주제어의 중의성 해소를 위한 Naïve Bayes 분류기 적용에 관한 연구

Application of a Naïve Bayes Classifier for Topic Word Sense Disambiguation

유현숙, 정영미(연세대학교 문헌정보학과)
Yu Hyeon-Sook, Chung Young-Mee
Dept. of Library and Information Science, Graduate School of Yonsei University

단어의 의미 중의성을 해소하는 것은 자연언어처리의 중요한 문제 중의 하나이다. 특히 문서의 주제어가 중의성을 가질 때, 이 문서는 부적합한 범주에 속하게 되어 정보검색시 잡음을 일으키는 원인이 되기도 한다. 그러므로, 본 논문에서는 문서를 대표하는 주제어의 의미 중의성을 해소하기 위해 주변 문맥자질을 고려하는 방법을 모색한다. 이를 위해 자연언어처리의 통계적 방법으로 문서 범주화에 많이 사용되는 Naïve Bayes 분류기를 중의성 해소에 적용하고, 그 결과 얻어진 중의성 해소 성능을 평가한다.

1. 서론

문서는 그 내용을 대표하는 주제어로 표현될 수 있다. 그러나, 주제어가 중의성을 가지는 경우, 이 문서가 속하게 될 범주의 결정은 단일 의미의 주제어를 가지는 문서들의 경우보다 복잡하게 된다.

그렇지만, 문서 범주화와 관련된 대부분의 기존 연구방법들은 실험집단 문서들을 단지 불용어를 제거한 단어들로 표현할 뿐, 색인어 특히 주제어 자체가 지니는 의미적 중의성은 고려하지 않았다.

따라서, 본 논문에서는 문서를 대표하는 주제어의 의미 중의성을 해소하기 위해, 자연언어처리의 통계적 방법을 적용한 실험을 수행하고 그 결과를 분석하였다. 이 실험에서는 통계적 연상추론을 통해 전체 문맥을 고려하는 Naïve Bayes 분류기가 중의성 해소에 사용되었으며, 하나의 문서를 문맥단위로 보고 5개 주제어에 대한 중의성 해소와 그에 따른 문서 범주화가 이루어졌다.

2. 중의성 해소

중의성 해소의 문제는 자연언어처리에서 매우 중요한 사항이다. 실제로, 자연어를 사용하는 정보검색 시스템이 여러 성능평가 실험에서 비교적 검색효율이 낮은 것으로 밝혀졌는데, 이는 질의와 문서의 구문 및 단어 중의성이 너무 많은 잡음을 생성해 내기 때문이다(Smeaton 1999, 101).

단어의 의미 중의성 해소란 다의적인 단어의 특정 용법 중 어떤 의미가 문맥내에서 사용되었는가를 결정하는 작업이다. 예를 들어, 명사 '배'는 과일, 교통수단, 신체일부 등의 의미를 가지는데, 문장 "나는 배를 먹었다."에서 '배'가 과일류의 의미로 사용되었음을 알아내는 것이다(서희철 1999).

효과적인 단어의 중의성 해소는 정보검색시 의미별 정보검색을 가능하게 하고, 기계번역에서는 대역어 선정에 도움을 주며, 의미태깅, 담화분석과 같은 여러 자연언어처리 응용 프로그램에서 활용가능하다. 그러므로, 문서를 대표하는 주제어의 중의성이 제거되어 이용자가 원하는 의미만

을 문서의 범주로 할당하는 중의성 해소는 정보 검색시스템의 효율을 높이는데 기여할 것이다.

단어의 의미 중의성을 해소하는 여러 방법 중에서도, 연상에 기반한 중의성 해소는 순수히 공기빈도 분석만을 통하여 단어의 의미 연상을 발견하는 것으로 일정한 문맥범위 내에서 적합한 의미를 찾는 과정이다.

예를 들어, "table"의 경우, 학습단계에서 일정한 문맥창을 설정하여 "table"이 "책상"이라는 의미로 사용되는 문맥에서 공기한 단어 출현 수와 "표"라는 의미로 사용되었을 때 각각의 공기 단어가 두 문맥에서 나타나는 조건 확률을 구한다. 여기에 새롭게 "table"을 포함한 문장이 주어졌을 때, "책상"과 "표" 2개의 의미에 대해 문맥내 각 공기단어의 조건확률을 계산하고, 그 값이 큰 쪽의 의미를 취하는 것이다(황도삼 외. 1999). 이 원리는 본 연구에서 Naïve Bayes 분류기에 적용되어 하나의 문서를 문맥범위로 한 문서 주제어의 중의성 해소 실험에 사용된다.

3. 중의성 해소 실험

3.1. 실험집단 및 중의어 선정

중의성 해소의 성질은 단어 의미 중의성 해소 시스템을 학습시키는 데 어떤 자료를 활용하였는가에 따라 어느 정도 민감하게 변화한다. 본 논문에서는 범주가 부여된 텍스트 말뭉치인 통계 학습집단을 중의성 해소에 사용하였다. 실험에 사용된 문서들은 연세대학교 문헌정보학과에서 구축한 KFCM-CL 중 1992년 7월(1일부터 4일까지)에 게재된 조선일보, 동아일보, 한국일보의 국제 및 경제분야 기사 340건이다. 이 기사들을 대표하는 주제어 중 중의성을 띄는 단어 5개를 선정하고 각 주제어의 중의적 의미가 속하는 범주별로 기사들을 수집한 결과, 180건의 문서가 전체 실험집단으로 구성되었다. 이를 학습집단과 검증집단의 비율이 3:1이 되도록, 학습집단 137문서, 검증집단 43문서로 구분하였다.

실험 대상으로 선정된 중의어들은 사전 분류된 전체 KFCM-CL 340건 기사내에서 여러 범주에 출현한 주제어들로 정의되었으며, 이와 같이 선정된 중의어들은 '**교육**', '**기업**', '**북한**', '**자동차**', '**자원**'의 명사 5개이다. 실험집단에 할당된 범주는 『기사자료표준분류표』('92년도판)의 대분류 단위로만 제한하였으며, 5개 중의어들은 각각 둘 또는 세 범주에만 속할 수 있는 것으로 한정하였다.

<표 1>은 본 논문의 실험에 선정된 중의어 및 실험집단에 대한 분석 결과이다.

중의어	선정된 범주 수	실험문서 수		문맥 자질 수	
		학습	검증	학습	검증
교육	2	12	4	2128	752
기업	3	45	15	3579	1995
북한	2	24	8	2816	2072
자동차	2	28	8	2729	832
자원	2	28	8	2987	1328
합계 (평균)	5범주 사용됨	137 문서	43 문서	(2848)	(1396)

<표 1> 중의어 및 실험집단 분석 결과

3.2. 중의성 해소를 위한 Naïve Bayes 분류기 적용

단어 의미 중의성 해소를 위한 분류 알고리즘은 사용되는 학습집단에 따라 통계 학습알고리즘과 비통계 학습알고리즘으로 나뉜다. 통계 학습알고리즘은 학습을 시킬 각 중의어의 정확한 문맥상 의미정보가 사전 부여된 학습집단이라는 것을 가정한다.

본 논문에서는 여러 통계 학습알고리즘 중, 통계적 자연언어처리 접근방법으로 Gale et al.(1992)의 Naïve Bayes 분류기를 사용하였다. Naïve Bayes 분류기의 개념은, 방대한 문맥내에서 중의적인 단어 주위의 내용어들만을 살펴보는 단어집합(*bag of words*) 모델로 공기빈도 정보를 이용한 통계적인 연상추론을 내재하고 있다. 각 내용어는 중의적인 단어의 어떤 의미가 문맥에서 사용되었는가에 대한 유용한 정보를 잠재적으로 제공하는 것으로 간주된다. 그러므로, Naïve Bayes 분류기는 특별한 자질선정을 하지 않는 대신, 모든 자질들로부터 증거(evidence)를 결합한

다. 이같은 효율성과 다수의 자질 증거결합 능력으로 인해, Naïve Bayes 분류기는 기계학습에 많이 사용되고 있다.

문맥자질을 고려하여 단어의 의미적 중의성을 해소하는 Naïve Bayes 결정규칙은 다음과 같다 (Manning and Schütze 1999, 237).:

Decide s'if s'

$= \text{argmax}_{sk}[\log P(s_K) + \Sigma_{v_j \text{ in } c} P(v_j \mid s_k)]$

이는 문서의 주제어로 표현되는 중의어를 w라 할때, 출현한 문서 c_i를 하나의 문맥범위로 하여 중의성을 해소한 다음 적합한 범주 s_k에 할당하기 위해 c_i의 문맥자질 v_j를 살펴보는 것으로, 자세한 Naïve Bayes 분류기는 <그림 2>에 제시되어 있다.

```
1 comment: Training
2 for all senses sk of w do
3  for all words vj in the vocabulary do
4        P(vj | sk) = C(vj, sk) / C(sk)
5  end
6 end
7 for all senses sk of w do
8        P(sk) = C(sk) / C(w)
9 end
10 comment: Disambiguation
11 for all senses sk of w do
12   score(sk) = logP(sk)
13 for all words vj in the context window c do
14   score(sk) = score(sk) + ΣlogP(vj | sk)
15 end
16 end
17 choose arg max_sk score(sk)
```

<그림 2> 단어 의미 중의성 해소에 사용된 Naïve Bayes 분류기

<그림 2>에서, $C(v_j,s_k)$는 중의어 w가 학습집단에서 의미 s_k로 문맥자질 v_j와 함께 출현한 수이고, $C(s_k)$는 학습집단에서 중의어 w가 의미 s_k로 출현한 수, $C(w)$는 중의어 w의 출현 총수이다. 범주기호가 부여된 학습집단으로부터 $P(v_j \mid s_k)$와 $P(s_k)$확률이 구해지고, 이 확률은 스무딩(smoothing)이 함께 적용되어 검증집단 문서 주제어의 중의성을 해소한 의미범주 할당에 사용된다.

4. 중의성 해소 실험결과 평가 및 분석

주제어의 의미 중의성을 제거한 중의성 해소 실험결과 평가를 위해, 평균 정확도(accuracy), 평균 정확률(precision), 평균 재현율(recall) 척도를 사용하였다. 5개 중의어에 대한 검증집단의 중의성 해소 실험결과 평가는 <표 2>에 제시되어 있다.

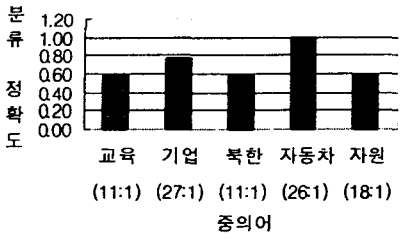
중의어	문서범주	평균 정확도	평균 정확률	평균 재현율
교육	정치(100)	0.58	0.5	0.25
	문화(600)			
기업	경제(200)	0.77	0.93	0.48
	산업(300)			
	국제(900)			
북한	정치(100)	0.58	0.5	0.25
	국제(900)			
자동차	경제(200)	1	1	1
	산업(300)			
자원	경제(200)	0.58	0.5	0.25
	산업(300)			
평균		0.7	0.57	0.45

<표 2> 5개 중의어의 중의성 해소 실험 결과 평가

본 실험에서, Naïve Bayes 분류기를 사용하여 문서 주제어의 의미 중의성을 해소한 결과는 특정한 양상을 보이는 것으로 나타났다.

실험문헌집단 선정시 학습집단과 검증집단의 비율이 3:1 정도가 되도록 구성하였지만, 각 기사의 실제 길이가 매우 상이하고 사용되는 자질 수가 다양하여 실험의 일반화가 어려웠다. 따라서 학습집단과 검증집단의 자질 수 비율에 따른 중의성 해소 성능에 차이가 있는가를 살펴보는 것이 의미있을 것이다.

<그림 3>을 보면, 학습집단과 검증집단의 자질 수 비율이 26:1에서 27:1 사이의 중의어 '**기업**', '**자동차**'의 범주화 성능이 비교적 높음을 알 수 있다. 이보다 더 낮은 자질 수 비율은 중의성 해소에 요구되는 자질을 결여한 것으로 보여진다.



<그림 3> 학습집단과 검증집단의 자질 수 비율에 따른 Naïve Bayes 분류기의 중의성 해소 정확도 성능 비교

마지막으로, 지금까지의 실험결과 분석을 기반으로 단어의 의미 중의성 해소에 적용된 Naïve Bayes 분류기의 성능 향상을 위한 몇 가지 향후 연구과제를 생각해 보고자 한다.

① 추출되는 공기어에서 잡음을 일으키는 단어의 영향을 가능한 줄이는 방법을 적용할 수 있다. 이를 위해 범주의 각 공기어에 대해 가중치를 부여한다.

② Naïve Bayes 분류기를 일반화시킨 로그선형 모델 접근방법을 적용할 수 있을 것이다. 로그선형 모델은 모든 자질들을 독립적으로 학습시키는 대신, 상호 의존적인 자질끼리 묶어 하위집합을 구성하는 것이다.

③ Naïve Bayes 분류기에 시소러스-기반 방법을 함께 사용하는 접근방법을 들 수 있다. Yarowsky(1992) 알고리즘의 경우, 단어들이 만약 시소러스 범주 t_i에 속하는 문맥에서 보다 자주 출현한다면 이 단어들을 범주 t_i에 첨가한다.

④ 통계적인 단어 의미 중의성 해소에 언어학적인 접근방법을 첨가하는 것으로, 연상관계에 기반하는 Naïve Bayes 분류기 방법에 구문정보나 선택제한 정보를 통합하여 사용할 수 있다.

5. 결론

단어 의미 중의성 해소는 중의적인 단어에 그 문맥에서 사용된 최적의 의미 범주를 할당하는 것이 궁극적인 목적이라 할 수 있다. 따라서, 본 논문에서는 하나의 문서를 문맥범위로 주제어의 의미 중의성을 제거하기 위해, 문서 범주화에 많이 사용되는 Naïve Bayes 분류기를 중의성 해소 실험에 적용해 보았다. 그 결과 70%의 평균 정확도 및 67%의 평균 정확률과 45%평균 재현율을 얻었다.

Naïve Bayes 분류기는 적용이 용이하므로 매우 효과적인 알고리즘이다. 그러므로, 향후 연구를 통해 이 분류기의 중의성 해소 성능을 보다 향상시킬 수 있다면, 이는 앞으로 정보검색시 질의어 및 검색되는 문서의 잡음을 줄이고 이용자의 검색결과 만족도를 증진시키는 기초가 될 것이다.

참고문헌

서희철, 이호, 백대호, 임해창. 1999. "유사어를 이용한 단어 의미 중의성 해결." **제11회 한글 및 한국어 정보처리 학술대회, 인간과 기계와 언어**. pp.304-309.

황도삼, 최기선, 김태석 공역. Makoto Nagao 저. 1999. **자연언어이해**. 서울:홍릉과학출판사.

Gale, W.A., Church, K.W. and Yarowcky, D. 1992. "A Method for Disambiguating Word Senses in a Large Corpus." *Computers and the Humanities* 26: 415-439.

Manning, C. D. and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Ch.7: Word Sense Disambiguation.

Smeaton, A.F. 1999. Using NLP of NLP Resouces for Information Retrieval Tasks. (In) Strzalkowski, T. (ed) 1999. *Natural Language Information Retrieval*. Kluwer Academic Publishers. pp.99-112.

Yarowsky, D. 1992. Word-sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Copora. In *COLING* 12: 454-460.