백오프 통계정보를 이용한 미등록어 포함 복합명사의 분해

박재한[0] 김명선 노대욱 나동열
연세대학교 정보기술학부 언어처리연구실
notlimit@empal.com

Segmenting Korean Nominal Compounds with an Unknown Morpheme Using Back-off Statistics

Jae-Han Park, Myoung-Sun Kim, Dae-Wook Rho, Dong-Yul Ra
Div. of Information Technology, Yonsei University

요 약

본 논문에서는 백오프 통계 정보를 이용하여 일반적인 복합명사 뿐만 아니라 외래어 미등록어를 포함한 복합명사도 잘 분해하는방법을 제안한다. 본 시스템은 입력으로 형태소분석기가 내주는 많은 분석 후보들을 받는다. 단음절 명사를 포함한 분석 후보도 포함되므로 입력 분석 후보의 수는 대단히 많게 된다. 본 모듈의 주요 작업은 이 중에서 가장 좋은 분석후보를 선택하는 것이 된다. 미등록어가 포함된 경우 이에 부합되는 분석 후보를 잘 선택하는 시스템의 개발을 목표로 한다. 이를 위해서 본 시스템에서 사용하는 주요 정보는 단어간 어휘 바이그램 통계정보이다. 또한 외래어 미등록어의 인식 정확성을 높이기 위해 음절 바이그램 정보도 이용한다. 통계정보는 대량의 품사 태깅 말뭉치에서 추출하였다. 데이터 부족 문제를 해소하기 위해서 우리는 백오프(back-off) 평탄화(smoothing) 기법을 이용하였다. 미등록어가 포함된 복합명사의 분석 후보의 수를 줄이기 위한 기술도 연구하였다.

1. 서 론

한국어의 경우 복합명사의 분해 문제는 매우 중요하다. 그 이유는 영어보다 더 다양하게 복합명사의 생성이 가능하며 구성 명사 사이에 띄어쓰기가 생략된 경우가 대부분이다. 그리고 구성 명사로 미등록어 가 많이 포함되기 때문이다. 이러한 복합명사 분해 문제는 많이 연구되었으나 미등록어를 포함한 경우 그 정확도가 만족스럽지 못하였다.

복합 명사 분해에서 기본적인 방법은 복합명사를 사전에 등록된 명사의 열로 분해하는 것이다. 그러나 가능한 분석 열 즉 분석 후보가 매우 많다는 것이 문제이다. 특히 본 시스템은 형태소분석 및 태깅 시스템의 내부 모듈로 이용되는 것을 목표로 하므로 형태소 분석기가 내주는 많은 분석 후보를 입력으로 받는다. 특히 한자어의 영향으로 한국어에서는 단음절 명사가 매우 많다. 이것은 분석후보의 수를 폭발적으로 증가시킨다.

본 복합명사 분해 시스템은 복합명사 내에 미등록어가 포함된 경우도 처리하는 것을 목표로 한다. 미등록어와 명사들이 합쳐서 복합명사를 이룬 경우 가능한 분해 후보는 매우 다양하며 매우 많은 후보가 생성된다. 이 경우 정답인 후보를 선정하는 문제는 매우 어려우며 지금까지 국내에서 많은 연구가 수행되었으나 만족할 만한 결과를 얻지 못하였다. 특히 복합명사 안에 외래어 고유명사가 미등록어로 사용된 경우가 매우 많으며 이의 성공적인 분석은 어려운 문제로 인식되어 왔다. 본 논문에서는 이러한 문제에 대처하는 기술에 대한 연구를 소개한다.

이러한 문제를 해결하기 위하여 본 연구에서는 통계 바이그램 정보를 가장 주요한 정보로 이용한다. 품사 바이그램 정보뿐만 아니라 특히 단어간 어휘 바이그램 정보를 최대한 활용하여 정확한 선택을 하도록 하였다. 이러한 통계 정보는 많을수록 좋으므로 대량의 품사 태깅 말뭉치에서 추출하여야 하며 현재의 시스템은 1000만 어절의 품사태깅 세종말뭉치를 이용하였다.

어휘 통계 정보를 주요 정보로 이용하기 때문에 대량의 훈련 말뭉치를 이용함에도 불구하고 많은 데이터 부

족문제가 발생하였으며, 이 문제에 대한 해소가 없이는 성공적인 시스템의 개발이 어려움을 알게 되었다. 이 문제에 대한 해결을 위하여 우리는 Collins[10]가 제안한 것과 유사한 back-off 기법을 사용하였으며 만족스런 결과를 얻을 수 있었다.

본 논문에서 복합 명사를 구성할 수 있는 품사는 명사, 의존명사, 접두사, 접미사로 보았다. 어떤 복합명사 어절은 마지막 부분이 조사로 분석될 수 있는 경우도 있다. 이 경우 본 시스템은 마지막 부분을 조사로 보고 분해한 후보 한 개와 마지막 부분을 조사로 보지 않고 분해한 경우에 대한 한 개의 후보를 생성하여 이 두 가지를 결과로 출력하도록 하였다. 그 이유는 이 둘 중의 최종 결정은 태깅 시스템에게 맡기기 위해서이다.

실험 결과 미등록어가 포함되지 않은 복합명사에 대하여 99.3% 이고 외래어 미등록어를 포함한 복합명사들 만에 대해서는 95.11% 의 정확도를 보였다. 이 두 가지를 통합한 전체적인 실험에 대해서는 99.20% 로서 기존의 시스템과 비교하여 향상된 성능을 보였다.

2. 관련 연구

기존 연구에서는 복합 명사 분해는 규칙을 이용하여 분해하거나 통계 정보를 이용하는 방법, 그리고 두 가지 모두 이용하는 방법으로 나눌 수 있다.

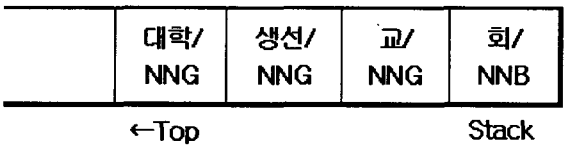
강승식(98)은 형태소 분석 결과로 추정된 복합명사를 단위 명사로 분해하는 방법으로 네 개의 분해 규칙과 두 가지의 예외 규칙을 사용하여 분해 후보들에 대해 가중치를 부여함으로써 최적 후보를 선택하는 알고리즘을 제시하였다. 이 방법은 규칙과 맞지 않는 예에 대해서는 오류를 범할 수 있다. 윤보현(97)은통계 정보와 선호 규칙을 이용하여 복합명사 분해를 하고 미등록어가 포함된 경우 휴리스틱을 이용하였다. 미등록어일 경우 휴리스틱을 이용함으로써 일정 범위 내의 경우만 처리할 수 없는 한계를 드러내었다. 심광섭(97)은 네 가지 유형의 음절간 상호 정보를 이용하는 분해 방법을 제시하고 있다. 분해 후보 생성을 위해 음절 바이그램을 이용하였다. 이현민(2000)은 복합명사 분해에 있어 일반적인 좌에서 우 방향 분해 대신 우에서 좌 방향으로 분해하는 방법을 제안하였다. 1음절 명사를 포함 하지 않아 분석 영역의 한계를 들어냈고 음절수가 길어지면서 정확률이 현저히 떨어졌다. 미등록어에 대한 대처 방법이 제시하지 못하였다. 김재훈(2003)은 사례기반 기계학습방법을 이용하고 음절 단위로 복합명사를 분리하는 방법을 제시하고 있다.

3. 미등록어 포함 복합명사 분해후보 생성 기법

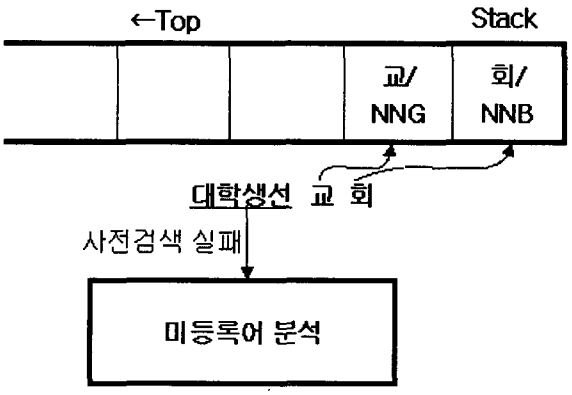
일반적으로 복합명사 분해의 대상은 등록어만으로 구성된 형태소의 집합이다. 하지만 이 논문에서는 미등록어를 포함한 형태소의 집합도 분해 후보로 간주하였다. 이 절에서는 복합명사 분해 후보의 생성과정을 일반적인 경우와 미등록어를 포함한 경우에 대해 예제를 통해 간략히 언급하겠다.

3.1 일반적인 복합명사 분해후보 생성

등록어 형태소만으로 구성된 예로서 "대학생선교회"의 분해 후보 생성과정을 살펴보겠다. 우선 형태소의 분석 방향은 우측에서 좌측으로 자소 단위로 이루어진다. 현재 얻어진 형태소를 사전검색을 취해 성공 시에는 스택에 넣고, 실패 시에는 현재 만들고 있는 형태소에 자소를 늘려 나간다. 진행된 분석이 좌측 끝 음절에 도달하고, 모든 형태소가 등록어로서 스택에 넣어진 상태이면 스택의 형태소 열이 하나의 분해 후보가 된다.



[그림 1] "대학생선교회"의 분석 예



[그림 2] 미등록어 분석 호출 예

그렇지 않고 좌측의 마지막 형태소가 사전검색에 실패한 경우에는 미등록어 분석으로 보내진다. 이 경우에는 다수의 후보가 생성될 수 있다.

이와 같이 좌측 끝 음절까지 분석이 되고, 스택이 등록어 만으로 구성되거나, 미등록어 분석이 종료되는 시점에는 다음 분해 후보의 생성을 위해 스택의 상위 형태소를 꺼낸다([그림 1]의 경우는 상위 두 개의 형태소를 꺼내고, [그림 2]의 경우에는 상위 한 개의 형태소를 꺼낸다). 이후 꺼내어 진 형태소를 위에서 취한 방식과 동일하게 우측에서 좌측으로 형태소 분석을 시도한다.

이러한 과정을 스택이 empty 상태에 도달 할 때까지 반복한다. 위 예제에서는 미등록어를 포함한 후보가 존재하지만, 등록어 만으로 구성된 분석열의 형태소간 바이그램 확률의 곱이 임계값을 초과하여 좋은 것으로 나타나므로 미등록어를 포함한 후보들은 모두 제거된다.

대학/NNG + 생선/NNG + 교/NNG + 회/NNB
대학/NNG + 생/XP + 선교/NNG + 회/NNB
대학생/NNG + 선교/NNG + 회/NNB
대학/NNG + 생선/NNG + 교회/NNG
대학/NNG + 생/XP + 선교회/NNG
대학생/NNG + 선교회/NNG

위의 예제는 위의 과정을 통해 생성된"대학생선교회"의 분해 후보 리스트이다.

3.2 미등록어를 포함한 경우의 분해후보 생성

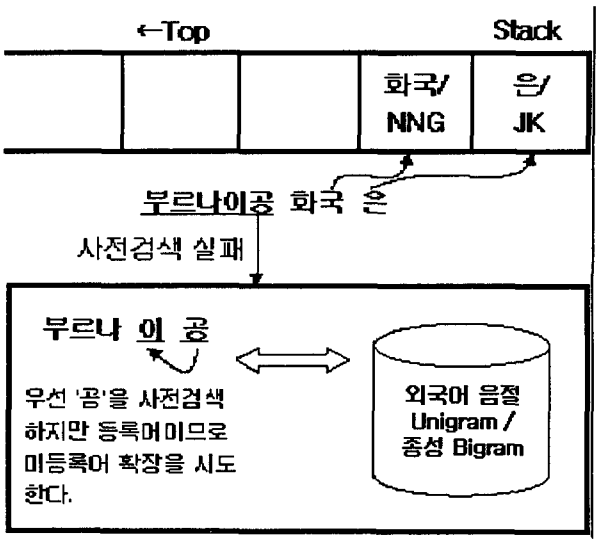
여기서는 주로 외국어 미등록어에 초점을 맞췄다. "부르나이공화국은"을 예로서 미등록어를 포함한 분해 후보의 생성과정을 살펴보겠다.

우선 3.1에서 제시한 방법에 의해 형태소 분석이 진행이 되고, 특정 단계에서 가장 좌측의 형태소가 사전검색 실패로 인해 [그림 2]와 같이 미등록어 분석으로 보내지게 된다.

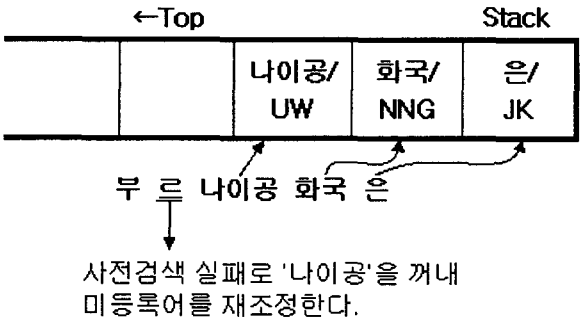
이렇게 미등록어 분석으로 보내진 형태소는 우측에서 좌측으로 음절단위의 미등록어 탐색을 시도한다. 처음 한 음절을 사전검색하고 성공시에는 등록어이므로 좌측으로 미등록어 확장을 시도한다. 이때 외국어 음절 유니그램과 종성 바이그램 정보를 이용해 미등록어 확장의 진위여부를 판단한다. 확장이 불가능하다면 미등록어 분석은 종료된다.

사전검색 실패시에는 미등록어라는 품사를 부여해 스택에 넣고, 나머지 형태소를 사전검색을 통해 품사를 부여한 후 스택에 넣는다.

만약 미등록어가 스택에 들어간 상태에서 나머지 형태소의 사전검색 실패시에는 스택에서 미등록어를 꺼낸 후 다시 미등록어의 확장을 시도한다. 좌측 끝 음절까지 분석이 되었다면 스택의 형태소 열은 하나의 분해 후보가 된다.



[그림 3] 미등록어 확장 예



[그림 4] 미등록어 재조정 예

이후 다음 분석을 위해 스택의 상위 두 개의 형태소를 꺼내고, 위의 방식과 동일한 분석을 시도한다. 이러한 과정은 미등록어 분석으로 보내진 형태소 전체가 미등록어로 나오거나 미등록어 확장이 실패한 경우까지 진행된다.

다음은 위의 과정을 통해 생성된 "부르나이공화국은"의 분해 후보 리스트이다. (UW : 미등록어 품사임).

부르나/UW + 이공/NNG + 화국/NNG + 은/JK
부/NNG + 르나이공/UW + 화국/NNG + 은/NNG
부르나이공/UW + 화국/NNG + 은/NNG
부르나이/UW + 공화국/NNG + 은/JK
부르나/UW + 이공/NNG + 화/XSN + 국은/NNG
부르나이/UW + 공화/NNG + 국은/NNG
⋮

이 예에서 미등록어를 포함한 분석은 46개 정도의 후보가 생성된다. 이후 모든 후보를 대상으로 미등록어와 좌,우 형태소간의 바이그램 정보를 통해 적절치 못한 6개의 후보는 제거되었다. 위의 리스트는 최종 결과 중 일부만을 보인 것이다. 복합명사 분해 모듈은 위 결과를 받아 최적의 한 후보를 정답으로 선택한다.

4. 정답 후보 선택을 위한 통계 정보의 이용

분해된 후보 중에서 가장 좋은 후보를 선택하기 위해 단어 바이그램 확률을 이용한 백오프 평탄화 기반 방법을 사용하였다. 다음과 같이 복합명사 후보가 형태소 어휘 mi와 품사 ti들의 열로 구성되어 있고 L개의 형태소로 분리된 후보라 하자.

m1/t1	m2/t2	·· ··	mL-1/tL-1	mL/tL

우리는 다음과 같이 정의된 복합명사확률을 후보간의 비교 및 선택의 척도로 이용한다 :

$$P(m_1/t_1, m_2/t_2, ..., m_{L-1}/t_{L-1}, m_L/t_L)$$

$$= \left[P(m_1/t_1 \mid \#) * \{ \prod_{i=2}^{L} P(m_i/t_i \mid m_{i-1}/t_{i-1}) \} * P(\# \mid m_L/t_L) \right]^{\frac{1}{\lambda}}$$

#는 공백, ?는 임의의 형태소가 올 수 있음을 나타낸다. 1/λ는 기하평균을 취하기 위한 지수이다. 분해된 후보 중에는 미등록어가 포함되어 있을 수 있고 그렇지 않을 수도 있다. 모든 후보는 같은 수식에 의해 계산된 확률값으로 경쟁을 하게 된다.

4.1 시작 확률의 계산

이것은 형태소가 어절의 첫 형태소로 나타날 확률을 말한다.

$$P(m_1/t_1 \mid \#) = \alpha * \frac{C(\#, m_1/t_1)}{C(\#, ?)} + (1-\alpha) * \frac{C(\#, t_1)}{C(\#, ?)}$$

우변의 첫번째 항에서는 형태소의 어휘와 품사를 같이 보며 데이터 부족 현상에 대비하여 두번째 항에서는 품사만을 보는 바이그램으로 백오프(back off) 하였다. 통계 정보는 세종 코퍼스에서 품사가 명사, 수사, 조사, 접두사, 접미사로 부착된 단어를 중심으로 추출한다.

$C(\#, m_1/t_1)$: m1/t1가 어절의 첫 형태소로 출현한 카운트 즉 횟수 (t1 이 UW 인 경우 이 카운트는 0 임.)

$C(\#, ?)$: 코퍼스 안의 어절의 총 수.

$C(\#, t_1)$: 특정 품사 t1 이 어절의 첫 형태소로 나타난 카운트.

4.2 중간 확률

$$P(m_i/t_i \mid m_{i-1}/t_{i-1})$$

$$= \alpha \frac{C(m_{i-1}/t_{i-1}, m_i/t_i)}{C(m_{i-1}/t_{i-1}, ? \sim \#)} + (1-) \frac{C(m_{i-1}/t_{i-1}, t_i)}{C(t_{i-1}, ? \sim \#)} \frac{C(t_{i-1}, m_i/t_i)}{C(t_{i-1}, ? \sim \#)} + (1-)(1-) \frac{C(t_{i-1}, t_i)}{C(t_{i-1}, ? \sim \#)}$$

위의 중간확률은 어절 내의 두 형태소 사이의 바이그램 확률을 말한다. ?~#는 공백이 아닌 임의의 형태소를 말한다. 위 식의 우변은 백오프 기법에 의하여 점차 완화된 바이그램을 이용하고 있다.

$C(m_{i-1}/t_{i-1}, m_i/t_i)$: 연속한두 형태소 mi-1/ti-1, mi/ti의 어휘 및 품사를 모두 본 출현 횟수. (둘 중 하나라도 미등록어이면 이 카운트는 0임.)

$C(m_{i-1}/t_{i-1}, t_i)$: 두 형태소가 연속해서 나왔을 때 이전 형태소의 어휘(mi-1)와 품사(ti-1), 현재 형태소의 품사(ti)를 본 출현 횟수. (앞이 미등록어이면 0 임.)

$C(t_{i-1}, m_i/t_i)$: 두 형태소가 연속해서 나왔을 때 이전 형태소의 품사(ti-1)와 현재 형태소의 어휘(mi)와 품사(ti)로 출현한 횟수. (뒤가 미등록어이면 0임.)

$C(m_{i-1}/t_{i-1}, ? \sim \#)$: 어휘 mi-1, 품사 ti-1 인 형태소가 임의의 형태소 앞에 출현한 카운트. (앞이 미등록어이면 0임.)

$C(t_{i-1}, t_i)$: 두 형태소가 연속해서 나왔을 때 두 형태소 품사 쌍 카운트.

$C(t_{i-1}, ? \sim \#)$: 품사(ti-1) 인 형태소가 임의의 형태소 앞에 나온 횟수.

4.3 끝 확률

$$P(\# \mid m_L/t_L) = \alpha * \frac{C(m_L/t_L, \#)}{C(m_L/t_L, ?)} + (1-\alpha) * \frac{C(m_L/t_L, \#) + C(t_L, \#)}{C(m_L/t_L, ?) + C(t_L, ?)}$$

위는 형태소 열의 마지막 형태소에 대한 확률이다.

$C(m_L/t_L, \#)$: 형태소(mL/tL) 다음이 공백인 경우의 형태소 어휘(mL)와 품사(tL) 카운트.

$C(m_L/t_L,?)$: 형태소(mL/tL)가 나온 총 카운트.

$C(t_L,\#)$: 품사(tL) 다음에 공백이 나온 카운트.

$C(t_L,?)$: 품사(tL)가 나온 총 카운트.

4.4. 확률값의 조정

각 확률은 백오프 기법을 도입함으로써 여러 항의 합으로 나타낸다. 각 항 사이의 가중치 조절을 위한 계수가 α 와 이다. 이들은 0 이상 1 이하의 값을 취하는데 이의 결정은 현재로서는 실험을 통하여 결정한다. 보통 앞의 항을 강조하기 위해서 매우 1에 가까운 값을 가진다.

그리고 후보가 1 음절 형태소를 가진 경우 이에 벌점을 부과하기 위해서 1음절이 나타난 수만큼 1/13을 곱해 주어 전체 확률 값을 낮추었다. 그 이유는 1음절 명사를 가진 후보는 정답이 되기 어렵고 1음절의 경우 2 음절 이상의 명사보다 코퍼스에 더 많이 나타나며 잘못 태깅된 1음절 명사가 많기 때문이다.

이렇게 하여 계산된 확률은 분해후보 내에 형태소 의 수가 많을수록 확률 값을 많이 곱하게 된다. 확률 값은 1보다 작은 값이기 때문에 전체확률 값을 낮아지게 한다. 이것은 정확한 후보 선택에 오류를 야기한다. 그래서 기하 평균을 취하여 많은 형태소 들로 나눠진 경우에도 공정한 경쟁을 할 수 있게 하였다

형태소 열의 수	기하 평균 λ 값
2개 이하	λ = (L+1)
3개 or 4개	λ = (L+1) - 0.5
5개 이상	λ = (L+1) 1.0

L은 분해 후보의 형태소 열의 길이이고 1을 더해 주는 이유는 시작 확률, 중간 확률, 끝 확률을 곱해 주기 때문에 형태소 열의 수보다 한번 더 곱해지기 때문이다. 위와 같은 방법으로 분해 되는 것을 예를 통하여 보자. "대학생선교회"가 6개로 분해 되지만 그 중에서 확률값이 가장 높은 두 가지 후보만 보겠다 :

A : 대학+생선+교회

B : 대학생+선교회

"대학"이란 명사가 "대학생"보다 시작확률이 좋지만 "대학+생선"과 "생선+교회"가 어휘 바이그램으로 나타난 적이 없어 결국 백오프 품사 바이그램에 의한 확률의 기여를 받는다. 그러나 이것은 낮은 확률값이 되게 한다. 결국 "대학생+선교회"가 선택되게 된다.

A 후보열	확률(log)	B 후보열	확률(log)
P(#|대학)	-2.94	P(대학생|#)	-3.28
P(생선|대학)	-7.27	P(선교회|대학생)	-8.20
P(교회|생선)	-7.45	P(#|선교회)	-0.02
P(#|교회)	-0.10		
기하평균 λ	3.5	기하평균 λ	3
확률	-7.07	확률	-5.820

4.5 미등록어의 확률

미등록어를 포함한 복합명사의 경우 많은 분해 후보가 생성되는데 대부분은 미등록어로 간주되는 형태소를 포함하고 있다. 물론 이 중에서 한 후보만이 올바른 미등록어를 가진 것이다. 결국 미등록어로 간주되는 형태소의 좋고 나쁨을 나타내는 것이 필요한데 우리는 미등록어 확률 $P(uk)$로 이를 나타낸다. 이것은 미등록어를 구성하는 음절들의 바이그램 확률들의 곱으로 계산한다.

$$P(uk) = P(s_1, s_2, ..., s_{L-1}, s_L)$$

$$= \prod_{i=1}^{L} P(s_i \mid s_{i-1})$$

$$P(s_i \mid s_{i-1}) = \alpha \frac{C(s_{i-1}, s_i)}{C(s_{i-1}, ?)} + (1-\alpha) \frac{C(s_{i-1}, ?) + C(?, s_i)}{\sum_{all} C(s_{i-1}, ?)}$$

$C(s_{i-1}, s_i)$: 두 음절이 연속해서 나온 바이그램 횟수.

$C(s_{i-1}, ?)$: 특정 음절(si-1)이 나온 카운트.

$C(?, s_i)$: 특정 음절(si)이 나온 카운트.

$\sum_{all} C(s_{i-1}, ?)$: 코퍼스의 음절 수.

미등록어의 음절 바이그램 통계 정보는 외래 고유명사를 모은 말뭉치에서 음절 바이그램과 유니그램을 구하였다. 미등록어가 길면 길수록 손해를 보기 때문에 곱한 수만큼 기하평균을 취하여 계산을 하게 된다. 미등록어의 α 는 0.999로 하였을 때 가장 높은 정확도를 보였다.

결국 미등록어를 가진 복합명사 분해 후보의 확률은 다음과 같이 P(uk)를 원래 확률에 곱하여 구한다.

$$P'(m_1/t_1, m_2/t_2, ..., m_{L-1}/t_{L-1}, m_L/t_L)$$

$$= P(m_1/t_1, m_2/t_2, ..., m_{L-1}/t_{L-1}, m_L/t_L) * \begin{pmatrix} \text{미등록어인 경우} : P(uk) \\ \text{등록어인 경우} : \delta \end{pmatrix}$$

복합 명사 분해 후보들 가운데 미등록어가 들어간 후보에 $P(uk)$를 곱해줘서 분해 후보 결정에 영향을 주게 된다. 이때 미등록어가 없는 분석 후보에 대해서도 등록어 분해에 대한 가중치 δ을 곱해주게 된다. 이유는 미등록어가 들어간 후보에 대해 $P(uk)$값을 곱해짐으로써 미등록어가 포함된 후보와 그렇지 않은 후보 사이에 확률 차이가 너무 생기기 때문이다. 등록어만으로 구성된 분해후보에 곱해 주는 δ는 값을 변화시키면서 실험해본 결과 0.01이 좋은 결과를 낳았다.

"오렌지카운티"의 경우 "오렌지/NNG+카/NNG+운/NNG+티/NNG"의 등록어만으로 분석된 후보는 확률이 -6.2901 이고 를 곱하면 -8.2901이 된다. "오렌지/NNG+카운티/UK"로 미등록어를 포함한 후보는 P(uk) 를 곱하기 전에는 -2.8061 곱한 후에는 -6.5784이다. "오렌지카운티/UK" 후보는 전체가 미등록어인 경우로서 확률은 -3.7337 인데 P(uk)를 곱하면 -7.0439이다. 이 예는 미등록어 확률인 $P(uk)$의 확률과 δ가 곱해지지 않아도 원래의 복합명사 확률만 이용해도 제대로 분석되는 경우이다. 그러나 $P(uk)$와δ 값이 곱해지더라도 오류가 없다는 것을 볼 수 있다.

"애틀란타올림픽"은 "애+틀+란+타올+림+픽"으로 모 두 등록어로 된 분석이 가능하나 미등록어가 포함된 "애틀란타/UK + 올림픽"으로 제대로 분해하였다.

5. 실험 및 검토

본 논문에서 사용된 데이터 집합은 세종 코퍼스 1,000만 어절에서 448만개 명사와 복합명사를 분리해 놓은 것에서 통계 정보를 추출하였다. 백오프 확률을 계산할 때 미등록어로 나오는 경우 코퍼스에는 미등록어가 없으므로 품사가 고유명사인 것을 미등록어로 간주하여 통계 정보를 수집하였다.

[표 1] 복합명사 테스트 셋

데이터 집합		복합 명사 수
Test set A	세종코퍼스	2277
Test set B	웹 (미등록어 포함)	307
Test set C	부산대코퍼스	8502

5.1 등록어만으로 된 복합명사 실험(α, 값에 따라)

이번 실험은 α, 의 값이 얼마일 때 가장 높은 정확도를 나타내는지 찾기 위해 값을 변경하며 실험하였다. Test set A에서는 α이 0.999이고 이 0.9999999 일 때가 α이 0.999이고 이 0.99999일 때보다 더 높은 정확도를 보였다. 하지만 Test set C에서는 α이 0.999이고 이 0.99999일 때 가장 높은 확률을 보였다. 전체 확률상 α이 0.999이고 이 0.99999일 때 가장 높은 정확도를 얻을 수 있었다.

[표 2] Test set A 실험 결과

α		복합명사 수	오류수	정확률
0.99999	0.99999	2277	12	99.47
0.999	0.99999	2277	14	99.39
0.99	0.99999	2277	14	99.39
0.999	0.999	2277	17	99.25
0.999	0.9999999	2277	11	99.52
0.99999	0.9999999	2277	14	99.39

[표 3] Test set C 실험 결과

α		복합명사 수	오류수	정확률
0.99999	0.99999	8502	78	99.08
0.999	0.99999	8502	60	99.30
0.99	0.99999	8502	60	99.30
0.999	0.999	8502	64	99.25
0.999	0.9999999	8502	71	99.17
0.99999	0.9999999	8502	88	98.96

5.2 음절 수에 따른 실험 결과

실험 데이터는 7음절 이상인 복합명사 수가 적기 때문에 긴 복합명사에 대한 분해 정확도를 오차가 있을 수 있다.

[표6]과 [표7]에서는 분석후보에 1음절 형태소가 나타날 때마다 벌점으로 1/13을 곱하였다. 이는 1음절 명사로 분해되는 것을 가능하면 억제하기 위한 것이다. 1음절 명사는 형태소 분석 및 태깅에서 많은 문제를 야기하는데 특히 오태깅이 많이 발생하여 태깅에 오류를 일으키는 것이 관찰되었다. 우리는 접두사, 접미사, 1음절 명사에 벌점을 부과하였고 그 결과 성능이 향상됨을 관찰하였다.

[표 4] Test A (α = 0.999, = 0.99999)

음절수	복합명사 수	맞은 수	정확률(%)
4음절	1457	1452	99.65
5음절	665	659	99.40
6음절	125	123	98.40
7음절	23	23	100.0
8음절	7	7	100.0
총 합	2277	2273	99.43

[표 5] Test C(α = 0.999, = 0.99999)

음절수	복합명사 수	맞은 수	정확률(%)
4음절	5196	5187	99.83
5음절	1575	1552	98.72
6음절	923	917	99.35
7음절	407	398	97.79
8음절	204	198	97.05
9음절	114	113	99.12
10음절	55	52	94.55
11음절	22	20	90.09
12음절	5	5	100.00
13음절	1	1	100.00
총 합	8502	8443	99.30

[표 6] Test A (α = 0.999, = 0.99999)

1음절 가중치	복합명사 수	오류수	정확률
미적용	2277	14	99.39
적용	2277	14	99.39

[표 7] Test C (α = 0.999, = 0.99999)

1음절 가중치	복합명사 수	오류수	정확률
미적용	8502	125	98.53
적용	8502	60	99.30

5.3 미등록어에 대한 실험결과

미등록어 확률인 P(uk)의 영향력을 알아보기 위해 이 확률을 적용한 것과 적용하지 않은 것을 구분하여 실험 하였다.

[표 8] Test set B (α = 0.999, = 0.99999)

P(uk) 확률	미등록어 포함 복합명사 수	오류수	정확률
미적용	307	17	94.46
적용	307	15	95.11

복합명사 전체가 하나의 미등록어인 경우는 쉽게 해결된다. 한국어의 미등록어 인식에서 특히 어려운 점은 복합명사가 미등록어와 여러 명사의 결합으로 이루어진 경우이다. 우리는 이에 해당하는 예들을 수집하여 실험하였다. 우리 시스템이 성공적으로 분석한 예들을 일부 보이면 다음과 같다.

"지식+정보+허브/UK+사이트",
"애플/UK+컴퓨터+용",
"세브란스/UK+정신+건강+병원",
"사이버/UK+테러+대응+센터",
"몬터레이/UK+해양+연구소",
"골드위즈/UK+그룹",
"경영+관리+솔루션/UK"

5.4 검 토

등록어만으로 구성된 복합명사 분해 실험에서 문제가 되었던 것은 1음절 명사, 접두사와 접미사였다. 1음절이기 때문에 분석후보의 수를 늘리는 작용을 하게 된다. 이 문제를 해결하기 위해 가중치 1/13을 적용하여 해결하였다. 하지만 1음절 명사의 경우 코퍼스에 자주 나타나지 않았던 것은 분해 오류를 일으켰다. 예를 들면 "말+경계선"과 같은 경우 "말경+계선"으로 분석되는 후보가 약간의 차이로 선택되는 것을 볼 수 있었다.

접두사와 접미사가 같이 쓰이는 접사가 코퍼스에서 접두사로만 나오거나 접미사로만 나온 경우 문제가 되었다. 접두사와 접미사의 출현 빈도를 비슷한 경우 다른 형태소와의 관계를 보고 판단할 수 있었지만 전혀 다른 쪽으로는 쓰이지 않은 접사의 경우는 문제가 되었다.

코퍼스에 나타나지 않았던 어휘에 대해 백오프를 통하여 계산하였지만 자주 나타난 명사와 비교되는 경우 문제가 되었다. "시장+세분화"에서 "세분화"가 코퍼스상에 나타나지 않아 "시장세+분화"를 선택하는 오류를 발생하였다.

미등록어의 경우, 분석 후보가 하나의 미등록어로 분석된 경우가 있으면 중간 확률을 계산하지 않기 때문에 높은 확률을 갖는다. 그래서 분해 후보를 생성하는 과정에서 음절 바이그램과 어휘 바이그램을 이용하여 하나의 미등록어 분석을 제한하였다. 하지만 그 과정에서 정답 미등록어가 나오지 않는 경우가 발생한 경우와 전체 미등록어를 분해 후보로 생성하는 오류가 발생하였다. 이러한 오류는 정답 자체가 분해되지 않는 오류와 백오프에서 확률 값 차이가 커서 $P(uk)$를 곱해도 그 차이가 좁혀지기만 하고 더 커지지 않아 실패하였다. 그래서 하나의 미등록어로 분석된 경우 시작확률과 끝 확률의 α 를 0.99999로 조정하여 해결하였다.

미등록어가 다음에 오는 등록어의 앞 1음절을 포함하여 분석이 되는 경우 문제가 되었다. "뉴코아/UK+한의

원"의 경우 "뉴코아"가 미등록어인 경우인데 "한의원"의 앞 음절인 "한"을 미등록어로 포함시킨 경우이다. "뉴코아한/UK+의원"으로 분석된 후보는 "의원"이 코퍼스에 출현빈도가 "한의원"보다 높아 정답으로 선택된다. 이처럼 미등록어와 등록어로 분석된 경우에 등록어의 길이가 더 긴 명사가 코퍼스에서 출현 빈도를 보고 많이 나온 명사이면 2음절 명사보다 높은 확률을 갖게 하여 해결하였다. 하지만 코퍼스에 나온 어휘 카운트만 가지고 판단하여 자주 나타나지 않은 명사에 대해서는 오류가 발생 하였다.

또 미등록어가 출현빈도가 낮은 등록어를 포함하는 분석 후보가 문제가 되었다. "사이버/UK+시티+센터"는 "사이버"가 미등록어인 경우인데 명사 "시티"를 포함하여 분석된 "사이버시티/UK+센터"보다 낮은 확률값을 갖는다. P(uk)확률을 이용하였지만 "시티"의 출현빈도가 낮아 문제를 극복하지 못하였다.

본 논문은 등록어로 구성된 복합명사의 경우 분해 정확도가 평균 99.31%(99.39%, 99.30%)이고, 미등록어를 포함한 복합명사들에 대해서는 평균 95.1%라는 높은 정확률을 보였다. 전체적으로 본 시스템의 성능은 99.20%를 보였다. 이 정확도는 윤보현(1997)의 96.8%, 강승식(1997)의 97.95%에 비해 높은 좋은 결과로 생각된다.

6.결 론

본 논문에서는 분해할 복합명사가 항상 미등록어가 존재 할 수 있다고 보고 미등록어를 포함한 후보들과 등록어로만 구성된 후보들과 같이 경쟁하여 더 높은 확률값을 갖는 것을 선택하도록 하였다. 본 시스템은 대량의 품사 태깅 훈련 말뭉치에서 추출한 단어가 바이그램 확률 정보를 주요 정보원으로 사용한다. 그러나 어휘 쌍을 보게 되므로 데이터 부족 문제가 많이 발생한다. 이를 극복하기 위해서 백오프 평탄화를 이용하여 완화된 바이그램 정보를 이용하도록 하였다. 품사 태깅이나 복합명사 분해에서 많은 문제를 일으키는 단음절 명사, 접사를 처리하기 위하여 벌점 가중치를 곱하여 문제를 완화시킬 수 있었다. 미등록어에 대해서는 외래어 고유명사내에서의 음절 바이그램 정보를 이용하여 미등록어 확률을 구하여 이를 전체 확률에 반영하는 기법을 사용하였다. 실험 결과 기존의 시스템보다 높은 성능을 나타내는 것을 관찰하였다.

향후 연구 과제로는 미등록어와 등록어 분석 후보가 경쟁을 할 때 가중치 조절 기법이 좀 더 실험이 된다면 지금보다 더 좋은 결과를 얻을 수 있을 것으로 보인다. 그리고, 한국어 미등록어에 대해 현재는 고유명사 사전에만 의존하고 있지만 사전에 없는 한국어 미등록어를 인식할 수 있는 기법을 개발하여야 한다.

참고 문헌

[1] 강승식, "한국어 복합명사 분해 알고리즘", 정보과학회 논문지(B), 제 25권, 제 1호, pp.172-182, 1998

[2] 박봉래, 황영숙, 임해창, "유사 어절의 TAIL 패턴 분석에 기반한 미등록 명사 추정", 정보과학회 봄 학술발표 논문집 제23권 1호, pp.907-910, 1996

[3] 심광섭, "합성된 상호 정보를 이용한 복합명사 분리",정보과학회 논문지(B) 제24권 11호, pp.1307- 1317, 1997

[4] 양장모, 김민정, 권혁철, "언어 정보를 이용한 한국어 미등록어 추정", 정보과학회지 제23권 1호, pp.957-960, 1996

[5] 윤보현, 조민정, 임해창, "통계정보와 선호규칙을 이용한 한국어 복합명사의 분해", 정보과학회논문지(B) 제24권 제8호, 1997

[6] 이현민, 박혁로, "복합명사의 역방향분해 알고리즘",한글 및 한국어정보처리, 2000

[7] 장동현, 맹성현, "효율적인 색인어 추출을 위한 복합명사 분석 방법", 제8회 한글 및 한국어 정보처리 학술발표논문집, pp.32-35,1996,

[8] 최재혁, "음절수에 따른 한국어 복합명사 분리 방안", 제8회 한글 및 한국어 정보처리 학술발표논문집, pp.925-928, 1996

[9] 김재훈,이공주, "사례기반 학습을 이용한 음절기반 한국어 단어 분리 및 범주 결정", 정보과학회(B), 제 10권, 제 1호, pp.47-56, 2003

[10] Collins, M. 1996. "A new statistical parser based on bigram lexical dependencies," Proc. ACL'96, pp.184-191.