정보관리연구, vol.42, no.2 2011, pp.193-210

단어 중의성 해소를 위한 SVM 분류기 최적화에 관한 연구

A Study on Optimization of Support Vector Machine Classifier for Word Sense Disambiguation

이 용 구*
Yong-Gu Lee

차 례

초 록

이 연구는 단어 중의성 해소를 위해 SVM 분류기가 최적의 성능을 가져오는 문맥창의 크기와 다양한 가중치 방법을 파악하고자 하였다. 실험집단으로 한글 신문기사를 적용하였다. 문맥창의 크기로 지역문맥은 좌우 3단어, 한 문장, 그리고 좌우 50바이트 크기를 사용하였으며, 전역문맥으로 신문기사 전체를 대상으로 하였다. 가중치 부여 기법으로는 단순빈도인 이진 단어빈도와 단순 단어빈도를, 정규화 빈도로 단순 또는 로그를 취한 단어빈도 × 역문헌빈도를 사용하였다. 실험 결과 문맥창의 크기는 좌우 50바이트가 가장 좋은 성능을 보였으며, 가중치 부여 방법은 이진 단어빈도가 가장 좋은 성능을 보였다.

키 워 드

단어 중의성 해소, SVM, 의미 분류, 문맥창, 가중치 부여 기법

* 계명대학교 문헌정보학과 전임강사
(Instructor, Dept. of Library & Information Science, Keimyung University, yonggulee@kmu.ac.kr)
• 논문접수일자: 2011년 3월 22일
• 최종심사(수정)일자: 2011년 4월 1일
• 게재확정일자: 2011년 4월 6일

ABSTRACT

The study was applied to context window sizes and weighting method to obtain the best performance of word sense disambiguation using support vector machine. The context window sizes were used to a 3-word, sentence, 50-bytes, and document window around the targeted word. The weighting methods were used to Binary, Term Frequency(TF), TF × Inverse Document Frequency(IDF), and Log TF × IDF. As a result, the performance of 50-bytes in the context window size was best. The Binary weighting method showed the best performance.

KEYWORDS

Word Sense Disambiguation, SVM, Semantic Classification, Context Window Sizes, Weighting Methods

1. 서 론

인간의 언어를 컴퓨터가 처리하는 자연언어처리는 다양한 분야에서 사용되어 좋은 결과를 가져오고 있다. 일례로 스마트폰에서 음성을 통한 웹을 검색하는 기능도 이러한 자연언어처리를 적용하였다고 볼 수 있다. 이렇듯 자연언어 처리는 인간의 언어와 관련된 다양한 분야에서 반드시 필요한 기법 중에 하나이다.

자연언어 처리 기법 중에서 동형이의어나 다의어와 같이 한 단어가 여러 가지 의미를 가질 때, 그 단어가 어떤 의미로 사용되었는지 정확히 분석하는 과정을 단어 중의성 해소(Word Sense Disambiguation: WSD) 기법이라고 한다. 이 기법은 텍스트에 나타난 중의성 단어가 어떤 의미로 사용되었는가를 식별한다. 식별 방법은 크게 텍스트 범주화(text categorization)와 같은 지도학습(supervised learning) 분류와 클러스터링과 같은 비지도학습(unsupervised learning) 분류 방법이 주로 사용된다. 일반적으로 지도학습 기반의 중의성 해소 방법이 비지도 방법보다 더 좋은 성능을 보이고 있다.

지도학습 기반 방법에서 중의성 해소를 위해서는 중의성 단어에 대해 의미 범주 또는 태그가 붙은 학습용 언어자원이 필요하다. 이를 통해 의미 분류기를 생성하기 때문이다. 영어의 경우 십여 년 전부터 이러한 말뭉치 언어자원을 만들기 위해 꾸준히 노력하였으며, 그 결과 현재 대규모의 단어 중의성 해소용 말뭉치(corpus)를 가지고 있다. 대표적인 예로 SENSEVAL (http://www.senseval.org/)에서는 특정 영어 단어의 다양한 의미에 대해 태깅이 되어 있을 뿐만 아니라, 한 텍스트에 나타난 모든 영어 단어에 대해 정확한 의미를 수작업으로 분석하여 태그를 달아 두었다. 하지만 국내의 경우 대

표적인 몇 개의 중의성 단어만을 그 대상으로 의미 태깅이 되어 있으며, 그나마 실험문헌의 크기도 매우 작은 편에 속한다.

국내의 단어 중의성 해소 말뭉치의 열악함으로 인해, 이를 이용하는 지도학습 기반의 중의성 해소 방법을 한글에 적용하는데 걸림돌이 되고 있다. 이는 지도학습 분류인 텍스트 범주화 기법을 큰 규모의 한글 말뭉치에 대해 적용하는 것을 어렵게 하고 있다. 큰 규모의 말뭉치를 대상으로 체계적으로 수행되어 그 성능을 분석해 볼 필요가 있다.

따라서 이 연구에서는 의미가 태깅된 말뭉치를 사용하여 지지 벡터 기계(Support Vector Machines: SVM) 분류기를 구축하고, 다양한 문맥창(context window) 크기와 자질 가중치 기법을 적용하여 한글 중의성 단어에 대한 의미 분류 성능을 파악해 보고자 하였다. 또한 이러한 기법들이 개별 중의성 단어에 대해 어떻게 다른 영향을 미치는지 살펴보고자 하였다.

2. 단어 중의성 해소 선행연구

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.

단어 중의성 해소는 문헌집단에 출현한 단어에 대해 사전에서 보여지는 의미범주와 관계없이 그 단어의 용법(usage)에 따라 서로 다른 의미로 구분하는 작업(word sense discrimination)과 달리, 단어의 의미를 미리 정의된 의미범주로 부여하는 작업이다. 따라서 단어 중의성 해소는 (1) 텍스트 내 모든 단어에 대해 각 단어가 갖는 여러 의미들을 구분하는 단계와, (2) 각 출현 단어에 대해 적절한 의미를 부여하는 단계로 구성된다(Ide and Veronis 1998). (1) 단계에서는 미리 정의된 의미범주를 사용하는데, 이러한 의미범주는 사전에 등록된 의미목록이나 시소러스의 동의어처럼 상호 연관된 단어나 자질들로부터 확보한다. (2) 단계에서는 중의성 해소 대상 단어가 출현한 문맥(context), 즉 특정한 단어가 출현한 텍스트에 포함된 정보를 사용하거나 어휘사전, 시소러스, 백과사전 등의 다양한 외부 정보원을 사용하여 정확한 의미를 부여한다.

중의성 해소 알고리즘은 중의성 해소에 필요한 정보를 입수하는 방법에 따라 (1) WordNet과 같은 시소러스와 LDOCE(Longman Dictionary of Contemporary English)와 같

은 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).

또한 중의성을 해소하고자 하는 단어의 수준에 따라 문헌에 출현한 모든 단어를 대상으로 중의성을 해소하는 방법과, 특정 단어를 대상으로 중의성을 해소하는 방법으로 나눌 수 있다. 전자의 경우 주로 지식 기반 알고리즘을 이용하며, 후자의 경우는 자동분류 또는 범주화 영역으로 보아 말뭉치 기반 알고리즘을 이용한다. 이 연구에서는 후자인 말뭉치 기반 알고리즘에서 다루도록 하였다.

말뭉치 기반 중의성 해소 기법은 기계학습 분야에서 사용되는 자동분류(automatic classification)와 텍스트 범주화(text categorization) 방법을 이용한다. 일반적으로 문헌의 자동분류는 분류 알고리즘에 의해 대상물들을 유사한 패턴을 갖는 것끼리 모아 집단화하는 작업을 말하며, 텍스트 범주화는 사전에 분류된 학습문헌 집합에 근거하여 이미 정의되어 있는 주제범주들을 새로운 문헌에 배정하는 작업이다(정영미 2005). 즉 중의성 해소 기법에서 텍스트 범주화 기법을 이용한다면 단어의 의미를 범주로 간주하고 문헌을 특정 범주로 배정하듯이, 중의성 단어를 가장 적합한 의미범주로 분류한다고 보면 된다. 따라서 말뭉치 기반 중의성 해소 기법도 학습과정을 통해 분류기를 구축하는 것이 필요하며, 이 과정에 반드시 실제 분류 상태에 대한 정보를 담고 있는 학습 데이터와 의미범주가 필요하다.

말뭉치 기반 기법을 보다 세분하면, 의미 태그가 부착된 말뭉치를 이용하는 지도학습 알고리즘(supervised learning algorithm)과, 의미 태그가 부착되지 않은 말뭉치를 이용하는 비지도 학습 알고리즘(unsupervised learning algorithm)으로 구분할 수 있다. 실제 중의성 해소 실험에서 지도학습 알고리즘의 성능이 비지도학습 알고리즘의 성능보다 우수하게 나타나고 있다(Gale, Church, and Yarowsky 1993; Schutze 1998; Levinson 1999). 하지만 지도학습 알고리즘을 사용할 경우 다음과 같은 문제점이 있다. 우선 지도학습 알고리즘에서는 의미가 태깅된 학습용 말뭉치가 필요하다. 이러한 말뭉치를 구축하기 위해서는 수작업으로 의미 태깅하는 일이 쉽지 않으며, 비용도 많이 소요된다(Stevenson 2003; Ide and Veronis 1998). 또한 한글의 경우 미리 의미가 태깅되어 있는 이용가능한 말뭉치도 극소수라는 점이다.

말뭉치 기반 중의성 해소 기법의 선행연구를 살펴보면, 지도학습 알고리즘 분야에서 주로 나이브 베이즈 분류기와, 결정트리(decision tree)나 결정리스트(decision list) 분류기를 이용한다(Pedersen 2002). 나이브 베이즈 분류기는 베이즈 정리(Bayes' theorem)에 근거한

확률적 분류기로서 학습 데이터를 이용하여 문맥에 출현한 각 단어가 특정 의미범주를 대표할 확률을 계산한 다음, 중의성을 해소하고자 하는 단어의 문맥에 출현한 단어들을 단서어로 하여 중의성 단어의 의미범주를 예측한다. 결정트리 분류기는 결정트리로 표현되는 분류기로서 학습 데이터를 이용하여 서로 다른 의미범주를 식별하도록 만들어진 결정트리를 이용한다. 루트 노드에서 시작하여 하향식 탐색을 하며, 중간 노드들은 중의성을 해소하고자 하는 단어의 문맥을 이용하여 분기 규칙에 따라 다음 가지로 분기를 계속하여 잎 노드까지 따라가서 의미범주를 예측한다. 최종 잎 노드가 의미범주에 해당한다.

Gale, Church, Yarowsky(1992)는 나이브 베이즈 분류기를 이용하여 단어 중의성을 해소하고자 하였다. 이들은 학습용 말뭉치로 Canadian Hansard를 이용하여 duty, drug, land, language, position, sentence에 대한 단어 중의성을 해소한 결과 약 90%의 정확률을 보였다. 특히 이 연구와 Yarowsky(1993)의 연구에서는 중의성 단어의 의미는 이 단어가 출현한 텍스트와 연어에 의해 극히 제한된다는 사실을 발견하였다. "한 텍스트 한 의미"(one sense per discourse)와 "한 연어 한 의미"(one sense per collocation) 제약은 중의성 해소 알고리즘에서 유용하게 사용될 수 있다. "한 텍스트 한 의미"는 중의성 해소 대상 단어의 의미가 특정한 문헌 안에서 같은 의미로 사용될 수 있다는 일관성을 나타내며, "한 연어 한 의미"는 인접한 단어들인 연어에서 중의성 해소 대상 단어의 의미가 한 의미로 사용되는 경향을 보인다는 것을 말한다. "한 연어 한 의미" 제약 기법에서는 단어간 거리, 출현순서, 구문적 관계 등이 해소 과정에 반영된다. 다시 말해 이러한 제약을 반영한 알고리즘은 중의성 단어가 한 텍스트 안에서는 하나의 의미로만 사용되며, 특정한 연어 안에서도 하나의 의미만 갖는다는 사실을 이용한다.

Leacock, Towell, Voorhees(1993)는 나이브 베이즈 기법을 신경망과 문맥 벡터(context vector) 기법과 비교하기 위해 분류기를 구축한 후, 'line'에 대한 단어 중의성 해소 성능을 비교 평가하였다. 각 기법은 모두 70%의 정확도를 보였으며, 통계적으로 분류기간의 차이점을 발견할 수 없었다. Mooney(1996)는 문맥으로부터 단어의 중의성을 해소하기 위해 서로 다른 학습 알고리즘을 실험적으로 비교하였다. 이 연구에서 비교된 알고리즘은 나이브 베이즈 기법, 퍼센트론, 결정트리 그리고 kNN 기법 등이다. 이 연구 역시 'line'을 대상으로 중의성 해소 성능 평가를 하였으며, 나이브 베이즈 기법과 퍼센트론 기법이 가장 좋은 성능을 보인 것으로 나타났다.

결정리스트가 단어 중의성 해소 연구에 이용된 대표적인 연구는 Yarowsky(1995)이다. 이 연구는 "한 텍스트 한 의미"와 "한 연어 한 의미"라는 제약을 반영한 비지도학습 기반 알고리즘을 이용하였으며, 중의성 해소의 정확성이 96%에 달하는 높은 성능을 보였다. 중의성

단어의 의미를 해소하기 위해 적은 수의 연어 사례를 확인하여 그 연어에 쓰인 중의성 단어의 의미를 수작업으로 태깅하였다. 여기에서 적은 수의 연어 사례를 종자(seed)라 부른다. 이 연어 종자가 포함된 문맥을 연어에서 사용된 중의성 단어의 해당 의미로 태깅하였다. 결정 리스트를 이용하여 현재 태깅된 문맥을 토대로 새로운 연어를 추출함으로써 단어의 중의성을 해소하였다. 또한 수작업 태깅을 대체하기 위해 사전의 뜻풀이에 있는 단어나, WordNet에서 의미 범주별로 연어를 포함하는 문맥이나, 연어가 될 만큼 공기하는 고빈도어를 이용하였다.

최근 가장 좋은 성능을 보이는 SVM을 단어 중의성 해소에 적용한 연구로는 Lee와 Ng (2002)이 있다. 이 연구는 SENSEVAL-2와 SENSEVAL-1 데이터를 이용하여 이웃 단어의 품사 정보, 문맥, 지역적 연어, 구문적인 관계 등을 이용하여 단어 중의성을 해소하고자 하였으며, SVM을 비롯하여 다양한 분류기를 비교하였다. 이 연구에서 중의성을 해소하기 위한 다양한 정보원과 SVM 분류기를 사용하였을 때 가장 좋은 성능을 보이는 것으로 나타났다.

이외에도 지도학습 알고리즘은 다른 종류의 단일 분류기를 이용하는 연구(Ng and Lee 1996; Leacock, Miller, and Chodorow 1998; Mihalcea and Moldovan 2001)와 복수 개의 분류기를 조합하여 중의성 해소를 하는 연구들(Pedersen 2000; Florian and Yarowsky 2002)이 있다.

3. 실험 설계

3.1 실험 설계 및 실험문헌 집단

이 연구에서는 단어 중의성을 해소하고자 할 때, 가장 좋은 성능을 보이는 분류기 중에 하나인 SVM를 이용하여 최적의 의미 분류 성능을 가져오기 위한 다양한 실험을 수행하였다. 즉 분류기를 구축하기 위한 여러 조건 중에 성능에 영향을 미치는 주요한 요인인 문맥창과 자질(feature)에 대한 가중치 부여 방법을 다양하게 변화시켜 가장 좋은 의미 분류 성능을 가져오는 요인들을 파악하고자 하였다.

우선 중의성 해소 분류기에서 사용할 문맥창의 크기는 지역문맥 3개와 전역문맥 1개를 적용하였다. 지역문맥은 좌우 3단어, 한 문장, 그리고 좌우 50바이트 크기를 사용하였으며, 전역문맥으로 중의성 단어가 출현한 신문기사 전체를 대상으로 하였다.

이 실험에서 사용한 자질에 가중치를 부여하는 방법은 4가지를 적용하였다. 단순빈도의 경우 출현 여부만을 나타내는 이진 단어빈도(Binary)와 단어빈도(TF)를 사용하였으며, 정규화 빈도는 단어빈도 × 역문헌빈도(TF*IDF)와 로그를 취한 단어빈도 × 역문헌빈도(logTF*IDF)를 사용하였다.

일반 텍스트 문서화 기법에서처럼 학습 문헌에 해당하는 학습 문맥의 수는 600개를 설정하고 의미 분류 대상인 검증 문맥은 200개로 하였다. 이는 일반적인 텍스트 문서화 기법에

서 학습 문헌과 검증 문헌의 비율을 3대 1로 하기 때문이다. 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 각각의 중의성 해소 단어에 대해 다른 학습문맥 집단과 검증문맥 집단을 10회 반복해서 추출하고 이를 대상으로 SVM 분류기를 구축하고 의미를 식별하여 평균 성능을 산출하였다. 또한 10개의 랜덤으로 추출된 집단(학습집단 + 검증집단) 하나하나에 대해 서로 다른 문맥창 크기와 가중치 기법을 적용하여 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.

〈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%
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 SVM 분류기

이 연구에서 사용한 SVM 분류기는 구조적 위험 최소화 원리를 이용하여 부정예제로부터 긍정예제를 분리해 낼 수 있는 결정면을 찾아내는 분류모형이다(Vapnik 2000). 이 분류기는 기존의 다른 분류기에 비해 가장 좋은 성능을 보여주는 것으로 평가되고 있다(Joachims 1998; Yang and Liu 1999).

SVM은 크게 선형 SVM과 비선형 SVM으로 나누고 비선형 SVM에서는 커널함수에 의해 만들어지는 비선형 결정함수를 이용하게 된다. 비선형 SVM의 기본적인 커널함수로는 다항식 커널함수, RBF 커널함수, sigmoid 등이 주로 사용된다(Vapnik 2000). 이러한 SVM 분류기를 사용하기 전에 해당 학습집단을 대상으로 몇 가지 결정해야 하는 파라미터가 있다. 그 이유는 이들 파라미터에 따라 분류성능이 달라질 수 있기 때문이다. 이러한 파라미터 중 대표적인 것이 마진폭과 분류 오류 사이에 타협점을 찾아주는 패널티 파라미터 C와 커널함수의 파라미터이다.

이 연구에는 사용된 실험집단이 중의성 해소 단어마다 달라 각각의 적합한 파라미터를 구하기보다, 9개중 3개의 단어에 대해 최적의 파라미터 값을 구하고 그것을 반영하여 분류기를 구축하여 성능을 평가한 결과 대부분은 차이가 없어 SVM 분류기에서 기본적으로 제공하는 파라미터를 그대로 사용하였다. 또한 사전실험에서 비선형 SVM보다 문헌 자동분류에 주로 쓰이는 선형 SVM이 더 좋은 성능을 보여 이를 본 실험에 사용하였다. 이 연구에서 전반적인 실험을 위한 프로그램은 Python을 사용하였으며 SVM 분류기는 LIBSVM(Chang and Jin 2001)을 사용하였다.

일반적으로 문헌 자동분류의 실험결과를 평가하기 위해 각 범주별로 정확도, 재현율, 정확률, 그리고 F1 척도를 사용한다. 또한 전체 범주의 성능을 평가하기 위해 평균 정확도, 평균 재현율, 평균 정확률, 마이크로평균 F1 척도와

매크로평균 F1 척도를 사용한다. 마이크로평균 F1 척도의 경우 개별 범주에 관계없이 전체 재현율과 정확률을 계산하여 이를 F1척도를 이용하여 계산하는 방법으로 고빈도 범주에 영향을 많이 받는 성능 평가 척도이며 범주화 성능평가를 위해 주로 사용되는 방법이다. 매크로평균 F1 척도의 경우 모든 범주에 대해 F1 척도를 구하고 이를 더한 다음 범주 수로 나누어 평균을 계산하는 방법으로 저빈도 범주에 영향을 많이 받는 성능평가 척도이다(Yang and Liu 1999). 이 연구에서는 마이크로평균 F1 척도를 사용하여 의미 분류 성능을 평가하였다.

4. 의미 분류 실험 결과

4.1 실험 개요

단어 중의성 해소 기법이 사용되는 환경을 살펴보면, 주로 이 기법을 통해 중의성을 해소하는 것으로 끝나는 것이 아니라 그 결과를 이용하여 다른 어플리케이션이나 프로그램에서 사용한다. 예를 들어 중의성 해소된 질의를 이용하여 보다 정교한 정보검색을 수행하거나, 문서의 범주화에 사용되어 더 좋은 결과를 가져오도록 한다. 물론 단어 중의성 해소 결과가 잘못되면 이를 이용하는 다른 기법에 많은 영향을 미치는 것이 사실이다. 따라서 이 연구에서 SVM 분류기를 이용하여 최적의 단어 중의성 해소 성능을 얻기 위해 크게 문맥창의 크기와 분류 자질의 다양한 가중치 기법을 적용하여 실험을 수행하였다. 즉, 9개의 중의성 단어와 두 가지 변수를 조합하여 나올 수 있는 총 144(9×4×4)번의 분류 실험을 모두 수행하였다. 그 결과, 이들 단어와 변수에 따른 의미 분류 성능을 〈부록 1〉에 제시하였다.

이들 분류 실험들을 문맥창과 가중치 부여 방법에 따라 요약 정리하기 위해, 9개 중의성 단어의 분류성능인 마이크로평균 F1 값을 〈표 3〉과 같이 평균 산출하였다.

〈표 3〉에서 마이크로평균 F1 값에 대해 9개 단어의 평균을 보면, 문맥창의 크기는 큰 그룹(좌우 50바이트, 전역)에서 좋은 성능을 가져왔

〈표 3〉 문맥과 가중치에 의한 평균 분류성능

(마이크로평균 F1)

문맥 \ 가중치	Binary	TF	TFIDF	LogTFIDF	평균
3W	0.9181	0.9182	0.6970	0.6514	0.7962
문장	0.9165	0.9128	0.7514	0.7109	0.8229
50B	0.9405	0.9399	0.7259	0.7544	0.8402
전역	0.9352	0.9288	0.7806	0.8979	0.8856
평균	0.9276	0.9249	0.7387	0.7536	0.8362

으며, 가중치 기법 중에서는 이진 빈도(Binary) 내지 단어빈도(TF)에서 가장 좋은 분류성능을 가져왔다. 가장 좋은 성능을 가져온 문맥창의 크기는 좌우 50바이트(50B)이고, 가중치 기법으로는 이진 단어빈도(Binary)로 나타났다. 또한 근소한 차이로 단어빈도를 사용한 분류방법(0.9399)과 문맥으로 신문기사 전체를 설정한 분류방법(0.9352)이 그 다음으로 좋은 성능을 보이는 것으로 나타났다.

좀 더 자세한 분석을 위해 문맥창의 크기에 따른 성능 평가와 가중치 기법에 따른 성능 평가를 다음에 제시하였다.

4.2 문맥창 크기의 성능 평가

단어 중의성 해소 연구는 기존의 문헌을 분류하는 텍스트 범주화와 달리 대상 단어의 의미를 분류한다. 이는 문헌 분류에서 해당 문헌에 나타난 모든 자질을 그 대상으로 삼을 수 있지만, 단어 중의성 해소에서는 문맥창의 크기를 다양하게 변화할 수 있다는 뜻이기도 하다.

단어 중의성 해소에서 문맥의 크기는 연어라는 또 다른 측면에서 중요함을 갖는다. 앞서 간단히 언급하였지만 연어를 이용한 경우 "한 연어 한 의미"라는 연어제약을 이용하면 중의성 단어에 대해 의미를 식별할 수 있다. 예를 들어, "주식 감자", "운동 경기", "연구 기간" 등은 연어로 인해 중의성 단어가 특정 의미로 사용되는 것을 확인할 수 있다.

기존 연구(Yaroskwy 1994; 정영미, 이용구 2005)에서 연어가 형성되면, 98% 이상의 중의성 해소 성능을 가져오는 것으로 나타났다. 이러한 연어 정보나 공기 정보를 이용할 때 나타나는 특성이 중의성 해소 성능을 어떠한 영향을 미치지는 문맥창의 크기를 통해 간접적으로 살펴볼 수 있을 것이다. 따라서 가장 좋은 성능을 보이는 이진 단어빈도 가중치 기법을 중심으로 문맥창의 크기에 따른 각 중의성 단어의 의미 분류 성능을 살펴보면, 〈표 4〉와 같다.

전체적으로 이 표를 보면, 앞서 〈표 3〉에서 제시하였듯이, 좌우 50바이트와 전역 문맥창 같이 비교적 문맥창의 크기가 클수록 좋은 분류 성능을 보여주었다. 좌우 50바이트의 문맥창이 평균 성능 0.9405로 가장 좋은 분류 성능

〈표 4〉 문맥에 따른 단어별 분류성능

(마이크로평균 F1, 이진 단어빈도 적용)

문맥	감자	경기	기간	신병	신장	연기	인도	지구	지원	평균
3W	0.9680	0.8864	0.9850	0.9520	0.9335	0.8599	0.8080	0.9270	0.9430	0.9181
문장	0.9654	0.8844	0.9744	0.9520	0.9330	0.8413	0.8240	0.9339	0.9404	0.9165
50B	0.9870	0.9355	0.9770	0.9430	0.9495	0.9159	0.8395	0.9630	0.9545	0.9405
전역	0.9930	0.9410	0.9765	0.9130	0.9145	0.9239	0.8440	0.9620	0.9485	0.9352
평균	0.9784	0.9118	0.9782	0.9400	0.9326	0.8853	0.8289	0.9465	0.9466	0.9276

을 보였다. 다만 이에 근소한 차이로 전역 문맥창, 좌우 3단어, 문장 순으로 분류 성능이 낮아지고 있는 것을 볼 수 있다. 경기와 연기의 경우 다른 단어에 비해 문맥창 크기에 따라 다소 차이가 보이지만, 전체적으로 문맥창의 크기에 따른 단어 내에서 성능 차이는 그리 크지 않은 것으로 나타났다.

이 결과를 9개의 단어별로 보면, 감자/경기/연기/인도 등은 기사 단위의 전역 문맥창에서 가장 좋은 분류 성능을 보였으며, 신장/지구/지원은 좌우 50바이트에서 가장 좋은 성능을 보였다. 또한 기간/신병의 경우 좌우 3단어에 가장 좋은 성능을 보였다.

기존의 선행연구(정영미, 이용구 2005)의 나이브 베이즈 분류기를 이용하여 중의성 해소를 한 결과에서는 감자/지구/경기/인도가 전역 문맥창에서, 나머지 단어들은 좌우 50바이트에서 가장 좋은 성능을 가져온 것과 비슷하다. 다만, 약간의 차이는 분류기의 특성에서 나오는 것으로 볼 수 있다.

특히 이 연구에서 좌우 3단어의 문맥창이 기간/신병에서 좋은 성능을 보였는데, 이들 단어의 SVM 분류기의 성능에는 연어가 많은 기여를 하는 것으로 보인다. 좀 더 자세히 이들 단어의 연어를 제시하면, 기간의 경우 '시간'을 나타내는 첫 번째 의미로서 매입/고시/선거/행사/만료 등이 연어로서 공기하였으며, '기본이나 중심이 되는 부분'인 두 번째 의미로는 반도체/산업 등이 제시되었다. 신병의 경우 '법적으로 구속되어 있거나 구속할 사람'의 의미로 확보/처리/요청 등의 연어가 제시되었으며, '새로 입대한 병사'의 두 번째 의미로는 교육/훈련 등이, '앓는 병'의 세 번째 의미로는 치료/이유가 제시되었다.

4.3 가중치 기법의 성능 평가

텍스트 범주화에서 자동 분류를 수행하기 위해서는 분류 대상인 텍스트나 문헌에 대해 색인 작업을 통해 문헌 표현(document representation)을 만들어야 한다. 이때 문헌 표현 내에 포함될 자질 또는 색인어를 선정하고, 선정된 자질에 대해 적절한 가중치를 부여한다. 전자를 자질 선정이라고 하며, 이 연구에서는 연구범위를 벗어나므로 자질 선정을 수행하지 않았다.

자질에 대해 가중치 부여는 학습문헌 집단에서 출현한 모든 단어 또는 자질 선정을 통해 획득된 단어를 대상으로 이루어진다. 구체적인 가중치 방법으로 정보검색 분야에서 많이 사용되는 단어빈도와 역문헌빈도 등이 있다. 다만 이들 방법에 대해 기존의 텍스트 범주화 연구들에서는 다양한 결론을 가져왔다. 심지어는 서로 다른 결론들이 존재하기도 한다. 단어 중의성 해소에 단어빈도와 역문헌빈도 중심의 가중치 부여 방법을 적용하고 이를 SVM 분류기를 사용하여 분류성능을 제시하고자 하였다. 앞서 제시된 최적의 문맥창 크기인 좌우 50바이트를 적용하여 이진 단어빈도(Binary), 단순 단어빈도(TF), 단어빈도 × 역문헌빈도(TFIDF),

그리고 로그를 취한 단어빈도 × 역문헌빈도(LogTFIDF) 등의 가중치 부여 방법을 적용하여 의미 분류를 수행한 결과 분류 성능은 〈표 5〉와 같았다.

〈표 5〉에서 보면, 이진 빈도인 Binary가 9개의 중의성 대상 단어에 대해 평균 0.9405로 가장 좋은 성능을 보였다. 다만 단순 단어빈도인 TF에 대해서도 성능 평균이 0.9399로 Binary와 많은 차이를 보이지 않고 있다. 하지만 나머지 기법들과는 상당히 많은 차이를 보이고 있다. 특히 TFIDF에서 대략 22% 낮아진 평균 성능을 보였다. 이는 정보검색에서 가중치 부여 방법과는 정반대의 결과를 보이는 것이며, 일반적으로 텍스트 범주화 기법에서는 이진 빈도와 나머지 빈도가 비슷한 결과를 보이거나(정영미, 임혜영 2000), 복잡한 빈도가 높은 성능을 보이는데 반해, 의미 분류에서는 이진 빈도와 단어빈도가 복잡한 빈도에 비해 매우 높은 성능을 보이는 것을 알 수 있다.

각각의 단어 수준에서 보면, 이진빈도(Binary)에서는 경기/신병/신장/연기/지원이 가장 좋은 성능을 보였으며, 단어빈도(TF)에서는 감자/기간/인도/지구가 가장 좋은 성능을 보였다. 가장 좋은 성능을 보인 단어는 감자(0.9890)이며, 가장 낮은 성능을 보인 단어는 인도(0.8420)로 나타났다. 경기/신장/지구의 경우 가중치 방법에 따른 의미 분류 성능에서 심한 편차를 보였다.

앞서 다양한 문맥창의 크기와 가중치 부여 방법을 SVM 분류기에 적용하였을 때, 9개의 중의성 단어에 대해 평균적으로 가장 좋은 성능을 보이는 문맥창의 크기는 좌우 50바이트이었으며, 자질의 가중치 부여 방법은 이진 빈도라고 제시하였다. 하지만 〈부록 1〉에서 보이는 전체 144가지의 실험에서 평균이 아닌 각각의 단어별로 가장 좋은 성능을 보이는 예를 추출하여 따로 표시하면 〈표 6〉과 같다.

〈표 6〉에서 보면, 앞서 제시된 것처럼 문맥창의 크기는 좌우 50바이트(50B)가 4개의 단어(신장/연기/지구/지원)에서 가장 좋은 의미 분류 성능을 보였으며, 그 다음으로 전역 문맥이 3개의 단어(감자/경기/인도)에서 가장 좋은 성능을 보였다. 그리고 앞서 설명하였듯이, 연어 효과로 인해 성능이 향상되는 것으로

〈표 5〉 가중치 부여에 따른 단어별 분류성능

(마이크로평균 F1, 좌우50바이트 문맥 적용)

가중치방법	감자	경기	기간	신병	신장	연기	인도	지구	지원	평균
Binary	0.9870	0.9355	0.9770	0.9430	0.9495	0.9159	0.8395	0.9630	0.9545	0.9405
TF	0.9890	0.9340	0.9800	0.9410	0.9420	0.9099	0.8420	0.9690	0.9520	0.9399
TFIDF	0.7915	0.6908	0.9755	0.7110	0.5705	0.5471	0.6525	0.7045	0.8895	0.7259
LogTFIDF	0.8425	0.6133	0.9700	0.7290	0.6690	0.6016	0.6850	0.7755	0.9035	0.7544
평균	0.9025	0.7934	0.9756	0.8310	0.7828	0.7436	0.7548	0.8530	0.9249	0.8402

〈표 6〉 단어별 최고의 분류성능

단어	문맥	가중치	Mi F1
감자	전역	Binary	0.9930
경기	전역	Binary	0.9410
기간	3W	TF	0.9865
신병	3W	TF	0.9540
신장	50B	Binary	0.9495
연기	50B	Binary	0.9159
인도	전역	Binary	0.8440
지구	50B	TF	0.9690
지원	50B	Binary	0.9545

보이는 기간과 신병이 좌우 3단어(3W)에서 가장 좋은 성능을 보였다.

가중치 부여 방법의 경우 이진빈도(Binary)가 6개의 단어(감자/경기/신장/연기/인도/지원)에서 가장 좋은 성능을 보였으며, 단어빈도(TF)가 나머지 3개의 단어(기간/신병/지구)에서 가장 좋은 성능을 보였다.

단어별로 보면, 가장 좋은 의미 분류 성능을 보인 감자(0.9930)이며 가장 낮은 성능을 보인 단어는 인도(0.8440)로 나타났다. 이는 단어 자체가 의미 분류에 어려움이 있을 수도 있으나 해당 단어의 의미 수에 의해 나타나는 현상으로 보인다. 이를 위해 9개 단어에 대해 모든 분류 실험의 마이크로평균 F1값을 의미 수별로 평균을 계산하여 분류 성능을 제시한 〈표 7〉을 보면, 2개의 의미를 갖는 단어(감자/지구)의 분류 성능은 0.9043으로 5개의 의미를 갖는 인도의 분류 성능(0.7522) 보다 매우 높은 것을 알 수 있다.

또한 전체적으로 볼 때, 의미 수가 적을수록 높은 분류 성능을 보이며, 반대로 의미 수가 많을수록 낮은 성능을 보이는 것으로 보인다. 이는 일반적인 텍스트 범주화에서 범주화의 성능이 분류해야 할 범주수가 많아짐에 따라 낮은 성능을 보이는 것과 같은 이치이다.

〈표 7〉 단어의 의미별 평균 분류 성능

의미 수	평균 Mi F1
2	0.9043
3	0.8786
4	0.7679
5	0.7522
평균	0.8362

5. 결 론

이 연구에서는 9개의 중의성 대상 단어에 대해 SVM 분류기를 이용하여 의미 분류를 하고 그 성능을 분석하였다. 보다 구체적으로 의미 분류를 위해 다양한 크기의 학습문맥 창을

적용하였으며, 자질의 가중치에 대해서도 여러 가지 방법을 적용하여 최적의 분류성능을 가져오는 방안을 연구하였다. 실험대상 집단으로 신문기사를 사용하였으며, 9개의 단어에 대해 수작업으로 의미를 태깅하였다.

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

첫째, 학습 문맥창의 크기가 클수록 의미 분류인 중의성 해소 성능이 높아졌다. 좌우 50바이트와 신문기사 전체가 문맥인 전역 문맥이 작은 크기의 문맥보다 더 좋은 성능을 보였다. 이 둘 문맥 사이에는 큰 차이를 보이지 않았으나 좌우 50바이트가 약간 더 좋은 성능을 보였다. 다만 몇몇 단어에서 가장 작은 문맥창 크기인 좌우 3단어에서 가장 좋은 성능을 보였는데, 이는 연어에 의해 중의성 해소가 영향을 미치는 것으로 생각된다.

둘째, 자질의 가중치 부여 방법에서는 단순 방법이 이진 빈도와 단어빈도가 더 좋은 의미 분류 성능을 보였다. 기존의 텍스트 범주화 기법에서 결과와 달리, 이들 방법이 더 복잡한 역문헌빈도를 이용한 방법보다 매우 큰 차이로 좋은 성능을 보였다.

셋째, 중의성 단어가 갖는 의미 수는 분류 성능에 영향을 미치는 것으로 나타났다. 즉 의미 수가 많아질수록 의미 분류 성능은 떨어지는 것으로 나타났다.

이 연구에서는 완벽하게 의미 분류기를 이용한 단어 중의성 해소 기법을 분석하지 못하였지만, 기존의 텍스트 범주화 기법과 비교해 보면 같은 측면도 있고 다른 측면도 있는 것으로 보인다. 같은 측면의 경우 의미 수와 문맥창의 크기에 따른 성능의 차이이며, 다른 측면은 가중치 부여 방법과 연어와 같은 단어 중의성 해소 기법의 고유한 부분이다. 다만 이 연구에서 제한점으로는 9개의 한정된 중의성 단어를 대상으로 SVM이라는 하나의 분류기를 사용하였기에 좀 더 많은 단어와 다양한 분류기에 적용하여 폭넓게 검증할 필요가 있다.

참고문헌

정영미. 2005.『정보검색연구』. 서울: 구미무역(주) 출판부.

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

정영미, 임혜영. 2000. SVM 분류기를 이용한 문서 범주화 연구.『정보관리학회지』, 17(4): 229-248.

Chang, C. and C. Lin. 2001. LIBSVM: a library for support vector machines. [cited 2011. 01. 30]. 〈http://www.csie.ntu.edu.tw/~cjlin/libsvm〉.

Debole, F. and F. Sebastiani. 2003. "Supervised term weighting for automated text categorization." *Proceedings of SAC-03, 18th ACM Symposium on*

Applied Computing, 784-788.

Florian, R., and D. Yarowsky. 2002. "Modeling Consensus: Classifier Combination for Word Sense Disambiguation." *Proceedings of EMNLP*, 25-32.

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

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

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

Joachims, T. 1998. "Text categorization with Support Vector Machines :Learning with many relevant features." *Proceedings of the 10th European Conference on Machine Learning*, 137-142.

Leacock, C., G. Miller, and M. Chodorow. 1998. "Using corpus statistics and WordNet relations for sense identification." *Computational Linguistics*, 24(1): 147-166.

Leacock, C., G. Towell, and E. Voorhees. 1993. "Corpus based statistical sense resolution." *Proceedings of the ARPA Workshop on Human Language Technology*, 260-265.

Lee, Y., and H. Ng. 2002. "An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation." *Proceedings of the 7th Conference on Empirical Methods in Natural Language Processing(EMNLP)*, Philadelphia, U.S.A., 41-48.

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*, Chania, Greece, 267-273.

Mihalcea, R., and D. Moldovan. 2001. "A Highly Accurate Bootstrapping Algorithm for Word Sense Disambiguation." *International Journal on Artificial Intelligence Tools*, 10(1-2): 5-21.

Ng, H. T. and H. Lee. 1996. "Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Examplar Based Approach." *Proceedings of the 34th Annual Meeting of the ACL, University of California*, California, U.S.A., ACL Press, 40-47.

Pedersen, T. 2000. "A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambigua-

tion." *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics*, 63-69.

Pedersen, T. 2002. "A Baseline Methodology for Word Sense Disambiguation." *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, 126-135.

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

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

Vapnik, V. N. 2000. *The nature of statistical learning theory*. 2nd ed. New York: Springer.

Yang, Y., and X. Liu. 1999. "A re-examination of text categorization methods." *Proceedings of the ACM SIGIR Conference on Research and Development in International Retrieval*, 42-49.

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

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

〈부록 1〉 문맥과 가중치 방법에 따른 단어별 의미 분류 성능

단어	문맥	가중치	Mi F1
감자	좌우3 단어	Binary	0.9680
		TF	0.9690
		TFIDF	0.6280
		LogTFIDF	0.6060
	한 문장	Binary	0.9654
		TF	0.9619
		TFIDF	0.8803
		LogTFIDF	0.7570
	좌우50바이트	Binary	0.9870
		TF	0.9890
		TFIDF	0.7915
		LogTFIDF	0.8425
	전역	Binary	0.9930
		TF	0.9890
		TFIDF	0.8570
		LogTFIDF	0.9615
경기	좌우3 단어	Binary	0.8864
		TF	0.8829
		TFIDF	0.6273
		LogTFIDF	0.4772
	한 문장	Binary	0.8844
		TF	0.8829
		TFIDF	0.6346
		LogTFIDF	0.5696
	좌우50바이트	Binary	0.9355
		TF	0.9340
		TFIDF	0.6908
		LogTFIDF	0.6133
	전역	Binary	0.9410
		TF	0.9180
		TFIDF	0.6453
		LogTFIDF	0.8289
기간	좌우3 단어	Binary	0.9850
		TF	0.9865
		TFIDF	0.9765
		LogTFIDF	0.9765
	한 문장	Binary	0.9744
		TF	0.9775
		TFIDF	0.9729
		LogTFIDF	0.9719

단어	문맥	가중치	Mi F1
기간	좌우50바이트	Binary	0.9770
		TF	0.9800
		TFIDF	0.9755
		LogTFIDF	0.9700
	전역	Binary	0.9765
		TF	0.9770
		TFIDF	0.9750
		LogTFIDF	0.9760
신병	좌우3 단어	Binary	0.9520
		TF	0.9540
		TFIDF	0.6720
		LogTFIDF	0.6140
	한 문장	Binary	0.9520
		TF	0.9370
		TFIDF	0.7760
		LogTFIDF	0.6590
	좌우50바이트	Binary	0.9430
		TF	0.9410
		TFIDF	0.7110
		LogTFIDF	0.7290
	전역	Binary	0.9130
		TF	0.9020
		TFIDF	0.8280
		LogTFIDF	0.8650
신장	좌우3 단어	Binary	0.9335
		TF	0.9330
		TFIDF	0.5815
		LogTFIDF	0.5315
	한 문장	Binary	0.9330
		TF	0.9285
		TFIDF	0.6735
		LogTFIDF	0.6005
	좌우50바이트	Binary	0.9495
		TF	0.9420
		TFIDF	0.5705
		LogTFIDF	0.6690
	전역	Binary	0.9145
		TF	0.9120
		TFIDF	0.6815
		LogTFIDF	0.8810

단어	문맥	가중치	Mi F1
연기	좌우3 단어	Binary	0.8599
		TF	0.8534
		TFIDF	0.5556
		LogTFIDF	0.4840
	한 문장	Binary	0.8413
		TF	0.8338
		TFIDF	0.5591
		LogTFIDF	0.5460
	좌우50바이트	Binary	0.9159
		TF	0.9099
		TFIDF	0.5471
		LogTFIDF	0.6016
	전역	Binary	0.9239
		TF	0.9114
		TFIDF	0.6438
		LogTFIDF	0.8869
인도	좌우3 단어	Binary	0.8080
		TF	0.8110
		TFIDF	0.6315
		LogTFIDF	0.5980
	한 문장	Binary	0.8240
		TF	0.8225
		TFIDF	0.6455
		LogTFIDF	0.6575
	좌우50바이트	Binary	0.8395
		TF	0.8420
		TFIDF	0.6525
		LogTFIDF	0.6850
	전역	Binary	0.8440
		TF	0.8430
		TFIDF	0.7180
		LogTFIDF	0.8125

단어	문맥	가중치	Mi F1
지구	좌우3 단어	Binary	0.9270
		TF	0.9290
		TFIDF	0.7100
		LogTFIDF	0.6970
	한 문장	Binary	0.9339
		TF	0.9339
		TFIDF	0.7557
		LogTFIDF	0.7317
	좌우50바이트	Binary	0.9630
		TF	0.9690
		TFIDF	0.7045
		LogTFIDF	0.7755
	전역	Binary	0.9620
		TF	0.9575
		TFIDF	0.7545
		LogTFIDF	0.9255
지원	좌우3 단어	Binary	0.9430
		TF	0.9450
		TFIDF	0.8905
		LogTFIDF	0.8780
	한 문장	Binary	0.9404
		TF	0.9369
		TFIDF	0.8646
		LogTFIDF	0.9048
	좌우50바이트	Binary	0.9545
		TF	0.9520
		TFIDF	0.8895
		LogTFIDF	0.9035
	전역	Binary	0.9485
		TF	0.9490
		TFIDF	0.9225
		LogTFIDF	0.9440