나이브 베이즈 분류기를 이용한 의미제약이 강화된 한국어 복합명사 의미 분석

이용훈○, 옥철영
울산대학교 전기공학부
yhsoft12@gmail.com, okcy@ulsan.ac.kr

A Semantic Analysis of Korean Compound Nouns with Enforced Semantic Constraints using a Naïve Bayes Classifier

Yong-Hoon Lee○, Cheol-Young Ock
Dept. of Electrical Engineering, University of Ulsan

요 약

본 논문에서는 사전 원어정보를 이용한 기존 방법에 나이브 베이즈 분류기를 추가로 이용하는 의미제약 기술에 대하여 소개한다. 의미제약은 의미 분석의 전처리 단계로서 부분적으로 중의성을 해소하여 입력된 복합명사의 분석 정확도 뿐만 아니라 전체적인 분석시간의 단축에도 큰 도움을 준다. 나이브 베이즈 분류기를 이용하는 방법은 사전의 의존성으로 인해 제약할 수 없는 2-gram을 대상으로 제약을 시도한다. 분류기를 위한 학습데이터는 의미 태깅된 기분석 2-gram사전을 이용하여 U-WIN의 관계정보와 사전 그리고 패턴들에 의해 생성된다. 원어정보로 해결하지 못하는 34.63%의 2-gram중 2.83%에 대해 추가로 제약에 성공 하였다.

주제어: 복합명사, 의미분석, 나이브 베이즈 분류기, 의미제약

1. 서 론

복합명사 의미 분석은 기계번역, 정보검색 등 자연어처리가 필요한 전반적인 분야에서 중의성 해소를 위한 필수적인 과정으로 그 중요성이 갈수록 증대되고 있다. 복합명사 의미 분석 방법 중 지식 기반(Knowledge-driven) 방식은 시소러스와 같은 언어 자원을 이용해 얻은 의미 단위로 유사도를 구하게 된다. 이 때 분해된 구성명사가 사전에 등재된 동형이의어나 다의어의 개수가 많을수록 그 유사도 비교 대상의 수가 많아지며 결과적으로 계산시간이 급격히 증가하며 의미 분석의 정확도도 떨어지게 된다.

이용훈[1]은 이러한 문제를 해결하기 위해 한국어 명사의 대략 60% 가량이 한자어를 포함한 외래어에서 유래 된 특성에 따라 원어정보를 이용하여 그 의미 영역을 축소시키는 방법을 제안하였다. 2개의 구성명사로 이루어진 복합명사는 사전에 하나의 표제어로 등재된 경우가 많으므로 제약 단위로 2-gram을 이용하였으며 합성된 2-gram이 사전에 등재되어 있고 원어정보를 포함하는 경우를 대상으로 수행하였다. 이는 정확히 의미를 제약할 수 있다는 장점이 있지만, 사전에 의존적인 여러 문제들로 인해 그 처리범위가 크지 못하다는 단점이 있다.

허정[2]은 어휘 의미 중의성의 정확률을 향상시키고, 연산 시간을 줄여 시스템의 효율성을 극대화하기 위해 복합명사 의미사전을 이용하였다. 복합명사를 구성하는 단일명사들이 서로 의미적 제약을 한다는 규칙에 따라 의미제약 관계를 이용하여 구축하였다. 한번 구축하면 여러 분야에 적용할 수 있다는 장점이 있으나, 수작업에 의해 구축되므로 비용이 많이 든다는 단점이 있다.

본 논문에서는 이러한 의미 제약의 범위를 확대시키기 위하여 지도학습 알고리즘의 하나인 나이브 베이즈 분류기(Naïve Bayes Classifier)를 추가로 적용해 정확률의 변화를 관찰 하였다. 본 논문은 2장에서 나이브 베이즈 분류기에 대해, 3장에서 구성명사 의미범위제약에 대해 기술한다. 4장에서는 이를 적용한 실험과 결과에 대한 분석을 수행하고, 5장에서는 결론과 향후 연구에 대해서 기술한다.

2. 나이브 베이즈 분류기

나이브 베이즈 분류기는 베이즈 정리에 기초하고 속성들 간의 독립성을 가정한 확률적인 모델이다. 매우 단순하지만 잘 알려진 전통적인 분류방법으로, 자연언어처리 분야에서 널리 사용되어 왔다. 이는 통계적인 알고리즘으로 학습문서의 통계 정보를 학습하고 이렇게 얻은 통계정보를 이용하여 입력 스트림으로부터 대상을 분류한

다[3]. 속성의 값으로 구성되는 데이터 (a_1, a_2, ..., a_n)가 주어졌을 때 분류기의 답에 해당하는 클래스 C_*는 식 (1)과 같이 가장 확률이 높은 부류로 결정 된다.

$$c_* = \arg max_{c_i \in C} P(C_i | a_1, a_2, \ldots, a_n) \quad (1)$$

식 (1)에 베이즈 정리(Bayes' Theorem)와 강한(naive) 독립 가정을 적용 하면 부류 C_{NB}를 결정 하는 확률 모델 나이브 베이즈 분류기는 식 (2)와 같이 정의된다.

$$\begin{aligned} C_{NB} &= \arg max_{c_i \in C} \frac{P(a_1, a_2, \ldots, a_n | c_i) P(c_i)}{P(a_1, a_2, \ldots, a_n)} \\ &= \arg max_{c_i \in C} P(a_1, a_2, \ldots, a_n | c_i) P(c_i) \\ &= \arg max_{c_i \in C} P(c_i) \prod_j P(a_j | c_i) \end{aligned} \quad (2)$$

3. 구성명사 의미범위제약

본 논문에서의 의미범위제약은 유사도 비교의 전처리 과정으로 유용한 자원을 이용하여 가능한 유효 후보들로 그 범위를 축소시키는 것을 의미한다. 이는 의미 조합의 복잡도와 유사도 비교에 소요되는 계산시간을 줄이기 위하여 중요한 과정이며 이에따라 정확도 역시 영향을 받게 된다.

3.1 사전 원어정보를 이용한 의미제약

이전 연구에서는 사전의 원어정보를 이용하여 유효 명사들로 그 후보를 고정하였다[1]. 이는 다음과 같은 조건을 만족하는 경우 대부분의 구성명사들이 올바르게 제약 될 수 있었다.

- 분해된 구성명사 2-gram의 사전 등재 여부
- 등재시, 원어 정보의 존재 여부
- 각 구성명사의 해당 원어정보를 가진 동형이의어 혹은 다의어 존재 여부

[그림 1]은 복합명사 '한국정당정치연구'의 의미제약 과정을 나타내며 [표 1]은 이에 따른 연산 횟수의 변화를 나타낸다.

표 1. 의미 범위 축소에 따른 연산 횟수 변화

	2-gram 유사도 비교	의미조합
의미제약 전	503	6,552
의미제약 후	107	72

하지만 이러한 과정은 원어정보를 이용해 의미 범위를 정확히 제약할 수 있다는 장점이 있으나 사전에 의존적이므로 다음과 같이 제약에 실패하는 경우들이 있다.

- 2-gram의 미등재
- 원어 정보의 부재
- 한자 코드의 차이
- 해당 구성명사와 2-gram의 원어 정보 표기 차이

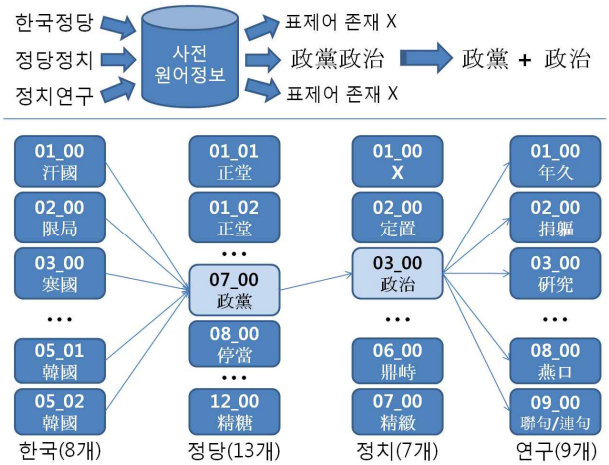

그림 1. 사전 원어정보에 의한 구성명사 의미제약

특히, 세번째 경우는 원어 정보가 한자(漢字)일 때 1음절의 한자가 2개 이상의 다른 코드로 기술된 경우가 다소 있어 합성된 2-gram의 원어정보를 가진 구성명사가 있음에도 불구하고 코드 비교시 다른 한자로 인식되어 제약에 실패 하였다. 예를들면 복합명사 '누적기록(累積記錄)'의 경우 사전에 등재되어 있으며 원어정보도 존재한다. 구성명사 누적의 경우 2개의 동형이의어가 존재하며 그 중, '누적_01(累積)'의 원어정보가 2-gram 누적기록의 그것과 같으므로 '누적_01'로 의미 제약이 되어야 하지만 사전 편찬시 다른 코드의 '누(累)'를 사용하였으므로 제약에 실패 하였다.

표 2. 표제어 '루이스'의 동형이의어별 원어정보

표제어	원어정보
루이스_01	Lewis, Matthew Gregory
루이스_02	Lewis, Gilbert Newton
루이스_03	Lewis, John Llewellyn
루이스_04	Lewis, Percy Wyndham
루이스_05	Lewis, Clarence Irving
루이스_06	Lewis, Harry Sinclair
루이스_07	Lewis, Clive Staples
루이스_08	Louis, Joe

또한 네번째 경우는 특히 라틴어 계열(영어권 언어)에서

발생하는 문제로 2-gram의 구성명사 원어정보가 약어로 기술되는 경우가 많아 비교할 수 없는 경우도 많이 있었다. 예를들면 복합명사 '루이스산'의 경우 원어정보는 'Lewis酸' 이며 구성명사 '루이스'는 사전에 8개의 동형이의어가 등재되어있다. 이중 'Lewis'만을 이용해 의미제약을 시도하면 [표 2]와 같이 일치하는 원어정보가 없으므로 제약을 할 수 없게 된다. 이처럼 원어 정보를 이용한 의미제약 방법은 사전에 전적으로 의존적이다는 단점이 있다.

3.2 나이브 베이즈 분류기를 이용한 의미제약

위의 문제점들은 원어정보를 이용한 의미 제약 방법이 사전에 의존적이기 때문에 생기는 문제들로써 이를 완화시키기 위해 추가적으로 나이브 베이즈 분류기를 이용한다.

3.2.1 학습데이터 생성

학습을 위해 학습데이터를 생성하는 과정은 [그림 2]와 같다. 첫째로, 분류를 위한 클래스로 의미 태깅된 2-gram명사 기분석 사전을 이용하여 중심 명사들을 추출한다. 여기서 중심명사는 가장 끝 어절의 명사를 의미한다.

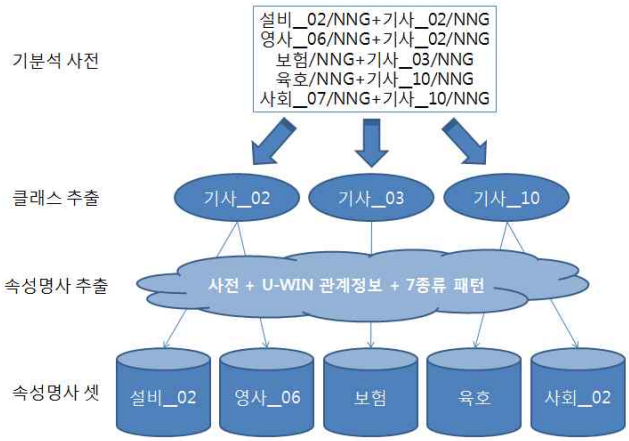

그림 2. 학습데이터 생성과정

둘째로, 입력된 인스턴스에 대한 확률적 비교를 위해 중심명사에 대한 각 공기명사의 속성명사 세트를 구성한다. 이를테면 중심명사 '기사_02'에 대한 공기명사 '설비_02'와 '영사_06'은 각각 하나씩의 관측 데이터 집합으로 구성되어 진다. 이를 위해 U-WIN의 관계정보와 사전정보들을 이용하며 [1]에서 데이터 부족 현상(Data Sparseness Problem)을 해소하기 위해 정의한 다음의 7가지 규칙과 패턴을 적용해 추출한다.

- 표제어의 뜻풀이
- 1차 하위어들의 뜻풀이
- 최상위어까지 존재하는 모든 상위어들의 뜻풀이
- 표제어의 동의어 관계인 표제어의 뜻풀이
- 표제어의 뜻풀이에서 추출된 명사류의 뜻풀이
- 표제어의 뜻풀이가 '~이르는(던) 말' 류 인 경우 그 대상명사(들)의 뜻풀이
- 표제어의 뜻풀이가 '~의 방언', '~의 잘못', '~의 옛말', '~을(를) 우리 한자음으로 읽은 이름', '~(으)로 순화', '~의 음역어' 인 경우 이 대상명사의 뜻풀이

이를 이용해 속성명사 세트를 만들고 [그림 3]과 같이 학습데이터로 이용한다. [그림 3]에서 기분석 사전에 의해 중심명사로 쓰인 '기사__02', '기사__10'이 클래스로 정의 됐으며 클래스 옆의 다의어들과 숫자는 7가지 규칙에 의해 추출된 속성 명사들과 그 빈도를 의미한다.

```
기사__02/NNG    데_01_00_02/1 예전_01_00_00/1 말_01_00_04/4 얼굴_01_00_01/1 물질적_00_00_00/1 몸_01_00_01/2 형체_00_00_00/1 문장_02_00_03/1 대상_11_00_01/1 여자_02_00_01/1 물건_00_00_01/3 구_05_00_01/1 때_01_00_05/1 단어_00_00_00/1
기사__02/NNG    활동_02_00_02/1 활동_02_00_01/4 현상_04_00_01/1 짓_01_00_00/1 일부분_00_00_00/1 의지_06_00_01/1 이상_12_00_01/1 병_04_00_01/4 동안_01_00_01/2 몸_01_00_01/2 성과_01_00_00/1 피부_02_00_00/1 기계_07_00_01/1 의료_02_00_00/1 대상_11_00_01/2 가지_04_00_01/1 기술_01_00_01/2 곳_01_00_01/1 정상적_00_00_00/1 일_01_00_10/1 시간_04_00_01/2 괴로움_00_00_00/1 머리_01_00_02/2 전신_01_00_00/1 조직_00_00_05/2 사람_00_00_07/1 부분_01_00_00/1 개체_02_00_01/2 생물체_00_00_00/1 기타_01_00_00/1 장소_05_00_00/2 일_01_00_01/5
기사__02/NNG    일_01_00_05/1 관련_00_00_00/3 시간_04_00_01/1 말_01_00_04/1 앞_00_00_02/1 현상_04_00_01/1 사물_10_00_02/1 여럿_00_00_00/1 차례_01_00_01/1 이상_05_00_01/1 기준_03_00_00/1 모임_01_00_00/1 의결_02_00_00/1 사람_00_00_01/1 좌우_01_00_01/1 행동_00_00_01/1 것_01_00_01/2 종류_02_00_01/1 관계_05_00_01/1 일_01_00_01/1
기사__10/NNG    종류_02_00_01/1 부문_06_00_00/2
기사__10/NNG    말_01_00_04/1 앞_00_00_02/1 여럿_00_00_00/1 짓_01_00_00/1 의지_06_00_01/1 상태_01_00_01/2 모임_01_00_00/1 질서_03_00_00/1 한데_01_00_00/1 사건_01_00_01/1 사람_00_00_07/1 사무_05_00_00/1 절차_02_00_00/1 것_01_00_01/3 종류_02_00_01/1 일_01_00_01/1
기사__10/NNG    시간_04_00_01/1 여럿_00_00_00/1 목적_03_00_01/2 사이_01_00_02/2 뒤_01_00_02/4 가을_01_00_00/1 하나_00_02_02/2 모임_01_00_00/2 천간_00_00_00/1 일_07_01_00/3 이때_00_00_00/2 사람_00_00_01/2 월_02_01_00/2 무_04_00_00/1 일_01_00_01/2
```

그림 3. 추출된 학습데이터의 예(클래스 : 기사)

3.2.2 나이브 베이즈 분류기를 이용한 의미제약

의미제약은 전체적으로 [그림 4]와 같은 과정으로 수행된다. 원어정보에 의한 두 과정이 모두 실패한 경우 최종적으로 나이브 베이즈 분류기를 이용해 제약을 시도한다. 본 논문에서 적용한 나이브 베이즈 분류기는 의미제약시 특정 의미집단을 가진 공기명사는 특정 클래스로 분류된다는 문제로 간주하여 적용할 수 있다. 따라서 입력된 2-gram의 중심명사가 학습데이터의 클래스로 존재하는 경우 앞 명사의 의미집합들을 비교하여 나이브 베이즈 분류기에 의해 최대사후확률(Maximum a Posteriori)을 만족하는 클래스로 제약되도록 하였다. 이렇게 만들어진 의미제약 모듈은 모든 2-gram을 탐색할 때까지 수행된다.

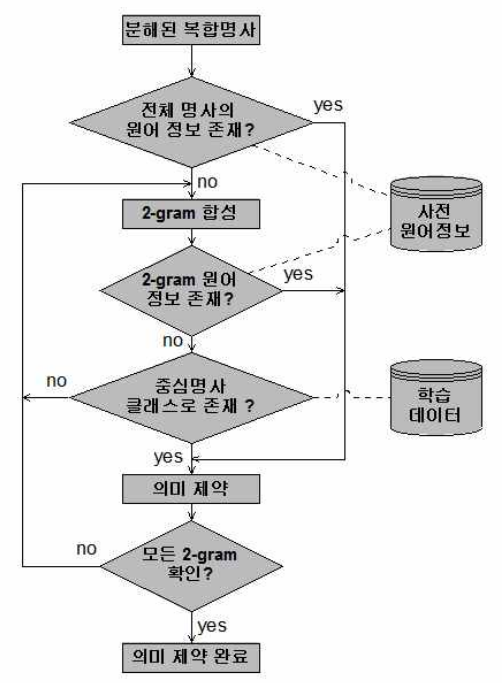

그림 4. 의미제약 알고리즘 순서도

4. 실험 및 결과

본 실험에 앞서 원어정보에 의한 제약 모듈만 적용된 이전 시스템의 정확도는 약 90.49%였다. 이 실험에서 의미 분별이 실패한 경우는 정답 세트의 오류(분해 및 태깅)로 인한 경우와(5.81%) 구성명사의 분해 오류(3.7%)인 경우가 있었다. 정답 세트의 오류는 수작업으로 수정하였고, 구성명사의 분해과정에서 위치별 명사 확률 계산에 필요한 학습 2-gram을 추가한 결과 [표 3]과 같이 정확도가 94.49%까지 상승하였다.

표 3. 기존 시스템 성능 실험 결과

	분석 명사	백분율(%)
정답 분류	38,476	94.49
오답 분류	2,240	5.51
총	40,716	100.00

세종말뭉치에서 추출한 의미태깅된 291,001개의 2-gram 복합명사 중 사전에서 속성명사 세트를 생성할 수 있는 유효한 대상으로 추려낸 결과 총 276,804개의 복합명사를 얻었으며 이를 이용해 학습데이터를 생성한 결과 943,996개의 학습데이터를 얻을 수 있었다. 하지만 학습데이터의 크기가 크고, 속성명사의 개수가 작은경우 같은 클래스의 다른 학습데이터와 중복된 경우가 많이 있어 속성명사의 개수가 일정 개수 이하인 경우 학습데이터에서 제거 하도록 하였다. 13개 이하인 경우를 제거한 경우가 가장 좋은 성능을 보였으며 그 결과 총 369,047개의 학습데이터로 약 61%가 축소되었다. 이를 이용해 수정된 의미 제약 모듈이 적용된 복합명사 의미 분석 시스템의 정확률 향상에 대한 실험을 하였다. 실험에 사용된 테스트 세트는 표준국어대사전에서 추출한 3음절 이상의 복합명사 40,716개를 사용했으며 그 결과는 [표 4]와 같다.

표 4. 제안된 시스템 성능 실험 결과

	분석 명사	백분율(%)
정답 분류	38,727	95.11
오답 분류	1,989	4.89
총	40,716	100.00

원어 정보만 사용한 제약 모듈에 비해 약 0.62%가 상승한 것을 알 수 있다. 이는 테스트 세트를 사전에서 추출했으므로 원어 정보에 의해 선 제약 되는 2-gram들이 많이 있었기 때문이다. 또한 제약된 2-gram들을 제약 방법별로 분석해 본 결과는 [표 5]와 같다.

표 5. 제약 방법별 분석 결과

	원어 정보	나이브 베이즈 분류기
총 2-gram	87,618	
제약된 2-gram	54,798	2,480
사용된 2-gram	483	1,426
제약 백분율(%)	62.54	2.83
총 제약율(%)	65.37	

총 2-gram은 테스트 세트에서 나올 수 있는 모든 2-gram 수를 의미하며, 이중 62.54%가 483개만으로 제약 됨을 알 수 있었다. 이는 특정 복합명사는 그 자체로 다른 복합명사 속에서 자주 사용됨을 나타낸다. 원어 정보에 의해 제약되지 못해 나이브 베이즈 분류기로 넘겨진 2-gram중 약 2.83%가 추가적으로 1,426개의 2-gram을 이용해 제약 되었으며 이로 인해 전체의 65.37%가 이 과정에서 제약되었음을 알 수 있다. 제약의 대상에서 제외된 34.63%의 2-gram은 원어 정보가 존재하지 않거나, 중심명사가 학습데이터에 그 클래스로 존재하지 않기 때문이며 이는 다음단계에서 2-gram들 간의 유사도 비교 후 유사도 별 체인 합성과정에 따라 최종적으로 태깅 되게 된다. 이뿐만 아니라 의미 제약 모듈은 의미 분석 시간에도 영향을 미치는데 이 실험에서는 [표 6]과 같이 약 1/3로 분석시간이 단축 되었음을 알 수 있다.

표 6. 복합명사 100개당 평균 의미 분석 속도

	제약 모듈 미사용	제약 모듈 사용
속도(초)	53	18

5. 결 론

본 논문에서는 원어정보를 이용한 기존 의미제약 방법에 나이브 베이즈 분류기를 추가로 이용하는 수정된 모델을 제시하였다. 원어정보를 이용한 방법은 사전에 의존적이기 때문에 정확한 2-gram 매칭에 의거해 사전에 등재되지 않은 표제어나 원어정보의 부재, 원어정보의 기술 방식 등에 따라 제약에 실패할 가능성이 크다. 위 실험에서는 테스트 세트를 사전에 등재된 복합명사에서 추출하여 실험 하였으므로 원어정보에 의해 제약된 경우가 많았으나 복합명사는 그 조합조건에 제한이 없기 때문에 사전에 존재하지 않는 일반 텍스트에서 추출한 복합명사나 신조어에 대한 분석시 제약에 실패할 확률이 크다. 이럴 경우 중심명사가 학습데이터 내의 클래스로, 구성명사가 사전에 표제어로 등재된 경우 의미 제약이 가능하며 실험 결과 보다 나이브 베이즈 분류기에 의한 제약율이 증가 할 것이다. 이 방법은 원어정보만 사용한 방법에 비해 0.62% 높은 95.11%의 정확률을 보였으며 본 논문의 실험을 통해 의미 제약 모듈이 부분적으로 중의성 해소에 성공해 정확도 뿐만 아니라 분석시간의 단축에도 도움을 주는 것으로 분석 되었다.

향후 연구로 구조 분해 단계에서 문제가 되어 의미 분석 결과에 영향을 미쳤던 미등록어, 접사 등에 대한 처리를 강화하고 더 다양한 종류의 학습데이터를 이용해 학습을 수행하는 연구가 이뤄진다면 더욱더 정확한 의미 분석 처리가 가능해 질 것이다.

참고문헌

[1] 이용훈, 옥철영, “의미기반 한국어 복합명사 분석”, 한국정보과학회 한국컴퓨터종합학술대회 논문집(C) pp.221-224, 2011.

[2] 허정, 장명길, “복합명사 의미사전을 이용한 동음이의어 중의성 해소”, 한국정보과학회 가을 학술발표 문집(Ⅱ) 제32권 제2호 pp.538-540, 2005.

[3] 손기준, 임수면, 박성배, 이상조, “동적인 문서 여과에서 나이브 베이즈 분류기와 코사인 유사 계수의 성능 비교”, 한국정보과학회 한국컴퓨터종합학술대회 논문집(B) pp.214-216, 2006.

[4] Andres Montoyo, Manuel Palomar, “Word Sense Disambiguation with Specification Marks in Unrestricted Texts", Database and Expert Systems Applications, Proceedings. 11th International Workshop, pp.103-107, 2000.