유사어를 이용한 단어 의미 중의성 해결

서희철, 이호, 백대호, 임해창
서울시 성북구 안암동 5가 1번지 고려대학교 컴퓨터학과
{hcseo, leeho, daeho, rim}@nlp.korea.ac.kr

Word Sense Disambiguation using Semantically Similar Words

Hee-Chul Seo, Ho Lee, Dae-Ho Baek, Hae-Chang Rim
Dept. of Computer Science and Engineering, Korea Univ.

요약

본 논문에서는 의미계층구조에 나타난 유사어 정보를 이용해서 단어 의미 중의성을 해결하고자 한다. 의미계층구조를 이용한 기존의 방법에서는 의미 벡터를 이용해서 단어 의미 중의성을 해결했다. 의미 벡터는 의미별 학습 자료에서 획득되는 것으로 유사어들의 공통적인 특징만을 이용하고, 유사어 개별 특징은 이용하지 않는다. 본 논문에서는 유사어 개별 특징을 이용하기 위해서 유사어 벡터를 이용해서 단어 의미 중의성을 해결한다. 유사어 벡터는 유사어별 학습 자료에서 획득되는 것으로, 유사어의 개별 정보를 가지고 있는 벡터이다.

세 개의 한국어 명사에 대한 실험 결과, 의미 벡터를 이용하는 것보다 유사어 벡터를 이용하는 경우에 평균 9.5%정도의 성능향상이 있었다.

1. 서론

단어 의미 중의성 해결은 다의어가 문장에서 이용된 의미를 찾아내는 작업이다. 예를 들어, 명사 '배'는 과일류, 교통수단, 신체일부 등의 의미를 가지는데, 문장 "나는 배를 먹었다."에서 '배'가 과일류의 의미로 이용되었다는 것을 알아내는 작업이다. 단어 의미 중의성 해결은 기계번역에서 대역어 선정에 도움을 주며[5], 정보 검색에서는 의미별 정보검색을 가능하게 한다[4].

단어 의미 중의성 해결을 위한 방법으로는 수작업으로 획득한 규칙을 이용하는 방법, 사전의 뜻풀이 말을 이용하는 방법, 의미계층구조에 나타난 유사어들을 이용하는 방법, 그리고 의미 부착된 말뭉치를 이용하는 방법이 있다[6].

본 논문에서는 의미계층구조를 이용하는 방법으로 다의어의 의미를 결정한다. 의미계층구조를 이용하는 방법은 의미계층구조에 있는 의미별 유사어들의 용례를 다의어의 의미별 학습 자료로 이용해서 학습한 후에 다의어의 의미를 결정한다. 의미계층구조를 이용해서 단어 의미 중의성 해결을 한 기존 연구로는 [1, 2, 8]이 있다. [2]는 의미계층구조로 시소러스(Roget thesaurus 1977)를 이용하며, 결정 목록(decision list)으로 의미를 결정했다. [1]은 의미계층구조로 WordNet을 이용하며, 다의어의 의미 중의성이 없는 친척단어(monosemous relatives)를 이용해서 추출한 용례를 학습 자료로 이용하고, TLC (Topical/Local Classifier)로 다의어의 의미를 결정했다. [8]은 의미계층구조로 한국어 의미계층구조[9]를 이용하며, 의미 중의성이 없는 단어만을 이용해서 학습 자료를 구한

후, 분류정보(classification information)로 다의어의 의미를 결정했다.

기존의 의미계층구조를 이용한 방법들은 의미 벡터를 이용해서 의미를 결정했다. 의미 벡터는 유사어들의 공통적인 특징만을 이용하고, 개별 특징은 고려하지 않는다. 본 논문에서는 유사어 개별 특징을 고려하는 유사어 벡터를 이용해서 의미를 결정하고자 한다.

2. 의미 벡터와 유사어 벡터

의미 벡터는 의미별로 분류된 학습 자료에서 학습한 후에 얻어지는 벡터이며, 유사어 벡터는 유사어별로 분류된 학습 자료에서 학습한 후에 얻어지는 벡터이다. 그래서 의미 벡터는 다의어의 의미수만큼 생기며, 유사어 벡터는 다의어의 유사어수만큼 생긴다.

그림 1은 의미 벡터와 유사어 벡터를 구하는 과정을 나타낸다.

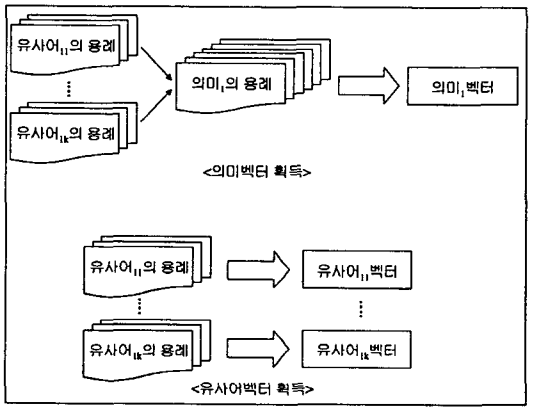

그림 1 의미 벡터와 유사어 벡터

의미 벡터는 의미별 유사어들을 모두 동일하게 취급하므로, 유사어들의 공통적인 특징만을 고려한다. 이로 인해 다의어의 의미 결정을 위한 자질값을 제대로 반영하지 못하는 경우가 있다. 예를 들어 '배'의 교통수단개념의 유사어에는 '보트', '비행기', '버스', '기차' 등이 있다. 이들을 이용해서 의미 벡터를 구한 경우, 동사 '타다'는 교통수단개념의 유사어들이 공통적으로 가지는 자질이므로 높은 자질값을 가지지만, 동사 '띄우다'는 교통수단개념의 유사어 중에서 '보트'와 '비행기'와만 공기하므로 낮은 자질값을 가진다. 즉, 동사 '띄우다'는 유사어들의 공통적인 자질이 아니므로 낮은 자질값을 가진다.

반면에 유사어 벡터를 이용하는 경우에는, '보트', '비행기', '버스', '기차'에 대한 유사어 벡터가 각각 존재하며, '보트'와 '비행기'의 유사어 벡터에서 동사 '띄우다'는 높은 자질값을 가진다. 그리고 의미별 유사어 벡터에서 동사 '띄우다'의 자질값의 합을 동사 '띄우다'의 자질값으로 이용하기 때문에 높은 자질값을 유지하게 된다.

그림 2는 의미 벡터를 이용한 의미 결정 과정을, 그림 3은 유사어 벡터를 이용한 의미 결정 과정을 나타낸다. 그림 2와 그림 3에서 사용된 다의어는 두 개의 의미를 가지며, 각 의미에 대해 두 개의 유사어를 가진다.

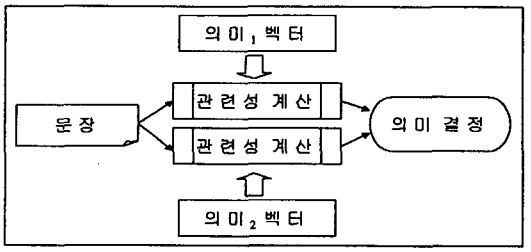

그림 2 의미 벡터를 이용한 의미 결정

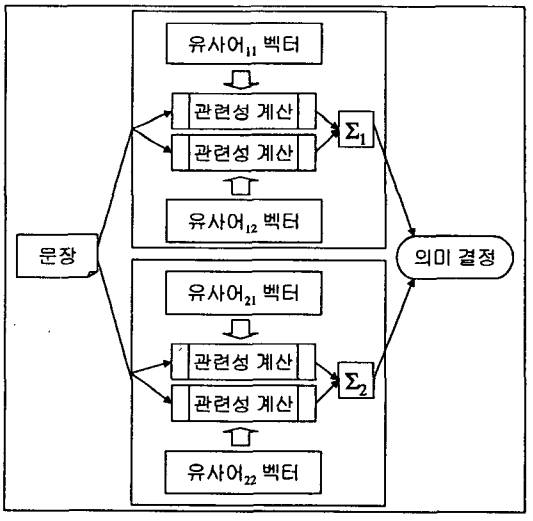

그림 3 유사어 벡터를 이용한 의미 결정

3. 실험

본 논문에서는 의미계층구조를 이용한 단어 의미 중의성 해결에서 유사어 벡터의 성능을 의미 벡터의 성능과 비교하는 실험을 한다.

3.1. 유사어 추출과 학습 자료 획득

본 논문에서는 한국어 의미계층구조[7]을 이용해서 유사어를 구한다. 의미계층구조에서 동일한 상위 개념을 가지면서 같은 레벨에 있는 단어들을 유사어로 간주하고, 유사어 중에서 의미 중의성이 없는 단어이면서 빈도가 100회 이상인 단어만을 이용한다. 그림 4는 의미계층구조에서의 유사어를 나타낸다.

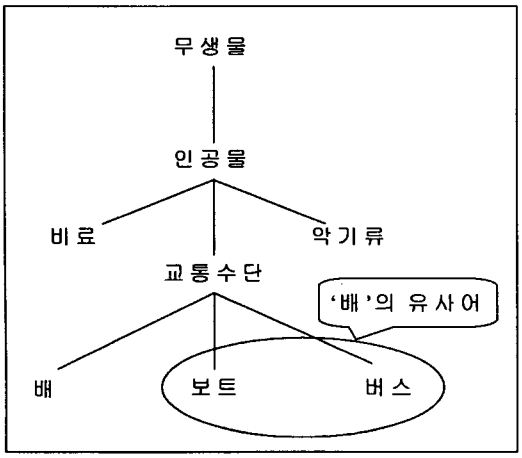

그림 4 의미계층구조에서의 유사어

품사 부착된 한국어 1000만 어절의 말뭉치에서 유사어의 용례를 추출해서 학습 자료로 이용한다.

3.2. 자질과 자질값

학습 자료에 나타난 어절과 형태소를 자질로 이용하며, 형태소의 경우에는 명사와 동사만을 자질로 이용한다.

벡터를 구성하는 자질값은 상호정보(mutual information)를 이용한다. 상호정보는 수식 (1)로 구한다[3].

$$I(x,y)=\log\frac{p(x|y)}{p(x)} \quad (1)$$

$I(x,y)$는 상호정보를, $p(x)$는 x가 발생할 확률을, $p(x|y)$는 y가 발생했을 때, x가 발생할 확률을 나타낸다. 본 논문에서 x는 자질로, y는 의미 벡터인 경우에는 의미로, 유사어 벡터인 경우에는 유사어로 이용한다.

3.3. 단어 의미 중의성 해결

본 논문에서는 수식 (2)를 이용해서 단어 의미 중의성을 해결한다.

$$\text{다의어의 의미}=\arg MAX_{\text{의미}_i} R_s(\text{문장},\text{의미}_i) \quad (2)$$

$R_s(\text{문장},\text{의미}_i)$[1]는 문장과 i번째 의미(의미_i)간의 관련성을 나타낸다. $R_s(\text{문장},\text{의미}_i)$는 의미 벡터에 대해서는 수식 (3)을, 유사어 벡터에 대해서는 수식 (4)를 이용해서 구해진다.

$$R_s(\text{문장},\text{의미}_i)=\sum_{k=1}^{n} c(f_k)\times v_i(f_k) \quad (3)$$

$$R_s(\text{문장},\text{의미}_i)=weight_i\times\sum_j R_{sw}(\text{문장},\text{유사어}_{ij}) \quad (4)$$

$$R_{sw}(\text{문장},\text{유사어}_{ij})=\sum_{k=1}^{n} c(f_k)\times v_{ij}(f_k)$$

$$weight_i=\frac{\log(|set(\text{의미}_i)\cap set(\text{상위}_i)|+1)}{\log(|set(\text{의미}_i)|+1)}$$

$$\text{유사어}_{ij}\in set(\text{의미}_i)\cap set(\text{상위}_i)$$

$$set(\text{의미}_i)=\{\text{유사어}|\text{의미}_i\text{의 유사어}\}$$

$$set(\text{상위}_i)=\{\text{유사어}|R_{sw}(\text{문장},\text{유사어})\text{값이 상위}_i\text{에 있는 유사어}\}$$

$c(f_k)$는 자질 f_k가 문장에서 나타난 횟수를 나타낸다. $v_i(f_k)$는 의미$_i$에 대한 자질 f_k의 자질값을 나타내며, 의미 벡터를 이용해서 구해진다. $R_{sw}(\text{문장},\text{유사어}_{ij})$[2]는 문장과 유사어$_{ij}$ 간의 관련성을 나타내며, 유사어$_{ij}$는 다의어의 i번째 의미의 j번째 유사어이다. $v_{ij}(f_k)$는 유사어$_{ij}$에 대한 자질 f_k의 자질값을 나타내며, 유사어 벡터를 이용해서 구해진다. $set(\text{상위}_i)$는 전체 유사어 중에서 문장에 나타난 다의어와 관련성이

1) s는 다의어의 의미를 나타낸다.
2) sw는 유사어를 나타낸다.

높은 상위 t개의 유사어들의 집합이며, *weight* 값을 정하는데 이용된다. $|set(상위_t)|$가 전체 유사어의 수와 같은 경우에는 모든 *weight*값이 1이 되므로, *weight* 값의 의미가 없다. 본 논문에서는 이런 경우를 유사어 벡터만을 이용한 것으로 간주하고, 그렇지 않은 경우를 유사어 벡터와 *weight*를 함께 이용한 것으로 간주한다.

3.4. 실험 결과

한국어 명사 '배', '밤', '고개'에 대해서 의미 중의성 해결 실험을 했다. 테스트 자료로는 수작업으로 의미 부착한 말뭉치를 이용했다. 다의어 정보와 테스트 자료에 대한 정보는 표 1에 있다.

표 1 다의어 정보와 테스트자료 정보

다의어	배	밤	고개
의미수	3	2	2
유사어수	31	14	23
테스트 문장수	3116	3856	2472
baseline	61.84%	96.52%	85.55%

'의미수'는 의미계층구조에서 구분한 다의어의 의미수, '유사어수'는 실험에 이용된 유사어수, '테스트 문장수'는 테스트를 위해 이용된 의미 부착된 문장의 수이다. 'baseline'은 가장 빈번하게 나타난 의미를 문장의 다의어 의미로 결정하는 경우의 정확률이다. 각 단어에 대해서 이용한 유사어는 표 2, 표 3, 표 4에 있다.

표 2 '배'의 유사어

의미	유사어
신체일부	가슴, 겨드랑이, 궁둥이, 어깨, 엉덩이, 옆구리 허리
교통수단	기차, 마차, 버스, 보트, 비행기, 수레, 인력거, 자동차, 자전거, 전차, 전철, 전투기, 지하철, 차, 택시, 트럭, 항공기
과일류	감, 대추, 사과, 수박, 참외, 토마토, 포도

표 3 '밤'의 유사어

의미	유사어
시간	가을, 겨울, 계절, 낮, 봄, 아침, 여름
과일류	감, 대추, 사과, 수박, 참외, 토마토, 포도

표 4 '고개'의 유사어

의미	유사어
신체	골치, 근육, 내장, 마음, 머리, 멱살, 목, 몸통, 살, 성격, 성질, 손발, 신경, 얼굴, 인격, 정신, 피부
물체(상태)	골짜기, 동굴, 벼랑, 산맥, 언덕, 절벽

의미 벡터와 유사어 벡터를 이용해서 미결정 비율에 대한 실험을 했다. 미결정은 의미 결정의 신뢰도가 낮은 경우에는 결정을 유보하는 것을 의미한다[8]. 본 논문에서는 의미 결정의 신뢰도를, 관련성(R_s(문장, 의미))이 높은 두 의미간의 관련성 비율을 이용해서 구한다. 미결정비율이 0%인 경우에는 모든 문장에 대해서 의미를 결정한 경우이다.

유사어 벡터에 대해서는 t값에 대한 실험을 했다. t값은 $set(상위_t)$에서의 t를 의미한다. t값이 크면, 많은 수의 유사어를 이용하지만, 다의어의 의미와 관련성이 낮은 유사어까지 이용하게 된다. 반면에 t값이 작으면, 적은 수의 유사어를 이용하지만, 다의어의 의미와 관련성이 높은 유사어만을 이용하게 된다.

그림 5는 미결정비율 0%에 대해서, 의미 부착을 한 결과이다. '의미벡터'는 의미 벡터를 이용해서 실험한 결과이다. '유사어벡터'는 유사어 벡터를 이용해서 실험한 결과이며, '유사어벡터+weight'는 유사어 벡터와 가장 좋은 결과를 나타낸 t로부터 얻어진 *weight*를 함께 이용한 실험의 결과이다. '이호(1999)'는 [8]에서의 실험 결과이다. '이호(1999)'는 의미계층구조를 이용해서 '고개'에 대한 실험은 하지 않았다. '의미벡터'가 baseline보다 낮은 이유

는, 기존의 다른 연구들에서 이용한 자질과는 달리 단순히 공기하는 단어만을 자질로 이용하기 때문이다. 그러므로, 자질의 성능을 향상시킨다면 '의미벡터' 뿐만 아니라, '유사어벡터'를 이용하는 경우에도 더 좋은 결과를 얻을 것으로 생각된다.

'밤'의 경우, 학습 자료로 이용되는 용례 빈도가 '밤'의 의미별 빈도를 제대로 반영하지 못할 정도로 baseline이 높고, 학습 자료에서 의미간의 비율(7.8:1)과 '밤'의 의미별 비율(27.7:1)이 너무 차이 나기 때문이다. 이로 인해, 빈도가 높은 의미가 빈도가 낮은 의미로 결정될 확률이 높게 된다.

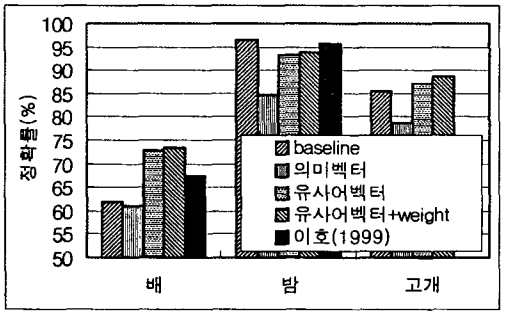

그림 5 실험 결과 (미결정비율=0%)

그림 6은 t값에 대한 실험 결과이다. t값이 커질수록 정확률도 증가하다가, 특정값 이상인 경우에는 약간 감소한다. 이는 이용한 수식의 특성상, t값이 커질수록 유사어수를 많이 가진 의미로 의미 결정될 가능성이 높기 때문이다.

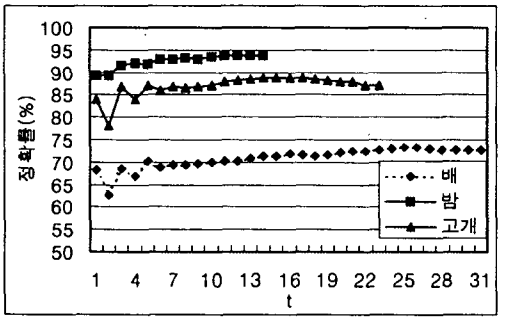

그림 6 t값에 대한 실험

그림 7, 그림 8, 그림 9는 미결정비율에 대한 실험을 한 결과이다. '밤'의 경우, '유사어벡터'와 '유사어벡터+weight'가 동일한 결과를 나타내기 때문에 '유사어벡터'만을 표시했다.

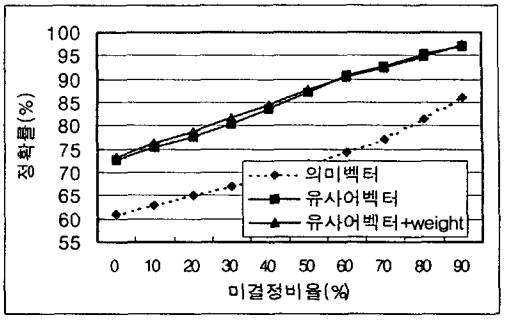

그림 7 미결정비율에 대한 '배'의 실험

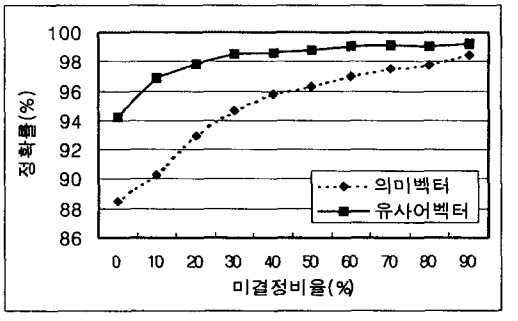

그림 8 미결정비율에 대한 '밤'의 실험

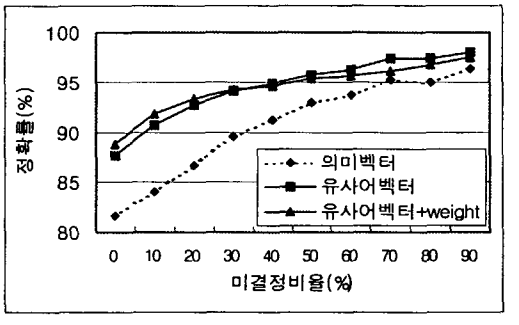

그림 9 미결정비율에 대한 '고개'의 실험

그림에서 나타나듯이, 세 단어 모두에 대해서 유사어 벡터를 이용한 경우가 의미 벡터를 이용한 경우보다 더 좋은 결과를 나타내고 있다.

6. 결론 및 향후 연구

본 논문에서는 의미계층구조에 나타난 유사어들을 이용해서 다의어의 의미를 결정했다. 기존의 의미계층구조를 이용한 방법에서 고려하지 않았던, 유사어 개별 특징을 고려해서 다의어의 의미를 결정함으로써 의미 벡터를 이용한 방법보다 평균 9.5%정도의 성능향상이 있었다.

앞으로는 자질의 성능을 향상시키는 방법에 대한 연구와 [8]에서 사전을 기반으로 동사의 유사어를 추출하는 방법을 이용해서 동사 의미 중의성 해결에 대해서도 연구를 할 계획이다.

참고 문헌

[1] Claudia Leacock, Martin Chodorow, and George A. Miller. "Using Corpus Statistics and WordNet Relations for Sense Identification.", Computational Linguistics, Vol 24, No. 1, pp.147-165, 1998.

[2] David Yarowsky, "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora", *Proceedings of the 14th International Conference on Computational Linguistics, COLING'92*, pp.454-460, Nantes, France, August 1992.

[3] Eugene Charniak, *Statistical Language Learning*, p.137 The MIT Press, Cambridge, Massachusetts London, England, 1993.

[4] Hinrich Schütze and Jan O. Pedersen, "Information Retrieval Based on Word Senses", *Fourth Annual Symposium on Document Analysis and Information Retrieval*, pp.161-175, Las Vegas NV, 1995.

[5] Hwee Tou Ng and Hian Beng Lee, "Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Examplar-Based Approach", *Proceedings on 34th Annual Meeting of ACL*, pp.88-95, 1996.

[6] Nancy Ide and Jean Véronis, "Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art", *Computational Linguistics*, Vol 24, No. 1, pp.1-40, 1998.

[7] 이호, *단어 의미 중의성 해결을 위한 분류 정보 모형*, 고려대학교 컴퓨터학과 박사학위논문, 1999.

[8] 조평옥, 옥철영, "한국어 명사 의미 계층 구조 구축", *제 9 회 한글 및 한국어 정보처리 학술대회 발표 논문집*, pp.129-135, 1997.

[9] 이수광, 조평옥, 안미정, 옥철영, 박재득, 박동인, "의미속성에 기반한 한국어 명사 의미 TAG에 관한 연구", *제 10 회 한글 및 한국어 정보처리 학술대회 발표 논문집*, pp.412-418. 1998.