

공기정보를 이용한 단어 의미 중의성 해결 방안

박요셉⁰, 김경임, 박혁로

전남대학교, 전자컴퓨터공학과

pys1249@ejnu.net, kyungim@ejnu.net, hyukro@jnu.ac.kr

Word Sense Disambiguation Method Using Co-occurrence Information

Yo-sep Park⁰, Hyuk-ro Park⁰

Chonnam University, Computer Science Department

요약

단어 의미 중의성은 자연언어처리 분야에서의 주요 관심 분야이다. 한국어에서의 단어 의미 중의성 문제는 다른 언어에 비하여 연구가 미흡한 상태이다. 기존 연구에서는 빈도 수에 기반한 공기 정보 벡터를 이용한 방법에서 처리되지 못하는 경우가 발생하였다. 또한 사전에 기반한 상위어 추출 시에 정형화된 형태가 아닌 경우에 어려움이 발생하였다. 본 논문에서는 상호정보량을 추가하여 공기 정보 처리 과정 시에 발생하는 오류를 최소화 하였다. 또한 대상 명사의 상위어 추출 문제를 해결하기 위해 어휘 지식 베이스를 적용하였다.

주제어: 단어 중의성 해결, 공기정보, 상호정보량, 어휘 지식 베이스

1. 서론

단어 의미 중의성이란 형태학적으로 같은 어휘들의 의미가 문맥에 따라 달라짐에 따라 발생하는 문제를 말한다. 자연언어처리에서 단어 의미 중의성(Word Sense Disambiguation : WSD)은 주요 관심 분야에 해당된다. 특히, 한국어 경우에는 많은 어휘들이 한자어에서 유래되어 동형이의어(homonym)의 개수가 많기 때문에 WSD 문제가 품사 및 구문적 중의성에 비하여 심각한 편이다. 그 자체가 하나의 완전한 작업은 아니지만 대부분의 자연언어처리 작업에서 반드시 필요한 작업이다. WSD는 기계번역과 정보검색 및 하이퍼 텍스트 탐색에서 매우 중요한 역할을 한다[1][2].

[11]에서 제시한 빈도 수에 기반한 공기정보 처리 시에 분석 되지 못하는 문제점이 있다. 또한 ‘우리말 큰사전(1997)’에 기반한 상위어 추출 오류가 발생하였다. 본 논문에서는 공기 정보 처리 시 문제점을 해결하기 위해 정보이론에서의 상호정보량을 추가하였다. 또한 상위어 추출 시 오류를 해결하기 위해 어휘 지식 베이스를 이용하였다.

본 논문의 순서는 다음과 같다. 2장에서는 관련 연구에 대해서 알아보고, 3장에서는 방법에 대해서 알아본다.

2. 관련 연구

단어 의미 중의성 해소 방안은 기본 접근 방식으로 규칙 대 규칙(rule-to-rule) 접근 방법과 독립 접근 방식으로 나눌 수 있다[1]. 첫째, 규칙 대 규칙 접근 방식은 잘못 형성된 의미 표현을 의미 분석 과정에서 함께 제거하는 방법이다. 선택제약(selectional restriction) 기반 방법이 이에 해당된다[3][4]. 둘째, 독립 접근 방식은 복합적인 의미 분석과는 독립적으로 수행하는 방법이다. 통계적 기반 WSD가 이에 해당된다[2].

통계적 기반의 WSD 방법은 학습시킬 데이터의 형태에 따라 세 가지 형태로 구분된다. 첫째, 레이블 처리된 학습 집합에 기반한 감독 중의성 해소 방법(Supervised

Disambiguation)이다[6,7]. 정보 이론(Information) 기반 중의성 해소[5]와 베이지안 분류(Bayesian Classification)가 이에 해당된다. 둘째, 학습 시 어떤 형태의 의미 태그된 데이터도 사용하지 않는 비감독 중의성 해소(Unsupervised Disambiguation)이다[8]. 셋째, 대규모의 중의성 해소를 위한 사전-기반 중의성 해소(Dictionary-Based Disambiguation)이다[9,10].

[11]에서는 문맥의 공기정보 벡터를 이용하여 한국어 명사의 의미 구분을 하였다. 사전을 이용하여 문맥의 단어들을 그들의 상위어 정보로 일반화시키고, 가중치 비교를 통해서 불필요한 정보를 제거한 후 SVD(Singular Vector Decomposition)기저벡터로 변환한다. 하지만 20 가지의 휴리스틱 패턴에 의한 상위어 추출 방법과 실험 집합의 형태와 분석에 관한 오류의 문제점이 있다.

[12]에서는 동형이의어 분별에 필요한 대량의 언어 자원 확보 및 선행 동형이의어 분별 모델의 개선하기 위해서 새로운 가중치 및 인접 어절에 대한 거리 가중치 적용 모델을 제안하였다. 가중치 값에 따라 동형이의어 분별에 영향을 주기 때문에 사용되는 정보를 정제 및 분류를 할 필요성이 있다.

[13]에서는 사전 뜻풀이의 특성을 이해하고 구문적 정보를 바탕으로 한 선택 기준을 제시하여 일반명사, 동사, 형용사, 부사를 격조사, 파생접미사, 전성어미 등과 함께 공기빈도, 어절 간 인접거리 정보 등을 WSD를 위한 의미정보에 확장해서 해결하고자 하였다. 추가적으로 사전-기반 중의성 해소 방안의 문제점인 자료 부족을 해결하기 위해서 ‘울산대학교 어휘 지능망(UOU-Word Intelligent Network : U-WIN)’ 시소스를 이용하였다.

3. WSD 해결 방법

3.1 공기정보 추출

공기정보를 추출하기 위하여 문맥의 크기를 대상 명사

를 포함하는 문장과 그 문장 전 후 2문장을 포함하였다. 이 문맥 크기 안에 있는 명사, 동사, 대상명사의 수식어, 대상명사의 지배동사를 구분하여 추출한다. 공기형태는 빈도 수를 기반으로 한 벡터로 나타냈다.

3.2 상위어 추출

[11]에서는 20여개의 패턴을 이용해서 ‘우리말 큰사전(1997)’을 통해 상위어를 추출하였다. 추출 시에 실패할 확률도 발생할 뿐만 아니라 성공하였지만 제대로 추출이 안 되는 경우가 발생하였다. 이를 해결하기 위하여 ‘울산대학교 어휘 지능망(U-WIN)’과 같은 어휘 지식 베이스를 이용하여서 오류를 최소화 하였다.

3.3 의미구분에 영향을 미치지 않는 단어 제거

추출한 단어 중에서 고유 명사나 대명사 같은 경우에는 의미 구분에 영향을 크게 주지 않기 때문에 이러한 불필요한 단어를 제거하기 위해서 TF/IDF를 사용한다.

$$T = \sum_i tf_{ij} \times \log(N/df_i)$$

T 는 변별력의 정도를 표현하는 값이며, i 라는 의미 일 때 j 라는 단어의 출현정도(tf_{ij})와 i 의 의미에서 j 라는 단어가 출현하는 샘플의 수(df_i)를 곱하여 표현한다. 하지만 지나치게 고유 명사나 대명사를 포함하는 경우나 저빈도에 해당되지만 의미 구분 중요한 정보를 처리하기 위해 정보이론의 상호정보량을 적용시켰다. 상호정보량이란 두 독립사건의 확률변수 X와 Y 사이의 의존 관계를 정량적으로 나타내는 것이다.

$$MI(x,y) = \log \frac{P(x,y)}{P(x) \times P(y)}$$

$$\approx \log \frac{Nf(x,y)}{f(x) \times f(y)}$$

여기서 $f(x)$, $f(y)$, $f(x,y)$ 는 각각 x 의 빈도, y 의 빈도이며 N 은 전체 뜻풀이 수이다.

3.4 추출된 공기정보 간의 유사도 비교

추출된 공기정보를 변별력이 있는 형태인 SVD로 변환하였다. 이는 공기 정보를 다차원 의미벡터로 합성하고 변환하는 과정에서 추출한 단어들을 구조적인 형태로 변화시킨다. 또한 차원 축소 효과를 통하여 보다 정규화되었다. 이를 이용해서 단어분포의 위치와 모양으로 표현한다.

$$sim(v,w) = \frac{\sum_{i=1}^n v_i w_i}{\sqrt{\sum_{i=1}^n v_i^2} \sqrt{\sum_{i=1}^n w_i^2}}$$

이 결과를 의미 벡터의 유사도 비교를 코사인 값으로 측정하여서 정답을 선택하였다. 여기서 i 는 훈련데이터에서 문맥을 표현하는 벡터의 축, y 는 훈련데이터에서 추출한 의미 벡터이고 w 는 평가데이터의 의미 벡터이다.

4. 결론

본 논문에서는 빈도 수에 기반한 공기정보를 이용해서

단어 의미 중의성 해결방안의 문제점을 해결하고자 하였다. 이를 위해서 정보이론의 상호정보량 도입하여 의미 구분에 중요한 정보를 추출하였다. 또한 사전에 기반한 상위어 추출 방법의 오류를 해결하기 위하여 어휘 지식 베이스를 적용하였다. 이를 통해서 WSD 문제 해결을 향상시켰다.

참고문헌

- [1] 자연언어처리, 김영택 외 공저, 2001.
- [2] 옥철영, 김준수, 옥은주, 이왕우, 이재홍, 최호섭, “한국어정보처리에서 동형이의어 중의성 해결 시스템 기술”, 정보통신부, 2002.
- [3] E. Kelly and P. Stone, “Computer Recognition of English Word Sense,” Amsterdam, The Netherlands: North-Holland, 1975.
- [4] S. F. Weiss, “Learning to disambiguate,” Information Storage and Retrieval, Vol. 9, pp.33-41, 1973
- [5] P. Brown, V. Della Pietra, S. Della Pietra and R. Mercer, “Word Sense Disambiguation using statistical methods”, Proceedings of the 29th Annual Conference of the Association for Computational Linguistics, pp. 264-270, 1991.
- [6] 허정, 옥철영, “사전의 뜻풀이에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템”, 한국정보과학회 논문지(소프트웨어 및 응용), Vol. 28, No. 9, pp. 688-698, 2001.
- [7] 이왕우, 이재홍, 이수동, 옥철영, “Bayes 정리에 기반한 개선된 동형이의어 분별 모델”, 제13회 한글 및 한국어 정보처리 학술대회, pp. 465-471, 2001.
- [8] 박성배, 장병탁, 김영택, “의미 부착이 없는 데이터로 부터의 학습을 통한 의미 중의성 해소”, 한국정보과학회 ‘2000 봄 학술 발표 논문집 B’, 제27권 1호, pp.330-332, 2000.
- [9] 송영빈, 최기선, “동사의 애매성 해소를 위한 시소리스의 이용과 한계”, 제12회 한글 및 한국어 정보처리 학술대회 발표논문, pp.255-261, 2000.
- [10] 이창기, 이근배, “의미 애매성 해소를 이용한 Word-Net 자동 매핑”, 제12회 한글 및 한국어 정보처리 학술대회 발표논문, pp.262-268, 2000.
- [11] 신사임, 이주호, 최용석, 최기선, “공기정보 벡터를 이용한 한국어 명사의 의미 구분”, 제13회 한글 및 한국어 정보처리 학술대회 발표논문집, p.472-478, 2001.
- [12] 김준수, 최호섭, 옥철영, “가중치를 이용한 통계 기반 한국어 동형이의어 분별 모델”, 정보과학회논문지(소프트웨어 및 응용), Vol. 30, No.11, pp.688-698, 2001.
- [13] 김준수, 옥철영, “정제된 의미 정보와 시소리스를 이용한 동형이의어 분별 시스템”, 정보처리학회논문지(B), Vol. 12B No.7, pp.829-840, 2005.
- [14] 김준수, 이왕우, 김창완, 옥철영, “상호정보량을 이용한 동형이의어 분별용 의미정보의 정제”, 한국정보과학회 2002 봄 학술 발표논문집(B), 제29권, 제1호, pp.460-462, 2002.