ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2015 March: 21(1): 1~13 http://dx.doi.org/10.13088/jiis.2015.21.1.01

사전과 말뭉치를 이용한 한국어 단어 중의성 해소

정한조
한국과학기술정보연구원 (KISTI), 첨단정보융합본부, NTIS 센터
(hanjo.jeong@kisti.re.kr)

박병화
한남대학교 경상대학 비즈니스통계학과
(bpark@hnu.ac.kr)

빅데이터 및 오피니언 마이닝 분야가 대두됨에 따라 정보 검색/추출, 특히 비정형 데이터에서의 정보 검색/추출 기술의 중요성이 나날이 부각되어지고 있다. 또한 정보 검색 분야에서는 이용자의 의도에 맞는 결과를 제공할 수 있는 검색엔진의 성능향상을 위한 다양한 연구들이 진행되고 있다. 이러한 정보 검색/추출 분야에서 자연어처리 기술은 비정형 데이터 분석/처리 분야에서 중요한 기술이고, 자연어처리에 있어서 하나의 단어가 여러개의 모호한 의미를 가질 수 있는 단어 중의성 문제는 자연어처리의 성능을 향상시키기 위해 우선적으로 해결해야하는 문제점들의 하나이다. 본 연구는 단어 중의성 해소 방법에 사용될 수 있는 말뭉치를 많은 시간과 노력이 요구되는 수동적인 방법이 아닌, 사전들의 예제를 활용하여 자동적으로 생성할 수 있는 방법을 소개한다. 즉, 기존의 수동적인 방법으로 의미 태깅된 세종말뭉치에 표준국어대사전의 예제를 자동적으로 태깅하여 결합한 말뭉치를 사용한 단어 중의성 해소 방법을 소개한다. 표준국어대사전에서 단어 중의성 해소의 주요 대상인 전체 명사 (265,655개) 중에 중의성 해소의 대상이 되는 중의어 (29,868개)의 각 센스 (93,522개)와 연관된 속담, 용례 문장 (56,914개)들을 결합 말뭉치에 추가하였다. 품사 및 센스가 같이 태깅된 세종말뭉치의 약 79만개의 문장과 표준국어대사전의 약 5.7만개의 문장을 각각 또는 병합하여 교차검증을 사용하여 실험을 진행하였다. 실험 결과는 결합 말뭉치를 사용하였을 때 정확도와 재현율에 있어서 향상된 결과가 발견되었다. 본 연구의 결과는 인터넷 검색엔진 등의 검색결과의 성능향상과 오피니언 마이닝, 텍스트 마이닝과 관련한 자연어 분석/처리에 있어서 문장의 내용을 보다 명확히 파악하는데 도움을 줄 수 있을 것으로 기대되어진다.

주제어 : 중의어 해소, 자연어처리, 결합 말뭉치, 벡터 공간 모델

논문접수일 : 2015년 2월 14일 논문수정일 : 2015년 3월 9일 게재확정일 : 2015년 3월 10일
투고유형 : 학술대회 우수 교신저자 : 박병화

1. 개요

요즘 화두가 되는 빅데이터 분석, 사회망 분석, 오피니언 마이닝의 중요한 근간이 되는 분야 중 하나가 텍스트 마이닝 및 자연어 처리(Natural Language Processing)분야이다. 특히 소셜네트워크 서비스, 블로그, 모바일 등에서 대량으로 생산되고 있는 비정형데이터를 처리하고 유용한 정보를 추출할 수 있는 빅데이터와 관련된 기술들은 자연어 처리 기술의 진보를 가속화하고 있다 (Kim and Kim, 2014). 자연어는 한 단어가 여러 가지 의미를 가지고 있는 어휘적 중의성 (ambiguity)을 내포하고 있으며 언어의 근본적인 특성이라 할 수 있다 (Agirre and Edmonds, 2007). 따라서 기계를 이용한 자동적인 자연어 처리 시 단어의 중의성으로 인한 문제는 완전한 자연어의 이해를 어렵게 하는 걸림돌이라 할 수 있다. 기존의 텍스트 마이닝 및 자연어 처리 방법론에서는 키워드(keyword)가 내포하는 의미가 아닌, 단순 키워드 자체를 처리 단위가 되는 자질

(feature)에 포함하여 처리를 한 이유로 단어의 중의성 문제를 안고 있다. 또한 보다 향상된 검색엔진을 위해서도 자연어 처리는 중요한 문제라 할 수 있는데 기존의 검색엔진은 단순 키워드 기반의 검색 방법을 사용하여왔고 이로 인해 사용자의 의도와 관련없는 검색결과를 제공하는 불편함을 제공하기도 하였다 (Kim and Kim, 2012).

단어 중의성 해소(WSD: Word Sense Disambiguation) 문제는 특정한 문장에서 단어의 쓰임에 의해 활성화된 단어의 의미를 결정하는 문제라 할 수 있다 (Agirre and Edmonds, 2007). 이에, 단어 중의성 해소를 위해 문장 또는 문단에서 사용된 키워드가 내포하고 있는 의미를 사전(dictionary), 분류체계(taxonomy), 및 언어 온톨로지(linguistic ontology)의 연관 단어들(예를 들어, domain term, hypernym, synonym 등)과 의미풀이말(gloss words)을 비교하여 공통된 단어의 수(co-occurred terms)에 의해 결정하는 방법과 의미 태깅된 예제 문장을 기반으로 한 지도학습(supervised learning)을 통해 결정하는 방법이 주로 이용되었다.

본 논문은 위의 의미풀이말과 말뭉치(corpus)를 동시에 사용하여 단어의 의미를 결정하는 단어 중의성 해소 방법론을 제시하고, 표준국어대사전에 정의된 연관 단어뿐만 아니라 예시로 정의된 예제 문장들을 세종 말뭉치에 병합하여 확장된 말뭉치를 사용함으로써 단어 중의성 해소 문제의 Precision과 Recall을 높이는 방법론을 제시한다. 본 논문은 다음과 같이 구성된다. 2 장에서는 WSD와 관련된 과거 연구 결과들을 살펴볼 것이다, 3 장에서는 본 논문에서 제안한 단어 중의성 해소 방법을 설명하며, 4 장에서는 실험 및 결과를 기술한다. 마지막 장에서는 결론 및 활용분야에 대하여 논의하도록 한다.

2. 관련연구

WSD를 위한 접근법에는 대략 세 가지로 나눌 수 있는데 지식 기반, 말뭉치 기반 지도 학습, 말뭉치 기반 자율 학습 방법이 있다 (Agirre and Edmonds, 2007).

2.1. 지식 기반 방법

말뭉치라 함은 사람들이 사용하는 문장들을 수집해 놓은 대용량의 문서 뭉치를 지칭하는데 (Choi and Park, 2013), 지식 기반 (knowledge based) 혹은 사전 기반(dictionary based)방법은 이러한 말뭉치를 이용하지 않고 단어 중의성을 해소하기 위해 주로 사전 (Lesk, 1996), 동의어 사전 (Yarowsky, 1992), 언어적 지식 기반을 이용한 방법들이 있다. 지식 기반 방법의 성능은 일반적으로 말뭉치 기반에 비해 떨어지나 (Roberto, 2009) 단어 중의성 해소를 위한 지식 기반 방법들은 다양한 언어자원을 활용하기 때문에 효율적이라 할 수 있다.

2.2. 말뭉치 기반 지도

말뭉치 기반 지도 학습(supervised corpus-based approach)은 대규모의 의미 부착 말뭉치를 사용함으로써 다른 WSD방법에 비해 더 좋은 성능을 보이지만 사람이 직접 의미정보를 태깅하여 만든 말뭉치를 이용하기 때문에 이러한 태깅한 말뭉치를 확보하는데 상당한 비용과 시간이 요구되며 자료 부족 문제(Hur and Ock, 2001; Kwon, 2010)를 가지고 있다.

2.3. 말뭉치 기반 자율 학습

말뭉치 기반 자율 학습(unsupervised corpus-

based approach)은 원시 말뭉치에서 직접 정보를 추출하는 것으로 말뭉치 구축이 용의하며 의미 부착된 말뭉치(sense-tagged corpus)를 필요로 하지 않는 방식 (Ock et al., 2002)으로써 말뭉치 내에서 중의어의 의미를 나타내는 클러스터를 자동으로 만들어 의미 구분을 시도한 Schütze (1999)와 Ng et al. (Ng et al., 2003)의 연구를 들 수 있다. 말뭉치 기반 자율 학습은 원시 말뭉치를 직접 사용하기 때문에 세밀한 의미 구분이 쉽지 않으며 정확도가 보장되기 어렵다.

일반적으로 영어를 이용한 단어 중의성 해소에 관한 활발한 연구가 진행되었으나 한국어와 영어의 언어적인 차이와 한국어에 대한 실용적인 의미계층망의 부재로 다양한 연구방법론을 직접 적용하기에는 어려움이 따른다 (Hur and Ock, 2001).

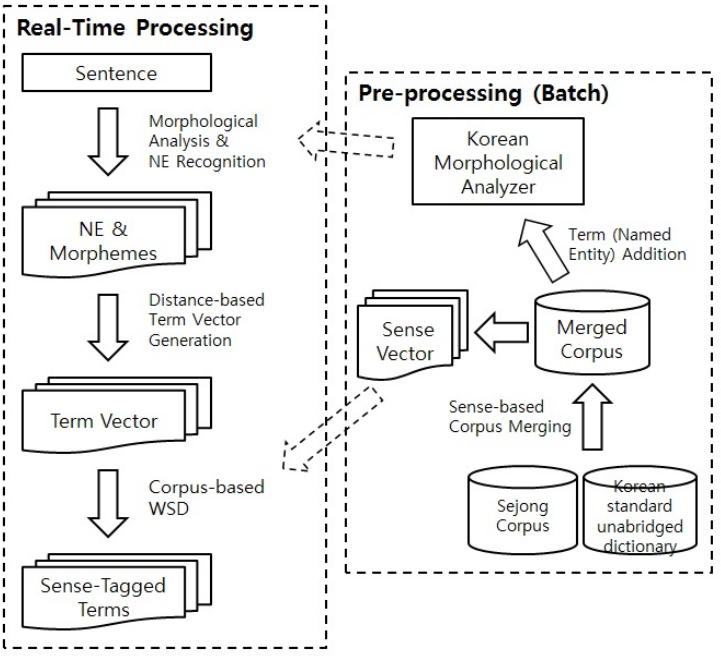
3. 결합 말뭉치 (Merged Corpus) 기반 WSD

3.1. 세종말뭉치, 표준국어대사전의 말뭉치의 결합

표준국어대사전은 국립국어원에서 한국어의 표준어 규정, 한글 맞춤법 등의 어문 규정을 준수하여 발행한 한국어 사전이며, 각 단어의 뜻과 센스 별 유의어, 정의, 속담, 용례들을 포함하고 있다. 본 논문은 말뭉치 기반의 WSD에 있어서 문제점인 말뭉치 부족을 현재 태깅된 세종말뭉치에 표준국어대사전에 정의된 속담, 용례 등 사전에 정의된 말뭉치들을 결합하여 말뭉치 부족 문제의 해결을 목적으로 하고 있다. 그리고 사전에 있는 이미 존재하는 문장들을 이용함으로써 시간과 노력이 많이 필요로 하는 매뉴얼 태깅 (manual tagging)의 문제점을 보완하며 매뉴얼 태깅보다 정확하고 유용한 문장들을 단어 중의성 해소에 사용할 수 있다.

세종말뭉치에 표준국어대사전의 문장들을 결합하기 위해 하나의 공통 인덱스를 사용해야 하는데 세종말뭉치의 경우 표준국어대사전에 정의된 센스 인덱스를 사용하여 태깅되어 있어서 별도의 맵핑 작업 없이 결합이 가능하다. 말뭉치에 포함하기 위해서 문장 형태의 데이터가 필요하며, 표준국어대사전에는 각 단어의 센스에 대한 정의, 속담, 용례 문장들이 존재한다. 이중 정의 문장은 해당 단어가 문장 내에서 일반적으로 쓰이는 문장이 아니라서 제외하고 속담, 용례 문장들만 말뭉치에 포함시켰다. 단어 중의성 해소의 주요 대상인 명사만을 대상으로 하였고 전체 명사 (265,655개)중에서 중의성 해소의 대상이 되는 중의어(29,868개)의 각 센스(93,522개)와 연관된 속담, 용례 문장 (56,914개)들을 말뭉치에 추가하였다. 특이한 점은 중의어의 비율이 전체 명사의 11.24%로 적은 편이나, 세종말뭉치 분석 결과 일반 문장에서 중의어가 차지하는 비중이 60%가 넘는 것으로 나타났다. 이는 자연어 문장 분석에서의 WSD의 중요성을 나타낸다고 할 수 있다.

마지막으로, 다음 장에 소개할 말뭉치 기반 WSD 알고리즘을 적용하기 위해, 문장을 Term Vector 형식으로 변환해야 하는데, 세종 말뭉치의 경우 센스 인덱스와 품사가 이미 태깅되어 있어 간단한 Parser를 가지고 WSD 알고리즘에서 사용될 주체 명사, 주체 명사의 센스, 의미풀이말로 사용 될 명사, 동사, 형용사 단어들을 문장에서 추출하여, 해당 주체 명사의 센스에 대해 Term Vector를 구성하였다. 표준국어대사전의



〈Figure 1〉 Merged Corpus-based WSD Process

속담, 용례 문장의 경우에도 한국어 형태소 분석기를 이용하여 형태소 분석을 한 후, 주체 명사의 센스 (즉, 속담, 용례 문장의 주체)에 대한 의미풀이말이 될 명사, 동사, 형용사 단어들만 뽑아내어 Term Vector를 구성하였다. <Figure 1>은 이러한 결합 말뭉치 기반 WSD의 전체 프로세스를 보여준다. 전처리 과정에서는 세종말뭉치 데이터와 표준국어대사전의 문장들을 결합하여 Merged Corpus를 구축하고 표준국어대사전의 센스 인덱스를 기반으로 하여 Sense Vector들을 생성한다. 문장의 형태소 분석 시 사용할 형태소 분석기의 성능을 높이기 위해 Merged Corpus에서 Sense Vector를 생성하는 데 사용된 Term들을 Named Entity에 추가하여 형태소 분석기의 NE 사전에 추가하여 확장한다. 전처리 과정 단계에서 확장된 형태소 분석기를 사용하여 실시간으로 문장이 입력되면 형태소 분석 및 NE Recognition 과정을 거쳐 Term들을 추출하고 이를 기반으로 입력된 문장을 Term Vector 형식으로 생성한다. 마지막으로, 입력 문장의 Term Vector와 전처리 과정에서 생성된 Sense Vector Model을 사용하여 다음 장에 소개 될 벡터 공간 모델 기반 WSD 방법으로 WSD를 수행하여 입력 문장의 Term들의 센스를 결정한다.

3.2. 벡터 공간 모델 (Vector-Space Model) 기반 WSD

3.2.1. Distance Weight 기반 센스 벡터 공간 모델

본 연구에서는 말뭉치를 이용하여 문장 내의 명사 단어들의 센스를 결정하는 것을 목표로 하고

있으며 효율적인 대규모 말뭉치 처리를 위해 벡터 공간 모델 (Lund and Burgess, 1996; Manning et al., 2008)을 이용한다. 일반적으로 벡터 공간 모델은 문서를 단어들의 벡터로 표현하는 방법으로 한 문서 안에서 한 단어가 여러 번 출현할 수 있고, 각각 다른 뜻으로 쓰일 가능성이 있으므로 다르게 쓰일 확률이 적은 문장 단위로 벡터 모델을 생성한다. 그리고 문장을 센스 단위로 처리하기 위해서 단어 벡터가 아닌 단어와 그 단어의 뜻의 조합을 인덱스로 가지는 센스 벡터를 이용한다. 추가로, 대규모 말뭉치를 효율적으로 처리하기 위해서 문장 안에서 명사의 뜻을 정의하는 데 영향을 많이 끼치는 기준 명사의 좌우 5개의 명사, 동사, 형용사만을 가지고 센스 벡터를 생성한다.

각 센스는 전체 문장에서 같이 쓰인 단어 또는 의미풀이 말들의 위치 값(distance)들을 기반으로 계산된 Distance Weight를 사용하여 벡터화 한다. 위치 값은 기준 센스에서 떨어진 단어 수의 역순으로 5에서 1까지 주어진다. 즉, 1 단어 떨어지면 5, 2 단어 떨어지면 4, 계속해서 5단어 떨어지면 1이 주어진다. 각 의미풀이 말들의 Distance Weight값들은 모든 문장에서 기준 센스와 같이 쓰인 경우의 위치 값의 누적값을 식(3)과 같이 구하고 하나의 기준 센스에 대해서 식(2)와 같이 정규화 한다.

$$V(s_i) = (ndw(gw_{i,1}), ndw(gw_{i,2}), ..., ndw(gw_{i,n})) \quad (1)$$

$$where \quad ndw(gw_{i,j}) = \frac{dw(gw_{i,j})}{\max_{j \in gw_{i,j}} dw(gw_{i,j})} \quad (2)$$

$$dw(gw_{i,j}) = \sum dist(gw_{i,j}) \quad (3)$$

위 식에서, $V(s_i)$는 센스 s_i를 표현하는 벡터를 의미한다. $gw_{i,j}$의 Distance Weight와 정규화된 Distance Weight를 나타낸다. $dw(gw_{i,j})$는 센스 j와 같이 출현한 문장 내의 모든 위치 값들의 단순 누적 값이다. 즉, Distance Weight는 위치 값뿐만 아니라 출현 빈도가 높을수록 커지게 된다.

3.2.2. 단어의 의미 결정

위의 센스 벡터 공간 모델을 사용하여 Training Set를 통해 나타난 중의성을 가지는 모든 명사에 대해 각 센스 별로 벡터 공간 모델을 생성한 후, 질의 문장이 주어지면, 질의 문장 중에 형태소 분석기를 통해서 명사를 뽑아낸다. 뽑아낸 각 명사들에 대해서 표준국어대사전을 통해 중의어 여부를 판단한 후, 중의성을 가지는 각각의 명사들에 대해서 식(4)와 같은 Naïve Bayes Classifier를 이용하여 센스를 결정한다. 즉, Naïve Bayes Model을 통해 Training Set로 센스 벡터를 구한 후, Maximum A Posteriori (MAP) Decision Rule을 통해 질의 문장내의 중의어의 센스를 결정한다.

식(4)에서 각 센스의 선험확률(prior probability)은 식(5)와 같이 단순하게 각 센스의 중의어의 다른 센스들과의 빈도수의 비율로 할당한다. 그리고 Likelihood는 Training Set에서 구한 각각의 센스 벡터에 대해서 질의 문장에서의 중의어 센스 벡터에 대한 Cosine Similarity를 식(6)과 같이 구한다. 마지막으로 질의 문장에서 각각의 중의어에 대해 이 과정을 통해 Posterior 값이 가장 큰 센스를 선택한다.

$$classify(gw_{i,1}, gw_{i,2}, ..., gw_{i,n}) = \arg\max_{s_i} p(s_i) p(gw_{i,1}, gw_{i,2}, ...gw_{i,n} \mid s_i) \quad (4)$$

$$p(s_i) = \frac{occur(s_i)}{\sum_{word(s_i)=word(s_k)} occur(s_k)} \quad (5)$$

$$p(gw_{i,1}, gw_{i,2}, ..., gw_{i,n} \mid s_i) = sim(V(s_i), V(s_q)) = \frac{\sum_j ndw(gw_{i,j}) \times ndw(gw_{q,j})}{\sqrt{\sum_j ndw(gw_{i,j}))^2} \times \sqrt{\sum_j ndw(gw_{q,j}))^2}} \quad (6)$$

위 식에서 $p(s_i)$는 센스 s_i의 선험적 확률을 나타내고, $occur(s_i)$는 전체 문자에서 센스 s_i가 출현된 전체 회수를 나타내며, $occur(s_k)$는 센스 s_i와 같은 단어(중의어)의 다른 센스들의 전체 출현 회수를 나타낸다. 마지막으로 $p(gw_{i,1}, gw_{i,2},..., gw_{i,n}|s_i)$는 센스 s_i일 때, 의미풀이말, $gw_{i,j}$들이 출현될 likelihood를 나타내면, 이는 센스 s_i와 WSD의 대상이 되는 질의 센스 s_q의 벡터들의 Cosine similarity로 대체하여 유추된다.

4. 실험 및 결과

실험은 품사 및 센스가 같이 태깅된 세종말뭉치의 문장 (약 79만 문장)과 표준국어대사전의 문장 (약 5.7만 문장)을 각각 또는 병합하여 교차검증 (Cross Validation)을 사용하여 수행하였다. 교차검증을 수행하기 위한 Test Set과 Training Set으로 나누기 위해 두 가지 방법을 사용하였다. 첫 번째 방법은 세종말뭉치의 문장을 순서대로 일련번호를 부여하여 모드 (Mod) 연산자를 사용하여 나머지가 각각 0~9인 10개의 10% Test Set를 만들고, Test Set를 제외한 각각의 남은 90%의 말뭉치로 10개의 Training Set를 만들어 각각의 Test & Training Set Pair에 대해 실험을 하였고, 두 번째 방법은 문장이 아닌 각 파일에 일련번호를 부여하여 Mod 연산자를 사용하여 10개의 Test & Training Set Pair를 만들어 실험을 수행하였다. <Table 1>은 문장을 기반으로 나눈 Test & Training Set Pair에 대한 결과를 나타내고, <Table 2>는 파일을 기반으로 나눈 Test & Training Set Pair에 대한 결과를 나타낸다. <Table 1>과 <Table 2>는 모두 10개의 Test Set의 평균값만 나타내는데, 10개의 Test Set의 결과가 두 방법 모두 거의 일치하였다. Precision은 평가된 중의어 중 정확하게 센스가 판단된 중의어의 비율로 식 (7)로 계산하였고, Recall은 Test Set의 전체 문장에서 나타난 중의어 중 평가 대상이 된 중의어의 비율로 식 (8)로 나타낸다.

$$Precision = \frac{\text{올바르게 판단된 중의어의 수}}{\text{평가된 중의어의 수}} \quad (7)$$

$$Recall = \frac{\text{평가된 중의어의 수}}{\text{전체 중의어의 수}} \quad (8)$$

결과에서 보듯이, 표준국어대사전의 경우 Precision 및 Recall이 모두 낮게 나왔다. 특히, Recall이 두드러지게 낮게 나왔는데, 중의어의 센스에 대한 예문이 전혀 없는 경우가 다수 존재하고, 세종말뭉치 문장에 비해 전체 문장 수가 7%에 불과한 점이 낮게 나온 이유 중의 하나로 추정된다. 추가로, Test Set 자체가 세종말뭉치의 하위집합(subset)이기 때문에, 상대적으로 세종말뭉치의 Recall이 높게 나온 것으로 추정된다. Precision의 경우에는, 마찬가지로 표준국어대사전의 말뭉치가 적은 이유도 있지만, 표준국어대사전의 경우 전문 인력이 태깅한 세종말뭉치와는 다르게 형태소 분석기를 통해 명사, 동사, 형용사를 자동으로 선택해서 이를 이용하였으므로, 형태소 분석 자체의 오류 때문에 정확도가 떨어졌을 것으

로 유추된다. 또한, Test Set 자체가 세종말뭉치의 하위집합임으로 선험확률이 상대적으로 Test Set 자체에 유용했을 가능성과 표준국어대사전의 경우 선험확률은 각 센스에 대한 예문의 수에 의존하게 되어 부정확할 가능성이 있다고 본다. 이를 뒷받침하는 결과로 <Table 1>의 세종말뭉치와 표준국어대사전의 병합 말뭉치의 성능 향상보다 <Table 2>의 병합 말뭉치의 성능 향상이 전체적으로 높은 것으로 볼 수 있다. <Table 1>의 경우는 같은 파일, 즉 동일 도서, 뉴스 등의 동일 저작자, 주제 등에서 Training Set와 Test Set으로 문장들이 분리된 것이고, <Table 2>의 경우는 하나의 소설, 뉴스 등의 컨텐츠의 모든 문장이 Test Set이나 Training Set이 됨으로 선험확률, 연관어 등이 서로 독립적이라 볼 수 있어 좀 더 무작위(random)한 실험 결과라고 볼 수 있다.

표준국어대사전을 이용한 결과가 세종말뭉치를 이용한 결과보다 좋지 않음에도 불구하고, 세종말뭉치의 Training Set와 표준국어대사전의 문장을 결합한 경우 세종말뭉치 단독으로 수행한 결과보다 모든 Test Set에 대해서 Precision과 Recall 모두 향상을 보인다. Recall의 향상은 Training Set에 세종말뭉치에 존재하지 않는 표준국어대사전의 중의어가 포함되게 되는 이유로 당연한 것으로 보이나, Precision의 경우에는 세종말뭉치 단독의 결과 자체도 상당히 높은 편인데, 모든 Test Set에서 일정하게 향상되었다는 점이 상당히 고무적이다. 추가적으로 <Table 1>과 <Table 2>의 실험에서 Recall 값이 전체적으로 감소했음에도, 즉 세종말뭉치의 영향력이 <Table 2>에서 감소했음에도 불구하고 Precision 값이 더 많이 향상된 점은 표준국어대사전의 말뭉치를 병합하는 것이 효과적이었다는 것을 보여준다. 특히, 표준국어대사전의 경우에는 각각의 센

〈Table 1〉 Results of Cross Validation (by Splitting Test & Training Sets Based on Sentences)

	Average Number of Sentences in Test Set	Average Number of Polysemic Words in Test Set	Average Number of Classified Polysemic Words	Average Number of Correctly Classified Polysemic Words	Precision (%)	Recall (%)
Korean Standard Dictionary Only	78936.2	283414	69508.8	59658.7	85.83%	24.53%
Sejong Corpus Only			267579.6	255479.9	95.48%	94.41%
Korean Standard Dictionary + Sejong Corpus			269527.1	258856.5	96.04%	95.10%
Improvement ((Korean Standard Dictionary + Sejong Corpus) / Sejong Corpus)	N/A				100.59%	100.73%

〈Table 2〉 Results of Cross Validation (by Splitting Test & Training Sets Based on Files)

	Average Number of Sentences in Test Set	Average Number of Polysemic Words in Test Set	Average Number of Classified Polysemic Words	Average Number of Correctly Classified Polysemic Words	Precision (%)	Recall (%)
Korean Standard Dictionary Only	80406.5	285000.8	70248.5	60304.1	85.84%	24.65%
Sejong Corpus Only			264651.8	250436.4	94.63%	92.86%
Korean Standard Dictionary + Sejong Corpus			266198.6	253837.2	95.36%	93.40%
Improvement ((Korean Standard Dictionary + Sejong Corpus) / Sejong Corpus)	N/A				100.77%	100.58%

스에 대한 예문의 수에 따라 선험확률이 결정되므로, 세종말뭉치와 같이 일상 문장들을 대상으로 Training 했을 때 중의어들의 각 센스에 대한 선험확률의 정확도가 부정확했을 가능성이 많아 선험확률에 대한 정확도 향상에는 기여한 것이 없어 보이나, 보다 연관이 많은 의미풀이말들을 Training Set에 포함시키는 효과를 주어 전체 정확도의 향상을 가져온 것으로 보인다.

5. 결론

본 논문에서는 한국어 문장에 대한 WSD의 성능을 향상시키기 위해, 기존의 말뭉치를 이용한 방법에 추가로 표준국어대사전의 예문을 자동으로 포함시켜서 수동(Manual) 태깅으로 인한 말뭉치의 부족과 부정확성 등을 개선하여, 자동으로 말뭉치를 확장하는 방법을 제시하였다. 추가적으로 일반적으로 쓰이는 단어 벡터 공간 모델을 센스에 적용하여 센스 벡터 공간 모델을 제시하였고, 이 모델에 지도 학습 모형인 Naïve Bayes Classifier를 적용하여 센스를 구별하는 방법을 제시하였다. 마지막으로 확장된 말뭉치에 대해서 다양한 Cross Validation을 수행하여 본 연구가 제시한 표준국어대사전을 이용한 말뭉치 확장, 센스 벡터 공간 모델과 Naïve Bayes Classifier를 이용한 WSD 방법을 테스트하였고, 단어 중의성 해소의 정확성과 재현율을 동시에 향상시킬 수 있었다.

본 연구는 키워드를 사용할 때 발생할 수 있는 모호성에 대한 해결 방안을 제시함으로써 키워드 기반의 인터넷 검색엔진의 검색 성능의 향상과 텍스트 마이닝, 자연어처리에 있어서 문장의 주제 및 내용을 보다 정확하게 파악하는데 기여할 수 있을 것으로 기대된다. Naïve Bayes Classifier는 널리 쓰이고 있으며 텍스트 분류 및 의료 진

단과 같은 실질적인 분야에서 효과가 있는 것으로 알려져 있다 (Al-Aidaroos et al., 2012; Besserve et al., 2007; Domingos and Pazzani, 1997; Kim and Park, 2012). 다만, Naive Bayes Classifier는 문제의 단순화, 효율성을 위해 실제와는 다르게 각각의 자질(Feature)들이 서로 독립적이라는 가정에 기초한다. 이러한 점은 본 연구의 제약이라 할 수 있으며 향후 문장 내 센스들의 모든 조합이나 부분적인 조합을 이용할 수 있는 방법이나 알고리즘을 적용하는 연구가 추가적으로 필요하다. 또한 구문분석을 통해 얻을 수 있는 단어간의 의존 및 수식관계를 추가적인 자질로 활용하면, WSD의 정확도를 높일 수 있을 것으로 보인다. 마지막으로, 세종사전과 표준국어대사전의 센스 인덱스를 맵핑해서 온톨로지 기반으로 구축된 세종말뭉치의 분류체계, 연관관계, 활용예제 등의 정보들을 활용하여 WSD의 성능을 향상시킬 수 있을 것으로 사료된다.

참고문헌(References)

Agirre, E. and P. Edmonds, *Word Sense Disambiguation: Algorithms and Applications*, Springer, 2007.

Al-Aidaroos, K. M., A. A. Baker, and Z. Othman, "Medical Data Classification with Naïve Bayes Approach," *Information Technology Journal*, Vol.11, No.9(2012), 1166~1174.

Besserve, M., L. Garnero, and J. Martinerie, "Cross-spectral Discriminant Analysis (CSDA) for the Classification of Brain Computer Interfaces," *Proceedings of the 3rd International IEEE/EMBS Conference on Neural Engineering*, (2007), 375~378.

Choi, Y. and J. Park, "The Need for Paradigm Shift in Semantic Similarity and Semantic Relatedness: From Cognitive Semantics Perspective," *Journal of Intelligence and Information Systems*, Vol.19, No.1(2013), 111~123.

Domingos, P. and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-one Loss," *Machine Learning*, Vol.29, No.2-3(1997), 103~130.

Hur, J. and C. -Y. Ock, "A Homonym Disambiguation System based on Semantic Information Extracted from Dictionary Definitions," *Journal of KIISE: Software and Applications*, Vol.28, No.9(2001), 688~698.

Kim, N. and J. Park, "Personal Information Detection by Using Naïve Bayes Methodology," *Journal of Intelligence and Information Systems*, Vol.18, No.1(2012), 91~107.

Kim, S. and G. Kim, "Ontology-based User Customized Search Service Considering User Intention," *Journal of Intelligence and Information Systems*, Vol.18, No.4(2012), 129~143.

Kim, S. and N. Kim, "A Study on the Effect of Using Sentiment Lexicon in Opinion Classification," Journal of Intelligence and Information Systems, Vol.20, No.1(2014), 133~148.

Kwon, M., "A Study on Word Sense Disambiguation Using GermaNet," *Dokohak*, Vol. 22(2010), 59~82.

Lesk, M., "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone," *Proceedings of the ACM-SIGDOC Conference*, (1986), 24~26.

Lund, K. and C. Burgess, "Producing High-dimensional Sematic Spaces from Lexical Co-occurrence," *Behavior Research Methods,*

Instrumentation, and Computers, Vol.28, No.2(1996), 203~209.

Manning, C. D., P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, New York, 2008.

Ng. H. T., B. Wang, and Y. S. Chan, “Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study,” *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, ACL*, (2003), 455~462.

Ock, C. -Y., E. -J. Ock, and W. -W. Lee, “Measuring Lexical Relationship of Co-occurrence Words in Modification Phrases,” *Hangul*, Vol.255 (2002), 129~154.

Roberto, N., "Word Sense Disambiguation: A Survey," *ACM Computing Surveys*, Vol.41, No.2(2009), 1~69.

Schütze, N., "Automatic Word Sense Discrimination," *Computational Linguistics*, Vol.24, No.1(1999), 97~123.

Yarowsky, D., “Word-sense Disambiguation Using Statistical Models of Roget’s Categories Trained on Large Corpora,” *Proceedings of COLING-92*, (1992), 454~460.

Abstract

Korean Word Sense Disambiguation using Dictionary and Corpus

Hanjo Jeong* · Byeonghwa Park**

As opinion mining in big data applications has been highlighted, a lot of research on unstructured data has made. Lots of social media on the Internet generate unstructured or semi-structured data every second and they are often made by natural or human languages we use in daily life. Many words in human languages have multiple meanings or senses. In this result, it is very difficult for computers to extract useful information from these datasets. Traditional web search engines are usually based on keyword search, resulting in incorrect search results which are far from users' intentions. Even though a lot of progress in enhancing the performance of search engines has made over the last years in order to provide users with appropriate results, there is still so much to improve it. Word sense disambiguation can play a very important role in dealing with natural language processing and is considered as one of the most difficult problems in this area. Major approaches to word sense disambiguation can be classified as knowledge-base, supervised corpus-based, and unsupervised corpus-based approaches.

This paper presents a method which automatically generates a corpus for word sense disambiguation by taking advantage of examples in existing dictionaries and avoids expensive sense tagging processes. It experiments the effectiveness of the method based on Naïve Bayes Model, which is one of supervised learning algorithms, by using Korean standard unabridged dictionary and Sejong Corpus. Korean standard unabridged dictionary has approximately 57,000 sentences. Sejong Corpus has about 790,000 sentences tagged with part-of-speech and senses all together. For the experiment of this study, Korean standard unabridged dictionary and Sejong Corpus were experimented as a combination and separate entities using cross validation. Only nouns, target subjects in word sense disambiguation, were selected. 93,522 word senses among 265,655 nouns and 56,914 sentences from related proverbs and examples were additionally combined in the corpus. Sejong Corpus was easily merged with Korean standard unabridged dictionary

* NTIS Center, Division of Advanced Information Convergence, Korea Institute of Science and Technology Information (KISTI)
** Corresponding author: Byeonghwa Park
70 Hannam-ro, Daedeok-gu, Daejeon 306-791, Korea
Tel: +82-42-629-7518, Fax: +82-42-629-8415, E-mail: bpark@hnu.ac.kr

because Sejong Corpus was tagged based on sense indices defined by Korean standard unabridged dictionary. Sense vectors were formed after the merged corpus was created. Terms used in creating sense vectors were added in the named entity dictionary of Korean morphological analyzer. By using the extended named entity dictionary, term vectors were extracted from the input sentences and then term vectors for the sentences were created. Given the extracted term vector and the sense vector model made during the pre-processing stage, the sense-tagged terms were determined by the vector space model based word sense disambiguation.

In addition, this study shows the effectiveness of merged corpus from examples in Korean standard unabridged dictionary and Sejong Corpus. The experiment shows the better results in precision and recall are found with the merged corpus. This study suggests it can practically enhance the performance of internet search engines and help us to understand more accurate meaning of a sentence in natural language processing pertinent to search engines, opinion mining, and text mining. Naïve Bayes classifier used in this study represents a supervised learning algorithm and uses Bayes theorem. Naïve Bayes classifier has an assumption that all senses are independent. Even though the assumption of Naïve Bayes classifier is not realistic and ignores the correlation between attributes, Naïve Bayes classifier is widely used because of its simplicity and in practice it is known to be very effective in many applications such as text classification and medical diagnosis. However, further research need to be carried out to consider all possible combinations and/or partial combinations of all senses in a sentence. Also, the effectiveness of word sense disambiguation may be improved if rhetorical structures or morphological dependencies between words are analyzed through syntactic analysis.

Key Words : Word Sense Disambiguation, Natural Language Processing, Merged Corpus, Vector Space Model

Received : February 14, 2015 Revised : March 9, 2015 Accepted : March 10, 2015
Type of Submission : Outstanding Conference Paper Corresponding Author : Byeonghwa Park

저 자 소 개

정 한 조

George Mason University에서 Information Systems 석사 및 Information Technology 박사를 취득하였다. LG 전자 CTO 부문 연구소에서 스마트 TV대상 BI 시스템 구축과 빅데이터 기반 앱 검색/추천 시스템, IoT 기반 스마트 기기 추천 시스템 등의 연구 및 개발에 참여하였고, 현재 한국과학기술정보연구원 (KISTI)에서 선임연구원으로 재직 중이다. 주요 연구 관심 분야는 빅데이터 기반의 데이터 마이닝, SNS 분석, 자연어 처리 및 검색/추천 시스템과 시맨틱 웹/온톨로지 Rule 기반 시스템, 기계학습 등의 Intelligence 기반 지식처리 시스템 등이다.

박 병 화

University of Arizona를 졸업하고 University of Nebraska-Lincoln에서 경영학 석사, George Mason University에서 Computational Sciences & Informatics 박사를 취득하였다. 현재 한남대학교 경상대학 비즈니스통계학과에 재직 중이며 연구 관심 분야는 생산 및 품질, 빅데이터 응용, 데이터 마이닝, 사회연결망분석 등을 이용한 학제간 연구에 관심이 많다.