http://dx.doi.org/10.3745/KIPSTB.2012.19B.1.063

U-WIN을 이용한 한국어 복합명사 분해 및 의미태깅 시스템

이 용 훈[†] · 옥 철 영[††] · 이 응 봉[†††]

요 약

본 논문에서는 통계기반의 복합명사 분해 방법과 어휘의미망(U-WIN)과 사전 뜻풀이에서 추출한 의미관계 정보를 이용하는 한국어 복합명사 의미 태깅 시스템을 제안한다.

본 시스템은 크게 복합명사 분해, 의미제약, 그리고 의미 태깅의 세 가지 부분으로 이루어진다. 분해과정은 세종말뭉치에서 추출한 위치별 명사 빈도를 사용하여 최적의 구성 명사 분해 후보를 선정하고 의미제약을 위한 구성 명사 재분해와 외래어 복원의 과정을 수행한다. 의미범위 제약과정은 유사도 비교의 계산량을 줄이고 정확도를 높이기 위해 원어 정보와 Naive Bayes Classifier를 이용해 가능한 경우 구성 명사의 의미를 선 제약한다. 의미 분석 및 태깅 과정에서는 bigram 구성 명사의 각 의미 유사도를 구하고 하나의 체인을 만들어가며 태깅을 수행한다.

본 시스템의 성능 평가를 위해 표준국어대사전에서 추출한 3음절 이상의 40,717개의 복합명사를 대상으로 의미 태깅된 테스트 셋을 구축하였다. 이를 이용한 실험에서 99.26%의 분해 정확도를 보였으며, 95.38%의 의미 분석 정확도를 보였다.

키워드 : 어휘의미망(U-WIN), 복합명사 분해, 의미제약, 나이브 베이즈 분류기, 의미유사도, 의미 태깅

Korean Compound Noun Decomposition and Semantic Tagging System using User-Word Intelligent Network

Yong-Hoon Lee[†] · Cheol-Young Ock[††] · Eung-Bong Lee[†††]

ABSTRACT

We propose a Korean compound noun semantic tagging system using statistical compound noun decomposition and semantic relation information extracted from a lexical semantic network(U-WIN) and dictionary definitions.

The system consists of three phases including compound noun decomposition, semantic constraint, and semantic tagging. In compound noun decomposition, best candidates are selected using noun location frequencies extracted from a Sejong corpus, and re-decomposes noun for semantic constraint and restores foreign nouns. The semantic constraints phase finds possible semantic combinations by using origin information in dictionary and Naive Bayes Classifier, in order to decrease the computation time and increase the accuracy of semantic tagging. The semantic tagging phase calculates the semantic similarity between decomposed nouns and decides the semantic tags.

We have constructed 40,717 experimental compound nouns data set from Standard Korean Language Dictionary, which consists of more than 3 characters and is semantically tagged. From the experiments, the accuracy of compound noun decomposition is 99.26%, and the accuracy of semantic tagging is 95.38% respectively.

Keywords : Lexical Semantic Network(U-WIN), Compound Noun Decomposition, Semantic Constraints, Naive Bayes Classifier, Semantic Similarity, Semantic Tagging

1. 서 론

복합명사는 중요한 의미를 나타낼 수 있는 성분으로 자연어 처리가 필요한 많은 분야에서 복합명사를 올바르게 분해하는 작업은 매우 중요하다. 이에 따라 한국어 복합명사에 대한 기존 연구는 대부분 구조 분해에 그 초점이 맞춰져 왔으며 활발히 연구되어 높은 성능을 보이고 있다.

복합명사 분해에 관한 초기 연구에는 크게 음절 길이에 따른 선호 분해 패턴들을 이용하는 방법과 통계데이터를 이용해 중의적 분해를 해결하는 방법들이 있다. 전자는 복합명사의 음절별 선호 분해 패턴을 이용해 차례로 분해를 시도한다. 이는 분해 속도 향상과 개념이 단순해 알고리즘의 적용이 쉽다는 장점이 있으나, 음절의 제한이 없는 복합명사의 결합 특성상 모든 복합명사를 대상으로 할 수 없다는 단점이 있다.

※ 이 논문은 2010년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2010-32A-H00006).
† 준 회 원 : 울산대학교 컴퓨터정보통신공학과 석사과정
†† 종신회원 : 울산대학교 컴퓨터정보통신공학과 교수(교신저자)
††† 정 회 원 : 충남대학교 문헌정보학과 교수
논문접수 : 2011년 12월 26일
수 정 일 : 1차 2012년 1월 25일
심사완료 : 2012년 1월 25일

최재혁(1996)은 복합명사의 띄어쓰기 오류를 교정하기 위해, 양 방향 최장일치법과 사전들을 이용해 형태소 분석을 하고 음절 수에 따른 복합명사를 처리하는 방법을 제안하였다[1]. 이는 어휘사전에 띄어쓰기 정보를 함께 수록해야 하며 음절 수에 따른 분해 형태들을 정의해야 한다는 단점이 있다.

강승식(1998)은 형태소 분석 결과로 추정된 복합명사를 단위명사들로 분해하는 방법으로 네 개의 분해규칙과 두 가지 예외규칙을 사용하여, 가능한 후보들을 생성하고 가중치를 부여해 최적 후보를 선택하는 알고리즘을 제안하였다[2].

통계데이터를 이용한 복합명사 분해 연구는 자연언어가 실세계에서 사용되는 용례들과 부속 정보를 포함하는 다량의 말뭉치를 분석하여 얻은 통계정보를 이용하여 중의성 문제를 확률적으로 해결하는 방법이다

윤보현(1995)은 복합명사 분석 시 중의성이 발생하는 문제를 해결하기 위해 말뭉치 분석을 통해 통계 정보를 추출하고 이를 이용하는 방법을 제안하였다[3]. 분석 시 사용하는 통계정보는 복합명사 내에서 단위명사가 중심어로 쓰인 빈도와 복합명사 구성 패턴 간의 통계적 선호 규칙을 이용한다.

J.T. Yoon(2001)은 복합명사의 문법적 특성 때문에 명사 간 통계적 연관성에 기반을 둔 분해 오류를 해결하기 위해, 언어와 통계적 지식을 기반으로 한 분해 모델을 제안하였다[4]. 분해 과정을 효율적으로 처리하기 위해 복합명사 내의 관계를 식별하였으며 관계기반의 문장 구조를 분석하였다.

또한, 모든 복합명사를 사전에 등록하는 것이 불가능하고 생성 조건에 제한이 없어 발생하는 신조어나 외래어, 고유명사와 같은 미등록어 처리에 대한 연구들도 수행됐다.

강유환(2004)은 기존 연구에서 미등록어에 대한 미처리나 휴리스틱을 사용하는 경우에 대한 문제점을 해결하기 위해, 단순 분해 위치 결정과 함께 미등록어를 사람 이름, 외래어, 지명 등과 같은 범주정보를 함께 제공하는 방법을 제시하였다[5]. 이를 위해 인명, 외래어, 지명에서 나타나는 음절의 다양한 출현 특성과 실마리 정보를 이용하였다.

강민규(2010)는 구성 명사에 미등록어, 1음절어, 접사 등이 포함된 경우에 발생하는 분해 중의성에 의한 오류를 해결하기 위해 재처리 과정을 통한 분해 오류 교정 기법을 제안하였다[8]. 특정 명사는 특정 위치에 자주 출현한다는 가정을 통해 처음-중간-끝의 형태로 1-gram 위치별 빈도데이터를 사용하고 두 구성 명사의 공기 빈도를 통해 간접적으로 연관성을 판단하기 위한 자료로 bigram을 말뭉치로부터 추출해 임계값에 따른 오류를 판단하고 교정을 수행한다.

이외에도 의미정보를 이용해 분해 중의성을 해결하는 연구도 수행되어 왔다[9,10,11]. 이는 의미 태깅이 아닌 의미정보 데이터를 이용해 분해 중의성을 가지는 명사 간의 의미 결합 가능성을 판단해 분해의 정확도를 높이기 위한 수단으로 제안되었다.

위와 같이 복합명사 분해에서 미등록어나 접사와 같이 분해 중의성을 유발하는 부분들에 대한 연구는 활발히 진행됐고 그 정확도 또한 높다. 하지만 정보의 의미가 중요해짐에 따라 자연어 처리의 여러 분야에서는 정확한 의미 분석이 필요하게 되었다. 의미 분석을 위한 기존연구는 주로 의미 집합을 구성하고 공유 명사의 개수를 이용해 계산된 유사도에 기반을 두어 이루어졌다.

Lesk(1986)는 기계가독형사전을 이용하여 영어로 쓰인 문장 내에서 단어의 의미를 자동으로 결정하기 위해 해당 단어의 중의적 의미들과 문장 내 주변 단어들의 뜻풀이 사이에 공통으로 쓰인 단어의 개수를 이용해 중의성을 해결하는 방법을 제안하였다[15]. 이는 정확한 매칭을 기반으로 수행되므로 자료부족현상(Data Sparseness Problem)이 심하다는 단점이 있다.

Cowie(1992)는 Lesk의 방법을 기반으로 최적화된 계산을 위해 시뮬레이티드 어닐링(simulated annealing)기법을 이용하여 문장 내에 존재하는 모든 어휘에 대해 중의성 해결을 수행하는 방법을 제안하였다[16].

David Yarowsky(1992)는 중의성을 해결하기 위해 Roget 시소러스를 이용한 통계적 의미 분류 기법을 제안하였다[17]. 이는 개념이 다른 어휘의 클래스는 주로 다른 문맥에 나타나고, 다른 어의는 다른 개념적 클래스에 속한다는 경향에 따라 각 범주들을 개념적 클래스로 간주하고 원시 말뭉치로부터 범주별 문맥을 생성한다. 이후, 상호 정보량을 통해 대표 범주 어휘들을 선택하고 중의성을 지닌 어휘를 전역 문맥의 어휘들을 대상으로 베이즈 규칙을 이용해 어휘의 의미를 분류한다.

본 논문에서는 기존 복합명사 분해 연구들에서 널리 사용된 음절별 선호 분해 패턴 방법이 가지는 음절 수 증가에 대한 문제점과 통계기반 방법에서의 구성 명사간 패턴 정보 추출 시 발생하는 자료부족 문제를 해결하고, 한국어 복합명사 내에서 출현하는 명사 구성 형태와 결합 특성을 반영하며 계산 복잡도를 줄이기 위해 위치별 unigram 명사 출현 빈도를 이용한 통계기반 복합명사 분해방법을 제안한다. 또한, 의미 분석에 사용될 후보 선택 시 분해 구조의 최적성을 포함하기 위해 미등록어와 등록어에 대한 가중치를 부여하였다. 이를 이용해 의미 분석 수행 전에 정확도 향상과 계산 횟수를 줄이기 위해 사전의 원어 정보를 이용하였으며, 사전에 등재되지 않은 어휘를 위해 지도학습 알고리즘인 Naive Bayes Classifier를 이용하였다. 분류기의 학습데이터 생성과 의미 분석의 유사도 비교를 위한 의미집단 구성에서 자료부족 문제를 해결하기위해 속성명사 추출을 위한 패턴 규칙을 적용하고 이를 U-WIN으로부터 추출한다.

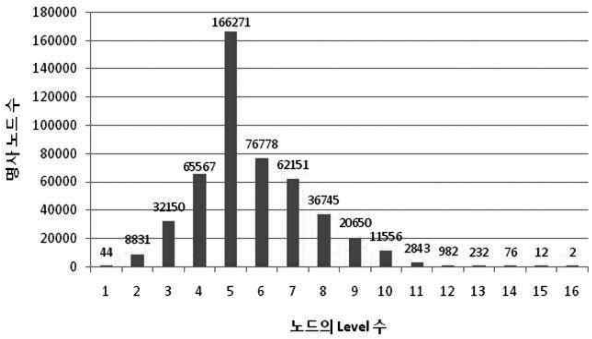
본 논문은 다음과 같이 구성되었다. 2장에서는 본 논문에서 제안한 구조 분해와 의미 분석을 위해 사용된 자원들에 대해 기술하고, 3장에서는 의미 분석의 전처리 과정으로 말뭉치를 이용한 통계기반의 복합명사 구조 분석 방법에 대해 제시한다. 4장에서는 구성 명사 의미 범위 축소를 위한 의미 범위 제약에 대해, 5장에서는 유사도 분석 결과에 따른 의미 태깅을 수행하는 의미 분석에 대해 제시한다. 6장에서는 제안된 시스템에 대한 성능을 평가하기 위한 실험을 수행한다. 마지막으로 7장에서는 본 논문의 결론과 앞으로의 연구 방향에 대해 논의한다.

2. 자 원

본 논문에서 제시하는 복합명사 의미 태깅 시스템은 복합명사의 구조 분해 및 의미 분석의 전반적인 부분에서 U-WIN, 고유명사 사전, 외래어 사전, bigram 빈도 사전 등을 이용한다. 이 장에서는 이러한 자원들을 설명한다.

2.1 U-WIN (User-Word Intelligent Network)[1)]

U-WIN은 한국어의 공통적이고 개별적인 속성을 바탕으로 한국인의 보편적인 인지 체계와 개념 관계를 파악하여 이를 어휘의 의미적/개념적 네트워크를 형성한 온톨로지적 의미망이라 할 수 있다[18]. 2012년 현재 48만여 명사에 대해 구축된 상태이다.

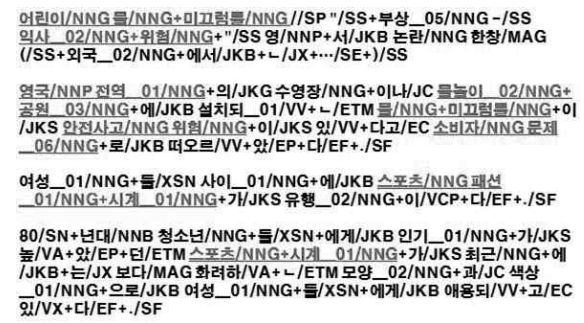


(그림 1) U-WIN의 level별 명사 노드 개수

2.2 위치별 명사 빈도 데이터

복합명사를 구성하는 명사들은 그 구조 내에서 특정 위치에 주로 나타나는 경향이 있다[8]. 이는 복합명사의 분해 및 합성 과정에서 발생하는 구조에 따른 결합 위치의 정보로서, 이를 이용한다면 위치에 따른 명사들의 확률을 부여하기 위한 좋은 자료가 될 수 있다.

위치별 명사 빈도 데이터는 세종말뭉치와 사전 뜻풀이에서 (그림 2)와 같은 연속적인 명사열을 "처음-중간-끝"의 위치별로 추출하였다. 그 결과 총 167,549개의 명사의 위치별 빈도 데이터를 얻었으며, <표 1>과 같이 특정명사는 주로 출현하는 위치가 다르며 그 빈도차가 큰 것을 알 수 있다. 예를 들어, 명사 "국제"는 처음 위치에는 412번의 높은 빈도로 출현했지만 명사열의 끝에는 한 번도 나오지 않은 것을 알 수 있다. 따라서 복합명사 분해 시 나뉜 후보 중 끝 부분에 "국제"가 나오는 경우는 잘못된 분해라는 것을 점수에 반영할 수 있다.


어린이/NNG 를/NNG+미끄럼틀/NNG //SP "/SS+부상__05/NNG -/SS
익사__02/NNG+위험/NNG+"/SS 영/NNP+서/JKB 논란/NNG 한창/MAG
(/SS+외국__02/NNG+에서/JKB+ㄴ/JX+…/SE+)/SS

영국/NNP 전역__01/NNG+의/JKG 수영장/NNG+이나/JC 를놀이__02/NNG+
공원__03/NNG+에/JKB 설치되__01/VV+ㄴ/ETM 틀/NNG+미끄럼틀/NNG+이
/JKS 안전사고/NNG 위험/NNG+이/JKS 있/VV+다고/EC 소비자/NNG 문제
__06/NNG+로/JKB 떠오르/VV+았/EP+다/EF+./SF

여성__01/NNG+들/XSN 사이__01/NNG+에/JKB 스포츠/NNG 패션
__01/NNG+시계__01/NNG+가/JKS 유행__02/NNG+이/VCP+다/EF+./SF

80/SN+년대/NNB 청소년/NNG+들/XSN+에게/JKB 인기__01/NNG+가/JKS
높/VA+았/EP+던/ETM 스포츠/NNG+시계__01/NNG+가/JKS 최근/NNG+에
/JKB+는/JX 보다/MAG 화려하/VA+ㄴ/ETM 모양__02/NNG+과/JC 색상
__01/NNG+으로/JKB 여성__01/NNG+들/XSN+에게/JKB 애용되/VV+고/EC
있/VX+다/EF+./SF

(그림 2) 말뭉치에서의 명사열 추출 과정

<표 1> 추출된 위치별 명사 출현 빈도의 예

명사	처음	중간	끝	총
나무	90	0	581	671
국제	412	17	0	429
운동	58	15	332	405
머리	79	3	323	405
제도	23	4	364	391
식물	79	3	297	379
자기	291	26	843	360
전기	266	25	39	330

2.3 외래어, 고유명사, 명사 bigram 사전

복합명사 분해 시 나타나는 분해 중의성은 미등록어나 외래어 등에 의해 나타나는 경우가 많다. 이에 따라 본 시스템에서는 분해정보로 활용할 수 있는 사전을 구축해 오류를 최소화한다.

외래어는 대부분 긴 음절로 이루어지므로 분해 중의성을 띈 후보에서 구성 명사가 한국어 명사로 분해되어 문제를 유발한다. <표 2>는 사전에서 원어 정보에 알파벳을 포함하는 명사들의 음절별 어휘수를 나타낸다. 2음절 이하는 구성 명사로 나뉠 확률이 희박하므로 3음절 이상을 대상으로 외래어 사전을 구성한다면 전체의 80.14%를 이용하여 오분해된 외래어의 복원을 시도할 수 있다. 외래어 사전은 표제어가 원어 정보로 알파벳을 포함한 경우를 전자사전으로부터 추출해 구축하였으며 총 8,327개의 데이터를 얻을 수 있었다.

고유명사 사전은 재분해 대상의 판별을 위해 이용되며 태깅된 말뭉치와 사전의 뜻풀이에서 4음절 이상의 고유명사만을 대상으로 구축하였으며 총 26,484개의 데이터를 얻을 수 있었다.

<표 2> 추출한 외래어명사의 음절별 개수

음절수	외래어 개수	비율(%)
1	165	1.59
2	1,899	18.28
3	2,946	28.35
4	2,562	24.66
5	1,642	15.80
6	773	7.44
7	289	2.78
8음절 이상	115	1.11
총	10,391	100.00

1) U-WIN 관련 Demo 페이지 : http://klplab.ulsan.ac.kr/Demo/ProjectDemo.php

명사 bigram 사전은 복합명사의 분해 후보 중 의미 분석을 위한 최적 후보 선택 시 가장 중요한 끝 두 어절의 분해 정확도에 따라 전체 오분해율이 감소하는 현상을 반영하기 위해 사용되는 자료이다. 명사 bigram 사전에 끝 두 어절이 존재하면 가장 확률이 높은 그 후보를 최적으로 선택해 오분해를 감소시키며, 이를 위해 인접한 bigram의 명사 출현 정보를 말뭉치로부터 추출해 구축한 결과 총 585,940개의 데이터를 얻을 수 있었다.

〈표 3〉 외래어, 고유명사, 명사 bigram 사전의 예

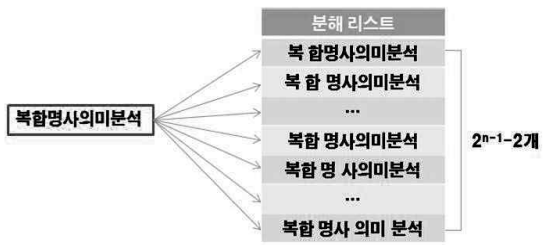
외래어 사전	고유명사 사전	명사 bigram 사전
틱소트로피	국방과학연구소	문예 창작
프리웨어	울산대학교	각종 풍속
실리콘	한국어정보학회	해당 의사
인디케이터	국제앰네스티	제작 지원국
…		

3. 복합명사 구조 분해

입력인 n음절 복합명사는 띄어쓰기가 되어 있지 않은 하나의 어절로 주어진다. 이는 의미 분석의 수행을 위해 의미 단위가 될 구성 명사로 나뉠 필요가 있다. 이 장에서는 의미 분석의 전처리 과정으로 수행되는 구조적 분해 방법과 학습 말뭉치에서 추출한 통계데이터에 의해 발생할 수 있는 오분해에 대해 후처리를 하는 방법에 대해 제시한다.

3.1 위치별 구성 명사 확률 부여

n 음절의 복합명사가 입력으로 주어졌을 때 분해 가능한 총 후보는 (그림 3)과 같이 모두 1음절로 나누어지는 경우와 분해되지 않은 경우를 제외한 총 $2^{n-1}-2$가지이다.



(그림 3) 가능한 모든 분해 후보

본 시스템은 입력 대상의 음절별 제한을 없애기 위해 위와 같이 분해 대상을 모든 후보로 삼고 입력 음절의 제한을 두지 않는다. 이를 위해 2장에서 언급한 위치별 명사 빈도를 사용한다. 모든 경우의 수로 나뉜 복합명사 분해 리스트 각 후보에 위치별 명사 빈도를 이용하여 확률과 가중치에 따른 순위를 매기고 이를 의미 분석의 후보 리스트로 사용한다. 구성 명사의 확률 부여 방식은 크게 위치별 빈도가 존재하는 경우와 그렇지 않은 경우의 두 가지로 나눈다.

$$L=\{L_b, L_m, L_e\}$$

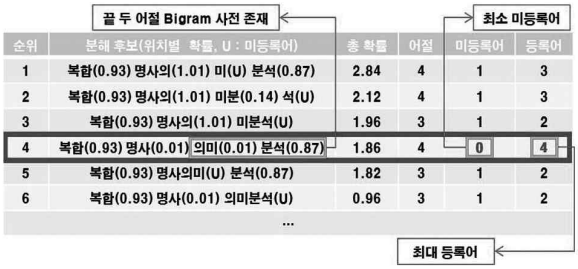
$$P(n)=\begin{cases} freq(n,L)/freq(n,T)+w_k, & \text{if } freq(n,L)>0 \\ w_u, & otherwise \end{cases} \quad (1)$$

명사의 위치별 빈도를 나타낼 L은 처음, 중간, 끝의 위치인 L_b, L_m, L_e로 이루어진다. 수식 (1)에서 명사 n의 위치별 확률 $P(n)$은 위치별 빈도 $freq(n,L)$이 존재하는 경우, 구성 명사가 해당 위치에 나타날 빈도 $freq(n,L)$을 명사의 전체 빈도 $freq(n,T)$로 나누어 해당 명사가 나타난 위치에 대한 출현 확률을 계산하고 등록어에 대한 가중치 w_k을 더하여 확률을 부여한다. 이에 반해 위치별 빈도가 존재하지 않는 경우는 구성 명사가 해당 위치에 출현하기에 적절하지 않음을 의미하므로 미량의 음수 값 미등록어 가중치 w_u를 부여해, 총 확률이 낮아지게 만들고 이로 인해 최적 후보 선택에서 멀어지도록 한다.

3.2 최적 후보의 선택

앞에서 분해된 후보 리스트에서 총 확률이 가장 높은 후보는 말뭉치에서 추출한 빈도가 의미하는 정답이지만 의미 태깅을 위한 유사도 분석 단위는 U-WIN 사전에 기반을 두므로 항상 정답이 될 수 없다. 따라서 구성 명사는 의미를 비교하기 위해 사전에 존재하는 표제어 형태로 분해되어야 한다. 최종 후보의 선택은 분해 후보들의 끝 두 어절, 어절 수, 미등록어 수, 등록어의 수와 같은 네 가지 조건을 이용하며 (그림 4)와 같이 확률 순으로 정렬된 리스트에서 1순위로 끝 두 어절 bigram의 사전 존재 여부, 2순위로 최소 미등록어 수, 3순위로 최소 어절 수, 4순위로 사전상 등록어 수의 최대를 만족하는 후보를 우선순위를 적용해 최적 후보로 선택한다. 여기서 1음절은 접두사나 접미사일 경우가 대부분이며 사전에 동형이의어 및 다의어의 개수가 많으므로 의미 분석 시 오분석의 확률과 분석 시간이 늘어난다. 따라서 이를 방지하기 위해 미등록어로 처리한다.

(그림 4)의 경우 가장 높은 확률을 가진 "복합-명사의-미-분석"은 말뭉치의 오류인 "명사의" 때문에 전체적인 확률이 높다. 이러한 이유는 등록어의 확률 계산공식에 의해 말뭉치에서 "명사의"가 총 한번 나왔으며 그 위치가 중간이기 때문이다. 이러한 경우도 최종 후보로서 배제되어야 한다.

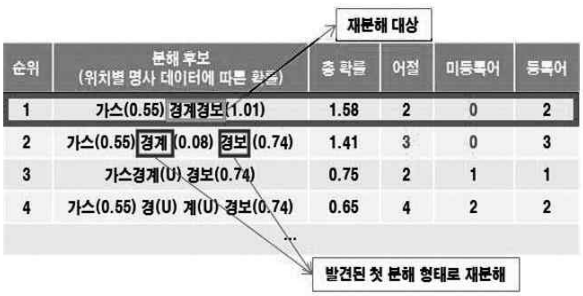


(그림 4) 의미 분석 후보 선택

위에서 언급한 네 가지 선택 조건을 순위별로 적용해 가장 만족하는 후보를 탐색한다. 미등록어 수가 가장 작은 경우는 유사도 비교를 할 수 있는 최적의 후보라는 의미이며, 어절 수와 등록어 수가 가장 많은 경우는 구조분해가 충분히 이루어졌음을 의미한다. (그림 4)의 경우 4가지 조건의 1순위인 끝 두 어절 bigram을 탐색해 본 결과 후보 1~3순위의 bigram인 "미 분석", "미분 석", "명사의 미분석"은 존재하지 않았다. 하지만 후보 4순위 "의미 분석"은 해당 사전에 존재하며 이를 만족하는 가장 높은 후보 순위이므로 이를 의미 분석을 위한 후보로써 최종 선택하게 된다. 만약 조건 1순위인 bigram "의미 분석"이 사전에 존재하지 않는다면 2순위 조건인 최소 미등록어 수를 탐색하여 이를 만족하는 후보 4순위가 최종으로 선택되게 된다. 만약 후보 4순위 이하에서 같은 최소 미등록어를 만족하는 경우가 2개 이상일 때, 3순위 조건인 최소 어절 수를 비교하고 또한 이 후보가 2개 이상인 경우 마지막으로 4순위 조건인 최대 등록어 수를 적용하게 된다. 최종적으로 선택된 후보 중 이를 가장 만족하는 최대 확률의 후보가 의미 분석을 위해 최종적으로 선택되게 된다.

3.3 분석 후보 재분해

말뭉치에서 추출한 위치별 명사 빈도 리스트에는 하나의 명사로 태깅된 복합명사들이 존재한다. 이러한 복합명사들은 단일명사보다 전체 및 위치별 출현 빈도가 낮음에도 높은 위치별 확률을 가지는 특성이 있다. 본 시스템에서는 구성 명사가 복합명사일 때 최소한의 단위로 나누어 유사도 비교 시 정확도 향상을 위한 비교 대상의 확장을 위해 재분해를 시행한다. 재분해는 구성 명사가 외래어와 고유명사가 아닌 4음절 이상으로 이루어진 1어절일 경우 이를 대상으로 수행하며, 최종 후보에 구성 명사로 재분해 대상이 포함된 경우를 확률 순으로 해당 미분해 명사가 미등록어를 포함하지 않는 가장 높은 확률의 분해된 형태로 재분해를 수행한다.



(그림 5) 재분해 대상의 탐색

(그림 5)는 복합명사 "가스-경계경보"에 대한 재분해 처리 과정을 나타낸 것이며 확률 선택에 의해 1순위인 "가스-경계경보"가 선택되었다. 하지만 "경계경보"가 말뭉치에서 총 1번 출현했으며 그 위치가 끝이므로 높은 확률이 부여되었다. 따라서 "경계경보"가 4음절의 외래어와 고유명사가 아닌 재분해 대상이므로 재분해를 수행하며, 미등록어가 존재하지 않고 최종적으로 세 가지 조건을 가장 만족하며 재분해 대상이 없는 후보인 "가스-경계-경보"가 선택되게 된다.

3.4 오분해된 외래어 복원

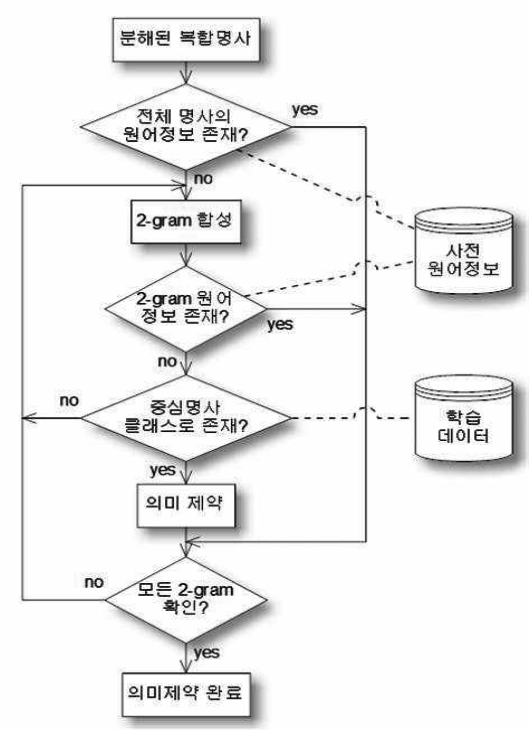
외래어 명사의 음절은 대부분 한국어 일반명사보다 길다. 따라서 외래어가 분해된 형태로 존재하는 후보의 구성 명사가 위치별 명사 빈도 리스트에 존재하는 경우 구성 명사의 개수가 많을수록 외래어로 계산된 확률보다 높아 오분해될 가능성이 많다. 이에 따라 본 시스템에서는 사전의 원어 정보를 포함한 표제어를 수집하여 외래어 사전을 구축하고 이를 이용해 오분해된 외래어의 복원을 수행한다.

(그림 6) 오분해된 외래어 복원 대상의 탐색

(그림 6)의 경우 정답인 "프로그램-유도"의 후보는 미등록어가 없는데도 명사 "프로"와 "그램"이 위치별 명사 리스트에 존재하여 그 계산된 총 확률값이 "프로-그램-유도"로 오분해된 후보보다 작다. 따라서 미등록어가 없고 총 확률이 가장 높은 "프로-그램-유도"가 최종으로 선택되게 된다. 외래어 복원의 단위는 오분해된 외래어의 음절이 대부분 길어서 최적 후보를 선택한 뒤 각 구성 명사에서 우방향으로 구성 명사들을 차례로 누적 합성하여 외래어 사전에서 찾으며 해당 어휘가 존재하는 경우 원형태로 복원한다.

4. 복합 후보 의미 범위 제약

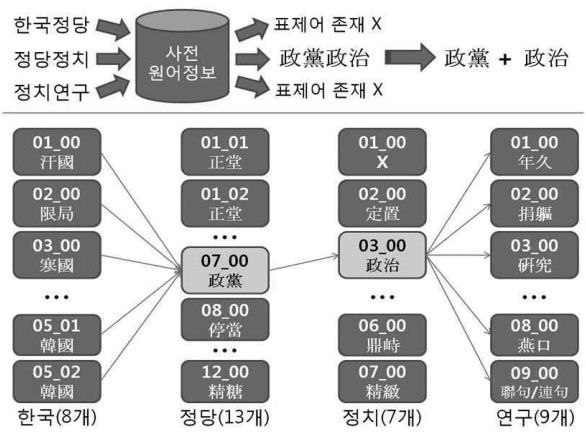
U-WIN의 개념 노드는 다의어 수준으로 구축되어 있어 관계정보를 이용한 명사의 추출 범위, 유사도 측정 대상, 의미체인의 결합수가 방대해진다. 따라서 유사도 측정 시 특정 bigram은 정답과 다른 의미의 높은 유사도 때문에 그 의미가 변질하여 정확도가 낮아질 수 있다. 따라서 의미 분석 수행 전에 각 구성 명사의 의미를 관련된 자원들을 이용하여, 정답이 될 수 있는 특정 의미의 유효 후보들로 제약할 방법이 필요하다. 이는 의미 조합의 복잡도와 유사도 비교에 소요되는 계산 시간을 줄이기 위한 중요한 과정이며 이에 따라 정확도 역시 영향을 받는다. 본 시스템에서는 구성 명사 bigram을 대상으로 원어 정보와 Naive Bayes Classifier를 이용한 의미 범위 제약을 시도한다. (그림 7)은 의미 범위 제약의 알고리즘 순서도이다.



(그림 7) 의미범위제약 순서도

4.1 사전 원어 정보를 이용한 의미제약

한국어 어휘에는 고유어, 한자어, 외래어와 그것들의 혼용어가 있는데 한자어는 그중에서 가장 큰 비중을 차지한다[19]. 또한, 2개 이상의 구성 명사를 가지는 복합명사는 구성 명사가 2개인 복합명사의 조합에 의해 만들어진 경우가 많다. 따라서 bigram을 기본단위로 사전에서 원어 정보를 가져올 수 있는 경우 그 의미의 범위를 해당 한자를 지닌 동형이의어들로 제약할 수 있다. (그림 8)의 복합명사 "한국정당정치연구"의 경우 최적의 분해 후보로 "2+2+2+2"의 패턴으로 분해가 되었으며 이에 따른 의미 집단의 유사도 비교의 총 횟수는 <표 4>와 같이 $8\times13+8\times7+8\times9+13\times7+13\times9+7\times9 = 503$ 번 이 며 ,



(그림 8) 의미범위 축소를 위한 원어 정보 사용의 예

의미조합이 가능한 결합의 경우의 수는 $8\times13\times7\times9=6552$ 가지나 생성 가능하다.

하지만 "정당정치"가 사전에 존재하고 원어 정보 "政黨政治"도 가지고 있으므로 "정당"과 "정치"는 원어 정보에 따라 "정당_07"과 "정치_03"으로 의미 제약이 되며 유사도 비교 수는 $8\times1+8\times1+8\times9+1\times1+1\times9+1\times9 = 107$ 번, 의미 결합수는 $8\times1\times1\times9=72$가지로 감소하게 된다. 따라서 원어 정보의 이용이 의미제약에 효과적임을 알 수 있다.

〈표 4〉 의미범위 축소에 따른 연산 횟수 변화

	bigram 유사도 비교	의미결합
범위 축소전	503	6552
범위 축소후	107	72

4.2 Naive Bayes Classifier를 이용한 의미제약

원어 정보를 이용한 의미제약에 성공한다면 정확히 의미를 제약할 수 있다는 장점이 있는 반면 사전에 의존적이므로, 다음과 같이 제약에 실패하는 여러 경우들이 있다.

① bigram의 미등재
② 원어 정보의 부재
③ 한자 코드의 차이
④ 해당 구성 명사와 bigram의 원어 정보 표기 차이

특히, 세 번째 경우는 원어 정보가 한자(漢字)일 때 1음절의 한자가 2개 이상의 다른 코드로 기술된 경우가 다소 있어 합성된 bigram의 원어 정보를 가진 구성 명사가 있는데도 코드 비교의 불가능으로 제약에 실패한다. 복합명사 "누적기록(累積記錄)"의 경우 "누적"은 2개의 동형이의어가 존재하며 그중, "누적_01(累積)"의 원어 정보가 bigram "누적기록"의 그것과 같으므로 "누적_01"로 의미 제약이 되어야 하지만 사전 편찬 시 다른 코드의 "누(累)"를 사용하였으므로 제약에 실패하였다. 또한, 네 번째 경우는 특히 라틴어 계열(영어권 언어)에서 발생하는 문제로 bigram의 구성 명사 원어 정보가 약어로 기술되는 경우가 많아 비교할 수 없는 경우도 많이 있다. 복합명사 "루이스산"의 경우 원어 정보는 "Lewis酸"이며 구성 명사 "루이스"는 사전에 <표 5>와 같이 8개의 동형이의어가 등재되어 있다.

〈표 5〉 표제어 "루이스"의 동형이의어별 원어정보

표제어	원어 정보
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

따라서 "Lewis"만을 이용해 의미제약을 시도하면 일치하는 원어 정보가 없으므로 제약을 할 수 없게 된다. 이처럼 원어 정보를 이용한 의미제약 방법은 사전에 전적으로 의존적인 단점이 있다. 따라서 이처럼 원어 정보에 의한 방법으로 제약에 실패한 경우 추가로 Naive Bayes Classifier를 이용해 의미 제약을 시도한다[20]. 수식 (2)의 $P(c \mid N_p)$는 속성값 N_p가 주어졌을 때 이 인스턴스가 클래스 c로 분류될 확률을 나타낸다. 따라서 이를 이용해 관측 데이터 셋이 속할 가장 유사한 클래스를 찾아 분류하는 문제로 가정해 의미제약에 적용할 수 있다. 속성값 N_p는 n개의 속성명사로 이루어진 다의어 단위의 의미집합을 의미하며 수식 (3)과 같이 학습데이터에서의 동형이의어 및 다의어 클래스 개수에 대한 사전확률(Prior) $P(c)$을 구하고, 독립가정에 의해 속성명사들이 해당 클래스에 속할 우도(Likelihood) $P(n_i \mid c)$를 계산해 최종적으로 이를 곱한 사후확률(Posteriori)을 구한다. 이렇게 각 클래스와 의미집합에 대해 구한 확률 중, 최대사후확률(Maximum a Posteriori)을 만족하는 클래스 c로 해당 명사를 분류한다. 이때 우도를 구하는 과정에서 클래스 내부에 특정 속성명사가 출현하지 않은 경우 조건부 확률이 0이 될 수 있으며, 이는 각 확률을 곱하게 되므로 전체 값이 0이 되게 된다. 이를 해결하기 위해 간단한 smoothing 기법으로 클래스 C에서의 속성명사 출현빈도가 0인 경우, $P(n_i \mid c)$을 $P(c)/N$으로 대체해 작은 값을 부여한다. N은 총 학습데이터의 수를 의미한다[21].

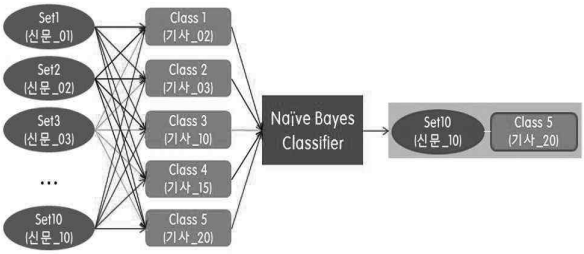
$$N_p = n_1, n_2, \ldots, n_{n-1}, n_n$$

$$P(c \mid N_P) = \frac{P(N_p \mid c)P(c)}{P(N_p)} \quad (2)$$

$$Classify(N_p) = \arg max_{c} P(C=c) \prod_{i=1}^{n} P(N_p = n_i \mid C=c) \quad (3)$$

입력된 bigram의 중심명사가 학습데이터의 클래스로 존재하는 경우 Naive Bayes Classifier에 의해 앞 명사의 의미집합들을 비교하여 의미제약을 시도하며 이는 모든 bigram을 대상으로 수행한다. (그림 9)는 이러한 의미제약 과정의 한 예이다. 앞 명사 "신문"의 모든 의미집합들을 학습데이터에 존재하는 뒷 명사 클래스 '기사'에 대해 사후 확률을 구한다. 그 결과 가장 높은 확률을 가지는 클래스인 "기사_20"에 "신문_10"이 분류 되었다.

의미제약을 수행하기 위해 우선 학습 과정이 필요하다. 빈도가 높은 복합명사는 각 구성 명사가 의미가 유사하고 그 의미를 제약한다는 가정을 가진다. 따라서 학습을 위한 데이터는 이를 만족하는 bigram의 의미가 제약될 확률이 높아진다는 정보를 가진다. 의미 태깅된 기분석 사전으로부터 가장 끝 어절 명사인 중심명사를 클래스로 정의하고 앞 명사와 관련된 속성명사들을 U-WIN의 관계정보와 사전을 통

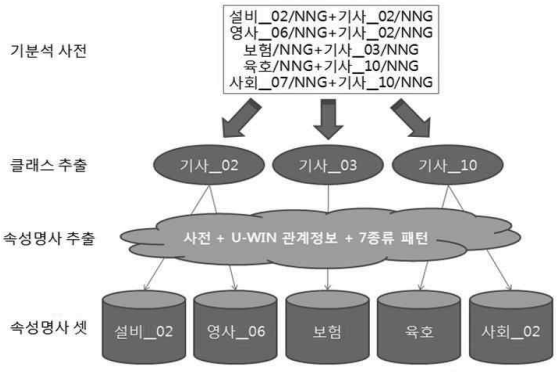


(그림 9) Naive Bayes Classifier를 이용한 "신문기사"의 의미제약 과정

해 다음의 7가지 종류로부터 추출하여 집합을 구성한다면 관측데이터의 자료부족 문제를 해결 할 수 있다.[22].

- 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"류인 경우 그 대상명사들의 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

또한, 이는 5.2절의 유사도 분석을 위한 표제어별 의미 집단의 벡터 구성과정에서 같은 문제를 해결하기 위해 속성명사 추출 시 적용한다. (그림 10)은 Naive Bayes Classifier를 위한 학습데이터의 생성과정을 나타낸다.



(그림 10) 학습데이터 생성과정

5. 복합명사 의미 분석

의미 분석은 본 논문의 궁극적 목표인 의미 태깅을 위한 과정이며, 제안하는 의미 태깅 방법은 bigram 간 유사도에 기반을 두므로 유사도 비교 알고리즘에 영향을 받는다. 이 장에서는 유사도 비교에 대한 과정과 이를 이용한 의미 결정방법에 대해 설명한다.

5.1 개념 노드 벡터 생성

의미 태깅을 위한 유사도 비교 방법은 Lesk에 의해 제안된 매칭 기반 방법을 이용한다[15]. 이 방법은 어휘의 의미를 결정하기 위해, 중의성 어휘들의 의미별 뜻풀이와 해당 어휘가 나타난 문맥 내의 어휘들의 뜻풀이에서 추출한 어휘 간의 공통된 개수를 이용하여 의미를 분별한다. 이 방법은 구현이 쉽고 다른 자원을 요구하지 않는다는 장점이 있으나 정확한 매칭에 기반을 두므로 자료부족 문제가 존재한다는 단점이 있다. 더욱이 한국어 사전은 뜻풀이가 매우 짧게 나타나는 경우가 많아 의미 관계 정보 추출에 한계가 있다. 따라서 본 시스템에서는 이를 해결하기 위해 어휘별 의미집단을 4.2절에서 정의한 7가지 종류의 규칙을 적용해 추출한다.

U-WIN의 개념노드는 다의어 수준으로 구축되어 있으므로 이를 단위로 하여 벡터를 생성하며 추출 대상은 일반명사와 고유명사로 각 명사의 출현빈도도 포함한다. 추출에 적용하는 7가지 규칙 중 1차 하위어는 표제어에서 나타나지 않거나 추상적인 경우를 위함이며, 1차로 한정한 이유는 2차 이상은 개수가 많으면 그 대표 명사의 의미가 변질할 수 있기 때문이다. 상위어들은 표제어와 1차 하위어에서 공유하는 개념이 없거나 하위어가 없는 경우 체인합성을 이용한 태깅 시, 실제 정답이 낮은 확률의 bigram일 때 미량의 확률을 부여해 이 역시 후보로 사용하기 위함이다.

동의어는 유사도의 계산방식이 정확한 매칭에 기반을 두므로 비슷한 의미이지만 형태가 달라 의미가 유사하더라도 그 값이 낮은 경우들이 있다. 이를 위해 동의어 관계 존재 시 그 뜻풀이도 포함한다. <표 6>과 같은 표제어는 뜻풀이에 동의어만을 명시하는 때가 있어 추출 범위가 더욱 한정적이다. 이와 같은 패턴이 뜻풀이에 있는 경우 패턴 명사에 관련 뜻풀이가 존재하므로 패턴 명사의 뜻풀이에서 명사류를 추출한다.

〈표 6〉 뜻풀이로 동의어를 가지는 표제어 "정당"

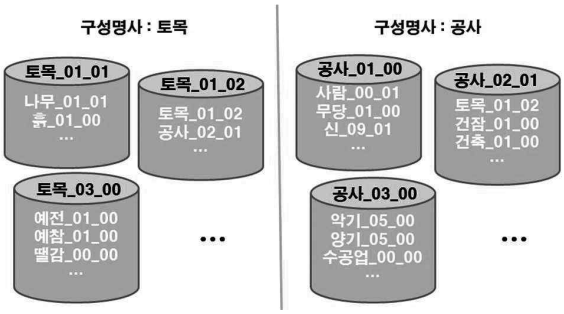
표제어	뜻풀이	동의어의 뜻풀이
정당_02	*주지09*(住持)(2).	절을 주관하는 중.
정당_03	*정전06*(正殿).	왕이 나와서 조회(朝會)를 하던 궁전. 경복궁의 근정전 창덕궁의 인정전이 있다.

<표 7>은 "~이르는 말"의 한 패턴인 "~아울러 이르는 말"의 예인 "당정_01"과 대상 명사인 "정당_07", "정부_08"의 뜻풀이이다. "당정_01"의 뜻풀이에서 {정당, 정부, 여당} 등의 명사가 추출 가능하나 개수와 그 빈도가 유사도 측정에 불충분하므로 대상 명사의 뜻풀이에서도 명사류의 추출을 시도해 {정치, 주의, 주장, 사람, 정권, 이상, 단체, 행정부} 등을 추가하여 자료부족 현상을 일부 해결할 수 있다. 이 밖에도 표준국어대사전의 "일러두기"에 존재하는 뜻풀이 패턴인 방언, 비표준어, 옛말, 외래어, 순화어, 음역어 등에 대해서도 같은 형식으로 추출한다.

〈표 7〉 "~아울러 이르는 말"의 예와 대상명사의 뜻풀이

표제어	뜻풀이
당정_01	*정당과 정부를 아울러 이르는 말.* 흔히 정당 중에서도 여당과 정부를 이르는 말.
정당_07	정치적인 주의나 주장이 같은 사람들이 정권을 잡고 정치적 이상을 실현하기 위하여 조직한 단체.
정부_08	행정부.

7가지 규칙을 적용해 추출한 다의어 벡터는 (그림 11)과 같이 하나의 벡터가 될 의미 집단으로 구성된다.



(그림 11) 개념 노드 벡터 구성도

5.2 bigram간 유사도 분석

생성된 벡터들을 대상으로 유사도를 구하기 위한 대표적인 유사도 측정 알고리즘으로는 코사인 계수(Cosine Coefficient), 다이스 계수(Dice's Coefficient), 중복도 계수(Overlap Coefficient), 자카드 계수(Jaccard's Coefficient) 등이 있다. 이들 4종류를 대상으로 본 시스템에 적용해 본 결과 성능에 큰 차이가 없었으나 가장 좋은 성능을 보인 자카드 계수를 사용한다.

자카드 계수는 두 개의 집합이 있을 때, 교집합의 크기를 전체집합의 크기로 나눈 것으로 정의되며, 벡터가 n개의 이진 속성으로 이루어진 경우 이 유사도 공식은 수식 (4)과 같이 간단한 단어의 출현 형태만으로 계산할 수 있게 된다. 수식 (4)에서 P(A, B)는 벡터 A와 B에 같은 어휘가 동시에 나타나는 경우, P(A, ¬B)는 벡터 A에만 해당 어휘가 존재하는 경우, P(¬A, B)는 그 반대의 경우의 어휘 수를 의미한다.

$$Jaccard(A,B) = P(A \cap B)/P(A \cup B) = \frac{P(A,B)}{P(A,B)+P(A,\neg B)+P(\neg A,B)} \quad (4)$$

5.3 가중치 부여

유사도 비교 시 의미 정보를 충분히 반영하지 못한 매칭 기반의 유사도 비교 결과에 따라 두 벡터에 대해 교집합의 값이 클수록 두 벡터 간의 유사도 값은 커지나, 공유 명사의 개수만을 이용하므로 속성명사의 분포에 따라 의도치 않은 bigram이 높은 유사도를 가질 수 있다. 이를 완화하기 위해 추가로 각 벡터의 특성을 이용하여 bigram 간의 의미

관계 중요도, 동형이의어의 어휘 대표성, 사용빈도에 따른 의미의 세분화에 대한 세 가지 조건을 가중치로 유사도에 더하여 단순 매칭에 따라 발생하는 문제점을 해결한다.

$$A = aw_1, aw_2, \ldots, aw_{n-1}, aw_n$$
$$B = bw_1, bw_2, \ldots, bw_{m-1}, bw_m$$

A와 B는 유사도 비교의 대상이 될 비대칭 이진 속성(asymmetric binary attribute) 벡터로, 이에 대한 첫 번째 가중치는 서로의 관계에 대한 가중치 W_r이며 이는 수식 (5)와 같다.

$$W_r = freq(A, bw) \times \alpha + freq(B, aw) \times \alpha \quad (5)$$

aw와 bw는 각각 A와 B명사의 속성명사를 의미하는 것으로 관계정보를 이용해 추출한 속성명사 속에서 직접적으로 의미 태깅된 비교대상 명사가 나왔을 때 즉, 뜻풀이에서 비교 대상 명사를 직접 언급한 횟수가 많을 때는 그렇지 않았을 때 보다 의미상으로 더 유사하다고 볼 수 있다. <표 7>에서 (a)의 "토목_01_02"는 "공사_02_01"의 속성명사 리스트에서 14번, "공사_02_01"은 "토목_01_02"의 속성명사 리스트에서 2번이 나왔으나 (b)는 "토목_01_02"만 "공사_07"에서 1번 나왔으므로 의미상으로 (a)에 비해 그 유사성이 낮다고 볼 수 있다. 따라서 수식 (5)를 계산해 가중치를 부여한다.

두 번째 가중치는 대표성에 관한 가중치 W_s 이며 이는 수식 (6)와 같다.

$$W_s = freq(A, aw) \times \beta + freq(B, bw) \times \beta \quad (6)$$

이는 하나의 표제어를 다의어 단위의 벡터로 구성할 때 얼마나 대표적으로 쓰이는지를 반영하기 위한 가중치이다. 관계정보에 의해 추출된 속성명사 속에서 직접적으로 의미 태깅된 대표 명사가 나오면 해당 명사는 포괄적인 개념의 의미로 여러 동형이의어 중 일반적으로 많이 사용되는 개념을 의미한다. <표 8>에서 (a)와 (b)에 둘 다 "토목_01_02"가 고루 사용되고 있다. 또한 총유사도 비교 리스트에서 "토목"의 다른 여러 동형이의어 및 다의어는 한 번도 나오지 않은 것으로 보아 대표성을 지닌다고 볼 수 있다. 따라서 수식 (6)를 계산해 가중치를 부여한다. W_r의 α는 서로 다른 의미 집단의 관계에 대한 가중치이므로 β보다 충분히 큰 값을 부여한다.

세 번째 가중치는 사용빈도에 따른 의미의 세분화에 대한 다의어 가중치 W_p이며 이는 수식 (7)과 같다.

$$W_p = (PolyCount(A_h) + PolyCount(B_h)) \times \gamma \quad (7)$$

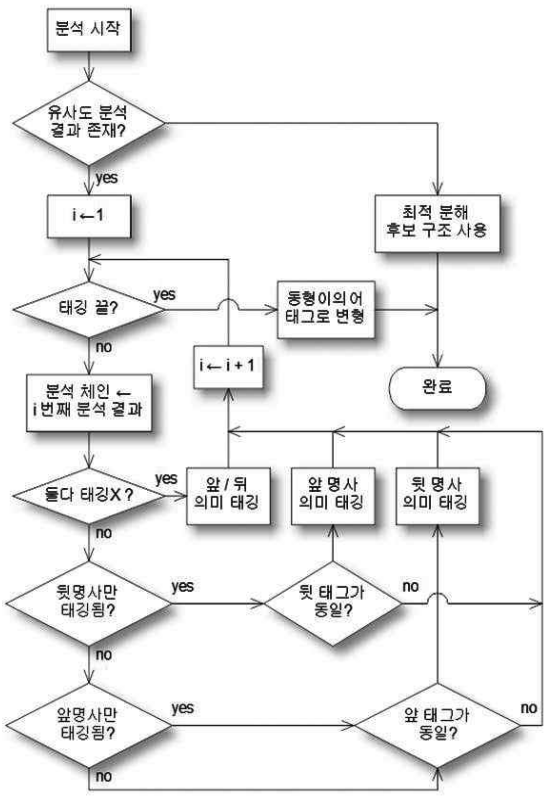
수식 (7)에서 A_h와 B_h는 각각 다의어 벡터 A와 B의 동형이의어를 의미하며 $PolyCount(A_h)$는 동형이의어 어휘 A_h의 다의어 개수를 의미한다. 일반적으로 한 어휘에 여러 개의 동형이의어가 존재할 때, 사용빈도가 높은 경우 다의어의 개수가 다른 동형이의어보다 같거나 많다. 그 이유는 다의어는 하나의 기본적인 의미가 확대되어 여러 뜻으로 쓰이는 단어이므로 다른 동형이의어에 비해 실제 사용빈도가 더 높아 세분되었기 때문이다. 따라서 수식 (7)의 가중치를 부여한다.

〈표 8〉 분석된 bigram 의미 체인의 공유 명사 및 빈도 리스트

bigram 체인	공유 명사 및 출현 빈도 (첫번째 명사/두번째 명사)
(a) 토목_01_02 / 공사_02_01	성과_01(1/1) 시간_04_01(1/2) 동안_01_01(1/3) 사람_00_07(1/1) 건축_01(2/10) 머리_01_02(1/2) 활동_02_01(2/4) 몸_01_01(1/2) 대상_11_01(1/2) 장소_05(1/2) 의지_06_01(1/1) 토목_01_02(4/14) 일_01_01(3/6) 공사_02_01(2/61)
(b) 토목_01_02 / 공사_07	사람_00_07(1/1) 건축_01(2/1) 활동_02_01(2/1) 토목_01_02(4/1) 일_01_01(3/3) 공사_02_01(2/3)

5.4 복합명사 의미 태깅

본 시스템에서의 의미 태깅은 의미 유사도 분석결과와 가중치를 적용한 최종 유사도 값을 기반으로 수행된다. 가장



(그림 12) 의미 태깅 알고리즘 순서도

순위	2-gram 유사도 결과		단계별 태깅 수행 결과
	앞 명사	뒷 명사	서울 / 경인 / 사무 / 서비스 / 직 / 노동 / 조합
1	서울_01_02	경인_02	서울_01_02(T) / 경인_02(T) / 사무 / 서비스 / 직 / 노동 / 조합
2	서비스_00_01	노동_03_01	서울_01_02 / 경인_02 / 사무 / 서비스_00_01(T) / 직 / 노동_03_01(T) / 조합
3	서울_01_02	사무_05	서울_01_02(C) / 경인_02 / 사무_05(T) / 서비스_00_01 / 직 / 노동_03_01 / 조합
4	서울_01_02	경인_01	이미 태깅됨
5	노동_03_01	조합_01_03	서울_01_02 / 경인_02 / 사무_05 / 서비스_00_01 / 직 / 노동_03_01(C) / 조합_01_03(T)
6	서울_01_02	조합_01_03	이미 태깅됨
7	사무_05	조합_01_03	이미 태깅됨
8	사무_05	조합_01_02	이미 태깅됨
9	사무_05	노동_03_01	이미 태깅됨
10	직_01_02	조합_01_05	다른 태그로 태깅 불가
11	서비스_00_02	직_06_02	다른 태그로 태깅 불가
...	...	...	...
38	직_06_01	조합_01_03	서울_01_02 / 경인_02 / 사무_05 / 서비스_00_01 / 직_06_01(T) / 노동_03_01 / 조합_01_03(C) (태깅 완료)
...	...	...	...
664	사무_06	조합_02	-

(그림 13) 의미 태깅 수행과정 (T : 태깅명사, C : 체인명사)

높은 확률을 가지는 1순위 bigram은 구성 명사에서 가장 핵심적인 역할을 하는 단위이며, 이는 전체적인 복합명사 의미 태깅의 방향을 제시해 태깅되지 않은 다른 구성 명사들도 이에 맞는 높은 유사도의 의미들로 체인을 형성해 태깅한다. 따라서 높은 유사도를 가진 bigram이라도 1순위 bigram과 체인을 형성할 수 없다면 해당 의미로 태깅되지 않는다. 유사도 분석의 단위는 다의어 단위이나 의미 태깅 후 실험은 동형이의어 단위로 수행한다. 그 이유는 유사도 분석 단계에서는 의미 단위를 최소화하여 비교 대상의 조합을 최대화하기 위함이다. 이에 따라 세분된 범위의 유사도를 얻을 수 있으므로 다의어 단위로 수행하며, 의미 태깅 후 실험은 정확률을 높이기 위해 세부개념으로 측정된 bigram을 이용하여 동형이의어 단위로 수행한다. 전체적인 의미 태깅은 (그림 12)와 같은 과정으로 이루어진다.

사전에 등재된 구성 명사가 존재하지 않을 시 벡터를 생성할 수 없으므로 구조 분해결과를 리턴하며 그렇지 않을 때 유사도를 순위별로 이용해 연쇄적인 체인을 구성하며 태깅을 수행한다. (그림 13)은 이러한 알고리즘을 이용한 복합명사 "서울경인사무서비스직노동조합"의 의미 태깅 과정이다.

복합명사 분해기에 의해 "서울-경인-사무-서비스-직-노동-조합"으로 나뉘었으며, 유사도 분석 결과에 따라 태깅을 수행한다. (T)는 태깅이 수행될 명사를 의미하며 1, 2순위 bigram의 경우 둘 다 태깅되지 않았으므로 태깅을 수행한다. 3순위의 경우 앞 명사 "서울_01_02"가 이미 태깅이 되어 있고 뒤 명사인 "사무_05"가 태깅이 되어 있지 않으므로 앞서 태깅된 앞 명사에 의해 체인으로 연결되어 뒤 명사가 태깅된다. 하지만 4순위의 경우, 높은 유사도를 가지고 있어도 이보다 더 높은 1순위 bigram에 의해 이미 태깅이 되었으므로 건너뛰게 된다. 마찬가지로 5순위, 38순위에 의해 체인으로 "조합_01_03"과 "직_06_01"이 태깅되게 된다. 38순위에서 모든 구성 명사에 태깅이 완료되었으므로 프로그램은 최종적으로 결과를 다의어 태그를 제거한 동형이의어 태그 형태의 "서울_01 / 경인_02 / 사무_05 / 서비스_00 / 직_06 / 노동_03/ 조합_01"를 결과로 출력한다. 가장 높은 의미의 bigram의 무조건적 태깅과 체인에 따른 합성과정에 의해 태깅은 빠른 수렴을 보인다.

6. 실험 및 평가

실험은 크게 첫째, 구조 분해 및 유사도 분석에 영향을 미치는 가중치에 대한 실험을 수행하고, 둘째, 의미범위 제약의 수행에 따른 정확도 변화에 대한 실험을 수행한다. 본 시스템의 성능평가를 위해 표준국어대사전에서 추출한 2음절 이상의 복합명사 40,717개를 대상으로 테스트 셋을 구성하였다. 테스트 셋의 의미 태깅은 본 연구실의 동형이의어 분별 시스템 UTagger[23]를 이용하였으며, 오분석된 태그는 수작업을 통해 교정하였다.

6.1 가중치 부여 실험

가중치에 따른 시스템의 성능측정과 최적 값을 구하기 위해 본 분해 부분에서 사용된 두 가지 가중치와 분석 부분에서 사용된 세 가지의 가중치에 대해 실험을 수행하였다.

6.1.1 구조 분해 가중치

3.1.1절에서 분해를 위한 위치별 구성 명사의 확률 계산 시 최적후보의 높은 확률값 부여를 위해 등록어와 미등록어에 대한 적절한 가중치가 필요하다. 가중치 결정에 대한 실험은 등록어 가중치 α에 대해 0.01씩 미등록어 가중치 β에 대해 -0.002씩 적용해 정확도의 변화를 관찰하였다. <표 9>는 이의 실험 결과이며, 등록어는 0일 때, 미등록어는 -0.010 이하의 가중치를 가질 때 최적의 구조 분해 성능인 99.26%를 보였다. 따라서 등록어의 가중치는 적용하지 않고 미등록어의 가중치는 -0.010 이하는 수렴하므로 -0.010으로 고정한다.

〈표 9〉 구조분해 시 등록어 및 미등록어의 가중치 결정을 위한 실험 결과
(W_k : 등록어, W_u : 미등록어, 단위 : %)

α \ β	0	-0.002	-0.004	-0.006	-0.008	-0.010	-0.012	-0.014
0	98.64	99.09	99.16	99.24	99.25	99.26	99.26	99.26
0.01	98.64	99.05	99.12	99.19	99.21	99.22	99.22	99.22
0.02	98.63	99.06	99.13	99.20	99.22	99.22	99.22	99.22
0.03	98.61	99.04	99.11	99.18	99.20	99.20	99.20	99.20
0.04	98.60	99.01	99.07	99.14	99.16	99.16	99.16	99.16
0.05	98.59	98.99	99.07	99.13	99.15	99.15	99.16	99.16
0.06	98.55	98.98	99.05	99.12	99.14	99.14	99.14	99.14
0.07	98.53	98.95	99.04	99.10	99.12	99.12	99.12	99.12
0.08	98.51	98.94	99.03	99.10	99.12	99.12	99.12	99.12
0.09	98.49	98.92	99.01	99.08	99.10	99.10	99.10	99.10
0.1	98.48	98.88	98.97	99.03	99.05	99.06	99.06	99.06

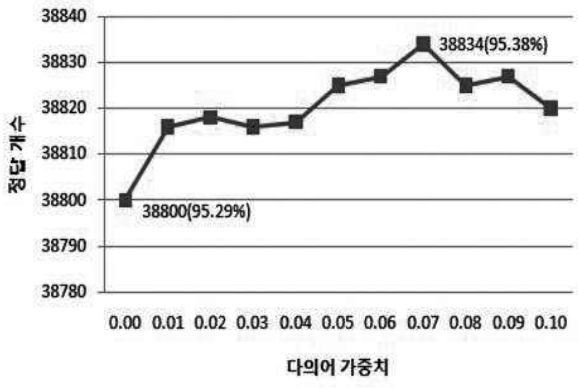
〈표 10〉 유사도 비교 시 가중치 W_r, W_s 에 따른 의미 태깅 실험 결과 오류 개수 (단위 : 개)

W_r \ W_s	0	0.02	0.04	0.06	0.08	0.10
0	3386	2020	2015	2029	2027	2031
1	2720	1917	1917	1928	1929	1930
2	2720	1922	1918	1920	1922	1925
3	2720	1922	1923	1924	1920	1922
4	2720	1923	1923	1928	1922	1922
5	2720	1921	1923	1927	1924	1923
6	2720	1921	1922	1929	1927	1924
7	2720	1919	1923	1929	1925	1928
8	2720	1918	1923	1928	1927	1927
9	2720	1918	1922	1928	1927	1926
10	2720	1918	1922	1928	1927	1928

6.1.2 유사도 분석 가중치

Bigram 유사도 계산에 사용된 자카드 계수는 단순히 공유 명사의 개수만을 이용하므로 관계유무는 고려하나 그 의미의 깊이는 고려하지 못한다. 이 때문에 생길 수 있는 단점을 보완하기 위해서는 추가적인 가중치가 필요하다. 5.3절에서 정의한 세 가지의 유사도에 대해 실험을 하였다. 실험을 위해 의미제약 방법은 원어 정보와 Naive Bayes Classifier를 둘 다 적용하였다. 가중치에 대한 실험은 관계 가중치 W_r에 대해 1씩, 대표성 가중치 W_s에 대해 0.02씩 적용해 변화를 관찰하였으며, 최적의 성능을 보이는 두 가중치에 다의어 가중치 W_p에 대해 추가로 0.01씩 적용해 결과를 측정하였다. <표 10>은 두 유사도 W_r, W_s에 대한 실험 결과이며, W_r이 1, W_s가 0.02~0.04의 가중치를 가질 때 의미 태깅 결과의 오류가 제일 적었으며, 최적의 가중치 덕분에 의미 분석이 3.61% 향상된 95.17%를 정확도를 보였다.

또한 W_p는 앞에서 실험한 두 개의 최적 가중치 값이 고정된 상태에서 (그림 14)와 같이 0.07일 때 그 오류율이 가장 낮았으며 0.09%가 향상된 95.38%의 정확도를 보였다.



(그림 14) 다의어 가중치에 따른 전체 실험 결과

6.2 의미제약 방법에 따른 실험

본 논문에서 제시한 두 가지 의미제약 방법이 성능에 미치는 영향을 알아보기 위해 실험을 하였다. 실험은 6.1절에서 최적의 성능을 보인 가중치로 고정한 뒤 의미제약을 수행하지 않은 경우, 원어 정보만을 적용한 경우, 원어 정보와 Naive Bayes Classifier를 함께 적용한 경우로 나누어 실험했으며 <표 11>은 그 결과이다.

〈표 11〉 의미제약 알고리즘에 따른 실험 결과

	의미제약 없음	원어 정보	원어 정보 + NB
정 답	30,281	38,783	38,834
백분율	74.37%	95.25%	95.38%

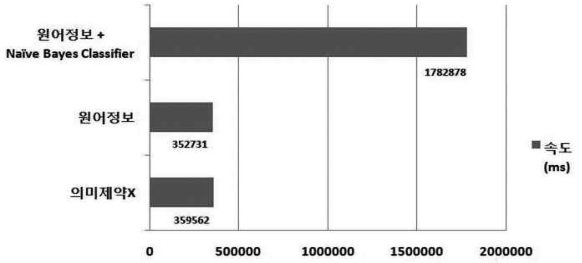
의미제약 알고리즘으로 원어 정보만을 이용한 경우, 적용하지 않은 경우보다 20.88%의 큰 정확도 향상이 있었다. 이는 사전에서 추출한 테스트 셋의 복합명사를 구성하는 bigram들이 원어 정보와 함께 사전에 존재할 확률이 높으며, 긴 음절의 복합명사는 짧은 음절의 복합명사들을 합성해 구성될 수 있음을 의미한다. 그러므로 <표 12>와 같이 447개의 bigram만을 이용해 42.52%인 37,257개가 제약에 성공되었으며 이후, 유사도 분석에 의해 높은 결과를 얻었다. 따라서 테스트 셋에 속하지 않은 복합명사들도 이 방법에 의해 높은 제약 성공률을 얻을 수 있음을 알 수 있다.

〈표 12〉 의미제약 방법별 분석 결과

	원어 정보	Naive Bayes Classifier
총 bigram	87,618	
제약된 bigram	37,257	1,283
사용된 bigram	447	1,185
제약율(%)	42.52%	1.46%
총 제약율(%)	43.98%	

Naive Bayes Classifier는 원어 정보를 이용해 의미제약에 실패한 경우 시도되는 것으로 분류기에 사용될 학습데이터를 위해 세종 말뭉치에서 추출한 의미 태깅된 복합명사 bigram 291,001개를 사전에서 속성명사 셋을 생성할 수 있는 유효한 대상으로 추려낸 결과 총 276,804개의 복합명사를 얻었으며, 이를 이용해 학습데이터를 생성한 결과 943,996개가 만들어졌다. 하지만 학습데이터의 크기가 너무 크고, 속성명사의 개수가 작을 때 같은 클래스의 다른 학습데이터들과 속성 명사들이 중복된 경우가 많아 속성명사의 개수가 일정 이하면 제거하였다.

실험결과 13개 이하인 경우를 제거했을 때 크기와 성능이 최적이었으며 그 결과 총 369,047개로 약 61%의 데이터가 축소되었다. 이를 이용해 실험한 결과 전체 bigram의 1.46%인 1,283개의 bigram만이 제약되었음을 알 수 있다. 이는 실험에 사용한 테스트 셋의 특성상 사전상의 원어 정보를 지닌 bigram이 많아 대부분이 원어 정보에 의해 선 제약되었기 때문이다. 말뭉치에서 추출한 복합명사의 경우 사전에 등재된 bigram이 존재하지 않을 때가 많으며 이 경우 Naive Bayes Classifier에 의해 제약될 확률이 높아진다. 또한, 의미 제약방법에 따른 수행시간에 대해서도 실험하였으며 이는 (그림 15)와 같다.



(그림 15) 의미제약 방법별 테스트 셋 분석속도

이 실험에 따르면 테스트 셋 복합명사 40,717개에 대해 의미제약을 수행하지 않을 때보다 원어 정보를 이용한 경우 정확도는 크게 향상되고 분석 시간도 약 6분에서 약 5분으로 줄어들어 의미제약으로써의 효율적인 방법으로 판단된다. 하지만 가장 높은 정확도를 보인 "의미정보 + Naive Bayes Classifier" 조합은 1,782,878ms가 소요되었으며, 이는 대용량의 학습데이터와 인스턴스별 확률 계산 때문이다. 정확도 향상률에 비해 계산 시간이 월등히 소요되므로 처리속도가 중요시 되는 시스템은 Naive Bayes Classifier의 적용 여부를 고려해보아야 한다.

6.3 분석 실패의 예와 원인

본 실험결과 최적의 분해성능을 보인 95.38%에 대한 오류 1,883개를 분석한 결과는 다음과 같다. 오류의 종류는 복합명사 분해 시 중의성을 보인 구조 분해 오류와 의미 분석 시 잘못된 태깅이 붙여진 의미 분석 오류로 나눌 수 있다. <표 13>은 이의 분석 결과이다.

〈표 13〉 오류의 종류에 따른 분석 결과

	구조 분해 오류	의미 분석 오류
개수	437	1446
백분율	19.17%	80.83%
총	1883(100%)	

구조 분해에 대한 오류는 전체 테스트 셋의 1% 내외로 그 종류는 다음과 같이 크게 네 가지로 나뉜다.

- 접사 - 산업 기지 개발 촉진법(촉진 법), 생산 성(생산성) 향상 운동
- 미분해 - 국민 사회주의(사회 주의) 독일 노동당, 자유주의(자유 주의) 경제
- 외래어 분해 - 라이 프니츠 볼프 철학
- 오분해 - 내산 소성 혐기성 생물

이 중, 접사에 의한 오분해가 가장 큰 비중을 차지했으며 정답 셋에서는 접사가 붙어 분해된 경우와 그렇지 않은 경우도 다소 있었다. 미분해의 경우 통계에 의해 재분해를 결정한다. 따라서 접사 분해규칙을 적용하고 미분해의 문제인 최소 단위를 정의해 정답 셋을 구축한다면 정확률이 더욱 향상될 수 있다. 외래어 분해의 경우 외래어 복원 부분이 있음에도 오분해 되었다. 이는 외래어 판별 및 복원을 구축한 사전을 통해서만 수행하므로 생기는 문제로 외래어만이 가지는 특성들을 반영해 적용한다면 이에 대한 처리가 가능할 것으로 보인다. 또한, 오분해의 경우 주로 1음절을 미등록어로 처리해 최적 후보를 선택하기 때문에 다른 분해 형태가 미등록어를 가지지 않는 경우 이 후보가 선택된다.

의미 분석에 대한 오류는 전체 테스트 셋의 3.5% 정도로 유사도가 높은 특정 bigram에 의해 의미 태깅이 잘못된 경우가 대부분이다. 이 bigram중에는 실제로 정답과 유사한 의미를 가진 명사의 속성명사들에 의해 태깅된 경우와 의미상으로 틀린 의미이나 유사도가 높아 태깅된 경우로 나뉠 수 있다. 두 경우 모두 가중치에 의해 잘못 태깅된 때도 있었다. 하지만 이는 최적 값이므로 달라진 가중치에 의해 오류율이 더 높아질 수 있어, 이는 바람직하지 못하다.

7. 결론 및 향후 연구과제

본 논문에서는 입력받은 n음절의 한국어 복합명사를 말뭉치에서 추출한 위치별 빈도 데이터를 이용해 분해하고 선택된 최적의 분해후보를 구성 명사별 의미 제약을 수행한 뒤, 유사도 비교결과를 이용해 의미 태깅을 수행하는 시스템을 제안하였다. 복합명사 분해는 대용량 말뭉치에서 추출한 위치별 빈도 데이터를 이용해 분해 시 명사의 출현 빈도뿐만 아니라 해당 명사의 위치 정보를 적용하므로 음절별 패턴 적용 방법에 따른 오류를 줄일 수 있었다. 또한, 최적의 후보 선택을 위해 외래어 복원과 재분해를 적용하였다. 분해 결과 최종 선택된 하나의 후보에 대해 의미 분석 시 복잡도와 정확도 향상을 위해 두 가지의 의미범위 제약 방법을 사용하였다. bigram으로 합성한 제약대상은 먼저 U-WIN 사전에 존재하는 원어 정보를 이용하며 이는 사전에 의존적이므로 정확한 매칭에 의한 제약에 실패한 경우 Naive Bayes Classifier를 이용해 확률에 따른 분류에 따른 제약을 시도하였다. 제약된 구성 명사들에 대해 7가지 규칙을 적용해 유사도 비교를 위한 의미 집단 벡터를 생성했으며 각 벡터는 속성명사 유무에 따른 비대칭 이진 속성을 나타내므로 자카드 계수를 이용해 유사도를 구할 수 있다. 하지만 이진 속성은 속성명사들의 추출 빈도에 따른 의미 깊이를 고려하지 못한다. 따라서 세 가지의 가중치를 적용해 이를 반영하였다. 이에 따른 최종 유사도 결과에 따라 의미 체인을 형성하며 의미 태그를 결정하였다. 성능 측정을 위해 표준국어대사전에서 추출한 3음절 이상의 40,717개의 복합명사를 대상으로 실험한 결과 구조분해는 99.26%의 정확도를 보였으며, 의미 태깅은 경우 95.38%의 정확도를 보였다.

본 논문에서 제시한 U-WIN 기반의 한국어 복합명사 분해 및 의미 태깅 시스템은 다음과 같은 장점이 있다.

첫째, 모든 분해 가능 경우를 분해 대상으로 삼고 사전의 뜻풀이와 대용량 말뭉치에서 추출한 위치별 명사 빈도 정보를 이용해 음절 및 분해 패턴의 제약을 없앨 수 있다.

둘째, 의미제약의 수행으로 bigram 유사도 분석대상을 크게 줄이고 의미 결합의 정확도를 향상할 수 있다.

셋째, 유사도 분석 시 문제가 되는 자료 부족 현상을 7가지 종류의 대상을 정의해 어휘망으로부터 추출해 완화할 수 있다.

하지만 다음과 같은 단점도 존재한다.

첫째, 구조 분해 시 접사에 대해 처리를 하지 않으므로, 말뭉치와 사전 뜻풀이에서 추출한 bigram 리스트에 분해형태가 존재하지 않으면 접사의 오분해 확률이 높아지며, 미등록어 또한 위치별 빈도데이터에 그 구성 명사가 없다면 제대로 인식될 수 있으나 존재할 때에는 오분해될 수 있다.

둘째, 의미 제약 시 Naive Bayes Classifier를 의미제약에 사용할 때는 대용량의 학습데이터에 의해 확률 계산 시 속도가 느려지는 단점이 있으며, 원어 정보만을 이용할 때는 사전에 의존적이므로 실패할 가능성이 있다.

현재 명사로 이루어진 문장만을 분석 대상으로 하지만, 접사에 규칙 및 사전적용에 따른 처리와 미등록어, Naive Bayes Classifier의 속도 향상에 대한 연구를 수행한다면 더욱 더 빠르고 정확한 의미 분석이 가능해 질 것이다.

참 고 문 헌

[1] 최재혁, “음절수에 따른 한국어 복합 명사 분리 방안”, 한국정보과학회 언어공학연구회, 제8회 한글 및 한국어 정보처리 학술대회 pp.262-267, 1996.

[2] 강승식, “한국어 복합명사 분해 알고리즘”, 한국정보과학회, 정보과학회논문지(B), 제25권 제1호, pp.172-182, 1998.

[3] 윤보현, 임희석, 임해창, “통계 정보를 이용한 한국어 복합 명사의 분석 방법”, 한국정보과학회 봄 학술발표논문집 제22권 제1호, pp.925-928, 1995

[4] J.T. Yoon, K.S. Choi, and M.S. Song, “Corpus-based approach for nominal compound analysis for Korean based on linguistic and statistical information.” In Proceedings of the 1999 Joint SIGDAT Conference on EMNLP/VLC. College Park, MD, pp.292-300, 2001.

[5] 강유환, 서영훈, “미등록어의 의미 범주 분석을 이용한 복합명사 분해”, 한국데이타베이스학회, 정보기술과 데이타베이스 저널 제 11권 제4호, pp.95-102, 2004.

[6] 임해창, 임희석, 윤보현, “자연어 처리 연구동향: 통계 기반의 자연어 처리”, 한국정보과학회지, 제12권, 제9호, pp.20-30, 1994.

[7] 박재한, 김명선, 노대욱, 나동열, “백오프 통계정보를 이용한 미등록어 포함 복합명사의 분해”, 한국정보과학회 언어공학연구회, 제16회 한글 및 한국어 정보처리 학술대회 발표자료집 제16권 제1호 pp.65-72, 2004.

[8] 강민규, 강승식, “한국어 복합명사 분해 오류 교정 기법”, 한국정

보과학회, 한국 컴퓨터 종합 학술 발표 논문집 제37권 제1호 (C), pp.254~259, 2010.
[9] 원상연, 김수남, 김광영, 남현숙, 권혁철, “한국어 문법검사기에서 의미정보를 이용한 복합명사의 분석제약”, 한국정보과학회 언어공학연구회, 제11회 한글 및 한국어 정보처리 학술대회 pp.288-293, 1999.
[10] 김도완, 이경순, 김길창, “의미관계와 문형정보를 이용한 복합명사 해석”, 한국정보과학회 언어공학연구회, 제11회 한글 및 한국어 정보처리 학술대회 pp.310-315, 1999.
[11] 강유환, 정천영, 서영훈, “명사의 의미 정보를 이용한 복합명사 분석의 중의성 해결”, 한국정보과학회 언어공학연구회, 제14회 한글 및 한국어 정보처리 학술대회 pp.171-175, 2002.
[12] 허정, 옥철영, “사전의 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템”, 한국정보과학회, 정보과학회 논문지, 소프트웨어 및 응용, 제28권 제9호 pp.688-698, 2001.
[13] 허정, 서희철, 장명길, “상호정보량과 복합명사 의미사전에 기반한 동음이의어 중의성 해결”, 한국정보과학회, 정보과학회논문지, 소프트웨어 및 응용, 제33권 제12호 pp.1073-1089, 2006.
[14] M. Lesk, “Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone,” In Proceedings of the 5th annual international conference on Systems documentation, pp.24-26, 1986.
[15] Cowie, J., L. Guthrie, J. Guthrie, “Lexical disambiguation using simulated annealing,” In Proceedings of COLING, 1992.
[16] Yarowsky D., “Word-Sense Disambiguation using Statistical Models of Roget's Categories Trained on Large Corpora,” In Proceedings of Coling-92, 1992.
[17] 최호섭(2007), “대규모 사용자 어휘지능망 구축과 활용”, 울산대학교 대학원 컴퓨터정보통신공학부 박사학위논문.
[18] 한국어의 한자어, 위키백과 - http://ko.wikipedia.org/wiki/한국어의_한자어
[19] 이용훈, 옥철영, “Naive Bayes Classifier를 이용한 의미제약이 강화된 한국어 복합명사 의미 분석”, 한국정보과학회 언어공학연구회, 제23회 한글 및 한국어 정보처리 학술대회 pp.102-106, 2011.
[20] Escudero, G., Màrquez, L., and Rigau, G. “Naive Bayes and exemplar-based approaches to word sense disambiguation revisited”. In Proceedings of the 14th European Conference on Artificial Intelligence (ECAI, Berlin, Germany), pp.421-425, 2000.
[21] 이용훈, 옥철영, “의미기반 한국어 복합명사 분석”, 한국정보과학회 한국컴퓨터종합학술대회 논문집(C) pp.221-224, 2011.
[22] UTagger, 2011년 국어정보처리시스템 경진대회 출품, 울산대학교 한국어처리연구실

이 용 훈

e-mail : yhsoft12@gmail.com
2010년 울산대학교 컴퓨터 · 정보통신공학부 (학사)
2012년 울산대학교 정보통신공학과 석사과정
관심분야 : 한국어정보처리, 복합명사, 의미분별

옥 철 영

e-mail : okcy@ulsan.ac.kr
1982년 서울대학교 컴퓨터공학과(학사)
1984년 서울대학교 컴퓨터공학과(석사)
1993년 서울대학교 컴퓨터공학과(박사)
1994년 러시아 TOMSK 공과대학 교환교수
1996년 영국 GLASGOW 대학교 객원교수
2007년~2008년 한국정보과학회 언어공학연구회 위원장
2007년 몽골국립대학교 IT대학 명예박사학위
2008년 국립국어원 객원연구원
1984년~현 재 울산대학교 컴퓨터정보통신공학부 교수
관심분야 : 한국어정보처리, 의미분별, 온톨로지, 지식베이스, 기계학습, 문서분류

이 응 봉

e-mail : eblee@cnu.ac.kr
1985년 성균관대학교 도서관학과(학사)
1992년 성균관대학교 문헌정보학과(석사)
1996년 성균관대학교 문헌정보학과 (정보학 박사)
2007년 미국 Univ. of Illinois at Urbana-Champaign 교환교수
2009년~현 재 한국문헌정보학회 부회장
2010년~현 재 충남대학교 사회과학연구소장
2010년~현 재 한국연구재단 자문위원
1996년~현 재 충남대학교 문헌정보학과 교수
관심분야 : 정보검색, 정보서비스, 데이터베이스 품질평가, 정보시스템 및 정보정책