻한다. 보다 자세한 설명은 3.2절 참조.

특히 본 논문에서는 저빈도어를 [세종1000만어절말뭉치]에서 빈도수 5 이하인 어휘들로 잠정 정의하여 이어지는 추후 논의를 전개하기로 한다.[2] 이러한 수치는 일단은 임의적이지만 선행연구 Weeber, Vos, and Baayen (2000)에 의존한 것으로, 무엇보다도 빈도수에 기반을 둔 어휘 분포도 상에서 저빈도쪽에 위치한 어휘들을 포함하는 것을 의미한다.[3]

[세종1000만어절말뭉치]를 모집단으로 하여 조사해본 결과, 빈도 1에서 5까지의 어휘들의 토큰은 전체 어휘 토큰의 58%를 상회한다. 이와 같이 저빈도어가 전체 토큰에서 많은 비율을 차지하는 경향은 선행연구에서도 보이고 있다.[4] 저빈도어가 전체 코퍼스 집단에서 대부분의 비율을 차지하는 것은 의미 중의성 해소연구에서 의미 있는 저빈도어를 추출하는 것에 대한 당위성과 근거를 부여하게 된다.

이러한 저빈도어가 확률기반의 통계작업에서는 무시되고 있는 편이다. 표준편차 기준 95% 신뢰수준에 근거한 연구에서는 나머지 5%에 들어 있는 요소들이 이론적으로 배제되기 때문에, 그 안에 포함되어 있는 요소들의 유의미성에 대한 논의 자체가 성립할 수 없다. 그러나 실제 언어현상에서는 중요한 의미적 정보가 저빈도어에 들어 있을 수 있다. 따라서 통계기반의 의미 중의성 해소 연구에서는 출현 빈도뿐만 아니라 의미 있는 저빈도어까지도 함께 추출하는 방식의 연구가 필요하다. 본 연구의 주안점은 그러한 요소들까지도 최대한도로 고려하는 알고리즘을 제시하는데 있다. 본 연구에서는 여러 연어정보 중에서 t-스코어를 활용하여 유의미한 공기어휘를 추출하고자 한다. 이 점은 3절에서 자세히 논의될 것이다.

2.2 기존 연구

2.2.1 통계적 검증. 저빈도어의 추출을 연구주제로 다루고 있는 Weeber, Vos, and Baayen (2000)은 여러 가지 통계적 검증을 이용하고 있다. 피셔 검증[5] (Fisher's test)을 이용해서 해당 검증의 유의도를 입증하고 로그 우도 (log likelihood)를 통한 통계적 의미 검증을 하고 있다.

Weeber, Vos, and Baayen (2000)의 또 다른 흥미로운 관찰은 유의미한 저빈도어 추출은 문맥적인 방향성과 별 연관성이 없다는 것이다. 이것은 의미 중의성 해소에 많은 기여를 했던 Yarowsky (1994) 등의 일련의 연구와 차이가 있다. 방향성 문제

[2] 그렇다고 어휘군이 "저빈도어"와 "고빈도어"로 양분되거나, 그 둘을 구분하는 어떤 절대적 수치가 반드시 설정되어야 한다는 의미는 아니다. 익히 알려져 있다시피, 빈도순에 따른 어휘의 분포는 불연속적인 분포가 아니라 연속적인 분포를 보이고 있기 때문이다. 다만 논의 진행의 편의상 저빈도어 기준 수치를 5로 설정한 것이다.

[3] 어휘들의 분포적 특성을 고려한 결과임.

[4] Weeber, Vos, and Baayen (2000)은 의학용어의 경우에 빈도가 5이하인 어휘가 대부분의 경우에 추출되어져야 할 중요한 용어들이며, 이에 대한 정보는 출현 빈도만을 고려한 경우에는 무시된다고 설명하고 있다. Weeber, Vos, and Baayen (2000)의 경우엔 알고리즘 자체가 추출 알고리즘이기 때문에 "저빈도어"에 대한 구체적인 기준이 제시될 수 있었다.

[5] 통계학에서 "test"가 한국어로는 "검증"이나 "검정"으로 번역되어 쓰이고 있다 (http://stat.anyang.ac.kr/sec/dic). 이 논문에서는 "검증"이란 용어로 통일하기로 한다. 기타 통계학 관련 용어도 앞의 사이트에 제공된 영한-한영 대조표를 활용하였다.

란 문맥을 다시 핵심어의 왼쪽 문맥과 오른쪽 문맥으로 나누어 볼 때 그러한 좌우 문맥의 역할이 서로 다르냐 다르지 않느냐 하는 문제이다. Yarowsky (1994)는 그 방향성이 의미 중의성 해소에서 중요한 차이가 있다고 본 반면에, Weeber, Vos, and Baayen (2000)은 유의미한 저빈도어 추출이 문맥의 좌우 방향성에 아무런 영향을 받지 않는다는 결론을 내리고 있다. 이와 같은 연구결과는 좌우 문맥을 직접적으로 고려하는 것이 유의미한 저빈도를 고려하는 의미 중의성 해소에 영향을 주지 않는다는 것을 뜻한다. 박병선 (2005)은 한국어의 경우에 문맥의 방향성에 따라서 연어정보에 차이가 있다고 주장한다. 본 연구에서 이러한 주장을 근거로 일부 검증을 시도해본 결과 유의미한 연어의 분포가 좌나 우에 쏠리지 않고 산개되는 현상을 발견할 수 있었다. 따라서 본 연구에서는 문맥의 방향성을 고려하지 않았다.[6]

2.2.2 기계 가독형 사전 (Machine readable dictionary). Stevenson (2003)은 기계 가독형 사전인 Longman Dictionary of Contemporary English (LDOCE)를 활용한 의미 중의성 해소 방법을 소개하고 있다. 이 방식도 적은 수의 사전적 정의를 활용해서 많은 양의 문맥을 처리하는 부트스트래핑 방식을 취한 연구라는 점에서 본 연구와 연구방식을 공유하고 있다. 또 Stevenson (2003)의 장점은 의미 분류에 활용되는 어휘들이 이미 기계 가독형 사전에 등재되어 있어서 재사용될 수 있다는 것으로, 적은 양의 사전적 어휘로 많은 분류 문맥들을 처리할 수 있다는 것이다.

그러나 이러한 방식은 문맥적인 특성을 반영하지 않는다는 점에서 본 연구와 차이가 있다. 특히 Stevenson (2003)의 연구가 적은 양의 사전적 정의를 기반으로 하는 것은 다양한 문맥적 특성을 반영하는데 문제점을 지닌다. 사전에 없는 미등록어가 의미 중의성 해소에 이용되는 것을 포착하지 못한다. 이에 반해서 본 연구는 해당 문맥의 특성을 나타내는 미등록어들을 사전에 새롭게 업데이트하는 효과가 있다. 이는 새로운 문맥이 주어지면 의미 있는 단어를 추출해서 업데이트하는 동적인 기능을 말한다. 자세한 것은 3.2 절에서 논의될 것이다.

3. 개념학습과 의미 중의성 해소 알고리즘

3.1 개념학습과 자극

3.1.1 긍정자극과 부정자극. Mitchell (1997)은 "개념"을 공통된 자질을 가진 자극들의 범주로 규정하고 있다. 예를 들어 '사각형'이라는 개념은 '네 개의 변'과 '네 개의 각'으로 표현할 수 있다. 따라서 사각형의 개념적 특징은 '사각형'의 특징들을 만족하는 자극들의 범주이다. 이런 자극들은 사각형의 크기나 색깔과는 상관없이 규정되는 것으로 '사각형'이라는 개념의 범주를 결정한다. 하나의 자극이 특정 개념 범주의 구성체라면, 그 자극은 '긍정적 자극'이며, 반대의 경우는 '부정적 자극'이다. 이런 관점에서 본

[6] 문맥의 좌우 방향에 따른 차이점에 대한 논의는 심사자 한 분의 지적에 따라 추가되었다.

다면, 어떤 개념을 규정한다는 것은 그 개념을 구성하는 자극들의 집합을 규정하는 것이며, 이는 '긍정적 자극'의 범주를 정의하는 것이다. 어떤 자극물로서 도형이 주어졌을 때, 그것이 사각형인지 아닌지를 판단하기 위해서는 그 개념을 구성하는 속성의 값이 '사각형'이라는 개념 조건에 부합하는지를 파악해야 한다. 이와 같이 개념을 범주화 하는 것은 자극공간(stimuli space)에서 '긍정적 자극'을 검색하는 것으로 설명될 수 있다. 다시 말하면, 하나의 개념을 형성하는 것도 자극공간을 검색해서 '긍정적 자극' 집합을 찾아내는 과정으로 정의될 수 있다.

3.1.2 연어현상과 자극. 긍정적 자극의 범주화는 어휘의 의미 분류에도 활용될 수 있다. [표 1]의 예는 [세종100만어절말뭉치]에서 추출한 것으로 의미 중의성을 지닌 '배'라는 단어를 포함하는 문장들이다. 동음이의어인 '배'를 어의(sense)에 따라 네 가지로 구분하고, 해당 어의로 쓰인 문장을 각각 하나씩 제시하였다.

의미	문장
운송수단	그러던 어느 날, 바다 저 먼 곳에서 한 척의 **배**가 섬 쪽으로 오는 것이 눈에 띄었습니다.
배수	그 가격이 실제 가격의 두**배**가 되는지, 10**배**가 되는지는 알 수 없는 일이다.
신체기관	거들은 **배** 부분을 둥글게 받쳐주는 덮개가 하나 더 있는 임부 전용과, 출산 후에도 몸매 보정용으로 입을 수 있는 웨이스트 니퍼형거들 두 종류가 있다.
과일	온 몸뚱이가 거무죽죽한 흙탕물에 철버덩 들어갔다가 솟구쳤을 땐, 돌이네 뒤곁 훤칠한 **배**나무가 탐스럽게 열렸던 귀여운 아기 똘**배**의 모습은 간 곳이 없었습니다.

[표 1] '배'라는 단어를 핵심어로 취하는 문맥의 예

[표 1]에서 보듯 '배'라는 단어는 '운송수단', '배수', '신체기관', '과일' 등의 여러 가지 의미로 각 문맥에 등장할 수 있다. 중의적인 표현의 어의를 구분하는데 있어서 각 어의별 문맥이 중요한 역할을 한다. 이것은 "분포가설"의 주장과도 일치하는 바, 중의적인 단어가 포함된 각각의 문맥에서 해당 어의를 식별하고 분류하는 일은 코퍼스 상에서 중의적인 단어와 공기하는 어휘들을 찾아냄으로써 가능하다.

[표 1]에서 운송수단으로서의 '배'라는 의미는 '그러던, 어느, 날, 바다, ...'라는 공기어휘들을 통해서 찾아낼 수 있다. 공기현상을 나타내는 어휘들은 연어의 특징을 나타내며, 해당 어의의 '긍정적 자극'을 나타낸다. 따라서 공기어휘들은 개별 어의에 대한 단서를 제공한다. (1)은 위 [표 1]에서 '배'와 공기하는 어휘들의 집합을 나타내는데,

조사나 의존어미 등 기능어를 제외한 단어들의 집합이다.[7]

(1) ㄱ. <그러던, 어느, 날, 바다, 저, 먼, 곳, 한, 척, 섬, 쪽, 오다, 눈, 띄다>

ㄴ. <그, 가격, 실제, 가격, 되다, 되다, 알다, 수, 없다, 일, 이다>

ㄷ. <거들, 부분, 둥글, 받치다, 덮개, 하나, 더, 있다, 임부, 전용, 출산, 후, 몸매, 보정용, 입다, 수, 있다, 웨이스트, 니퍼형거들, 두, 종류, 있다>

ㄹ. <온, 몸뚱이, 거무죽죽, 흙탕물, 철버덩, 들어갔다, 솟구치다, 때, 돌이네, 뒤곁, 훤칠하다, 나무, 탐스럽게, 열리다, 귀엽다, 아기, 똘, 모습, 가다, 곳, 없다>

일반적으로 문맥 의미 분류의 단서가 되는 어휘가 공기하는 경우에 문맥 의미 분류가 쉬워진다. 위 (1)의 어휘목록에서 '바다, 섬, 오다'라는 단어가 문맥에 있으면 '운송수단의 배'로 금방 의미 분류가 되고, '거들, 웨이스트, 니퍼형 거들'과 같은 단어가 공기하면 '신체의 배'로 분류된다. 공기하는 어휘를 모두 다 고려하기보다는 실제 영향을 미치는 공기어휘들로 한정해 준다면 고려할 대상이 그만큼 줄어들고 따라서 더 효율적이라 할 수 있다. 즉 해당 어의의 식별에 기여하지 않는 공기어휘들은 그 자리에 어떤 어휘가 오더라도 상관없다고 볼 수 있으므로, 그 자리를 '?'로 대치하면, 이제 (1)의 공기어휘 목록이 각각 (2)처럼 바뀌게 된다.

(2) ㄱ. <?,?,?, 바다,?,?,?,?,?, 척, 섬,?, 오다,?,?>

ㄴ. <?, 가격,?, 가격,?,?,?,?,?,?,?>

ㄷ. <거들,?,?,?,?,?,?,?,?,?,?,?, 몸매, 보정용, 입다,?,?, 웨이스트, 니퍼형거들,?,?,?>

ㄹ. <?,?,?,?,?,?,?,?,?,?,?,?, 나무, 탐스럽게, 열리다,?,?,?,?,?,?,?>

(2)와 같이 연어현상이 제한적인 경우에 어떤 제한적인 어휘들이 해당 문맥에 존재한다면 의미 분류가 가능해진다. 예를 들면 (2ㄱ)에서와 같이 <바다, 척, 섬, 오다>라는 단어만 문맥에 공기한다면 '운송수단의 배'로 분류될 수 있다. 다시 말하면, 14개의 어휘가 올 수 있는 문맥에서 4개의 어휘만으로도 의미 판별이 가능해 진다.

여기서 극단적인 두 가지의 경우와 그 사이에 존재하는 여러 가지의 경우를 생각해 볼 수 있다. 극단적인 경우로 첫째 어떤 값이 오더라도 범주의 결정에 영향을 미치지 못하는 경우와 둘째 어떤 값도 오지 못하는 경우를 상정해 볼 수 있다. (3)에서처럼 어떤 값으로도 채워질 수 있는 경우에는 속성들을 '?'로 표시하고, 어떤 값도 존재할 수 없는

[7] 세종말뭉치의 형태소 분류 중 격조사, 보조사, 접속조사, 그리고 어미로 태깅된 형태소를 제외한다.

경우에는 '∅'로 표시하기로 한다. 예를 들어, 하나의 범주 A를 구성하는 속성이 네 개가 있고, 해당 속성들은 'X'라는 값을 가진다고 하자. 그럴 경우, 가장 일반적, 포괄적 범주에 대한 정의는 모든 속성이 어떤 값이든 허용하는 (3ㄱ)의 경우이고, 가장 특정되고 제한적인 정의는 모든 속성이 어떤 값도 허용하지 않는 (3ㅂ)의 경우일 것이다. 그리고 이러한 양 극단 사이에는 속성의 값이 'X'나 '?'로 구성된 경우들이 놓일 수 있다. (3)에서는 그중 일부만 예시로 보인다.[8]

(3) ㄱ. <?, ?, ?, ?> ↑ 포괄적, 일반적
ㄴ. <X, ?, ?, ?>
ㄷ. <X, X, ?, ?>
ㄹ. <X, X, X, ?>
ㅁ. <X, X, X, X>
ㅂ. <∅, ∅, ∅, ∅> ↓ 제한적, 상세적

이러한 개념학습과 자극이 어휘 분포의 설명에도 확장될 수 있을까? 본 연구에서는 연어현상에서와 같이 공기하는 어휘들은 핵심어의 의미를 결정짓는 속성에 대한 '긍정적 자극'을 제공한다는 가설을 취하기로 한다. '운송수단의 배'의 경우 가장 일반적이고 포괄적인 경우는 '바다'와 같이 '운송수단의 배'를 연상하기에 충분한 단어가 하나라도 나올 경우이고, 가장 제한적이고, 특수한 경우는 '바다, 척, 섬, 오다'의 모든 단어가 나오는 경우가 될 수 있을 것이다.

(4) <?,?,?, 바다,?,?,?,?, 척, 섬,?, 오다,?,?> ↑ 제한적
<?,?,?, 바다,?,?,?,?,?, 섬,?, 오다,?,?>
<?,?,?, 바다,?,?,?,?,?,?,?, 오다,?,?>
<?,?,?, 바다,?,?,?,?,?,?,?,?,?,?>
<?,?,?,?,?,?,?,?,?,?,?,?,?,?,?> ↓ 일반적

(4)에 예시되어 있듯이, 제한적인 경우는 가장 많은 수의 공기어휘가 문맥에 등장한 경우를 말하고 가장 일반적인 경우는 가장 적은 수의 공기어휘가 문맥에 등장한 경우이다. 이와 같은 검색은 ($WORD_{바다} \cap WORD_{척} \cap WORD_{섬} \cap WORD_{오다}$)와 같은 어휘들의 부울 논리 검색식으로 변환될 수 있다.[9] 가장 제한적인 경우는 검색어로 산정된 모든 어휘가 검색식에 있는 검색의 경우이고, 가장 일반적인 경우는 검색어인 어휘가 하

[8] <∅, ∅, ∅, ∅>이나 <?,?,?,?>과 같은 가장 제한적이거나 가장 일반적인 경우는 현실적으로 존재하지 않는 특수한 경우이다. 아무것도 제약조건이 될 수 없으면, 제약의 대상도 없을 것이다. 그러나 이와 같은 특수 조건은 논리적인 정의를 위해서 필요하다.

[9] 이러한 방식의 부울 논리(Boolean logic)가 활용된 연구는 Mohammad and Pederson (2004)에서 발견되는데, 중의성을 가진 해당 어휘의 통사적 자질을 부울 논리 검색식에 활용한다. 예를 들어서 '$WORD_{go}$의 목적어 ∩ $WORD_{think}$의 주어'와 같이 통사적 정보를 이용하게 된다. 따라서 이와 같은 연산(computation)을 위해서는 해당 단어의 통사정보를 추출하기 위한 파서기의 적용이 불가피하다. 그러나 본 연구에서는 이러한 통사 정보 없이 형태적 정보만을 이용하였다.

나도 없는 검색의 경우일 것이다. 따라서 (4)와 같은 분포는 (5)와 같은 검색식으로 전환될 것이다.

(5)
$$
\begin{array}{ll}
(\mathrm{WORD}_{\text{바다}} \cap \mathrm{WORD}_{\text{칙}} \cap \mathrm{WORD}_{\text{섬}} \cap \mathrm{WORD}_{\text{오다}}) & \uparrow \text{ 제한적} \\
(\mathrm{WORD}_{\text{바다}} \cap \cancel{\mathrm{WORD}_{\text{칙}}} \cap \mathrm{WORD}_{\text{섬}} \cap \mathrm{WORD}_{\text{오다}}) & \\
(\mathrm{WORD}_{\text{바다}} \cap \cancel{\mathrm{WORD}_{\text{칙}}} \cap \mathrm{WORD}_{\text{섬}} \cap \cancel{\mathrm{WORD}_{\text{오다}}}) & \\
(\mathrm{WORD}_{\text{바다}} \cap \cancel{\mathrm{WORD}_{\text{칙}}} \cap \cancel{\mathrm{WORD}_{\text{점}}} \cap \cancel{\mathrm{WORD}_{\text{오다}}}) & \\
(\cancel{\mathrm{WORD}_{\text{바다}}} \cap \cancel{\mathrm{WORD}_{\text{칙}}} \cap \cancel{\mathrm{WORD}_{\text{점}}} \cap \cancel{\mathrm{WORD}_{\text{오다}}}) & \downarrow \text{ 일반적}
\end{array}
$$

제한적인 검색은 많은 검색의 조건을 갖고 있는 경우이고, 일반적인 검색으로 변하면 검색조건이 줄어들게 된다. 의미 중의성 해소에 이를 적용하면 다음과 같다. 처음에는 많은 검색어로 검색을 하지만, 검색이 일치하지 않으면 적은 수의 검색어로 검색조건을 완화해 가면서 검색을 한다. 어느 단계에서든 검색어가 해당 문맥에서 일치하게 되면 의미 중의성이 해소되게 된다.

3.2 알고리즘

[표 2]는 본 연구에서 제안하는 의미 분류 시스템의 알고리즘이다. 알고리즘을 단계별

```
1.   Comment: Training
2.   For all senses of s_i of w do
3.     Find the set of all words Q and R
         Q = {x | x ∈ s_i and x ∉ s_j}
         R = {y | y ∈ s_j and y ∉ s_i}
         ∀x, y t-scores of x, y > threshold of t-score
4.   End
5.   Comment: Disambiguating
6.   For all words v_k in the context window in the set Q and R
7.     For each attributive constraint word in the set Q or R
8.       If the constraint words are satisfied by v_j
9.       Then do this
10.        If the constraint words of s_i of Q is more specific than s_j of R
11.        Then choose the sense of s_i and mark the context window Q_annotated
12.      Else replace the next general constraints that are satisfied by s_i of Q or s_j of R
13.  End
14.  Comment: Updating
15.  For all words v_k in the Q_annotated
16.    Check if v_k is not in Q (or R)
17.    Then add v_k to Q (or R) if t-score of v_k > threshold of t-score
18.  End
19.  Comment: Repetition
20.  Goto Disambiguating and then Updating while no residual unannotated context window
```

[표 2] 알고리즘

로 설명하면 다음과 같다. 1단계는 부트스트래핑 방식을 이용한 방법으로 이 절차를 통해 적은 수의 의미 분류 데이터를 만들어 내어서, 많은 수의 실험 데이터에 적용하게 된다. 본 연구에서는 상대적으로 적은 규모인 100만 어절 학습 코퍼스를 통해서 얻어진

학습데이터를 1000만 어절 규모의 상대적으로 많은 규모의 실험 코퍼스에 적용하는 방식으로, 그 둘 사이에 1 대 10 정도의 비율을 유지하였다.

본 연구에서 활용된 코퍼스는 본래 어절별로 형태소 분석이 되어 있는 것이나, 약간의 가공 작업을 거쳐 문장단위 구분을 추가하였다. 문장경계를 무시하고 문맥을 고려하는 경우에는 일정 숫자의 문맥 어휘를 연구 대상으로 삼을 수 있겠으나[10] 문장경계를 최대 문맥으로 취한 경우에는 해당 핵심어가 들어 있는 문장들을 문맥으로 고려한다. 문헌 상 두 연구 사이에 의미 중의성 해소 작업의 성능에는 커다란 차이를 보이고 있지 않으므로 (Stevenson and Wilks, 2001), 본 연구에서는 일단 문장경계를 받아들여 후자의 방식을 취하였다. 이런 식으로 이미 형태소 분석이 되어 있는 학습코퍼스를 일단 문장 단위로 분리해 내고, 이어서 해당 핵심어가 있는 문장을 수집한다. 이 단계에서 모국어 화자들의 직관을 활용하여 중의적인 핵심어를 각각의 어의에 따라 분류하여 표기한다. 그런 다음 분류된 각 문장에서 주변 단어(내용어)를 추출해 낸 뒤에,[11] 그 단어들 중에서 서로 중복되지 않고 해당 어의에만 공기하는 어휘집합을 찾아낸다. 예를 들어서 '신부'라는 단어의 공기어휘 집합을 보면, '카톨릭 신부'와 '결혼식 신부'가 서로 중첩되지 않게 (5)와 같은 각각의 공기어휘 집합을 만들어 낸다 ([표 2]의 3번 줄).

(6) ㄱ. 신부는 연단에서 천천히 내려와서 신도들에게 말하였다. (카톨릭 신부)
<신부, 연단, 천천히, 내려오다, 신도, 말하다>

ㄴ. 천천히 내려와서 신부는 신랑에게 말하였다. (결혼식 신부)
<천천히, 내려오다, 신부, 신랑, 말하다>

(6)이 보여주는 것은 (6ㄱ)과 (6ㄴ)에 주어진 문장으로부터 각각 <> 안에 열거된 단어들을 추출해 내는 것이다. 그 다음 ㄱ과 ㄴ의 <> 안 명단을 비교해 보며 중복되지 않는 단어들을 찾아내는 작업을 한다. (6)에서는 중복되지 않는 단어에 밑줄을 그어 표시하였다. 그러한 어휘들은 각각의 어의에 종속되는 어휘집합으로 간주된다. 이러한 어휘들을 추출해서 '카톨릭 신부'와 '결혼식 신부'의 어휘 집합을 만들어 낸 후, 그 집합 내 각 어휘들의 t-스코어를 계산해서 임계치[12] 0 이상이 되는 어휘들을 추려낸다 ([표 2]의 3번 줄).

[표 2] 알고리즘의 2단계에서는 Mitchell (1997)에서 소개된 Find-Specific 알고리즘을 활용해서 의미 분류를 한다. Find-Specific 알고리즘은 기계학습 방식으로 검색조

[10] 핵심어와 함께 고려 대상이 되는 문맥을 '창'(window)이라고 한다. '창'의 크기를 고려하는 것에는 여러 가지 견해가 있다. Yarowsky (1992)에서는 좌우 50개 단어, 즉 100개 단어를 포함한 문맥을 창으로 보았지만, Yarowsky (1997)에서는 좌우 20개 단어, 즉 40개 단어로 보는 것이 적절하다고 주장하였다. 본 연구에서는 Yarowsky (1997)의 주장을 따라서 '창'의 크기를 40개 단어로 간주하는 경우를 논의하기로 한다. 관련된 논의는 6절에서 하기로 한다. 심사위원 중 한 분이 이점을 지적해 주었다.

[11] 연구에 활용된 어휘들은 형용사, 명사(일반명사, 고유명사), 동사, 부사와 같은 내용어(content word)만을 고려하였고 기능어(function word)를 무시하였다 (각주 7 참고). 이는 정보검색이나 기타 자연어처리에서 흔히 취하는 방식이다.

[12] 여기서는 threshold를 임계치로 번역하여 사용한다. 임계치는 critical value의 번역어로도 사용되고 있는 바, 여기에서는 그 두 용어에 차이를 두지 않는다.

건을 만족하는 자극들(stimulus)을 찾아내는 매우 간단한 구조로 개념학습 방식을 활용한다. [표 3]은 Mitchell (1997, 26쪽)의 Find-Specific 알고리즘을 인용한 것이다.

1.	For each attribute constraint a_i in h
2.	If the constraint a_i is satisfied by x
3.	Then do nothing
4.	Else replace a_i in h by the next more general constraint that is satisfied by x

[표 3] Mitchell의 Find-Specific 알고리즘

Find-Specific 알고리즘에 따르면, 주어진 자극이 제약조건(constraint)을 만족하면 검색에 성공하게 되며([표 3]의 2, 3번 줄), 제약조건을 만족하지 않으면 더 일반적인 제약조건으로 검색조건을 변경하게 된다([표 3]의 4번 줄). 이와 같이 개념의 범주를 결정하는 조건들을 검색하는 범주 검색의 방식으로 개념학습이 이루어진다. 개념학습의 자세한 내용과 절차는 3.1절에서 소개가 되었다. 이러한 Find-Specific 알고리즘에 따라 본 연구에서는 가장 상세한 제약조건, 즉 공기어휘 집단이 주어진 문맥에서 추출한 어휘 집단에 모두 포함이 되어 있는지 확인한다. 만약 포함이 되었다면 제약조건은 충족이 되므로 해당 핵심어는 분류가 되고, 의미 중의성은 해소된다([표 2]의 10, 11번 줄). 검색 결과가 도출되지 않으면 더 일반적인 제약조건으로 이동한다([표 2]의 12번 줄).

이러한 개념학습 장치는 부울 논리 연산인 'AND' 연산을 활용하는데, 이러한 부울 논리 연산은 확률 연산에 비해서 직접적으로 공기어휘를 고려할 수 있게 한다. 부울 논리 연산은 확률 연산으로 치환이 가능하다. $\text{WORD}_{\text{나오다}} \cap \text{WORD}_{\text{섬}}$이 검색식인 경우에, 확률연산은 '나오다'와 '섬'의 확률이 동시에 독립적으로 존재하는 독립연산으로 계산된다. 즉 P(나오다) ∩ P(섬)의 확률 연산과 같고, 이는 P(나오다)×P(섬)처럼 두 확률의 곱으로 연산된다. 본 논문에서 저빈도어는 1,000만 어절 기준으로 절대 빈도 1에서 5까지 어휘로 잠정 정의되었다. 따라서 개략적으로 각각 $\frac{1}{10^8}$에서 $\frac{5}{10^8}$정도의 아주 적은 확률을 취한다. 특히 독립가정으로 확률의 곱으로 연산 된다면, 매우 미미한 확률을 가진다. 따라서 저빈도어의 경우에 출현 빈도에 기반을 둔 확률 연산에 의존하면 확률 예측을 낮추는 역효과가 생겨난다. 이러한 방식은 저빈도어를 직접적으로 고려의 대상에서 제외하는 결과로 이어질 수도 있다. 본 연구에서 사용한 개념학습 장치는 이러한 확률연산이 아닌 부울 논리 방식을 활용한 검색이므로 낮은 확률적 고려가 없이 저빈도어를 직접적으로 고려하는 장점을 갖고 있다.

[표 2]의 알고리즘 2단계에서는 (6)과 같이 학습 코퍼스에서 추출된 해당 핵심어의 공기어휘 집합을 이용하여 실험 코퍼스에서 추출된 핵심어들의 어의를 구분한다. Find-Specific 알고리즘을 이용해서 해당 중심어의 공기어휘가 학습 코퍼스를 통해서 모인 어휘 집합들에 가장 근접하게 많은 어휘를 가진 문맥부터 분류하게 된다. [표 4]는 (6)에서 수집된 <연단, 신도>와 <신랑>으로 '카톨릭 신부'와 '결혼식 신부'를 의미 분류하

는 예를 보여준다.

번호	분류	실험문장	공기어휘
1	카톨릭 신부	연단을 뒤로 하고 신부는 신도들과 예배를 보았다.	<연단, 뒤, 신부, 신도, 예배>
2	결혼식 신부	신랑들에게 신부들은 항상 잔소리를 하게 마련이다.	<신랑, 신부, 항상, 잔소리, 마련이다>
3	?	당신을 오늘 최고의 신부로 만들어 줄 헤라	<오늘, 최고, 신부, 만들다, 헤라>
4	카톨릭 신부	신도들에게 신부님은 인기가 최고다	<신도, 신부, 인기, 최고>

[표 4] 의미 분류의 예

[표 4]에서 1번 데이터의 공기어휘 집합을 보면 (6ㄱ)에서 찾아낸 '카톨릭 신부'의 공기어휘 집합 <연단, 신도>를 포함하고 있고 따라서 검색조건을 충족하므로 중의성이 해소된다. 마찬가지로 2번 데이터는 (6ㄴ)에서 찾아낸 <신랑>을 포함하고 있으므로 '결혼식 신부'로 의미 분류된다. 이번에는 3번 데이터의 공기어휘 집합을 보자. <오늘, 최고, 신부, 만들다, 헤라>에는 (6)을 통해 추출된 어휘들이 들어있지 않다. 따라서 검색 조건이 만족되지 않아 중의성이 해소될 수 없다. 반면 4번 데이터의 경우에는, (6)에서 찾아낸 '카톨릭 신부'의 공기어휘 집합 <연단, 신도>를 다 포함하고 있지는 않으나, 그중 일부인 <신도>는 들어있다. 즉 <연단, 신도>라는 상세조건은 만족시키지 못하나 더 일반적인 제약조건인 <신도>를 포함하고 있다. 따라서 이 단계([표 2] 12번 줄)에서 '카톨릭 신부'로 의미 분류될 수 있다.

[표 2] 알고리즘 3단계에서는 2단계에서 의미 분류가 된 실험 데이터에서 새로운 공기어휘 집합을 추출해서 통계적으로 의미가 있는 어휘를 기존 공기어휘 집합에 추가한다. 예를 들어서 [표 2]의 2단계를 통해 의미 분류에 성공한 [표 4] 1, 2, 4에서 기존의 학습 데이터에 없는 어휘 집단인 <뒤, 예배, 인기, 최고>와 <항상, 잔소리, 마련이다>를 각각 추출한다. 각각의 어휘를 t-스코어로 검증하고 검증 결과 임계치 0이상인 경우에만 '카톨릭 신부'와 '결혼식 신부'의 각각의 기존 공기어휘 집합에 추가한다.

이 과정을 거치면 기존 공기어휘 집합은 늘어나게 되고, 이 늘어난 공기어휘 집합을 이용해서 남은 문맥을 처리하게 된다. [표 2] 알고리즘 4 단계에서 이러한 작업을 설명하고 있는데, 이 단계에서는 2단계와 3단계를 반복하게 되고, 분류할 데이터가 없을 때까지 반복하게 된다 ([표 2] 20번 줄). 이와 같이 약간의 학습 데이터로 많은 양의 실험 데이터를 처리하기 위해 실험 데이터를 학습 데이터로 추가하게 되는 방식은 Yarowsky

(1995)에 의해 제시된 바 있다.[13] 이 방식은 초기에는 실험 데이터를 약간만 분류하는데, 분류에 성공한 데이터는 다시 학습 데이터로 추가되게 되어서 학습 데이터는 늘어나게 되고, 이를 다시 실험 데이터에 적용하게 된다. 이와 같은 과정을 반복하게 되면 학습 데이터는 늘어나게 되고 늘어난 학습 데이터는 거의 전체 실험 데이터를 최종적으로 분류할 수 있게 된다.[14]

3.3 t-스코어를 통한 유의미성 검증

'긍정적 자극'을 유발하는 어휘들은 전체 코퍼스에서의 빈도와 정비례하지 않는다. 예를 들어서 '성직자의 신부'를 의미하는 '신부'라는 단어는 다른 연어보다 성직자와 관련된 어휘와 같이 등장하게 될 것이다. 따라서 성직자와 관련된 어휘들은 속성 값이 '긍정적 자극'의 속성 값으로 될 것이다. [세종100만어절말뭉치]에서 조사된 결과로는 '성직자 신부'라는 단어는 '김대건, 문익환,...'과 같은 성직자 인명과 공기하는 경우가 많았다.[15] 그런데 코퍼스 내에서 인명은 출현 빈도가 너무 낮아서 빈도 중심의 알고리즘에서는 직접 반영되기가 어렵다. 즉, '성직자 신부'의 경우 인명은 낮은 빈도수로 인해 '긍정적 자극'에 포함되기가 어렵다. 물론 빈도가 낮다하여 통계상 모두 무시된다는 뜻은 아니다. 빈도수가 낮은 경우라도 적합한 통계기법을 활용하면 유의미한 결과를 추출해 낼 수 있다. Church et al. (1992)에서는 의미 있는 공기현상이 통계적으로도 추출될 수 있다는 점을 보이고 있다. 이러한 통계적 기법을 통해서 '긍정적 자극'을 추출해 낼 수 있다.

본 연구에서는 우선 [세종100만어절말뭉치]를 대상으로 핵심어휘 '거리, 신부, 배'를 포함한 문장을 추출한 후, 주변 어휘의 연어값을 측정해 보았다. 연어값으로는 빈도, 예상빈도, z-스코어, t-스코어, 상대빈도를 측정하였다.[16] 해당 수식의 계산방법은 다음과 같다.

(7) ㄱ. $\text{예상빈도} = \frac{\text{전체텍스트에서의발생빈도}}{\text{전체텍스트어절수}} \times \text{범위내어절수}$

ㄴ. $\text{z-스코어} = \frac{\text{관찰빈도(공기빈도)} - \text{예상빈도}}{\text{표준편차}}$ [17]

ㄷ. $\text{t-스코어} = \frac{\text{관찰빈도(공기빈도)} - \text{예상빈도}}{\sqrt{\text{관찰빈도(공기빈도)}}}$

---

[13] 이러한 방식은 학습 데이터와 실험 데이터를 동시에 학습 및 분류하는 동기화 학습 방식과도 유사하다 (Chakrabarti, 2003). Yarowsky에서 언급된 방식은 엄밀하게 말하면 동기화 방식이라고는 말할 수는 없다.

[14] Yarowsky (1995, 191-192쪽)는 그러한 과정을 초기 상태, 중간 상태, 종료 상태로 나누어 그림으로 설명하고 있다.

[15] 특히 '김대건'과 같은 단어는 장르에 종속되지 않고 여러 곳에서 등장하고 있다. 다른 인명은 신문 기사와 같은 장르에서 많이 등장하는 반면에 '김대건'은 소설, 수필과 같은 산문이나 신문 기사 등 다양한 장르에 나타나고 있었다.

[16] 각각의 수식은 Barnbrook (1996)를 참조하였다.

[17] Barnbrook에서 언급되는 표준편차는 다음과 같다.
$\text{표준편차} = \sqrt{\text{범위내어절수} \times \text{발생확률} \times (1 - \text{발생확률})}$

ㄹ. 상대빈도 $= \log_2 \frac{\text{관찰빈도(공기빈도)}}{\text{예상빈도}}$

(7)의 네 가지 연어 값을 일부 자료에 적용해 본 결과 예상빈도 및 상대빈도는 신뢰도에 문제가 있다는 점이 경험적 검증 과정에서 드러났다. z-스코어의 경우 신뢰도 상의 문제는 없었으나 저빈도어의 의미를 지나치게 확대하는 경향이 있었다. 그리고 이러한 결과는 기존의 연구와도 일치하였다 (Barnbrook, 1996; 강범모, 2003). 따라서 그러한 문제가 드러나지 않는 t-스코어 검증을 채택하였다. 그리고 다른 연어값은 통계적인 분포를 표현하는 수치이지, 통계적 결정을 내리는 검증이 아니다. 따라서 통계학적인 측면에서도 t-스코어를 선택할 타당성이 있다.

t-스코어를 계산하는 방법은 다음과 같다. 해당 어휘 X가 전체 1,000만 어절에서 출현 빈도 10이고, 핵심어 Y가 포함된 전체 어절은 1,000어절이며, 출현 빈도는 4라고 가정해 보자. 우선 예상빈도는 $\frac{10}{10,000,000} \times 1,000$이므로 0.001이다. 그런데 1000어절 내 실제 출현 빈도는 4이고 이는 예상빈도 0.001에 비해 훨씬 큰 수치이므로 X는 핵심어 Y에 유의미한 어휘라고 볼 수 있다. 이 경우 t-스코어는 $\frac{4-0.001}{\sqrt{4}}$이므로 대략 1.8 정도가 된다. 이것은 t-분포표 상에서 유의도 0.05, 그리고 무한대 자유도가 기준일 때, t-스코어인 1.645를 상회하므로 영가설은 기각되게 된다. 여기서 영가설은 해당 어휘 X가 전체 1,000만 어절에서 출현하는 통계적 분포와 핵심어 Y가 포함된 1,000어절에서 출현하는 통계적 분포에는 차이가 없고, 따라서 X는 Y를 포함한 문맥에서 우연에 의한 연어관계라는 것을 말한다. 영가설이 기각되었으므로, X가 Y를 포함한 문맥에서 우연에 의한 연어관계가 아닌 유의미한 연어관계라는 사실이 입증되는 것이다.

통계적인 엄밀성에 비추어서 볼 때, t-분포는 신뢰도가 일정한 값 이상일 경우 영가설을 기각하게 된다. 이 기각 영역은 승인 영역에 비해서 매우 적은 확률적 분포를 가지는 신뢰도이므로, 신뢰수준이 매우 높다. 일반적인 t-분포에서 이러한 임계치는 0.05 수준이다. 즉 통계적 분포 상 95% 이상일 경우를 말하므로 상당히 높은 신뢰를 지니게 된다.

본 연구에서는 t-스코어의 임계치를 0 이상으로 산정하였다. 이 점은 기존 관례를 벗어나는 것으로 추가 설명이 필요하다. 우선 실험에서 -1, -0.5, 0, 0.5, 1, 2와 같은 수치로 어휘들을 검증하며 그 결과를 비교하여 보았다. 또한 각 수치별 어휘집합이 의미 분류에 어떻게 기여하는지를 살펴보고, 실제 노이즈 (noise)와 관련된 부분도 살펴보았다. 이러한 절차를 통해 검토해 본 결과, 0 이상이 가장 적절한 것으로 판단되었다. 즉 t-스코어가 0 이상일 경우에 해당 어휘들이 유의미한 공기어휘로 선택되었다. 이와 같은 임계치 0은 Zinsmeister and Ulrich (2003)에서도 연어관계 논의에서 경험적으로 설정된 수치이다.[18] 마찬가지로 Sun, Shen, and Tsou (1998)에서도 경험적 수치로서 0.00이 이용되고 있다. 특히 Hardcastle (2005)는 BNC 코퍼스를 통해서 유의미한 연

[18] "For the manual[ly] inspection we set a cut-off at t-score 0.00." (Zinsmeister and Ulrich, 2003, 936쪽)

어쌍을 t-스코어로 측정하였으나, 유의미한 연어관계가 엄격하게 적용되는 t-분포도 상에서 ($n = \infty$, $\alpha = 0.005$)는 추출되지 않는다고 밝히고 있다. Hardcastle (2005)는 t-스코어의 값은 통계적 엄밀성의 기준을 적용해서 입증하면 유의미한 연어관계를 정확하게 포착하는 정확도는 증가하나, 일반적으로 유의미한 연어관계들을 많이, 넓게 포착해야 하는 재현율은 떨어지는 것을 관찰하였다. 따라서 정확도와 재현율의 최적점을 찾는 것은 경험으로 가능할 것이다. 본 연구에서 검증을 통해 설정된 수치 0.00도 이러한 선행연구에 의해서 뒷받침되고 있다.[19]

유의미한 저빈도어가 실제 어떻게 고려되는가라는 질문을 할 수 있다. 최초의 학습 데이터는 100만 어절 기준에서 추출된 것이고 이 중 저빈도어는 50% 이상으로 전체 어휘 중 다수를 차지하고 있다. 본 연구의 알고리즘에서 업데이트라는 부분은 실험 데이터에서 수집된 어휘들 중 학습 데이터에 포함되지 않은 어휘를 t-스코어 검증을 거쳐서 다시 학습 데이터로 포함시키는 과정을 말한다. 이러한 과정은 실험 데이터의 유의미한 저빈도어를 고려하는 특성을 보이게 된다. 이와 같은 작업을 반복적으로 10회 진행했을 때, 저빈도어의 포함 비율은 증가하면서, [표 5]와 같은 정량적 접근을 보이고 있다.

반복 회차	전체 학습 어휘에 대한 저빈도어 비율
1회	14.3%
2회	35.8%
3회	45.0%
4회	37.9%
5회	59.9%
6회	70.2%
7회	70.4%
8회	70.5%
9회	70.7%
10회	70.9%

[표 5] 반복 회차당 전체 학습 어휘 대비 저빈도어 반영 비율

[표 5]에서 살펴본 바와 같이 학습 데이터는 전체 어휘 중 약 71%가 저빈도어로 구성이 된다. 학습 데이터는 해당 핵심어휘를 포함한 문맥에서 추출되는데, 이를 토대로

[19] 이 부분은 한 심사자가 지적한 아래와 같은 문제점에 대한 답변도 된다. "유의미성을 -1이상, -0.5이상, 0.5이상, 1이상, 2이상 등으로 나누어 고려하는데 이는 정확해 보이지 않는다. t-score의 값은 99%, 95% 신뢰도 범위내에서 각각 일정한 값 이상이어야 하는데 자유도에 따른 이 t-score의 값은 그 범위 내에 있지 않은 듯 하다." 이러한 지적은 임계치로 설정한 값 0이 기존의 t-분포표에도 나오지 않는 값이므로 무리스럽고 무의미하다는 지적으로 해석된다. 일반론적인 관점에서 그런 우려는 충분히 공감하나, 그러한 수치가 우선 앞에서 보았듯 기존의 비슷한 연어 연구에 의하여 뒷받침 되고 있고, 또한 실제 자료의 검증을 통해 얻어진 경험적 수치라는 점에서 임계치 0은 나름대로 의미가 있다는 것이 본 논문에서의 판단이다. 물론 이러한 수치는 추후 연구에 의해 더 검증되어야 할 것이다. 6절에 관련 논의 추가.

보면 유의미한 전체 어휘 중 대다수가 저빈도어로 구성되어 있다. 업데이트하는 과정을 통해서 저빈도어가 더 많이 고려되게 되는데, 이것은 해당 문맥적 특성이 반영된 저빈도어가 추출되기 때문이다.

4. 나이브 베이즈 구분자(Naive Bayes Classifier)를 활용한 정확도 검증

나이브 베이즈 구분자는 Brown et al. (1991)과 Gale, Church, and Yarowsky (1993)에서와 같이 의미 중의성 해소에 많이 활용되고 있는 알고리즘이다.[20] 이 방식은 베이즈 확률을 기반으로 한 것으로, 미리 수집된 확률을 검증이 요구되는 데이터에 적용하는 것이다. 본 연구에서 활용되는 공식은 (8)과 같다 (Mitchell, 1997, 177쪽).

$$
\begin{aligned}
(8) \qquad V_{\mathrm{MAP}} &= \arg\max_{v_j \in V} P(v_j \mid a_1, a_2, \ldots, a_n) \\
&= \arg\max_{v_j \in V} \frac{P(a_1, a_2, \ldots, a_n \mid v_j) P(v_j)}{P(a_1, a_2, \ldots, a_n)} \\
&= \arg\max_{v_i \in V} P(a_1, a_2, \ldots, a_n \mid v_j) P(v_j)
\end{aligned}
$$

나이브 베이즈 구분자 방식은 현 단계의 확률적 추론을 다음 단계의 확률적 추론에 적용하는 베이즈 확률을 변형한 것이다. 이 방식은 서로 다른 가능성에 대한 확률을 비교해서 더 높은 확률을 추론해내는 추론확률 방식이다. 본 논문에서는 이러한 나이브 베이즈 구분자를 적용해서 이행되는 현 단계 확률이 검증되는 것을 살펴보고 이후에 적용될 확률이 정확한지를 살펴보았다. 이것은 베이즈 확률추론의 '최대 이후 가설(Maximum a posteriori hypothesis)'을 활용하는 것이다 (Mitchell, 1997). 베이즈 추론은 현재 단계에서 관찰된 확률이 P(D)이고 이전 단계에 측정된 확률이 P(h)이라면 이후에 예측될 P(h|D)의 최고 확률을 예측하는 것이다. 현재 산출된 확률이 정확하다면 이후 확률도 최대로 정확하게 예측될 것이다. 따라서 현재 산출된 확률의 정확도가 이후 확률에 적용된다면 현재 산출된 확률의 정확도를 검증할 수 있을 것이다. 본 연구에서는 Find-Specific 알고리즘으로 산출된 결과가 정확한지를 입증하는데 나이브 베이즈 구분자를 활용하였다.

본 연구는 실험 데이터에서 추출된 유의미한 공기어휘 집합을 기 추출된 학습 공기어휘 집합에 업데이트하는 과정을 거친다. 만약 업데이트되는 어휘 집합이 많은 노이즈(noise)를 포함한다면 이는 연산이 잘못된 것일 것이다. 현 단계에서 예측된 데이터를 통해서 베이즈 추론을 적용하고, 그 결과 정확도가 낮아졌다면 현 단계의 연산이 잘못된 것으로, 많은 노이즈가 있는 것으로 판명된다.

이러한 문제점에 대해서 본 연구에서 Find-Specific 알고리즘을 적용한 후와 적용하기 전에 나이브 베이즈 구분자를 적용해서 정확도를 비교하여 보았다. Find-Specific

[20] Gale and Yarowsky (1992)에서 사용된 변형된 나이브 베이즈 구분자 알고리즘은 영어 의미 중의성 해소의 경우에 약 98%의 정확도를 보인다고 한다.

알고리즘이 현 단계의 확률을 잘 예측했다면 Find-Specific 알고리즘이 적용된 후에 정확도가 증가할 것이다. 여기에서 연구가설은 'Find-Specific 알고리즘을 적용한 후에, 즉 실험 데이터가 학습 데이터로 업데이트된 경우가 정확도가 가장 높다.'이다. 다시 말하면, 현 단계의 확률이 잘 예측되었으므로 이후 단계의 확률이 잘 예측될 것이 가설이 될 것이다. 왜냐하면 Find-Specific 확률을 적용하고 나이브 베이즈 구분자 확률을 적용한 경우가 '최대 이후 가설'을 최적화할 것이기 때문이다. 이러한 논증은 Mitchell (1997)에서 설명되고 있는 바, Find-Specific 알고리즘은 이전 단계의 확률을 더 정확하게 예측하므로 가설의 공간을 줄이게 되고, 다음 단계의 예측 확률을 증가시킬 것이다 (Mitchell, 1997, 162–163쪽). 따라서 추가되는 데이터가 정확하다는 것도 입증하게 될 것이다.

본 연구에서는 다음 세 가지 서로 다른 경우에서 각각의 정확도를 측정해 보았다. 나이브 베이즈 구분자만 적용한 경우와 Find-Specific 알고리즘만 적용한 경우, 그리고 Find-Specific 알고리즘을 적용하고 나이브 베이즈 구분자를 적용한 경우이다. 실제 실험 결과 Find-Specific 알고리즘을 적용한 후가 나이브 베이즈 구분자 실험에서 가장 정확도가 높은 것으로 나타났다. 이런 결과는 6절에서 논의가 될 것이다.

5. 실험

본 연구에서는 실험에 쓰일 단어를 '신부, 거리, 배'로 설정하였다. 중의성 해소가 필요한 어휘를 선정할 때는 해당 어휘의 어의들의 분포를 살펴보아야 한다.[21] '신부, 거리, 배'는 해당 단어가 하나의 어의 당 사용량이 80% 이상을 넘지 않는다는 것을 확인하였다. 실제 의미 분류에서 '신부, 거리, 배'는 다음과 같은 사전적 의미를 참조하였다.[22] [표 6]은 100만 어절을 통해서 조사된 어휘별 어의의 사용 비율이다.

반면 너무 낮은 비율로 쓰인 어의는 부트스트래핑 방식의 적용에서 학습 공기어휘 집합을 선정하는데 문제가 있었기 때문에 배제하였다. 사용 비율이 20% 이상일 경우에 실험에 쓰일 데이터로 선정하였다. 또한 실제 사전과 코퍼스의 용례와는 차이가 있다. 따라서 [표 6]에서처럼 실제 의미 있는 빈도를 갖는 어의를 대상으로 설정하였다. 예를

[21] 강범모 (2005)는 98% 이상의 동음이의어가 제1어의의 사용이 해당 동음이의어의 전체 사용중 99% 이상을 차지한다고 설명한다. 실제 코퍼스에서 조사된 바로는 전체 동음이의의 사용 중 최다 빈도 어의 하나만이 전체 사용량의 99%를 차지한다고 한다. 그리고 이런 경우에 전체 동음이의어 중 98%가 해당한다고 한다. 이와 같은 경우는 동음이의 구분이 불필요하다. 그러나 강범모 (2005)의 연구에 따르면 연구의 대상이 되는 동음이의어가 전체의 2%에 해당 할지라도 전체 표제어를 중심으로 볼 때, 2,000–3,000여개가 된다. 따라서 동음이의를 해소하는 것은 언어학적으로도 인공 지능적으로도 중요한 의미를 지닌다.

[22] 두산동아 새 국어사전(2005) 참조. 2004년 국립국어원 발간 표준국어 대사전에는 더 많은 수의 어의를 포함하고 있다. '배'의 경우에도 고어의 경우를 포함시키고, '성씨' 및 '모리배'와 같이 무리를 나타내는 형태소를 포함하고 있었다. 이와 같은 현상은 어의가 확장될 경우에 구분이 되는 단어를 찾을 수 있는가 하는 문제와 연결이 된다. 이러한 문제에 대해서 Find-Specific 알고리즘이 명확히 구분되는 구분자를 찾을 수 있는가에 대한 질문이 심사자들 중에서 지적이 되었다. 부트스트래핑 방식을 이용한 의미 중의성 해소는 실제 초기 구분자를 찾기 위한 초기 데이터가 일정한 양이 필요하다. 아주 없거나 거의 없는 데이터를 추출해서 얻을 수 있는 초기 데이터는 거의 미미하기 때문에 이 경우는 향후 연구의 대상이 될 수는 있어도 본 논문의 연구과제와는 거리가 있다.

어휘	해당 어의	사용 비율
신부	카톨릭 신부	20%
	결혼식 신부	80%
거리	길거리	45%
	내용이 될 만한 것 (국거리, 반찬거리)	3%
	제시한 수가 될 만한 것	0%
	오이나 가지를 세는 단위	0%
	사이간의 거리, 인간관계 사이의 거리	52%
배	위장	23%
	짐승이 새끼를 낳거나 알을 까거나 하는 횟수를 세는 단위	0%
	배아, 새끼	0%
	선박	44%
	배나무의 열매, 과일	7%
	갑절 또는 곱절	26%

[표 6] 100만 어절에서 나타난 '신부, 거리, 배'의 해당 어의들의 사용 비율

들어, '신부'의 경우에 '카톨릭 신부'와 '결혼식 신부', '거리'의 경우에 '길거리', '사이간의 거리', 그리고 '배'의 경우 '위장', '선박', '갑절'과 같은 어의가 선정이 되었다.

연구 자료의 수집은 다음과 같은 절차를 거쳤다. 앞에서도 일부 언급되었듯이, 우선 형태소 분석이 된 세종 말뭉치에서 문장단위로 코퍼스를 정리하고, 해당 핵심어가 있는 문장만을 수집하였다.

(9)
2BT_0010006690	두	두/MM
2BT_0010006700	배,	배/NNG+,/SP
2BT_0010006710	세	세/MM
2BT_0010006720	배,	배/NNG+,/SP
2BT_0010006730	예전보다	예전/NNG+보다/JKB
2BT_0010006740	몇	몇/MM
2BT_0010006750	배나	배/NNG+나/JX
2BT_0010006760	열심히	열심히/MAG
2BT_0010006770	살거라구요.	살/VV+ㄹ/ETM+거/NNB +이/VCP+라구요/EF+./SF

(10) ㄱ. 두 배, 세 배, 예전보다 몇 배나 열심히 살거라구요.

ㄴ. 두/MM 배/NNG+,/SP 세/MM 배/NNG+,/SP 예전/NNG+보다/JKB 몇/MM 배/NNG+나/JX 열심히/MAG 살/VV+ㄹ/ETM+거/NNB

+이/VCP+라구요/EF+./SF

(9)는 실제 형태소 분석된 세종 말뭉치의 예이며, 이 자료를 토대로 (10)에서처럼 문장 단위 자료가 두 가지 형태로 산출된다. 형태소 구분이 된 경우와 형태소 구분이 되지 않은 경우이다. 문맥에서 핵심어 추출하는 방식(KWIC: Key Word In Context)으로 핵심어가 포함된 문장을 추출하였다. 형태소가 분석이 된 (10ㄴ)과 같은 문장과 (10ㄱ)과 같은 형태소 분석이 되지 않은 문장을 함께 추출하였다. 학습 데이터와 실험 데이터 모두 문장 단위로 수집되었으며, 실험 데이터와 학습 데이터의 비율은 약 10 대 1을 유지하였다. 실험 데이터는 1,000만 어절에서 학습 데이터는 100만 어절에서 추출되었다.

부트스트래핑 방식을 적용하기 위해서 학습 데이터로 추출이 된 데이터를 모국어 화자의 직관을 이용해서 의미 분류작업을 하였다. 의미 분류에는 3명의 대학원생이 참여하여, 동일한 문장에 두 사람씩 동일한 종류의 분류작업을 하도록 하였다. 해당 어휘에 대한 [표 6]의 의미분류를 기준으로 하여 주어진 문장 내 핵심 어휘가 어떤 어의를 가지는지를 표시하는 방식으로 [표 7]과 같은 분류 결과를 얻었다.

대상 어휘	분류 문장	의미 분류
배	몹시 **배**가 고팠던지 후루룩거리며 남자는 맛있게 식사를 했다.	신체의 배
신부	이재룡씨가.. 상원이가 완전히 잡혀있는 것 같은데, 여 **신부** 친구들 나와서 노래하라 그래.	결혼식 신부
거리	두희와 말순, 삼**거리**에 다다른다.	길거리

[표 7] 언어 직관에 따른 의미 분류의 예

실제 의미 분류에서 화자들이 서로 일치하지 않는 데이터가 5% 내외로 산출되었다. 이런 데이터는 재검토를 통해서 다시 분류를 하고, 검토를 통해서 분류가 되기에 충분하지 않은 공기어휘를 포함한 경우와 화자들의 직관이 서로 일치하지 않는 경우는 실험에서 제외하였다. 이어지는 단계에서는 별도 프로그램을 작성하여 해당 어의에 대한 공기어휘를 수집하였다.[23] 나머지 실험 절차는 3.2절의 알고리즘 설명과, 4절의 나이브 베이즈 구분자를 활용한 정확도 검증에 대한 설명에 이미 자세히 소개되었다.

6. 논의

학습 데이터를 통해서 수집된 어휘 중 통계적으로 의미가 있는 어휘는 t-스코어가 0 이상이었을 경우다. t-스코어가 0보다 적을 경우에는 의미 분류에 아무런 도움이 되지 않는 어휘가 많았다. t-스코어가 0.5나 1 이상이면 너무 적은 양의 어휘가 수집이 되어서

[23] 모든 작업은 Perl 프로그램을 활용해서 수행하였다.

의미 분류에 적절하지 않았다. 즉, 임계치를 0으로 설정한 경우에 최적의 결과를 얻을 수 있었고, 의미 분류에 도움을 주는 어휘들이 적합한 규모로 수집되었다.

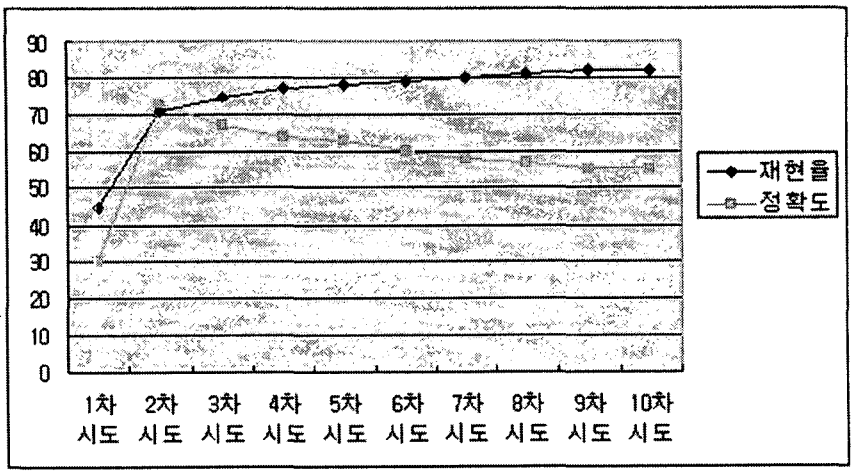
의미 있는 단어들은 대부분 타입 당 빈도가 1 정도이어서 본 작업에서 출현 빈도만을 고려하는 것은 적절하지 않았다. 전체 모집단에서 저빈도가 다수를 차지하듯이 의미 분류의 경우에도 저빈도어가 많은 영향을 미쳤다. 예를 들어서, '카톨릭 신부'의 경우 '성당, 성직자, 스님'과 같이 분류에 직접적인 영향을 미치는 어휘들은 빈도가 1-3이었다. 이 경우에 t-스코어는 0 이상이며, 0.5이상이나 1에 근접한 수치를 보이는 어휘도 많았다. '카톨릭 신부'의 경우 인명을 나타내는 고유명사가 자주 발견되었다. 균형 코퍼스가 신문기사와 같이 시사성을 지니는 경우가 포함되어 있어서 고유명사를 많이 포함하고 있었다.

Find-Specific 알고리즘을 10차례 반복적으로 적용한 경우에는 저빈도어가 전체 요소 중 많은 부분을 차지하게 된다. 이는 재현율이 증가하면서, 실험 데이터의 어휘들을 학습 데이터로 업데이트하는 과정에서 많은 저빈도어가 추가되는 것을 의미한다. 이것은 새로운 의미 분류의 대상이 되는 문맥에서 추출된 저빈도어가 학습 데이터로 추가되기 때문이었다.

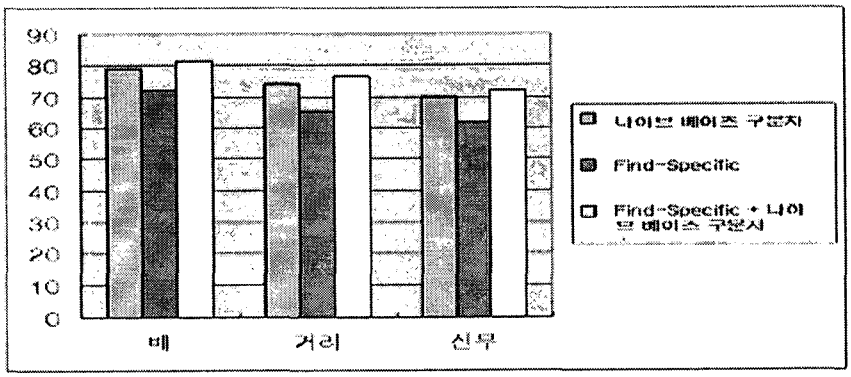
Find-Specific 알고리즘을 반복적으로 적용한 후 정확도와 재현율을 살펴보면, 적용 횟수가 증가할수록 재현율은 증가하나 정확도가 떨어지는 것을 볼 수 있었다 ([그림 1]). 1차 시도는 순수하게 학습 데이터에서 수집된 데이터로 정확도나 재현율은 매우 낮다. 그러나 계속 횟수가 늘어나면서 정확도나 재현율은 증가한다. 특히 2차 시도에서 재현율과 정확도가 급격히 증가하였다. 반면 3차 시도 이후부터는 정확도가 점점 낮아지면서 횟수가 반복할수록 어느 정도 일정한 지점에 수렴되는 것으로 나왔다. 이것은 증가되는 학습 데이터에 노이즈가 끼어들기 때문이다. 한편 재현율은 시도 횟수가 늘어남에 따라 점점 증가하지만, 2차 시도에서만 급격히 증가했을 뿐 약간씩 증가하면서 역시 일정한 지점으로 수렴되었다.

4절에서 논의한 바와 같이 Find-Specific 알고리즘을 적용하기 전후에 각각 나이브 베이즈 구분자를 적용하여 정확도를 측정 해 보았다 ([그림 2]). 전체적으로 Find-Specific 알고리즘과 나이브 베이즈 구분자를 함께 적용한 경우가 나이브 베이즈 구분자를 단독으로 적용한 경우보다 약 3-4%의 정확도의 상승이 있었다.

이 경우에 Find-Specific 알고리즘과 나이브 베이즈 구분자 알고리즘을 적용한 경우가 정확도가 제일 높은 것은, 베이즈 확률의 '최대 이후 가설'로 뒷받침된다. '최대 이후 가설'을 다시 설명하자면, 현 단계의 확률이 정확히 측정된다면 이후 단계의 확률도 정확히 예측된다는 것인데, Find-Specific 알고리즘을 적용한 것의 현 단계 확률이 정확히 예측이 되어서, 이후 단계인 베이즈 확률 적용에서 향상이 있었다는 것이다. 연구 결과 학습 데이터로 추가되는 어휘들이 '최대 이후 확률 가설'에 부합하는 특성을 보인다. 다시 말하면, 실험 집단에서 Find-Specific 알고리즘을 통해 분류된 공기어휘는



[그림 1] 1차 – 10차 시도별 Find-Specific 알고리즘의 정확도와 재현율 측정



[그림 2] 정확도 비교 측정 결과

t-스코어가 0이상인 경우에 학습 데이터로 추가되는데, 이 때 추가되는 공기어휘들은 노이즈가 적고 신뢰할 수준이라는 의미가 된다.

세종 코퍼스는 훈련된 모국어 화자들이 언어직관을 활용해서 직접 형태소 분석을 하거나 자동 형태소 분석을 언어직관을 통해서 확인한 것으로 대체로 정확한 편이라 할 수 있다. 그러나 이번 연구를 통해 일부 오류도 발견되었는바,[24] 본 연구에서 수집된 데이터에서 대략 2-3% 정도의 오류를 발견할 수가 있었다. 이러한 오류로 인해 본 연구에서는 오류점검 절차를 거치기에 필요 이상의 반복적인 실험을 해야 하는 등 많은 시간이 실험에서 추가로 소요되었다. 한 가지 흥미로운 점으로는 Find-Specific 알고리즘을

[24] 세종 말뭉치 태그의 정확도에 대한 별도의 연구는 이미경 외 (2005) 참고.

통해 이러한 태깅 오류 중 일부를 발견할 수 있었다. 상세 제약조건으로 검색하는 경우 해당 공기어휘를 발견할 수 없어서 의미 분류를 하지 못하는 경우가 발생하였는데, 이를 통해 태깅 오류를 발견하게 되었다.

(11) ㄱ. 성장 미 로버트 **배**로 교수 특별 대담 박영철 금융 (... 배_명사+로.....)

ㄴ. 금주 표정에 **배**역 그래서, 번호 가르쳐 주었어? (... 배_명사+역.....)

ㄷ. 바구니 들고 올라오는 **배**여사 (... 배_명사+여사.....)

형태소 분석 오류는 위 (11)에 예시되어 있다. 모두 명사 '배'로 태깅이 되어 있기 때문에, 해당 문맥들은 실험 데이터인 '배'가 포함된 문맥으로 분류된다. 따라서, 위 (11)에 있는 문맥들은 실험 절차상 데이터 선정에서 배제될 이유가 없다. Find-Specific 알고리즘을 적용하면 해당 문맥에 명사 '배'의 의미를 지니는 공기어휘들이 포함되어 있지 않으므로, (11)의 예들은 의미 분류 될 수가 없고, 오류로 인식되게 된다. 이와는 달리, 확률을 이용하여 구분하는 경우에는, 확률 연산을 통해서 일률적으로 구분하기 때문에 무조건적으로 의미 구분이 이루어지게 된다. 확률방식을 이용하는 나이브 베이즈 구분자에서는 (11)의 예와 같은 분석 오류가 발견되지 않는다. 이처럼 데이터 상의 오류가 감지되지 않고 정상 데이터로 분류 된다면, 정확도와 재현율 산정에 문제가 되기 때문에 정확한 실험의 결과를 측정 시에 문제를 발생시킨다.

본 연구에서는 문맥을 문장 단위로 설정하였는데, 문장 당 평균 26.5개의 어절이 들어 있었다. Yarowsky (1997)의 연구에서는 좌우 문맥 40여개의 단어를 분류에 사용하는데 비해, 적은 수의 어휘를 활용하였다. 문장 단위 문맥은 (12)와 같은 상황에 보다 효과적이다. (12)는 좌우 문맥을 고려하는 방식의 문제점을 나타낸다.

(12) ㄱ. 영식이는 $배_1$가 아파서 $배_2$를 타고 바다로 갔다.

ㄴ. 영식이는 $배_1$가 아팠다. $배_2$를 타고 바다로 갔다.

(12)의 예에서 $배_1$과 $배_2$는 각각 '신체 기관의 배'와 '운송수단의 배'이다. 만약 문장구분을 하지 않는다면 (12ㄱ)과 (12ㄴ)은 차별화 될 수가 없다. 반면 본 연구에서처럼 문장 경계를 고려한다면 (12ㄴ)에서는 '아팠다'가 $배_1$의 공기어도 되면서 동시에 $배_2$의 공기어도 되는 문제가 생기지 않을 것이다. Yarowsky (1993)의 "하나의 연어현상에서 하나의 의미 (one sense per collocation)"는 의미 중의성 해소에서 하나의 문맥 내에서 중의적 단어는 하나의 의미만을 갖는다는 제약 조건이다. 이 조건에 따르면 문맥이 주어진 (12ㄱ)과 (12ㄴ)에서의 '배'는 하나의 어의만 갖는다. 따라서 (12ㄴ)의 경우에도 한 가지 의미로만 분석될 것이나 그것이 (12ㄱ)과 달리 해석되지 않는다. 따라서 문장경계를 고려한 의미 중의성 해소는 (12)와 같은 장점을 갖게 된다.

또한 부울 논리를 이용한 검색에 기반을 둔 Find-Specific 알고리즘은 적은 수의 어휘가 포함된 문맥의 경우에 언어직관과 유사한 결과를 산출해 낼 수 있었다. 언어 직관적으로 의미구분은 어휘의 개수보다는 충분한 단서를 제공하는 어휘가 공기할 경우에 가능하다. (13)과 같은 문장에서는 화자의 직관으로는 중의성 해소가 어렵다.

(13) ㄱ. **신부**는 부를 수 있다.

ㄴ. 그러나 **신부**는 서성거리지 않는다.

ㄷ. **신부** 뿐이다.

(13)의 예에서 '카톨릭 신부'와 '결혼식 신부'라는 어의로 구분할 만한 아무런 단서가 없다. 나이브 베이즈 구분자를 적용할 경우에는 의미 구분은 확률적인 연산에 의해서 이루어지므로, 각각의 의미의 구분은 공기 어휘의 확률적 수치에 근거해서 결정된다. 이것은 주어진 맥락만으로는 의미구분이 불가능한 상황에서 억지로 의미구분을 하므로, 화자의 언어직관과 맞지 않는다. 이 경우에는 더 많은 문맥의 고려를 통하여 의미 구분을 시도하는 것이 더 자연스러운 방식이라고 하겠다. Find-Specific 알고리즘은 부울 논리 검색 기반으로 충분한 단서가 주어질 때 의미구분을 하고 있고, 이런 처리 방식은 화자의 직관에 부합된다.

Find-Specific 알고리즘은 유의미한 저빈도어를 고려하기 때문에 인명(人名)과 같은 고유명사가 의미 중의성 해소에 이용될 수 있다. (14)에서 '신부'의 두 가지 어의인 '결혼식 신부'와 '카톨릭 신부'는 문맥에 포함된 인명을 통해 직관적으로 쉽게 구분된다.

(14) ㄱ. 박홍 **신부**께서 — 우리학교 송자도 그랬잖아

ㄴ. 탤런트 홍학표 **신부**

즉, '신부'는 고유명사인 '박홍'과 '홍학표'를 통해서 의미 구분이 가능한데, 그러한 고유명사는 대체로 저빈도로 출현한다. 또한 인명은 새로운 문맥을 처리할 때마다 추가로 반영되어져야 하는데, 본 논문에서 제시된 Find-Specific 알고리즘을 활용한 동적 문맥 처리 기능은 새로운 문맥 추가 시에 의미 있는 저빈도어를 활용할 수 있다.

7. 결론

본 연구에서는 통계적으로 의미 중의성 해소에 단서가 되는 공기어휘들을 추출한 뒤 개념학습을 활용하는 Find-Specific 알고리즘을 적용했다. 유의미한 단어 추출에는 연어 값 측정 방식인 t-스코어를 적용하였다. 실제 코퍼스 상에는 많은 어휘들이 저빈도어이며, 이 방식으로 추출된 어휘들도 대부분이 저빈도어로 구성이 된다. 또한 본 연구에서는 문맥을 통해서 추출된 유의미한 어휘를 추가하는 방식을 이용하고 있는데, 이러한 방식 역시 저빈도어를 적절히 활용한다.

자연어처리를 포함한 인공지능의 문제인 "지식 습득의 병목 현상"에 대해서 본 연구는 부트스트래핑 방식을 제안하였다. 이 방식을 위해서 Yarowsky (1995)의 연구 방식을 이용, 검증된 공기어휘만을 다시 학습 공기어휘 집합에 추가하는 방식을 취하였다. 이러한 방식을 통해서 추가되는 어휘에 대한 검증은 베이즈 확률의 "최대 이후 가설"을 활용하였다. 검증 방식으로 나이브 베이즈 구분자를 적용했고, 그 결과 실제 정확도가 증가하였다. 결론적으로 추가되는 데이터는 노이즈가 많지 않고 신뢰할만한 수준임을 확인할 수 있었다.

본 연구는 초기 학습 데이터를 수집하기 위해서 모국어 화자의 언어직관을 활용하였다. 적은 양의 학습 데이터는 많은 양의 데이터를 처리할 수 있는 장점이 있으나, 일부나마 의미 분류에 화자의 직관이 활용됨으로써 완전 자동화에 도달하지는 못했다. 이와 같은 연구의 한계에 대해서 향후 연구에서는 언어직관 없이 분류되는 비감독 기반에 대한 연구가 필요하다. 비감독기반의 방식은 모국어 화자들의 의미 분류가 없이, 코퍼스만으로 학습하는 방식이다. 이 방식은 Schuetze (1998) 등에서 논의되고 있거니와, 감독기반과 다르게 의미 분류된 코퍼스가 전혀 필요하지 않다. 주어진 텍스트만으로 기계학습을 통해서 의미 중의성을 해소하게 된다. 이와 같은 방식은 정확한 데이터를 수집하는데 많은 시간과 노력이 드는 지식병목 현상을 해결 할 수 있는 장점이 있으나 실제 정확도가 감독기반보다 떨어지는 단점이 있다. 하지만 비감독방식으로의 발전은 지식병목현상의 해결을 위해서 필수적인 방향이라고 하겠다.

또한 향후 연구 방향으로 미관찰된 데이터에서 정확성을 높이기 위한 알고리즘의 연구 및 개발이 필요하다. 다시 말하면, 실험데이터가 미관찰된 데이터를 포함할 경우 관찰 오류를 범할 수 있는데, 이를 해결하기 위해서 미관찰된 데이터를 다룰 수 있는 새로운 알고리즘이 필요하다. 즉, 실험 데이터가 늘어나면 늘어날수록 노이즈를 지닌 단어가 늘어나게 되고, 필요 없는 단어들이 끼어들게 된다. 이러한 노이즈를 낮출 수 있는 알고리즘의 발견과 적용도 향후 필요한 연구 중의 하나다. 이런 방식을 위해서 Expectation-Maximization (EM) 알고리즘과 같은 기존 해결방식의 적용도 필요할 것이다 (Manning and Schuetze, 2000).

<참고문헌>

Barnbrook, Gilmore. 1996. *Language and Computers: A Practical Introduction to the Computer Analysis of Language.* Edinburgh University Press, Edinburgh.

Brown, Peter, Stephen Pietra, Vincent Della, and Robert Mercer. 1991. Word sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting*, pp. 264–270, Berkeley. CA. Association for Computational Linguistics.

Chakrabarti, Soumen. 2003. *Mining the Web.* Morgan Kaufmann Publishers, San Francisco.

Church, Kenneth, William Gale, Patrick Hanks, and Donald Hindle. 1992. Using

Statistics in Lexical Analysis. In U. Zernik (ed.), *Lexical Acquisition: Exploring On-Line Resources to Build a Lexicon.* Hillsdle: LEA., pp. 115–164.

Gale, William, Kenneth Church, and David Yarowsky. 1993. A method for disambiguating word senses in a large corpus. *Computers and the Humanities* 26, 415–439.

Gale, William and David Yarowsky. 1992. Discrimination Decisions for 100,000-Dimensional Spaces. In A. Zampoli, N. Calzolari, and M. Palmer (eds.), *Current Issues in Computational Linguistics: In Honour of Don Walker.* Kluwer Academic Publishers, pp. 429–450.

Hardcastle, David. 2005. What is an "interesting score?"; An Evaluation of Measures Using the Distributional Hypothesis to Derive Cooccurrence Scores from the British National Corpus. *Corpus Linguistics 2005.* Birmingham.

Harris, Zellig. 1964. Distributional Structure. In J. Fodor and J. Katz (eds.), *The Structure of Language.* Pritence Hall, pp. 33–49.

Ide, Nancy and Jean Vernois. 1998. Introduction to the Special Issue on word Sense Disambiguation: The State of the Art. *Computational Linguistics* 24.1, 1–40.

Manning, Christopher and Hinrich Schütze. 2000. *Foundations of Statistical Natural Language Processing.* The MIT Press.

Mitchell, Tom. 1997. *Machine Learning.* McGraw-Hill Company.

Mohammad, Saif and Ted Pederson. 2004. Combining Lexical and Syntactic Features for Supervised Word Sense Disambiguation. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL).*

Schütze, Hinrich. 1998. Automatic Word Sense Discrimination. *Computational Linguistics* 24.2, 97–123.

Stevenson, Mark. 2003. *Word Sense Disambiguation.* CSLI Publication. Stanford University Press.

Stevenson, Mark and Yorik Wilks. 2001. The Interactions of Knowledge Sources in Word Sense Disambiguation. *Computational Linguistics* 27.3, 321–349.

Sun, Maosong, Dayang Shen, and Benjamin Tsou. 1998. Chinese Word Segmentation without Using Lexicon and Hand-crafted Training Data. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING '98*, pp. 1265–1271.

Weeber, Marc, Rein Vos, and Harald Baayen. 2000. Extracting the Lowest-Frequency Words: Pitfall and Possibilities. *Computational Linguistics* 26.3, 301–317.

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

Yarowsky, David. 1993. One Sense Per Collocation. In *Proceedings of ARPA Human Language Technology Workshop*, pp. 266–271, Princeton, NJ.

Yarowsky, David. 1994. Decision lists for lexical ambiguity resolution: application to accent restoration in Spanish and French. In *Proceedings of 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pp. 88–95, Las Cruces, NM.

Yarowsky, David. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics(ACL-95)*, pp. 189–196, Cambridge, MA.

Yarowsky, David. 1997. Homograph disambiguation in Text-to-speech Synthesis. In J. Santen, R. Sproat, J. Olive, and J. Hirschberg (eds.), *Progress in Speech Synthesis*, pp. 159–174, New York. Springer.

Zinsmeister, Heike and Heid Ulrich. 2003. Identifying predicatively used adverbs by means of a Statistical Grammar Model. In Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 conference*, pp. 932–939. UCREL, Lancaster University.

강범모. 2003. *언어, 컴퓨터, 코퍼스 언어학*. 고려대학교 출판부.

강범모. 2005. 동음이의어의 사용양상. *어학연구* 41.1, 1–29.

박병선. 2005. *한국어 계량적인 연구방법론*. 도서출판 역락.

이미경·정한민·성원경·박동인. 2005. 품사표지 부착 말뭉치 검증. *제17회 한글 및 한국어 정보처리 학술대회*, 145–150쪽.

접수 일자: 2006년 4월 20일
게재 결정: 2006년 7월 18일