# 코퍼스 기반 의미체계와 의미 별 공기정보를 이용한 비지도식 의미구분

신사임[O] 최기선
한국과학기술원 / 전문용어 언어공학 연구센터 / 언어자원은행
{miror[O], kschoi}@world.kaist.ac.kr

# Word Sense Disambiguation using corpus based sense distribution and collocation

Saim Shin[O] Key-Sun Choi
KAIST / KORTERM / BOLA

**요 약**

본 논문은 원시코퍼스에서 추출한 동음이의어의 의미 별 공기정보를 사용한 비지도식 의미구분 시스템의 구축을 제안한다. 대용량 원시코퍼스에서 추출한 의미체계를 기준으로 의미구분을 수행하였기 때문에 비현실적인 의미체계에 의한 문제점을 해결하였고, 원시코퍼스에서 추출한 공기정보로 데이터 획득비용과 부족문제를 해소하였다. 실험을 통해 의미체계의 현실화와 비지도식 훈련데이터 추출방법이 의미구분의 성능향상에 기여함을 보였다.

## 1. 서 론

문서에서 동음이의어의 정확한 의미를 결정하는 의미구분은 자연언어의 처리 및 이해에서 성능을 좌우하는 중요하고 어려운 문제이다.

의미구분 시스템의 성능을 제한하는 요소는 크게 두 가지이다. 첫째는 기준이 되는 사전의 문제이다. 의미구분에 사용하는 의미체계와 실제 코퍼스에서 사용하는 의미체계에 차이가 존재한다. 사전에 등재된 의미 중에는 실제로 코퍼스에 등장하지 않는 의미 및 구분할 필요가 없는 애매한 의미들이 존재한다. 이런 문제점들은 높은 성능의 의미구분 시스템 구축을 위한 훈련데이터의 획득을 어렵게 한다. 둘째는 훈련을 위한 정확한 의미구분 데이터 획득의 어려움이다. 훈련에 필요한 의미구분 코퍼스는 높은 구축 비용뿐만 아니라 의미편중 현상까지 더해져서 심각한 데이터 부족문제를 해결해야 한다. 이의 극복을 위해 대량의 의미구분 데이터 대신 양질의 개념체계 및 사전 같은 언어자원들을 이용하기도 한다 [1]. 그러나, 이들의 구축 비용 또한 높고 구축한 언어자원의 적용 가능한 범위에도 한계가 있기 때문에 데이터부족 문제를 완전히 해결하지 못한다.

본 논문에서는 비교적 구축 및 획득이 용이한 원시코퍼스를 자동 군집화 방법으로 의미구분 시스템에 적용하였다. 대용량 원시코퍼스에서 추출한 의미체계를 기준으로 의미구분을 수행하였으므로 의미구분 시스템에서 비현실적인 의미체계에 의해 나타나는 성능저하를 해결하였고, 원시코퍼스에서 추출한 공기정보를 훈련데이터로 사용하여 데이터 획득에 대한 비용과 부족현상을 해소하였다.

## 2. 기존 연구

2004년 SENSEVAL3[1]에 참여한 의미구분 시스템들 역시 1장에서 지적한 문제들을 포함하고 있다.

제한된 대상 동음이의어들만을 의미구분하는 Lexical Samples Task에 출전한 [2,3] 시스템은 적은 양의 훈련 데이터에서 추출한 공기정보를 기반으로 군집화 방법 및 통계와 확률계산 등으로 데이터부족 문제를 해결하였다. 이 밖에도 [4]의 시스템은 여러 기계학습방법을 적용하여 결과를 통합하는 투표방식으로 적은 양의 데이터로 최적의 결과를 얻고자 하였다. 이들 시스템들은 여전히 대상 동음이의어에 대하여 적은 양이더라도 양질의 의미구분 코퍼스를 필요로 하고 있다. 그러므로, 의미구분 대상 단어를 늘여갈수록 새로운 단어들에 대한 의미구분 데이터를 계속 공급해 주어야 시스템을 지속적으로 확장할 수 있는 단점이 있다.

문서의 가능한 모든 실질어들의 의미를 태깅하는 Allwords Task에 발표된 시스템들은 데이터부족 문제의 해결을 위해 평가데이터의 의미체계와 연계되어 있는 언어자원의 의미정보들을 여러 가지 분석으로 필요한 특징정보들을 추출하여 의미구분에 사용하고 있다 [5, 7]. 그러나, 이러한 시스템 역시 양질의 언어자원이 같은 의미체계를 평가 대상문서와 공유한다는 고비용의 작업 이전제되어 있다.

## 3. 원시 코퍼스 기반 의미구분

[1] http://www.senseval.org/

### 3.1. 의미체계 결정 및 의미 별 공기정보 추출

[6]은 대용량 코퍼스의 공기정보를 통하여 코퍼스에서 실제 사용되고 있는 의미체계를 추출하였다.

[6]에서 정의한 공기정보는 같은 문장에서 함께 등장하고 있는 모든 실질어들의 쌍이다. 원시코퍼스에서 추출한 동음이의어의 공기정보에는 대상 단어가 여러 의미로 사용되는 경우의 공기정보가 혼재되어 있다. [6]은 이와 같은 동음이의어의 공기정보를 각 공기정보의 공기정보 패턴의 유사도 비교를 수행하는 자동군집화 방법으로 군집화하였다. 같은 의미에서 사용하는 공기정보는 유사한 문맥에서 함께 등장하기 때문에, 공기정보 패턴 또한 유사하다. 결론적으로, 의미애매성을 가지고 있는 동음이의어의 공기정보를 같은 의미로 사용되는 공기정보들의 군집으로 추출할 수 있다. 이 군집의 체계는 코퍼스에서 등장하는 실용적인 의미체계를 반영한다. 본 연구에서는 [6]에서 제안한 방법으로 코퍼스에서 추출한 동음이의어의 의미체계를 의미구분에 사용하여 사전의 수동구축 비용을 줄이고, 실제 문서의 의미체계와 사전 의미체계의 차이에서 오는 성능저하를 최소화 하였다. 동음이의어의 공기정보를 통한 의미체계 결정 과정은 그림 1과 같다.

또한 의미체계 결정 과정에서 동음이의어 공기정보의 군집화 결과로 의미애매성이 해소된 군집들을 추출할 수 있다. 이 군집은 의미애매성에 유용한 의미 별 의미정보들을 포함한다. 본 연구에서는 그림 1의 군집화 과정을 통해 얻은 의미 별 공기정보 군집들을 의미구분 시스템의 훈련데이터로 사용한다. 의미 별 공기정보는 원시코퍼스에서 통계적인 군집화 방법으로 추출하였기 때문에, 별도의 의미구분 학습데이터가 필요하지 않은 비지도식 시스템이다.

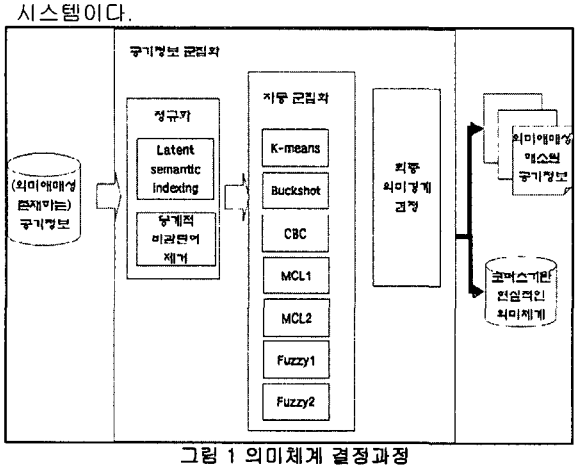


그림 1 의미체계 결정과정

### 3.2 . 의미 별 공기정보 기반 비지도식 의미구분

추출한 의미체계와 비지도식으로 구축한 의미 별 공기정보의 의미구분의 기여도를 평가하기 위해, 수작업 의미체계로 의미 구분한 기존의 의미구분 코퍼스를 제안하는

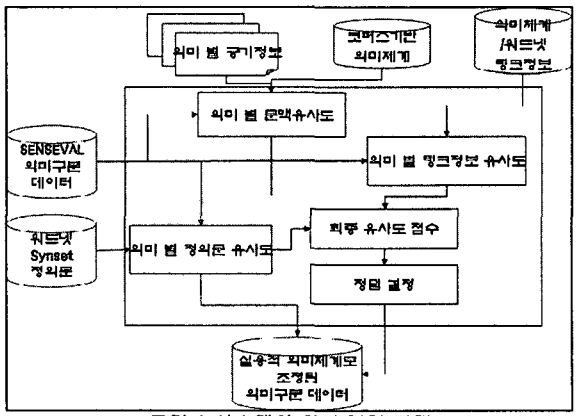


그림 2 시스템의 의미 결정 과정

코퍼스 기반 의미체계로 다시 의미 구분하여 보았다. 이 작업으로 현실적인 의미체계를 사용하여 의미구분을 수행하는 경우 의미구분의 성능향상 정도와 자동 추출한 의미 별 공기정보가 의미구분에 어느 정도 기여를 하는지 평가할 수 있다.

비지도식 의미구분에 사용한 의미정보들은 다음과 같다.

- 자동추출 의미체계와 수작업 의미체계 링크정보 (C)
  [6]은 코퍼스에서 추출한 의미체계의 각 의미가 기 구축된 의미체계의 어느 의미를 포함하는지 연결하는 작업을 수행하였다. 이 작업은 추출한 의미체계를 기 구축된 언어자원과 연결하여 의미정보를 확장하는데 유용하다. 기존의 의미체계로 태깅한 의미구분 코퍼스를 새로운 의미체계로 변환하여 사용하는 경우 사용한다.
- 수작업 의미체계 정의문 (D)
  기 구축 의미체계의 정의문 정보는 C를 통하여 의미구분에 필요한 의미정보 추출에 사용할 수 있다.
- 자동 추출한 의미 별 공기정보
  의미 별 군집에 속한 공기정보는 중요도에 따라 3 부류로 다시 세분한다. 이 과정의 중요도는 각 공기정보의 문맥백터와 군집의 중심벡터와의 코싸인 유사도로 결정하였다 [6].
  - 대표정보 (H): 각 군집을 대표하는 중심과의 유사도가 0.7 이상이고 상위 5위 안의 공기정보
  - 일반정보 (M): 전체 군집요소에서 대표정보와 저급정보를 제외한 나머지 공기정보
  - 저급정보 (L): 군집에 속하지만 중심과의 유사도가 0.1 이하인 군집의 특성을 정확히 반영하지 않는 공기정보

열거한 의미정보들은 식 (1)-(3)에 의해 통합된다. $x$는 의미구분 대상 단어이고 *corpus*는 공기정보를 추출하는 대상 코퍼스이다. 각 문맥 $c_{nx}$에서 동음이의어 $x$의 의미는 각 문맥에서 추출한 실질어 벡터와 의미 별 군집원소의 벡터와의 *score*값을 계산하여 가장 큰 *score*를 보이는 의미 $x_k$로 결정한다. *score*값의 계산은 앞서 제안한 의미정보들과 문맥벡터의 유사도를 각각의 가중치를 적용하여 합산한다

[2] http://www.cis.upenn.edu/~treebank/home.html
[3] http://www.cogsci.princeton.edu/~wn/

$$answer(c_{nx}, corpus) = \sum_{x_k \in V} \max(score(c_{nx}, C_x, corpus)) \quad (1)$$

$$score(c_{nx}, C, corpus) = \alpha \cdot score(c_{nx}, C_{x_k}, corpus) + \beta \cdot score(c_{nx}, H_{x_k}, corpus) + \chi \cdot score(c_{nx}, M_{x_k}, corpus) + \delta \cdot score(c_{nx}, L_{x_k}, corpus) + \varepsilon \cdot score(c_{nx}, D_{x_k}, corpus) \quad (2)$$

$$(\alpha + \beta + \chi + \delta + \varepsilon = 1)$$

$$score(x_k, Y, c) = \frac{|Y_{x_k} \cap c_{nx}|}{|Y_x \cap c_{nx}|} \quad (3)$$

## 4. 평가 및 토의

평가를 위하여 SENSEVAL2의 영어 Lexical Samples Task 데이터를 제안한 방법과 의미체계로 변형하여 정확도를 평가하여 보았다. 평가집합의 정답은 기존의 의미구분 결과를 본 논문에서 제안한 코퍼스 기반 의미체계로 수동으로 변환하여 사용하였다. SENSEVAL2의 평가집합은 WordNet 1.6과 연결되어 있으므로 본 논문의 평가에서 기 구축된 개념체계는 WordNet을 사용하였다. 의미구분에 사용할 군집 별 공기정보는 Penn Tree Bank[2]의 공기정보에서 추출하여 사용하였다. 이 외에도, WordNet[3]과의 연계정보와 WordNet Synset의 정의문을 의미정보로 사용하여 3장에서 설명한 방법으로 의미구분하여 결과를 비교하였다.

표 1은 각 의미정보를 사용한 의미구분의 정확도이다.

| I | M | O | L | D | IM | IO | IL |
|---|---|---|---|---|---|---|---|
| 94.04 | 99.45 | 42.68 | 100 | 65.9 | 99.83 | 99.83 | 100 |
| ID | MO | ML | MD | OL | OD | LD | … |
| 99.85 | 99.94 | 100 | 99.79 | 100 | 77.66 | 100 | |

**표 1 의미정보 조합에 따른 의미구분 정확도**

SENSEVAL에서 경쟁한 시스템들 중 영어 Lexical Samples Task의 최상위권 성능이 85% 정도인 것을 감안하면, 표 1의 결과는 전반적으로 이보다 나은 성능을 보인다. SENSEVAL에서 사용한 의미체계의 단어 당 평균 의미수가 약 3.25개 이고, 본 논문에서 사용한 의미체계의 평균 의미 수는 3.13개이다[6]. 평균 의미개수의 차이가 크지 않음에도 성능이 향상된 주원인은 의미체계 조정에 의해 의미체계와 코퍼스의 의미와의 차이가 줄어서 코퍼스에서 추출한 데이터를 정확하게 사용하여 의미구분에 적용할 수 있었기 때문이다. 이 결과를 통해 코퍼스의 의미를 반영하도록 사전의미를 재조정 하는 것만으로도 의미구분 시스템의 성능향상을 꾀할 수 있다는 것을 알 수 있다.

또한, 원시코퍼스에서 군집화 방법으로 추출한 공기정보를 적용하여 의미구분 과정의 기여도를 평가하였다. 대표정보 I만을 이용하여 의미구분을 하여도 94%의 성능을 보이는데, 이 결과는 정답결정에 사용한 공기정보의 크기가 매우 작은데도 불구하고 높은 성능을 보이고 있다. 즉, 비지도식으로 추출한 대표정보가 의미구분 과정에 중요한 원소들을 정확하게 추출하였음을 알 수 있다. 일반정보 M만을 사용한 경우가 I보다 좋은 성능을 보이는 이유는 I는 각 단어 당 5개의 상위 정보만을 사용한 반면, M에서 적용하는 공기정보는 훨씬 크기 때문에 공기정보 데이터의 문맥과의 일치율이 높기 때문이다. 낮은 정확도를 보이는 D와 O는 의미 별 공기정보와 결합한 ID/MO/MD 의 결과에서 성능이 보완되는 것을 볼 수 있다. 그러므로, 의미 별 공기정보의 적용이 의미구분의 성능향상에 기여한다는 것을 알 수 있다. 또한, 기존 시스템들이 많이 사용하는 정의문 정보 D의 경우도 의미 별 공기정보를 통하여 더 정확하게 사용할 수 있었다..

기 구축 의미체계와 연계한 L의 결과는 따로 수작업이 거의 필요 없는 좋은 결과를 보여준다. 또한, L의 추출은 [6]에서 자동으로 이루어지는 결과에서 획득한다. 이 실험결과를 통해 새로운 의미체계를 위한 의미데이터 구축을 위해 기존의 의미구분 데이터의 의미체계를 조정하여 사용하는 것도 저비용의 정확한 데이터 구축 방법임을 도출할 수 있다.

## 5. 결론 및 향후 연구

본 논문은 원시코퍼스에서 자동 군집화 방법을 통해 추출한 동음이의어의 의미 별 공기정보로 비지도식 의미구분 시스템의 구축을 제안한다. 또한, 의미체계와 코퍼스의 의미와의 차이조정이 의미구분의 성능향상에 기여한다는 것을 실험으로 증명하였다.

향후 연구는 온톨로지 의미정보를 기반으로 하는 기존의 모든 단어 의미구분 시스템에 의미 별 공기정보를 적용하여 모든 단어 의미구분 시스템을 개발하는 것이다. 온톨로지 기반 의미구분 방법은 의미 별 공기정보 방법의 적용률을 올리고, 온톨로지 기반 의미구분 시스템의 정확률을 의미 별 공기정보를 통하여 정확률을 향상시킴으로 상호 보완적으로 의미구분 시스템의 성능 향상이 가능할 것이다.

## 참고문헌


[1] Saim Shin, Juho Lee, Yongsoek Choi, Key-Sun Choi. 2001. "Word Sense Disambiguation Using Vectors of Co-occurrence Information", Natural Language Processing Pacific Rim Symposium.

[2] Peter D.Turney. 2004. "Word Sense Disambiguation by Web Mining for Word Co-occurrence Probabilities", ACL SENSEVAL3 Workshop

[3] Tom O'hara, Robecca Bruce. 2004. "Class-based Collocations for Word-Sense Disambiguation", ACL SENSEVAL3 Workshop

[4] Eneko Agirre, David Martinez. 2004. "The Basque Country University system: English and Basque tasks", ACL SENSEVAL3 Workshop

[5] L. Villarejo, L. M'arquez, E. Agirre, D. Mart'ınez, B. Magnini, C. Strapparava, D. McCarthy, A. Montoyo, A. Su'arez. 2004. "The "Mining" System on the English Allwords Task", ACL SENSEVAL3 Workshop

[6] Saim Shin, Key-Sun Choi. 2004. "Automatic clustering of collocation for detecting practical sense boundary", ACL

[7] 신사임, 이주호, 배희숙, 김장희, 최기선. 2004. "대용량 코퍼스의 의미정보 획득을 위한 시소러스의 기반 의미 구분 시스템", 한국 인지과학회 학술대회