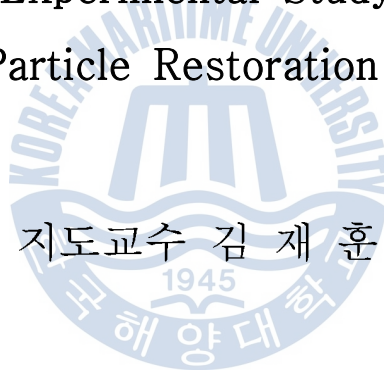


공학석사 학위논문

한국어 격조사 복원에 대한
실험적 고찰

Experimental Study
on Case Particle Restoration in Korean



2013년 8월

한국해양대학교 대학원

컴퓨터공학과

황보천

본 논문을 황보천의 공학석사 학위논문으로 인준함.

위원장 박 휴 찬 (인)

위 원 류 길 수 (인)

위 원 김 재 훈 (인)



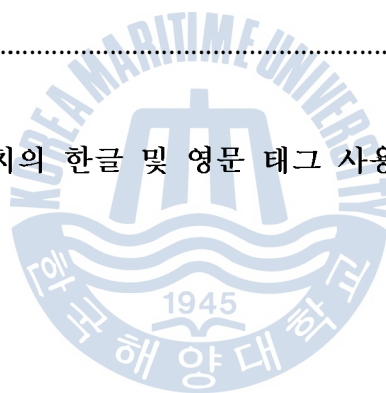
2013년 7월 31일

한국해양대학교 대학원

목 차

List of Tables	iv
List of Figures	v
Abstract	vi
1. 서 론	1
2. 관련 연구	
2.1 ETRI 구문구조 부착 말뭉치	3
2.2 Conditional Random Fields	4
2.3 의존관계 분석	6
3. 격조사 복원시스템	
3.1 개요	7
3.2 ETRI 말뭉치 추출기	8
3.3 격조사 복원 자질 추출기	
3.3.1 격조사 복원 자질집합	11
3.3.2 격조사 복원 자질 추출기	15
3.4 격조사 복원기	20
4. 실험 및 평가	
4.1 실험환경	23

4.2	체언과 용언 사이의 거리 분석	24
4.3	격조사 복원 성능	
4.3.1	실험방법	25
4.3.2	격조사 복원 성능분석	28
4.4	거리별 격조사 복원 성능분석	
4.4.1	전체 거리별 성능분석	29
4.4.2	거리 1과 2의 성능분석	30
4.5	자질별 중요도 분석	33
5.	결론 및 향후 연구	38
	참고문헌	40
	부록 A 원시 말뭉치의 한글 및 영문 태그 사용 예	24



List of Tables

Table 1	ETRI 구문구조 부착 말뭉치 통계	3
Table 2	MA_BUF_N	6
Table 3	PC_BUF_N	7
Table 4	MA_BUF_N	7
Table 5	PC_BUF_N	7
Table 6	실험대상 격조사 표	24
Table 7	추출 말뭉치 전체 개수 및 비율	25
Table 8	체언과 용언 사이의 거리별 격조사의 개수	26
Table 9	격조사 복원 성능 평균(거리: 1 ~ 25)	92
Table 10	거리별 격조사 복원 성능 평균(거리: 1 ~ 25)	03
Table 11	격조사 복원 성능 평균(거리: 1 ~ 2)	13
Table 12	거리별 격조사 복원 성능 평균(거리: 1 ~ 2)	23
Table 13	보격조사 10-차 교차검증 결과(거리: 1)	33
Table 14	자질별 격조사 복원 성능(macro-average)	73
Table 15	자질별 격조사 복원 성능(micro-average)	73

List of Figures

Fig. 1 ETRI 구문구조 부착 말뭉치의 일부	5
Fig. 2 격조사 복원시스템	7
Fig. 3 ETRI 말뭉치 추출기	9
Fig. 4 자질집합	14
Fig. 5 격조사 복원 자질 추출기	20
Fig. 6 격조사 복원기	21
Fig. 7 격조사가 생략된 실험 말뭉치	21
Fig. 8 격조사 복원결과	22
Fig. 9 학습기 CRF++ 실행화면	72
Fig. 10 분류기 CRF++ 실행화면	82
Fig. 11 거리별 격조사 복원 성능 평균(거리: 1 ~ 2)	23
Fig. 12 자질별 격조사 복원 성능(macro-average)	43
Fig. 13 자질별 격조사 복원 성능 평균(macro-average)	53
Fig. 14 자질별 격조사 복원 성능(micro-average)	53
Fig. 15 자질별 격조사 복원 성능 평균(micro-average)	63

Experimental Study on Case Particle Restoration in Korean

Hwang-Bo, Cheon

Department of Computer Engineering
Graduate School of Korea Maritime University

Abstract

This thesis is an experimental study on case particle restoration in Korean. The case particles in Korean sentences are omitted frequently. The omitted particles cause ambiguity in syntactic attachment and decrease performance of syntactic analysis. In this thesis, we restore the omitted case particles using machine learning techniques and suggest the most proper features for case particle restoration. The system for restoring omitted particles can be one component in the parsing system and also can be used for indexing terms in information retrieval. We have done experiments on several experimental settings and have observed the results. For the experiments, we have used ETRI syntactic tree-tagged corpus. The correct restoration rate of the system is 81.11 in accuracy of omitted case particles. We have observed that nouns and verbs, themselves, are very important features for restoring case particles.

KEY WORDS : Syntactic analysis, Machine Learning,
Case particle restoration



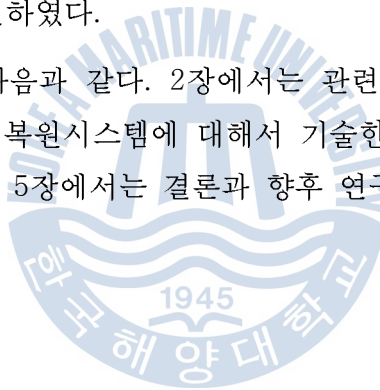
제 1 장 서 론

최근 정보기술의 발달로 스마트폰, 태블릿 PC 등 모바일 디바이스(mobile device)의 확산과 페이스북(Facebook)이나 트위터(Twitter) 같은 소셜 네트워크 서비스(Social Network Service, SNS) 이용자의 급증으로 다양한 종류의 데이터가 급속히 생성되고, 유통되며, 저장되고 있다. 이러한 대량의 데이터를 기반으로 가치 있는 데이터를 추출하는 정보검색 기술들이 부각되고 있다. 정보검색은 이용자의 정보요구에 적합한 문서를 찾아주는 것이다. 즉, 정보검색 시스템은 이용자들이 찾을 수 있는 문서를 미리 수집하고, 그 내용을 분석하여 이용자들이 쉽게 찾을 수 있도록 색인하며 색인된 문서와 이용자의 정보요구가 일치하는 문서를 찾아주는 시스템이다. 정보검색시스템에 대한 연구는 1960년대 초부터 시작되어 꾸준히 발전해 왔으며 인터넷의 발달로 정보획득에 대한 요구가 증가하면서 정보검색에 대한 중요성도 더불어 증가하게 되었다. 오늘날에는 너무나 많은 정보들이 인터넷에 존재하므로 얼마나 정확한 문서를 찾느냐가 관건이 된다. 하지만 한국어의 형태소 분석에서 나타나는 중의성 문제와 웹에서의 언어파괴 현상이나 무의미한 말들과 의문형인지 청유형인지 구분하기 어려운 어미까지 생겨나 통용되고 있어 한국어 정보검색시스템 개발에서 해결해야 할 문제들이 산재해 있다고 할 수 있다. 또한 한국어 문장에서 조사의 생략이 자주 발생됨에 따라 형태소 분석에서 나타나는 중의성 문제를 가중시키고 구문분석의 복잡도를 높일 뿐만 아니라 구문분석 오류의 원인이 되어 생략된 조사를 복원하는 문제도 정보검색시스템 개발에 있어서 해결해야 될 과제로 남아 있다. 따라서 본 논문에서는 한국어 문장에서 자주 발생하는 생략된 조사를 복원한다면 중의성 해소와 정보검색시스

템의 성능을 향상할 수 있다는 점에 착안하여 한국어 문장에서 생략된 조사를 복원하는 격조사 복원시스템을 제안한다.

본 논문에서는 기계학습 방법으로 생략된 조사를 복원하기 위한 ETRI 구문구조 부착 말뭉치를 분석하였으며, 한국어 문장에서 자주 발생하는 생략된 조사를 복원하는 실험을 하였다. 실험결과 체언과 용언 사이의 거리가 멀어지면 멀어질수록 조사 복원의 성능은 저하된다는 것을 알 수 있었다. 또한 기계학습을 이용한 실험에서 추출한 자질들의 조합 구성방식에 따라 격조사의 복원 성능이 다르게 나타난다는 것도 확인되었다. 실험을 반복적으로 수행하면서 격조사 복원 성능이 우수한 자질집합을 발견하였으며, 자질집합 중에서 체언과 용언이 격조사 복원 성능에 중요한 역할을 한다는 것도 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대하여 기술하고, 3장에서는 격조사 복원시스템에 대해서 기술한다. 4장에서는 실험 및 평가에 대해서 논하고, 5장에서는 결론과 향후 연구에 대하여 기술한다.



제 2 장 관련 연구

2.1 ETRI 구문구조 부착 말뭉치

ETRI 구문구조 부착 말뭉치는 (김재훈 등, 2005) 형태소 분석, 구 묶음 분석과 다양한 언어현상을 고려하여 문장의 올바른 구조를 파악하는 단계인 구조분석에서 복문 구조분석을 위한 복문을 구성하는 단문들 간의 구조정보까지를 포함하여 2005년도에 구축하였다. 구축된 말뭉치는 구조분석에서 용언간의 구조분석에 활용할 수 있으며, 기타 구조분석을 위한 정보추출에도 활용이 가능하다. 기계번역에서 활용되는 자연언어 분석과 한국어 형태소 분석, 구 묶음 분석, 구문 분석의 성능을 개선하는데도 활용될 수 있다. 또한 기계번역에서의 성능향상은 물론 정보검색의 색인어 추출 성능개선에도 활용될 수 있다. ETRI 구문구조 부착 말뭉치의 통계치를 보면 Table 1과 같이 전체 문장 수는 101,602개이고, 문장 당 평균 어절 수는 21.6개, 어절 당 평균 형태소 수는 1.98개, 전체 품사 수는 71개이다.

Table 1 ETRI 구문구조 부착 말뭉치 통계

구	분	통	계(단위:개)
문장 수		101,602	
문장 당 평균 어절 수		21.6	
어절 수		2,290,914	
어절 당 평균 형태소 수		1.98	
총 형태소 수		4,375,332	
전체 품사 수		71	

ETRI 구문구조 부착 말뭉치의 각 문장은 Fig. 1과 같이 메타정보, 문장, 형태소 분석, 구 묶음, 의존구조 분석으로 구성되어 있다. 메타정보는 문장번호(Sentence No), 상태(Status), 제작자(worker), 수정자(corrector), 제작일자(m_time), 문장설명(Comment)으로 표현하고, 형태소 분석은 두 줄로 표현되어 있는데 첫 번째 줄은 어절을 표현하고, 두 번째 줄은 형태소 분석 결과를 표현한다. 구 묶음은 형태소 분석 결과에서 구 묶음이 형성되면 ‘_’로 연결하여 구 묶음을 표현하였다. 의존구조 분석은 세 줄로 표현되어 있는데 첫 번째 줄은 구 묶음 번호이고 두 번째 줄은 용언이 위치한 중심어 번호를 표현하고 세 번째 줄은 형태소와 의존관계를 표현한다. 본 논문에서는 한국어 문장에서 생략된 조사를 복원하기 위하여 기계학습 자질로 ETRI 구문구조 부착 말뭉치에서 자질을 추출하여 실험에 사용한다.

2.2 Conditional Random Fields

기계학습 모델 중 하나인 Conditional Random Fields(CRF)는 주로 조건부 확률을 최대화하는 비방향성 그래프 모델로 순차적 데이터를 세분화하고 레이블을 할당하는데 적합한 확률 모델이다(Lafferty, McCallum, & Pereira, 2001). CRF는 순차적 라벨링 작업량(Lafferty, McCallum, & Pereira, 2001; Pinto, McCallum, Wei & Croft, 2003)를 줄이는 성능을 보였고, (김학수, 2007)은 영역 분류 모델에 적용하였고, (김형기, 이광국, 김희율, 2010)은 궤적군집화를 이용한 혼잡 영상의 이동 객체 검출, (이창기 등, 2006)은 세부 분류 개체명 인식에 적용하였다. 또한 (김재훈, 김형철, 2010)은 대용어 해소(anaphora resolution) 문제와 같은 복잡한 문제에도 적용하였다. 본 논문에서는 CRF 모델을 적용하여 격조사 복원 성능이 우수한 자질집합을 찾고, 체언과 용언 사이의 거리에 따라 조사 복원 성능이 어떻게 변하는지를 확인한다.

메타정보

Sentence No.: 170142, Status: 20
worker: skylake99@hanmail.net, corrector: , m_time: 2005-10-16 22:06:07
Comment:

문장

임 사장은 한국 문학 전문 출판의 길을 포기하지 않겠다며, 실용서에 눈을 돌린 독자들이 언젠가는 다시 돌아올 것이라고 거듭 다짐하고 기대한다.

형태소 분석

임	임[인명고유명사]
사장은	사장[용언불가능보통명사]+은[일반보조사]
한국	한국[지명고유명사]
문학	문학[용언불가능보통명사]
전문	전문[용언가능보통명사]
출판의	출판[용언가능보통명사]+의[관형격조사]
길을	길[용언불가능보통명사]+을[목적격조사]
포기하지	포기하[일반동사]+지[종속연결어미]
않겠다며,	않[일반동사]+겠[미래시제선어말어미]+다며[종속연결어미]+,[킴마기호]
실용서에	실용서[용언불가능보통명사]+에[부사격조사]
눈을	눈[용언불가능보통명사]+을[목적격조사]
돌린	돌리[일반동사]+ㄴ[관형사형전성어미]
독자들이	독자[용언불가능보통명사]+들[복수접미사]+이[주격조사]
언젠가는	언젠가는[지시시간부사]
다시	다시[성상상태부사]
돌아올	돌아오[일반동사]+ㄴ[관형사형전성어미]
것이라고	것[기타보조용언]+이[긍정지정사]+라고[종속연결어미]
거듭	거듭[지시시간부사]
다짐하고	다짐하[일반동사]+고[대등연결어미]
기대한다.	기대하[일반동사]+ㄴ다[평서형종결어미]+,[문미기호]

구문

임_사장[인명고유명사]+은[일반보조사] 한국_문학_전문_출판[용언불가능보통명사]+의[관형격조사] 길 [용언불가능보통명사]+을[목적격조사] 포기하[일반동사]+지_않[기타보조용언]+겠[미래시제선어말어미]+다며[종속연결어미]+,[킴마기호] 실용서[용언불가능보통명사]+에[부사격조사] 눈[용언불가능보통명사]+을[목적격조사] 돌리[일반동사]+ㄴ[관형사형전성어미] 독자_들[용언불가능보통명사]+이[주격조사] 언젠가는[지시시간부사] 다시[성상상태부] 돌아오[일반동사]+ㄴ_것_이[기타보조용언]+라고[종속연결어미] 거듭[지시시간부사] 다짐하[일반동사]+고[대등연결어미] 기대하[일반동사]+ㄴ다[평서형종결어미]+,[문미기호]

의존구조 분석

1	14	임_사장[인명고유명사]+은[일반보조사] (주격)
2	3	한국_문학_전문_출판[용언불가능보통명사]+의[관형격조사]
3	4	길[용언불가능보통명사]+을[목적격조사]
4	14	포기하[일반동사]+지_않[기타보조용언]+겠[미래시제선어말어미]+다며[종속연결어미]+,[킴마기호]
5	7	실용서[용언불가능보통명사]+에[부사격조사]
6	7	눈[용언불가능보통명사]+을[목적격조사]
7	8	돌리[일반동사]+ㄴ[관형사형전성어미] (주격)
8	11	독자_들[용언불가능보통명사]+이[주격조사]
9	11	언젠가는[지시시간부사]
10	11	다시[성상상태부사]
11	14	돌아오[일반동사]+ㄴ_것_이[기타보조용언]+라고[종속연결어미]
12	14	거듭[지시시간부사]
13	14	다짐하[일반동사]+고[대등연결어미]
14	0	기대하[일반동사]+ㄴ다[평서형종결어미]+,[문미기호]

Fig. 1 ETRI 구문구조 부착 말뭉치의 일부

2.3 의존관계 분석

의존관계 분석은 어절 사이의 의존구조(dependency structure)의 관계를 분석하는 것이다. 한국어에서 의존구조 자체를 분석하는 연구는 매우 활발하게 진행되었으나(서광준, 1993), 의존관계 분석에 대한 연구는 그다지 연구되지 않았다(Chung, 2004). (서광준, 1993)에서는 구문 분석 과정에서는 의존문법을 기반으로 조사와 어미를 구분하여 각각의 어절을 지배소와 의존소로 나누고 이들 간의 의존관계를 밝혔다. (Chung, 2004)은 구문분석 과정에서 의존소 및 지배소의 의존관계 분석을 위하여 코퍼스에서 통계 정보를 추출하여 각 어휘들 간의 의존관계를 분석하였다. (Kim, Park & Lee, 2007)는 파스트리 커널을 이용하여 의존관계에 있는 절들 간의 대상을 분석하였고, 절들 간의 어휘 정보 및 거리 같은 표면적 정보를 사용하여 절들 간의 의존관계를 분석하였다. 본 논문은 생략된 격조사를 복원함으로써 의존관계 분석에 많은 도움을 줄 수 있을 것이다. 예를 들면 “밥 먹었다.”를 “밥-을 먹었다.”로 복원함으로써 문장의 그 의미가 더욱 명확해질 뿐 아니라 구문분석의 오류도 크게 줄일 수 있을 것이다.

제 3 장 격조사 복원시스템

3.1 개요

본 논문에서 제안하는 격조사 복원시스템은 Fig. 2와 같이 ETRI 구문구조 부착 말뭉치에서 형태소 분석, 구류음, 의존구조 분석 결과를 추출하는 ETRI 말뭉치 추출기, 각 문장의 중심이 되는 체언을 중심으로 격조사 복원을 위한 자질을 추출하는 격조사 복원 자질 추출기, 격조사 복원 자질 추출기에서 추출한 학습 말뭉치를 실험 데이터로 사용하여 한국어 문장에서 자주 생략되는 조사를 복원하는 격조사 복원기로 구성된다.

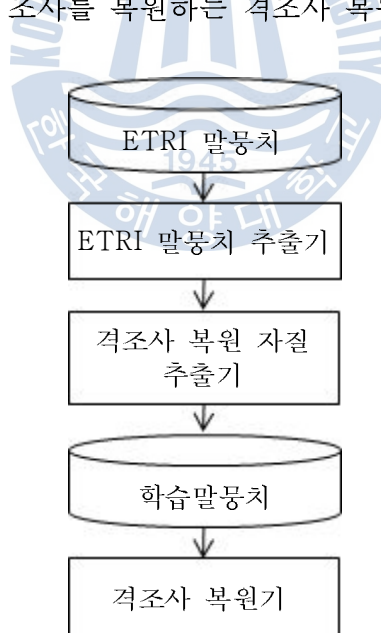


Fig. 2 격조사 복원시스템

3.2 ETRI 말뭉치 추출기

ETRI 말뭉치 추출기는 ETRI 구문구조 부착 말뭉치에서 메타정보와 문장을 제외한 형태소 분석, 구뭉음, 의존구조 분석 결과를 추출하는 것이다. ETRI 말뭉치 추출기는 Python으로 구현되었으며, 구체적인 추출과정은 Fig. 3에서 보이고 있다. Fig. 3에서 보듯이 “dep-tree.sentence”에 들어 있는 ETRI 구문구조 부착 말뭉치의 정보를 한 줄씩 입력 받아서 처리한다. 구체적인 예를 설명하면 아래와 같다. 메타정보인 “====중략====”, “Sentence No.: 170142, Status: 20”, “worker: sky-lake99@hanmail.net, corrector: , m_time: 2005-10-16 22:06:07”, “Comment:”, “-----중략-----”등은 반복적인 과정을 통해서 제거한다. 다음으로 문장 “임 사장은 한국 문학 전문 출판의 길을 포기하지 않겠다며, 실용서에 눈을 돌린 독자들이 언젠가는 다시 돌아올 것이라고 거듭 다짐하고 기대한다.”도 본 논문에서는 사용하지 않으므로 제거한다. 그 외 형태소 분석, 구뭉음, 구문구조 분석 결과는 추출 대상이므로 ETRI 말뭉치 추출기를 통해서 수집한다. 먼저 형태소 분석 “사장은 사장 [용언불가능보통명사]+은[일반보조사]”를 입력받으면 문자열 분리함수 “split()”을 사용하여 어절 “사장은”과 형태소 분석 “사장[용언불가능보통명사]+은[일반보조사]”를 공백으로 분리하고, 리스트 자료형¹⁾ “MA_BUF”에 저장한 후 다음 Line의 Text를 읽는다. 다음 Line의 Text가 형태소 분석이면 앞의 과정을 반복적으로 수행하면서 리스트 자료형인 “MA_BUF”에 저장한다.

1) 리스트(lists) : Python 내장 자료형(built-in types)으로 []를 사용하며, 임의의 객체를 저장하는 집합적 자료형이다. 각 자료는 순서를 가지고 있고, 순서에 따라 접근이 가능하다.

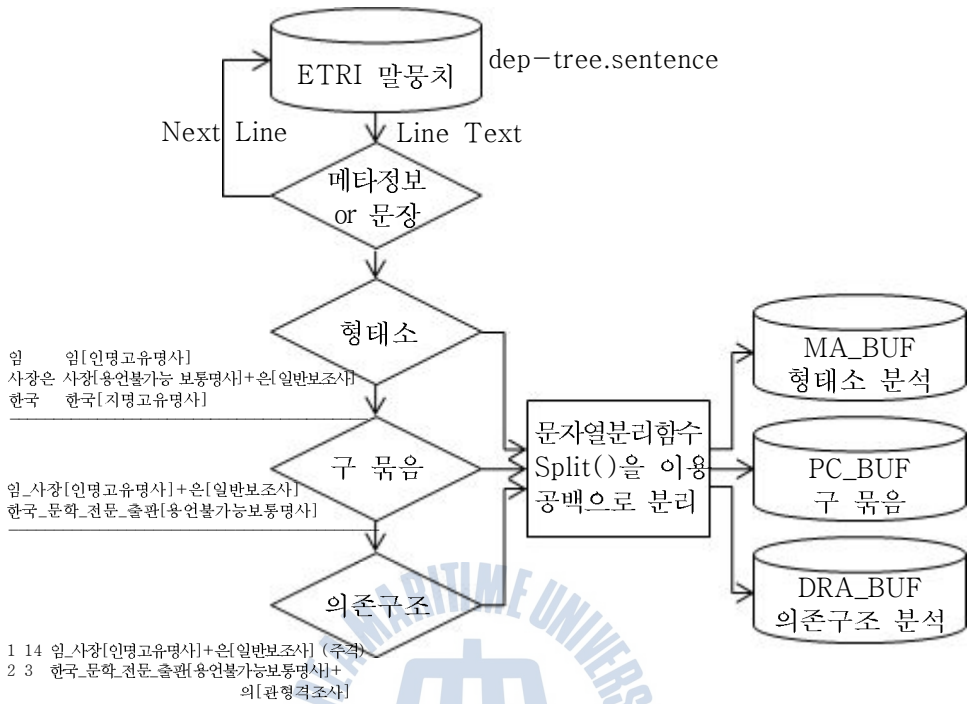


Fig. 3 ETRI 말뭉치 추출기

예를 들면, 리스트 자료형인 “MA_BUF”에는 다음과 같이 저장된다.

사장은, 사장[용언불가능보통명사]+은[일반보조사]

다음으로 구묶음 “입_사장[인명고유명사]+은[일반보조사] 한국_문학_전문_출판[용언불가능보통명사]+의[관형격조사] 길[용언불가능보통명사]+을[목적격조사] 포기하[일반동사] ...중략...”를 입력받으면 문자열 분리함수 "split()"을 사용하여 구묶음 “입_사장[인명고유명사]+은[일반보조사]”, “한국_문학_전문_출판[용언불가능보통명사]+의[관형격조사]” 등 각각을

공백으로 분리하고, 리스트 자료형 “PC_BUF”에 저장한 후, 다음 Line의 Text를 읽는다. 다음 Line의 Text가 구 묶음이면 앞의 과정을 반복적으로 수행하면서 리스트 자료형인 “PC_BUF”에 저장한다.

예를 들면, 리스트 자료형인 “PC_BUF”에는 다음과 같이 저장된다.

임_사장[인명고유명사]+은[일반보조사],
한국_문학_전문_출판[용언불가능보통명사]+의[관형격조사],
길[용언불가능보통명사]+을[목적격조사],
.....중략.....

마지막으로 의존구조 분석 “1 14 임_사장[인명고유명사]+은[일반보조사] (주격)”를 입력받으면 문자열 분리함수 "split()"을 사용하여 구 묶음 번호 “1”, 중심어 번호“14”, 구 묶음 “임_사장[인명고유명사]+은[일반보조사]”를 공백으로 분리하고, 리스트 자료형 “DRA_BUF”에 저장한 후 다음 Line의 Text를 읽는다. 다음 Line의 Text가 의존구조 분석이면 앞의 과정을 반복적으로 수행하면서 리스트 자료형인 “DRA_BUF”에 저장한다.

예를 들면, 리스트 자료형인 “DRA_BUF”에는 다음과 같이 저장된다.

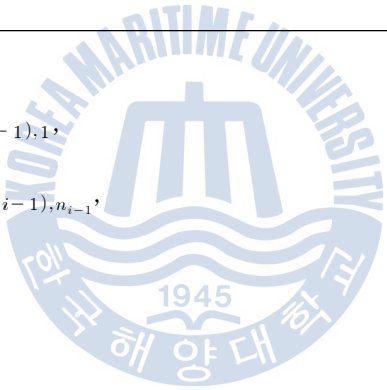
1, 14, 임_사장[인명고유명사]+은[일반보조사]

위와 같이 말뭉치 추출 작업을 반복적으로 수행한 후, ETRI 말뭉치 추출기의 출력은 격조사 복원 자질 추출기의 입력이 된다.

3.3 격조사 복원 자질 추출기

3.3.1 격조사 복원 자질집합

본 논문에서 적용하는 자질집합은 생략된 조사를 포함하는 체언 w_i 를 기준으로 체언 앞에 오는 어절의 첫 형태소 $m_{(i-1),1}$ 와 끝 형태소 $m_{(i-1),n_{i-1}}$, 체언 뒤에 따라오는 용언 w_j ($j > i$)의 첫 형태소 $m_{j,1}$ 와 끝 형태소 m_{j,n_j} , 체언과 용언 사이의 거리 $d(w_i, w_j) = j - i$ 이다. 한 문장이 n_i 개의 형태소 $m_{i,1} m_{i,2} \dots m_{i,n_i}$ 로 구성되고 그 형태소의 품사가 $t_{i,1} t_{i,2} \dots t_{i,n_i}$ 라면 자질벡터(feature vector)는 아래와 같다.



$$\begin{aligned}
 < m_{(i-1),1}, t_{(i-1),1}, \\
 & m_{(i-1),n_{i-1}}, t_{(i-1),n_{i-1}}, \\
 & m_{i,1}, t_{i,1}, \\
 & m_{j,1}, t_{j,1}, \\
 & m_{j,n_j}, t_{j,n_j}, \\
 & d(w_i, w_j) >
 \end{aligned}$$

예를 들어, 입력된 문장 중 “한국 문학 전문 출판의 길을 포기하지 않겠다며,”의 구 묶음 결과가 “한국_문학_전문_출판[용언불가능보통명사]+의[관형격조사] 길[용언불가능보통명사]+을[목적격조사] 포기하[일반동사]+지_않[기타보조용언]+겠[미래시제선어말어미]+다며[종속연결어미]+,[컴마기호]”라고 가정하면 자질벡터는 아래와 같이 된다.

$< m_{(i-1),1}, t_{(i-1),1},$	출판, 용언불가능보통명사,
$m_{(i-1),n_{i-1}}, t_{(i-1),n_{i-1}},$	의, 관형격조사,
$m_{i,1}, t_{i,1},$	길, 용언불가능보통명사,
$m_{j,1}, t_{j,1},$	포기하, 일반동사,
$m_{j,n_j}, t_{j,n_j},$	다며, 종속연결어미
$d(w_i, w_j) >$	1

위 예의 자질벡터 기준으로 설명하면, “U00:%x[0,0]”는 문장의 중심이 되는 체언 앞에 오는 어절의 첫 형태소 “출판”이고, “U01:%x[0,1]”은 첫 형태소의 품사 “용언불가능보통명사”이며, “U02:%x[0,2]”는 체언 앞에 오는 어절의 끝 형태소인 “의”이고, “U03:%x[0,3]”는 끝 형태소의 품사 “관형격조사”이다. “U04:%x[0,4]”는 문장의 중심이 되는 체언 “길”이고, “U05:%x[0,5]”는 체언의 품사 “용언불가능보통명사”이다. “U06:%x[0,6]”는 체언 뒤에 따라오는 용언의 첫 형태소 “포기하”이고, “U07:%x[0,7]”는 용언의 첫 형태소의 품사 “일반동사”이다. “U08:%x[0,8]”는 체언 뒤에 따라오는 용언의 끝 형태소 “다며”이고, “U09:%x[0,9]”는 용언의 끝 형태소의 품사 “종속연결어미”이다. “U10:%x[0,10]”은 문장의 중심이 되는 체언과 용언 사이의 거리 “1”이다. 각각의 자질들 중 상관관계가 높다고 판단되는 자질들을 “U12:%x[0,1]/%x[0,4]”, “U13:%x[0,1]/%x[0,5]”, “U14:%x[0,3]/%x[0,4]”, “U15:%x[0,3]/%x[0,5]”, “U16:%x[0,4]/%x[0,6]”, “U17:%x[0,4]/%x[0,7]”, “U18:%x[0,5]/%x[0,6]”, “U19:%x[0,5]/%x[0,7]”, “U21:%x[0,1]/%x[0,3]/%x[0,4]”, “U22:%x[0,1]/%x[0,3]/%x[0,5]”,

“U24:%x[0,4]/%x[0,7]/%x[0,9]”, “U25:%x[0,5]/%x[0,7]/%x[0,9]”와 같이 묶어서 구성한다.

U00:%x[0,0]	출판
U01:%x[0,1]	용언불가능보통명사
U02:%x[0,2]	의
U03:%x[0,3]	관형격조사
U04:%x[0,4]	길
U05:%x[0,5]	용언불가능보통명사
U06:%x[0,6]	포기하
U07:%x[0,7]	일반동사
U08:%x[0,8]	다며
U09:%x[0,9]	종속연결어미
U10:%x[0,10]	1
U12:%x[0,1]/%x[0,4]	용언불가능보통명사/길
U13:%x[0,1]/%x[0,5]	용언불가능보통명사/용언불가능보통명사
U14:%x[0,3]/%x[0,4]	관형격조사/길
U15:%x[0,3]/%x[0,5]	관형격조사/용언불가능보통명사
U16:%x[0,4]/%x[0,6]	길/포기하
U17:%x[0,4]/%x[0,7]	길/일반동사
U18:%x[0,5]/%x[0,6]	용언불가능보통명사/포기하
U19:%x[0,5]/%x[0,7]	용언불가능보통명사/일반동사
U21:%x[0,1]/%x[0,3]/%x[0,4]	용언불가능보통명사/관형격조사/길
U22:%x[0,1]/%x[0,3]/%x[0,5]	용언불가능보통명사/관형격조사/용언불가능보통명사
U24:%x[0,4]/%x[0,7]/%x[0,9]	길/일반동사/종속연결어미
U25:%x[0,5]/%x[0,7]/%x[0,9]	용언불가능보통명사/일반동사/종속연결어미

본 논문의 실험에서 사용될 자질집합은 Fig. 4와 같다.

```
# Unigram
U00:%x[0,0]
U01:%x[0,1]
U02:%x[0,2]
U03:%x[0,3]
U04:%x[0,4]
U05:%x[0,5]
U06:%x[0,6]
U07:%x[0,7]
U08:%x[0,8]
U09:%x[0,9]
U10:%x[0,10]

U12:%x[0,1]/%x[0,4]
U13:%x[0,1]/%x[0,5]
U14:%x[0,3]/%x[0,4]
U15:%x[0,3]/%x[0,5]
U16:%x[0,4]/%x[0,6]
U17:%x[0,4]/%x[0,7]
U18:%x[0,5]/%x[0,6]
U19:%x[0,5]/%x[0,7]

U21:%x[0,1]/%x[0,3]/%x[0,4]
U22:%x[0,1]/%x[0,3]/%x[0,5]

U24:%x[0,4]/%x[0,7]/%x[0,9]
U25:%x[0,5]/%x[0,7]/%x[0,9]
```

Fig. 4 자질집합

3.3.2 격조사 복원 자질 추출기

격조사 복원 자질 추출기는 ETRI말뭉치 추출기에서 추출한 형태소 분석(MA_BUF), 구 묶음(PC_BUF), 의존구조 분석(DRA_BUF)에서 격조사 복원을 위한 자질들을 추출하는 것이며, Python으로 구현되었다. Fig. 5와 같이 문장의 중심이 되는 체언을 기준으로 실험에 사용할 자질을 추출하는 과정이다. 격조사 복원 자질 추출기는 자질을 추출하기 위한 사전 작업으로 형태소와 품사를 분리하는 단계와 자질을 추출하는 두 단계로 구분하여 실행된다. 먼저 형태소와 품사를 분리하는 작업으로 ETRI말뭉치 추출기에서 추출한 형태소 분석 “MA_BUF”와 구 묶음 “PC_BUF”를 문자열 분리함수 “split(‘+’)”을 사용하여 분리한 후 형태소와 품사를 분리한다.

예를 들어, MA_BUF(형태소 분석)에 저장되어 있는 “[[임, 임[인명고유명사], [사장은, 사장[용언가능보통명사]+은[일반보조사]],중략.....]”에서 어절은 제외하고 형태소만 아래와 같이 분리하여 리스트 자료형 “MA_BUF_N”에 저장한다.

```
[ [임],          [인명고유명사],          임,          0 ],  
[ [사장, 은], [용언가능보통명사, 일반보조사], 사장, 은, 0 ],  
.....중략.....
```

PC_BUF(구 묶음)도 위 MA_BUF과 동일하게 형태소와 품사를 아래와 같이 분리하여 리스트 자료형 “PC_BUF_N”에 저장한다.

[[임_사장, 은], [인명고유명사, 일반보조사],
 임_사장은, 0],
 [[한국_문학_전문_출판, 의], [용언가능보통명사, 관형격조사],
 한국_문학_전문_출판의, 0],
중략.....

다음은 “MA_BUF_N(형태소 분석)”과 “PC_BUF_N(구 묶음)”의 상호
 각 배열의 위치를 계산하여 “MA_BUF_N”에 “PC_BUF_N”의 각 배열
 을 저장한다.

예를 들어, “MA_BUF_N”의 배열이 Table 2, “PC_BUF_N”의 배열이
 Table 3과 같다고 가정하면 Table 4와 같이 “MA_BUF_N”에는
 “PC_BUF_N”의 배열 “1, 2, 3”이 저장되고 Table 5와 같이
 “PC_BUF_N”에는 “MA_BUF_N”의 배열 “5, 6, 7”이 저장된다.

Table 2 MA_BUF_N

배열	MA_BUF_N에 저장된 List 값
5	[출판, 의], [용언불가능보통명사, 관형격조사], 출판의, 0
6	[길, 을], [용언불가능보통명사, 목적격조사], 길을, 0
7	[포기하, 지], [일반동사, 종속연결어미], 포기하지, 0

Table 3 PC_BUF_N

배열	PC_BUF_N에 저장된 List 값
1	[한국_문학_전문_출판, 의], [용언불가능보통명사, 관형격조사], 한국_문학_전문_출판의, 0
2	[길, 을], [용언불가능보통명사, 목적격조사], 길을, 0
3	[포기하, 지않, 겠, 다며], [기타보조용언, 미래시제선어말어미, 종속연결어미], 포기하지않겠다며, 0

Table 4 MA_BUF_N

배열	MA_BUF_N에 저장된 List 값
5	[출판, 의], [용언불가능보통명사, 관형격조사], 출판의, 1
6	[길, 을], [용언불가능보통명사, 목적격조사], 길을, 2
7	[포기하, 지], [일반동사, 종속연결어미], 포기하지, 3

Table 5 PC_BUF_N

배열	PC_BUF_N에 저장된 List 값
1	[한국_문학_전문_출판, 의], [용언불가능보통명사, 관형격조사], 한국_문학_전문_출판의, 5
2	[길, 을], [용언불가능보통명사, 목적격조사], 길을, 6
3	[포기하, 지않, 겠, 다며], [기타보조용언, 미래시제선어말어미, 종속연결어미], 포기하지않겠다며, 7

다음은 자질을 추출하는 단계이다. 먼저 구 묶음 “PC_BUF_N”을 읽어서 배열에 격조사가 없으면 Next Array로 Skip하여 다음 배열을 읽고, 격조사가 있으면 문장의 중심이 되는 체언과 체언의 품사, 격조사를 추출한다.

예를 들면, 아래의 배열을 각각 읽었다고 가정하면 첫 번째는 본 논문에서 적용하는 격조사가 아니기 때문에 Skip하고 목적격조사가 있는 배열에서 자질을 추출한다.

[출판, 의], [용언불가능보통명사, 관형격조사], 출판의, 1
[길, 을], [용언불가능보통명사, 목적격조사], 길을, 2

자질 추출결과는 다음과 같다.

길, 용언불가능보통명사, 목적격조사

두 번째로 체언 앞에 오는 어절의 첫 형태소와 품사, 끝 형태소와 품사는 “MA_BUF_N”에서 자질을 추출한다. 추출은 “MA_BUF_N”에 저장된 “PC_BUF_N”의 배열의 위치를 참조하여 “MA_BUF_N”의 배열의 위치를 찾는다.

예를 들면, 체언 앞에 오는 어절이 아래와 같다고 가정한다.

[출판, 의], [용언불가능보통명사, 관형격조사], 출판의, 1

체언 앞에 오는 어절의 자질 추출결과는 다음과 같다.

출판, 용언불가능보통명사, 의, 관형격조사

세 번째로 체언과 용언 사이의 거리를 계산하여 마지막 자질을 추출한다. 체언과 용언사이의 거리는 “DRA_BUF”에 저장되어 있는 구 묶음 번호와 중심어 번호를 참조하여 자질을 추출한다.

예를 들면, “PC_BUF_N”에 있는 배열의 위치가 “DRA_BUF”에 아래와 같이 저장되어 있다고 가정하면, 체언과 용언 사이의 거리는 구 묶음 번호 ‘3’과 중심어 번호 ‘4’의 차인 ‘1’을 자질로 추출한다.

3 4 길[용언불가능보통명사]+을[목적격조사]

위와 같이 문장의 중심이 되는 체언의 앞 어절 첫 형태소와 품사, 끝 형태소와 품사, 체언과 체언의 품사, 체언의 격을 결정하는 격조사의 품사, 체언 뒤에 오는 용언의 첫 형태소와 품사, 끝 형태소와 품사, 체언과 용언 사이의 거리 등 추출한 자질은 "TRIAN_BUF"에 다음과 같이 저장한다.

출판, 용언가능보통명사, 의, 관형격조사, 길, 용언불가능
보통명사, 포기하, 일반동사, 다며, 종속연결어미, 1, 목적
격조사

격조사 복원 자질 추출기 Fig. 5는 위에서 설명한 것과 같이 하나의 문장을 반복적으로 수행하여 추출한 자질들을 "TRAIN_BUF"에 순차적으로 저장한다. 격조사 복원 자질 추출기는 위 작업을 문장의 끝을 만날 때까지 반복적으로 수행하고 완료되면 "TRAIN_BUF"에 저장된 자질들을 학습말뭉치 "TrainingData.txt" 파일에 저장하고 종료한다.

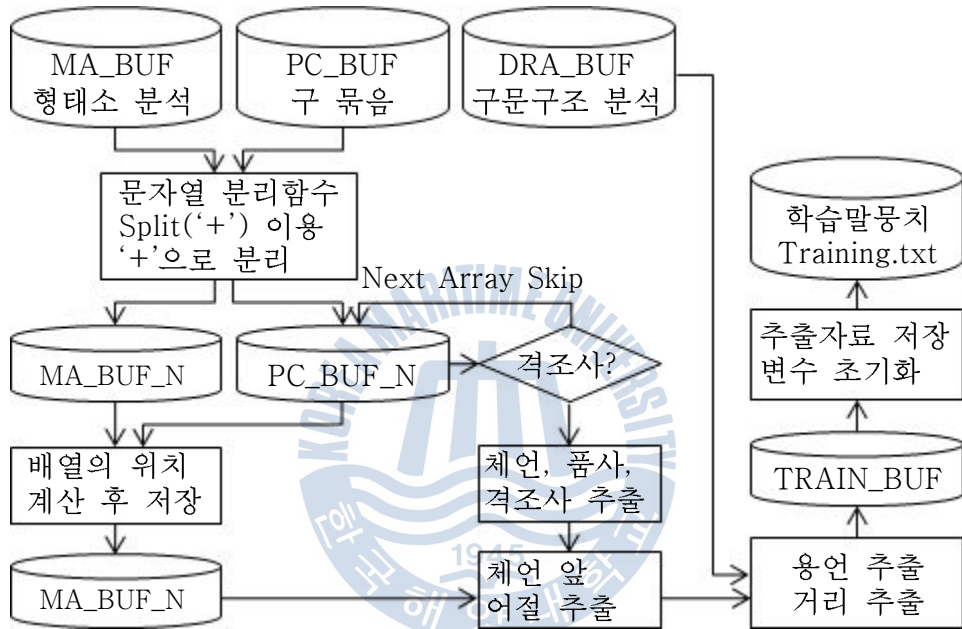


Fig. 5 격조사 복원 자질 추출기

3.4 격조사 복원기

격조사 복원 자질 추출기에서 추출한 학습말뭉치(Training.txt)의 90%는 학습말뭉치(training corpus)로 10%는 실험말뭉치(test corpus)로 분리하여 격조사 복원기의 실험에 활용한다. 격조사 복원기는 Fig. 6과 같이 학습말뭉치를 기계학습기(learner)로 학습시키면 모델(model)이 생성되고 모델과 격조사가 생략된 실험말뭉치를 입력받아 분류기(classifier)가 실행된다. 분류기는 실험말뭉치에서 생략된 격조사를 복원한 후 결과를 출력한다.

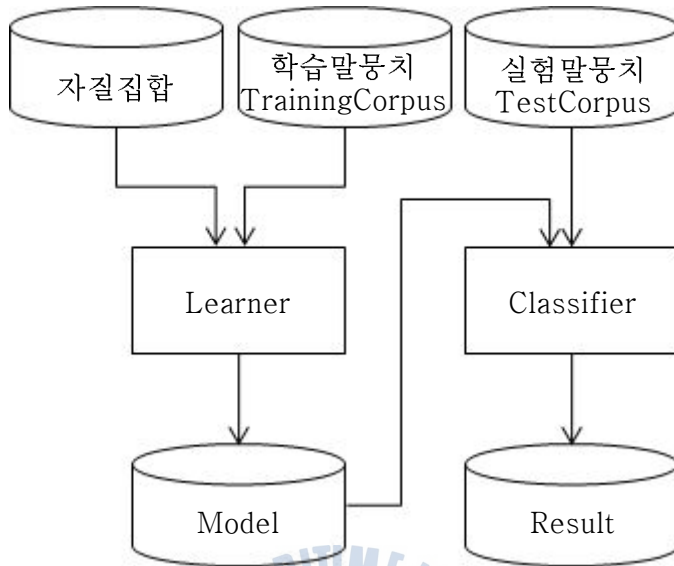


Fig. 6 격조사 복원기

예를 들면, 격조사가 생략된 실험말뭉치가 Fig. 7과 같다면 분류기는 실행 후 생략된 격조사를 복원하여 Fig. 8과 같이 출력한다.

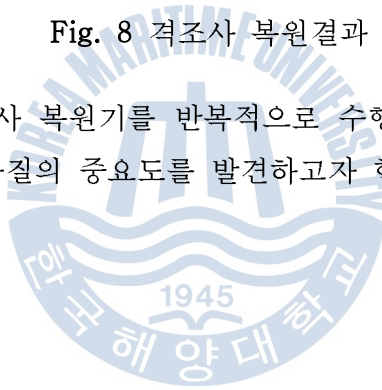
$< m_{(i-1),1}, t_{(i-1),1},$	있,	정상형용사,
$m_{(i-1),n_{i-1}}, t_{(i-1),n_{i-1}},$	는,	관형사형전성어미,
$m_{i,1}, t_{i,1},$	처지,	용언불가능보통명사,
$m_{j,1}, t_{j,1},$	이,	긍정지정사,
$m_{j,n_j}, t_{j,n_j},$	라면,	종속연결어미
$d(w_i, w_j) >$	1	

Fig. 7 격조사가 생략된 실험 말뭉치

$< m_{(i-1),1}, t_{(i-1),1},$	있,	정상형용사,
$m_{(i-1),n_{i-1}}, t_{(i-1),n_{i-1}},$	는,	관형사형전성어미,
$m_{i,1}, t_{i,1},$	처지,	용언불가능보통명사,
$m_{j,1}, t_{j,1},$	이,	긍정지정사,
$m_{j,n_j}, t_{j,n_j},$	라면,	종속연결어미
$d(w_i, w_j)$	1	
복원결과 >		부사격조사

Fig. 8 격조사 복원결과

본 논문에서는 격조사 복원기를 반복적으로 수행하여 최상의 자질집합과 자질집합에서 각 자질의 중요도를 발견하고자 한다.



제 4 장 실험 및 평가

4.1 실험환경

본 논문에서는 한국어 문장에서 자주 생략되는 조사를 복원하는데 있어 필요한 자질집합과 자질추출, 격조사 복원 성능, 자질집합이 격조사 복원 성능에 미치는 영향을 밝혀내는 실험을 수행하고 그 결과를 평가한다. 조사는 주격, 서술격, 관형격, 목적격, 보격, 부사격, 호격조사가 있으나, 본 논문에서는 체언과 용언 사이에 있는 격조사만을 대상으로 실험한다. 따라서 체언에 붙어서 그 체언을 서술어 역할을 하도록 하는 서술격조사와 체언과 체언사이에 있고 조사가 생략될 경우 구 묶음 과정에서 복합명사로 인식되어 하나의 덩어리로 묶이는 관형격조사는 본 논문의 범위를 벗어나기 때문에 실험에서 제외한다. 본 논문에서는 Table 6과 같이 앞의 체언을 주어가 되게 하는 “이, 가, 께서, 에서”등의 주격조사, 앞의 체언을 타동사의 목적어가 되게 하는 “을, 를”등의 목적격조사, 앞의 체언을 보어가 되게 하는 “이, 가”등의 보격조사, 앞의 체언을 부사어가 되게 하는 “에, 에서, 에게” 등의 부사격조사, 독립어로서의 호칭이 되게 하는 “야, 여, 아, 이여, 시여”등의 호격조사만을 대상으로 실험한다.

본 논문에서는 ETRI 말뭉치 추출기와 격조사 복원 자질 추출기는 Python v3.3.2²⁾으로 구현하였고 실험은 기계학습 방법으로 CRF++³⁾이라는 도구를 사용하였다. 모든 실험은 10차 교차검증(10-fold cross validation) 방법으로 평가한다.

2) <http://www.python.org/download/>

3) <http://crfpp.googlecode.com/svn/trunk/doc/index.html#format>

Table 6 실험대상 격조사 표

종 류	조사의 예	비 고
주격조사	이, 가, 께서, 에서	앞의 체언을 주어가 되게 하는 조사
목적격조사	을, 를	앞의 체언을 타동사의 목적어가 되게 하는 조사
보격조사	이, 가	앞의 체언을 보어가 되게 하는 조사
부사격조사	에, 에서, 에게	앞의 체언을 부사어가 되게 하는 조사
호격조사	야, 여, 아, 이어, 시어	독립어로서의 호칭이 되게 하는 조사

4.2 체언과 용언 사이의 거리 분석

ETRI 구문구조 부착 말뭉치의 전체 101,602 문장 중에서 실험과정에서 발견된 오류문장을 제외하고, Table 7과 같이 101,565 문장에서 격조사가 있는 체언을 중심으로 543,306개의 <체언 앞 어절, 체언, 용언, 거리 - 클래스> 쌍을 추출하였다. 추출한 각 조사의 비율을 보면, 전체 자질 543,306개 중 주격조사의 자질이 119,055개로 21.90%이고, 목적격조사의 자질이 171,537개로 31.37%이고, 보격조사의 자질이 54,344개로 10.00%이고, 부사격조사의 자질이 198,318개로 36.50%이고, 호격조사의 자질이 52개로 0.01%로 나타났다. 추출된 543,306개의 자질을 Table 8와 같이 체언과 용언 사이의 거리별로 분석한 결과 호격조사를 제외한 주격조사, 목적격조사, 보격조사, 부사격조사는 체언과 용언의 거리가 3 이하에 많이 위치한다는 것을 알 수 있다. 보격조사의 경우 체언과 용언의 거리가 1 이상은 자질의 수가 현격하게 적게 나타났다.

Table 7 추출 말뭉치 전체 개수 및 비율

종 류	개 수	비 율
주격조사	119,055	21.90
목적격조사	171,537	31.57
보격조사	54,344	10.00
부사격조사	198,318	36.50
호격조사	52	0.01
합 계	543,306	101,565문장

4.3 격조사 복원 성능

4.3.1. 실험방법

격조사 복원 자질 추출기에서 추출한 학습 말뭉치(Training.txt)를 10차 교차검증을 위한 분리작업을 선행한다. “Training.txt”의 90%는 학습말뭉치인 “TrainingCorpus.txt”로 10%는 실험 말뭉치인 “TestCorpus.txt”로 각각 분리하여 생성한다. 격조사 복원 성능이 가장 우수하다고 평가된 자질집합과 “TrainingCorpus.txt”를 Fig. 9과 같은 학습기를 실행하면 “Model.txt”파일이 생성된다. 학습기 실행명령어는 다음과 같다.

```
$ crf_learn Template TrainingCorpus.txt Model.txt
```

Table 8 체언과 용언 사이의 거리별 격조사의 개수

거리	주격조사	목적격조사	보격조사	부사격조사	호격조사
1	59,563	131,635	54,230	97,846	10
2	24,371	25,442	84	37,180	6
3	11,842	6,838	11	19,992	8
4	7,130	2,981	5	11,396	3
5	4,484	1,595	1	7,214	2
6	2,996	906	2	4,938	2
7	2,160	532	5	3,699	1
8	1,557	394	2	2,712	3
9	1,151	270	1	2,200	1
10	897	177	2	1,807	0
11	688	169	0	1,560	2
12	578	150	0	1,389	3
13	413	106	0	1,284	1
14	355	99	1	1,237	0
15	280	100	0	1,087	4
16	228	47	0	950	2
17	161	43	0	738	1
18	100	29	0	502	1
19	51	14	0	273	1
20	26	2	0	173	1
21	16	2	0	72	0
22	5	2	0	45	0
23	1	2	0	14	0
24	1	2	0	9	0
25	1	0	0	1	0
합계	119,055	171,537	54,344	198,318	52

```

[hong8e@nlp PostPosition]$ ls
Template TestCorpus.txt TrainingCorpus.txt
[hong8e@nlp PostPosition]$ crf_learn Template TrainingCorpus.txt Model.txt
CRF++: Yet Another CRF Tool Kit
Copyright (C) 2005-2012 Taku Kudo, All rights reserved.

reading training data:
Done!18.48 s

Number of sentences: 1
Number of features: 4503080
Number of thread(s): 4
Freq: 1
eta: 0.00010
C: 1.00000
shrinking size: 20
iter=0 terr=0.63591 serr=1.00000 act=4503080 obj=623385.19606 diff=1.00000
iter=1 terr=0.61117 serr=1.00000 act=4503080 obj=481675.07384 diff=0.22732
iter=2 terr=0.45236 serr=1.00000 act=4503080 obj=424147.15656 diff=0.11943
iter=3 terr=0.45842 serr=1.00000 act=4503080 obj=391871.17284 diff=0.07610

```

Fig 9 학습기 CRF++ 실행화면

생성된 Model.txt” 파일과 실험 말뭉치 “TestCorpus.txt”파일을 Fig. 10과 같이 분류기를 실행하면 “TestCorpus.txt”파일에 복원된 격조사를 부착한 “Result.txt” 파일이 생성된다. 분류기 실행명령어는 다음과 같다.

```
$ crf_test -m Model.txt TestCorpus.txt
```

```

[hong8e@nlp PostPosition]$ ls
Model.txt Template TestCorpus.txt TrainingCorpus.txt
[hong8e@nlp PostPosition]$
[hong8e@nlp PostPosition]$ crf_test -m Model.txt TestCorpus.txt
상당하 [성상형용사] ㄴ [관형사형전성어미] 비중 [용언불가능보통명?
自? [목적격조사] 화엘? 고 [대응연결어미] 1 [목적격조사]
중속 [용언가능보통명사] 의 [관형격조사] 문제 [용언불가능보통?
自? [목적격조사] 화엘? 면서 [대응연결어미] 1 [목적격조사]
계급 [용언불가능보통명사] 계급 [용언불가능보통명사] 구조 [용언가?
매糶自? 구현되 [일반동사] 는 [관형사형전성어미] 1 [
주격조사] [주격조사]
표상 [용언가능보통명사] 의 [관형격조사] 체계 [용언가능보통명?
? [보격조사] 旋ㄴ ㄹ? 며 [대응연결어미] 1 [보격조사]
동시에 [지시시간부사] 동시에 [지시시간부사] 그것 [지시대명사] 생산되 [
일반동사] [주격조사] [관형사형전성어미] 1 [부사격조사]
생산되 [일반동사] ㄴ [관형사형전성어미] 생산물 [용언불가능보통?
自? 보 [일반동사] ㄴ다 [평서형종결어미] 1 [부사격?
떨? [목적격조사]
교수 [용언불가능보통명사] 의 [관형격조사] 연구 [용언가능보통명?
? [주격조사] 旋ㄴ ㄹ? ㄴ데 [종속연결어미] 2 [주격조사]
연구 [용언가능보통명사] 가 [주격조사] 대표적 [용언불가능보통?
自? [보격조사] 旋ㄴ ㄹ? ㄴ데 [종속연결어미] 1 [보격조사]
증가하 [일반동사] 어 [종속연결어미] 사회 [용언불가능보통명사] ?
흘나? ? [목적격조사] 고 [대응연결어미] 1 [주격조사]

```

Fig. 10 분류기 CRF++ 실행화면

4.3.2. 격조사 복원 성능분석

체언과 용언 사이의 거리를 고려하지 않고, 실험하여 Table 9와 같은 결과를 확인했다. 전체 격조사 복원 성능은 약 77.02%로 나타났고, 보격 조사가 94%, 목적격조사가 84%, 부사격조사 72%, 주격조사 64%의 순으로 격조사 복원 성능을 확인했다. 호격조사의 경우 자질의 수가 52개로 매우 작기는 하지만 복원성능이 0%로 확인된 것이 특이하다고 할 수 있다.

Table 9 격조사 복원 성능 평균(거리: 1 ~ 25)

구 분	Macro-Average	Micro-Average	총개수	평균개수	
				정답	오답
주격조사	64.79	64.76	119,055	7,710.10	4,195.40
목적격조사	84.72	84.72	171,537	14,532.00	2,621.20
보격조사	94.65	94.51	54,344	5,136.00	298.40
부사격조사	72.96	72.94	198,318	14,464.40	5,367.40
호격조사	0.00	0.00	52	0.00	5.20
전 체	77.02	77.02	543,306	41,842.5	12,487.6

4.4 거리별 격조사 복원 성능분석

4.4.1. 전체 거리별 성능분석

체언과 용언 사이의 거리별 격조사 복원 성능을 실험하고, Table 10과 같은 결과를 확인했다. 격조사 복원 성능은 주격조사의 경우 체언과 용언 사이의 거리가 1일 때 67.39%, 2일 때 64.17%, 3일 때 60.48% 등 거리별 약간의 성능차이는 있었으나, 큰 변화는 없었고, 목적격 조사의 경우, 거리가 1일 때는 91.98%의 성능을 보이다가 거리가 2일 때 72.57%, 3일 때 49.03% 등으로 체언과 용언 사이의 거리가 1씩 멀어짐에 따라 격조사 복원 성능이 현격하게 낮아지는 현상을 확인했다. 보격조사의 경우 체언과 용언 사이의 거리가 1일 때는 복원 성능이 94.78%로 매우 우수한 것으로 확인되었으나 2 이상일 경우에는 Table 9에서도 알 수 있었듯이 자질의 개수가 현격하게 작아졌고 격조사 복원 성능도 0%로 확인되었다. 부사격 조사의 경우 주격조사와는 반대로 체언과 용언 사이의 거리가 1일 때

71.93%, 2일 때 67.65%, 3일 때 72.79%, 4일 때 74.64% 5일 때 76.65% 등으로 거리가 멀어지면 멀어질수록 복원성능이 우수한 것으로 확인되었다. 주격조사와 부사격조사의 경우 체언과 용언 사이의 거리에 무관하게 거의 동등한 수준의 복원성능이 확인되었고, 목적격조사와 보격조사의 경우 체언과 용언 사이의 거리가 인접해야만 복원 성능이 우수하다는 것을 확인했다.

Table 10 거리별 격조사 복원 성능 평균(거리: 1 ~ 25)

Ma : Macro-Average, Mi : Micro-Average

거리	주격조사		목적격조사		보격조사		부사격조사		호격조사	
	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi
1	67.39	67.30	91.97	91.98	94.85	94.71	71.92	71.82	00.00	00.00
2	64.13	64.09	72.57	72.55	00.00	00.00	67.62	67.64	00.00	00.00
3	60.45	60.44	49.03	49.02	00.00	00.00	72.78	72.78	00.00	00.00
4	62.26	62.29	39.50	39.42	00.00	00.00	74.64	74.63	00.00	00.00
5	62.74	62.76	34.81	34.73	00.00	00.00	76.63	76.63	00.00	00.00
6이상	63.05	63.02	26.25	26.31	00.00	00.00	79.81	79.81	00.00	00.00

4.4.2. 거리 1과 2의 성능분석

전체 거리별 격조사 복원 성능분석에서 체언과 용언 사이의 거리가 2이상일 경우, 격조사 복원 성능이 현격하게 낮아진다는 것을 확인했기 때문에 체언과 용언 사이의 거리가 1과 2인 자질만을 대상으로 격조사 복원 성능이 어떻게 변하는지 실험한다. ETRI 구문구조 부착 말뭉치의 101,565 문장에서 체언과 용언 사이의 거리가 1과 2인 자질만을 대상으로 430,367개의 자질을 추출하여 4.1.1절의 실험방법으로 실험을 반복적으로 수행하

였다. Table 11과 같이 격조사 복원 성능 평균을 보면 주격조사의 경우 67.19%로 거리를 제한하지 않았을 때보다 약 3%의 성능이 향상되었고, 목적격조사는 88.83%로 약 4%의 성능이 향상되었고, 보격조사는 94.45%로 약 0.1%의 성능이 저하되었으며, 부사격조사는 70.99%로 약 2%의 성능이 저하되었다. 호격조사의 경우 자질의 수가 16개로 매우 작기는 하지만 16개 모두 격조사 복원이 되지 않았다. 체언과 용언 사이의 거리가 1과 2인 격조사 복원 성능은 79.68%로 4.3.2절의 격조사 복원 성능보다 약 2%의 성능이 향상되었음을 확인했다.

Table 11 격조사 복원 성능 평균(거리: 1 ~ 2)

구 분	Macro-Average	Micro-Average	총개수	평균개수	
				정답	오답
주격조사	67.19	67.14	83,934	5,634.90	2,758.30
목적격조사	88.83	88.83	157,077	13,952.70	1,754.80
보격조사	94.51	94.37	54,314	5,125.30	305.90
부사격조사	70.99	70.94	135,026	9,578.60	3,923.90
호격조사	0.00	0.00	16	0.00	1.60
전 체	79.68	79.68	430,367	34,291.5	8,744.5

Table 12와 같이 거리별 격조사 복원 성능 평균을 보면, 주격조사의 경우 거리가 1일 때 69.31%이고, 2일 때 62.06%로 약 7%정도 복원 성능차이가 있고, 목적격조사의 경우 거리가 1일 때는 92.20%의 매우 높은 복원 성능을 보이다가 2일 때 71.35%로 약 21%의 차이가 났고, 보격조사는 거리가 1일 때는 94.65%로 전체 격조사 복원 성능 중 가장 우수하게 나타난 반면, 거리가 2일 때는 자질의 수가 현격하게 작아진 점도 있지만 0%로 나타난 것이 특이하다. 부사격조사는 거리가 1일 때는 73.05%이고 2일 때

65.65%로 약 8%의 성능저하가 나타났다. 호격조사의 경우에는 전체 거리별 격조사 복원 성능에서와 동일하게 복원이 되지 않았다.

Table 12 거리별 격조사 복원 성능 평균(거리: 1 ~ 2)

Ma : Macro-Average, Mi : Micro-Average

거리	주격조사		목적격조사		보격조사		부사격조사		호격조사	
	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi
1	69.31	69.23	92.20	92.21	94.65	94.51	73.05	72.94	00.00	00.00
2	62.06	62.03	71.35	71.35	00.00	00.00	65.65	65.66	00.00	00.00

위에서 설명한 거리별 격조사 복원 성능을 막대그래프로 표현하면 Fig. 11과 같다.

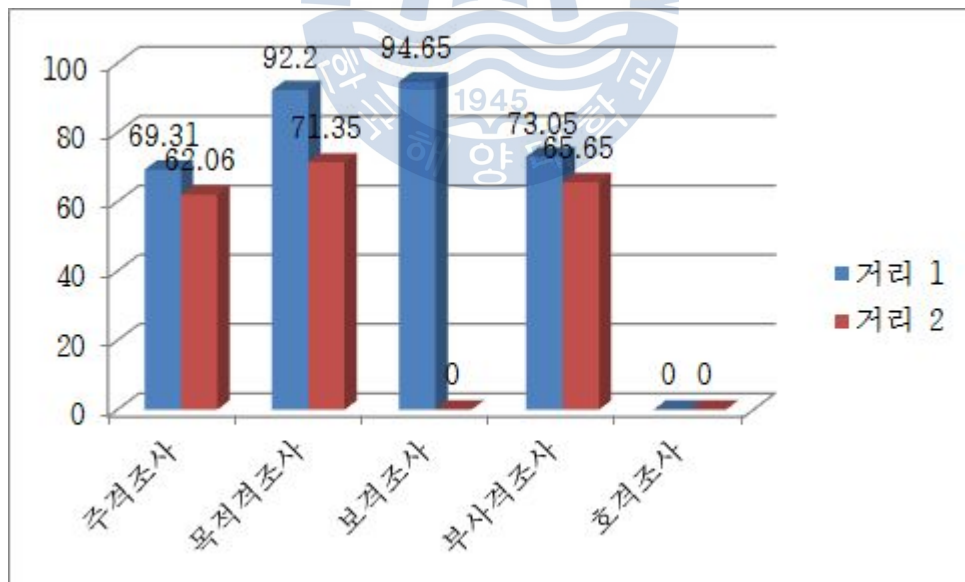


Fig. 11 거리별 격조사 복원 성능 평균(거리: 1 ~ 2)

실험에서 격조사 복원 성능이 높은 보격조사를 10차 교차검증 방법으로

평가한 결과는 Table 13과 같다. 실험 2, 실험 6, 실험 9에서는 격조사 복원 성능이 97%이상으로 확인되었고 실험 3, 실험 4는 격조사 복원 성능 평균 94.65%보다 낮은 90.74%와 90.34%로 확인되었다.

Table 13 보격조사 10-차 교차검증 결과(거리: 1)

구 분	자질개수	정답	정답비율	오답	오답비율
실험 1	4,539	4,276	94.21	263	5.79
실험 2	4,836	4,763	98.49	73	1.51
실험 3	5,400	4,900	90.74	500	9.26
실험 4	6,391	5,774	90.35	617	9.65
실험 5	5,724	5,297	92.54	427	7.46
실험 6	5,430	5,283	97.29	147	2.71
실험 7	5,661	5,329	94.14	332	5.86
실험 8	5,668	5,412	95.48	256	4.52
실험 9	5,126	5,023	97.99	103	2.01
실험 10	5,453	5,196	95.29	257	4.71
합 계	54,228	51,253	94.65	2,975	5.35

4.5 자질별 중요도 분석

자질별 중요도 분석은 위 Fig. 4의 자질집합을 대상으로 각 자질들이 격조사 복원 성능에 어떠한 영향이 있는지를 확인하는 실험이다. ETRI 구문 구조 부착 말뭉치의 101,565 문장에서 체언과 용언 사이의 거리가 1과 2인 자질만을 대상으로 430,367개의 자질을 추출하여 4.1.1절의 실험방법으로 실험을 반복적으로 수행하였다. Fig. 12의 자질별 격조사 복원 성능(macro-average)과 Fig. 13의 자질별 격조사 복원 성능 평균(macro-average), Fig. 14의 자질별 격조사 복원 성능(micro-average)과 Fig. 15의 자

질별 격조사 복원 성능 평균(micro-average)에서 보면, 자질 1에서 자질 4까지는 체언 앞에 오는 어절의 첫 형태소와 품사, 끝 형태소와 품사이고, 자질 5와 자질 6은 체언과 체언의 품사이다. 자질 7에서 자질 10까지는 체언 뒤에 오는 용언의 첫 형태소와 품사, 끝 형태소와 품사이다. 자질 11은 체언과 용언 사이의 거리이다. 실험결과를 분석해 보면, 체언 앞에 오는 어절에 해당하는 자질은 자질집합에서 제외되어도 격조사 복원 성능에는 거의 변화가 없는 것을 확인하였다. 반면 체언에 해당하는 자질 5가 생략될 경우, 보격조사의 경우에는 격조사 복원 성능에는 변화가 없고 목적격조사의 경우에는 약간의 변화는 있었지만, 부사격조사와 주격조사의 경우에는 격조사 복원 성능이 많이 저하되는 것을 확인할 수 있다. 체언의 품사에 해당하는 자질 6이 생략될 경우에도 격조사 복원 성능에는 큰 변화가 없었고 용언에 해당하는 자질 7을 생략할 경우에는 보격조사를 제외한 주격, 목적격, 부사격 조사는 10% 이상 격조사 복원 성능이 저하된 것을 알 수 있다. 자질 8에서 자질 11까지는 자질집합에서 제외되어도 격조사 복원 성능의 변화는 거의 없는 것으로 나타났다.

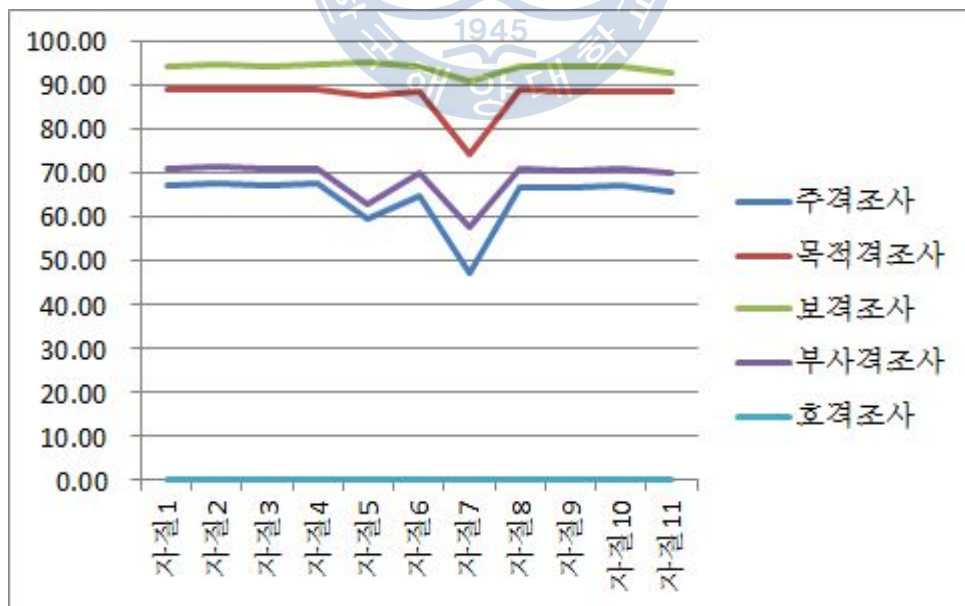


Fig. 12 자질별 격조사 복원 성능(macro-average)

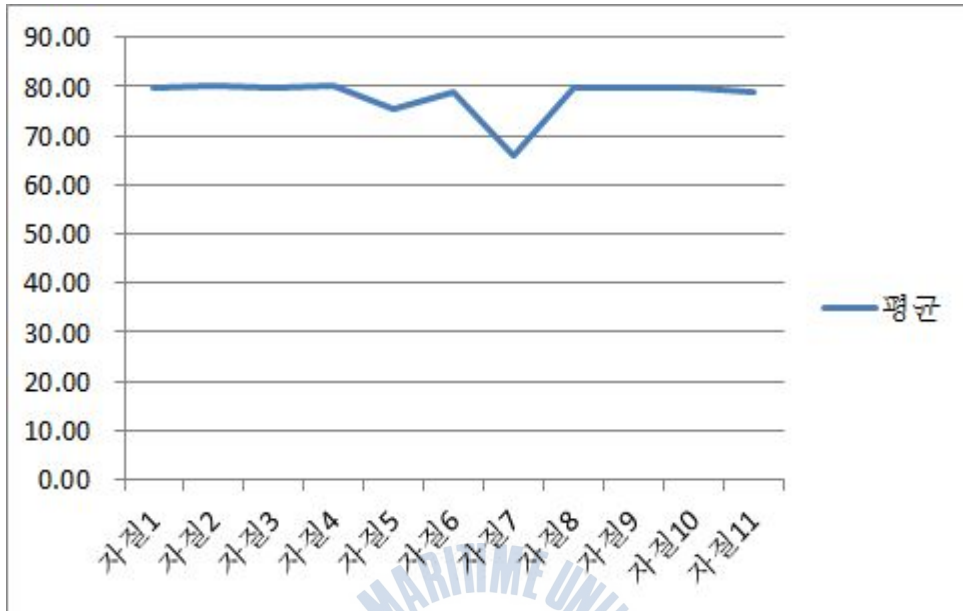


Fig. 13 자질별 격조사 복원 성능 평균(macro-average)

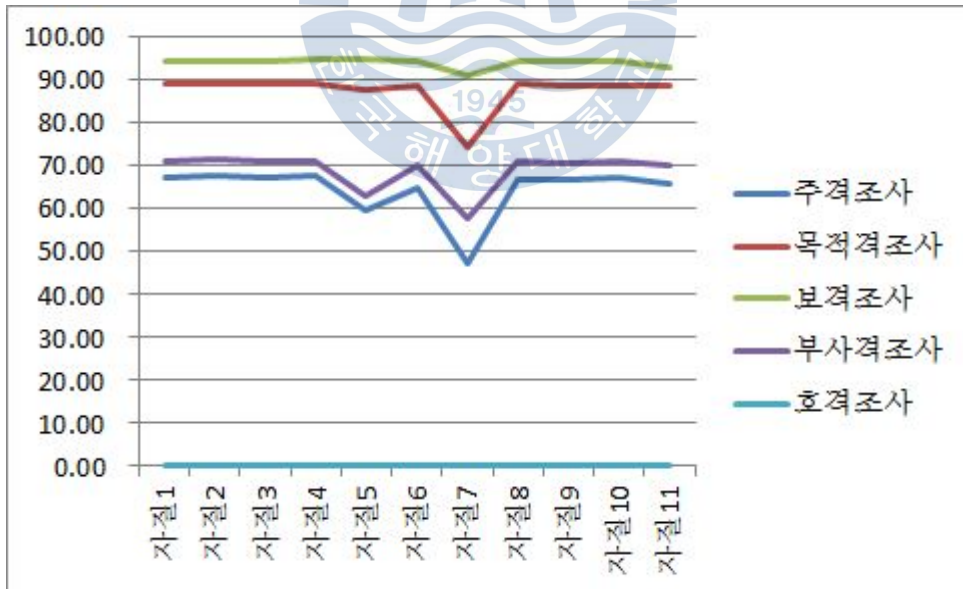


Fig. 14 자질별 격조사 복원 성능(micro-average)

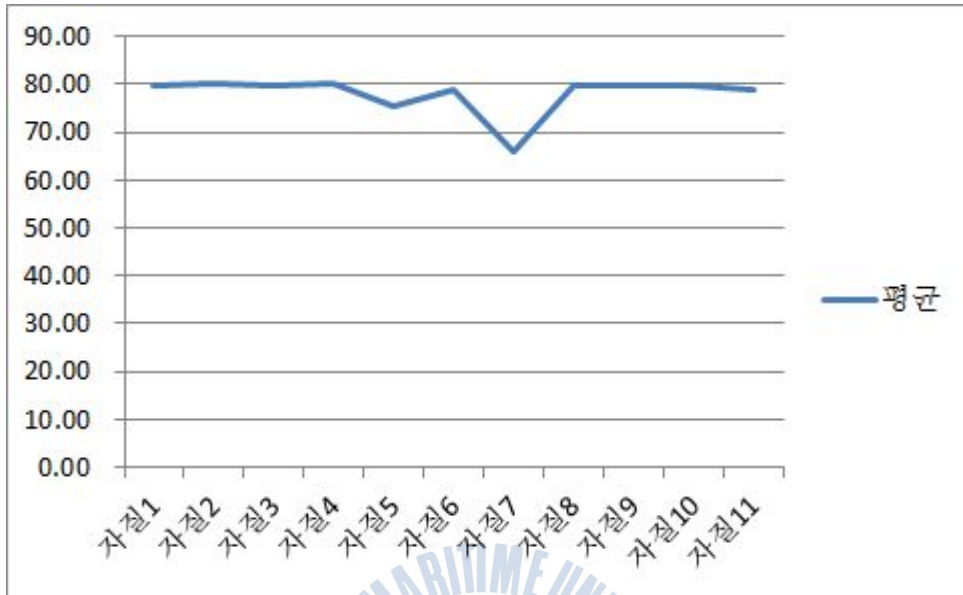


Fig. 15 자질별 격조사 복원 성능 평균(micro-average)

위에서 살펴본 바와 같이 각 자질별 격조사 복원 성능 실험에서 성능이 가장 저하된 자질이 격조사 복원 성능에 가장 큰 영향이 있다고 판단할 수 있다. 실험결과인 Table 14, 15을 분석해 보면 각 자질들 중 7번째 자질인 용언이 66%의 성능으로 가장 낮은 격조사 복원율로 확인되었으며, 다음으로 5번째 자질인 체언이 75%로 확인되었다. 또한 Table 14에서 보면 주격조사와 목적격조사의 경우, 체언 앞에 오는 어절의 첫 형태소 자질1과 품사 자질2, 끝 형태소 자질3과 품사 자질4가 Table 11의 자질집합에 포함된 평균 결과 값 67.19%와 88.83%보다 포함되지 않은 평균 결과 값이 조금 높게 확인되었다. 보격조사와 부사격조사의 경우에도 자질집합에 자질이 포함된 평균 결과 값 94.45%와 70.99%보다 포함되지 않은 평균 결과 값이 조금 높게 확인되었다. 체언 뒤에 따라오는 용언의 첫 형태소 품사 자질8, 끝 형태소 자질 9, 품사 자질 10, 체언과 용언의 거리 자질 11도 자질집합에 포함되었을 때와 제외되었을 때의 격조사 복원 성능 차이는 거의 없으므로 자질 1부터 4까지, 자질 8부터 11까지는 격조사 복원 성능에는 영향이 없으므로 격조사 복원 중요도가 낮다고 할 수 있다.

Table 14 자질별 격조사 복원 성능(macro-average)

구 분	주격 조사	목적격 조사	보격 조사	부사격 조사	호격 조사	평균
자질1	67.39	89.02	94.57	70.99	0.00	79.79
자질2	67.82	89.17	94.62	71.34	0.00	80.05
자질3	67.16	88.89	94.50	70.88	0.00	79.66
자질4	67.57	89.09	94.73	71.22	0.00	79.95
자질5	59.85	87.76	95.09	63.06	0.00	75.44
자질6	64.90	88.82	94.51	70.08	0.00	78.95
자질7	47.42	74.59	91.11	57.79	0.00	66.06
자질8	66.76	88.93	94.48	71.00	0.00	79.63
자질9	66.80	88.78	94.47	70.79	0.00	79.52
자질10	67.29	88.86	94.48	71.13	0.00	79.75
자질11	65.79	88.74	92.81	70.17	0.00	78.91

Table 15 자질별 격조사 복원 성능(micro-average)

구 분	주격 조사	목적격 조사	보격 조사	부사격 조사	호격 조사	평균
자질1	67.39	89.02	94.51	70.98	0.00	79.80
자질2	67.82	89.17	94.56	71.34	0.00	80.06
자질3	67.15	88.89	94.44	70.88	0.00	79.67
자질4	67.57	89.09	94.67	71.22	0.00	79.96
자질5	59.85	87.77	95.03	63.07	0.00	75.45
자질6	64.90	88.83	94.46	70.08	0.00	78.96
자질7	47.43	74.59	91.11	57.79	0.00	66.07
자질8	66.75	88.93	94.43	71.00	0.00	79.64
자질9	66.80	88.78	94.42	70.78	0.00	79.53
자질10	67.30	88.86	94.42	71.14	0.00	79.76
자질11	65.78	88.75	92.76	70.17	0.00	78.92

제 5 장 결론 및 향후 연구

본 논문은 한국어 문장에서 자주 조사가 생략되어 발생하는 중의성을 해소하기 위한 방안으로 문장에서 생략된 조사를 복원하여 구문분석기에 전달함으로써 정보검색기의 성능을 향상시킨다는 점에 착안하여 자질집합이 격조사 복원 성능을 좌우하고 자질집합 중 각 자질들이 격조사 복원 성능에 어떤 영향이 있는지를 실험을 통해 확인해 나가는 격조사 복원시스템에 대하여 기술하였다. 실험에는 형태소 분석, 구 묶음, 의존구조 분석이 된 ETRI 구문구조 부착 말뭉치를 활용하였으며, ETRI 말뭉치 추출기와 격조사 복원 자질 추출기는 Python으로 구현하여 ETRI 구문구조 부착 말뭉치의 전체 101,602 문장 중에서 오류문장을 제외한 101,565문장에서 격조사가 있는 체언을 중심으로 543,306건의 <체언 앞 어절, 체언, 용언, 거리 - 클래스>쌍을 추출하였다. 모든 실험결과의 신뢰도를 높이기 위하여 10차 교차검증 방법을 사용하여 실험결과를 평가하였으며, 격조사 복원기를 반복적으로 수행하여 격조사 복원 성능이 가장 우수한 자질집합을 찾았다. 격조사 복원 자질 추출기에서는 추출한 말뭉치의 90%는 학습 말뭉치로 10%는 실험 말뭉치로 파일을 각각 생성하고, 생성된 말뭉치는 학습기를 수행한 후, 모델을 생성하고 분류기를 실행하는 실험을 반복적으로 수행한 결과 한국어 문장에서 생략된 조사는 체언과 용언 사이의 거리가 가까우면 가까울수록 복원율이 높다는 것을 확인하였다. 체언과 용언 사이의 거리가 1과 2로 제한하여 추출한 말뭉치로 실험한 결과 최고 81.11%의 격조사 복원 성능을 확인하였다. 또한 자질집합 중 각 자질들이 격조사 복원의 성능에 어떤 영향이 있는지를 실험한 결과 각 자질들 중 용언이 66.01%, 체언이 75.38%로의 성능으로 가장 낮은 순으로 확인되었

다. 특이한 점은 주격조사와 목적격조사의 경우 체언 앞에 오는 어절의 첫 형태소 자질1과 품사 자질2, 끝 형태소 자질3과 품사 자질4가 자질집합에 포함된 평균 결과 값보다 포함되지 않은 평균 결과 값이 우수한 것으로 나타났다. 이는 체언 앞에 오는 어절에 해당하는 자질들은 자질집합에 포함되지 않는 것이 조사 복원 성능이 더 우수하다는 것을 증명한다. 또한 보격조사와 부사격조사의 경우에도 자질집합에 자질이 포함된 평균 값과 포함되지 않은 평균값의 차이가 거의 없으므로 주격조사와 목적격조사와 같이 체언 앞에 오는 어절을 자질집합에서 제외하는 것이 좋을 것으로 판단한다. 다만 체언 앞에 오는 어절이 자질집합에 포함되지 않을 경우 조사의 복원 성능은 우수할 수 있겠으나 문장에 있어서의 중의성이 해소가 될지는 실험을 통해 연구가 되어야 할 것으로 판단된다.

향후 연구로는 본 논문에서 발견한 자질집합 중 체언 앞에 오는 어절을 제외한 격조사 복원시스템 구현하여 격조사 복원 성능을 향상시키는 방향과 체언 앞에 오는 어절이 생략되었을 경우 문장에서의 중의성 해소에 어떤 영향이 있는지에 대한 연구가 진행된다면 정보검색시스템에서의 구문 분석기의 부하를 크게 감소할 수 있을 것으로 기대한다.

참고문헌

- 김재훈 등, 2005. 구문구조 부착 말뭉치 구축, 모비코앤시스메타(주).
- 김재훈, 김형철, 2010. CRF를 이용한 대응어 결정 시스템, *한국마린엔지니어링학회 공동학술대회 논문집*, pp. 433~434.
- 김학수, 2007. Conditional Random Fields를 이용한 영역 행위 분류 모델, *한국인지과학회*, 제18권, 제1호, pp. 1~14.
- 김형기, 이광국, 김희율, 2010. Conditional Random Fields 구조에서 궤적군 집화를 이용한 혼잡 영상의 이동 객체 검출, *한국멀티미디어학회*, 제13권 제8호, pp. 1128-1141.
- 이창기 등, 2006. Conditional Random Fields를 이용한 세부 분류 개체명 인식, *한국정보과학회 언어공학연구회 학술발표 논문집*, pp. 268-272.
- 서광준, 1993. 어절 사이의 의존관계를 이용한 한국어 구문 분석기, 한국과학기술원 석사학위 논문.
- Chung, H., 2004. *Statistical Korean Dependency Parsing Model based on the Surface Contextual Information*, 고려대학교 박사학위 논문.
- Kim, S-S., Park, S-B., & Lee, S-J., 2007. Analyzing Dependencies of Korean subordinate Clauses using parse tree kernels, In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics*, pp. 218-228.
- Lafferty, J., McCallum, A., & Pereira, F., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, In

Proceedings of the 18th International Conference on Machine Learning, pp. 282-289.

Pinto, D., McCallum, A., Wei, X., & Croft. W. B., 2003. Table extraction using conditional random fields. In *Proceedings of the ACM SIGIR*.



부록 A 원시 말뭉치의 한글 및 영문 태그 사용 예

한글태그	영문태그	축약태그	형태소계	사용 예제
인명고유명사	NNA	N	N+	한스[인명고유명사]+는[일반보조사]
기관고유명사	NNB			통계청[기관고유명사]+에서는[부사격조사]
지명고유명사	NNC			한반도[지명고유명사]+는[일반보조사]
서적고유명사	NND			삼강행실도[서적고유명사]+는[일반보조사]
사건고유명사	NNJ			2차대전[사건고유명사]+을[목적격조사]
기타고유명사	NNE			EU협정[기타고유명사]
외국어	NNK			SI[외국어]+시창[용언불가능보통명사]
용언불가능보통명사	NNF			규모[용언불가능보통명사]+가[주격조사]
용언가능보통명사	NNI			정보화[용언가능보통명사]+에[부사격조사]
단위성의존명사	NDA			4[숫자수사]+명[단위성의존명사]+의[관형격조사]
기타의존명사	NDF			것[기타의존명사]+이[보격조사]
인칭대명사	NPA			당신[인칭대명사]+의[관형격조사]
지시대명사	NPB			그[지시대명사]+들[복수접미사]+을[목적격조사]
주격조사	POA			P
목적격조사	POB	용어[용언불가능보통명사]+를[목적격조사]		
부사격조사	POC	것[기타의존명사]+으로[부사격조사]		
보격조사	POD	해방[용언가능보통명사]+이[보격조사]		
관형격조사	POE	남자[용언불가능보통명사]+의[관형격조사]		
호격조사	POH	정숙[인명고유명사]+이[호격조사]		
접속조사	POJ	때[용언불가능보통명사]+와[접속조사]		
일반보조사	POK	육신[용언불가능보통명사]+은[일반보조사]		
사동보조용언	EVE	Y	Y+	부담지[일반동사]+우[사동보조용언]+게[종속연결어미]
피동보조용언	EVF			채우[일반동사]+어지[피동보조용언]+ㄴ[관형사형전성어미]
기타보조용언	EVK			고민하[일반동사]+게하[기타보조용언]+ㄴ[관형사형전성어미]

한글태그	영문태그	축약태그	형태소경계	사용 예제		
서수사	NUA	U	U+	일곱[서수사]		
양수사	NUB			3[숫자수사]+천[양수사]		
숫자수사	NUC			3[숫자수사]+천[양수사]		
접두사	FPA	F	F+	제[접두사]+2[숫자수사]+차[단위성의존명사]		
인명접미사	FSA			도미[인명고유명사]+네[인명접미사]+뿐만[일반보조사]		
지명접미사	FSB					
보통명사형접미사	FSC					
복수접미사	FSD			사람[용언불가능보통명사]+들[복수접미사]+이[주격조사]		
숫자형접미사	FSE			30[숫자수사]+여[숫자형접미사]		
기타접미사	FSF			40[숫자수사]+년[단위성의존명사]+간[기타접미사]		
지시동사	VBA			B	B+	그러[지시동사]+기[명사형 전성어미]
일반동사	VBB					위하[일반동사]+어서[종속연결어미]
지시형용사	VAA			V	V+	아니[지시형용사]+라[종속연결어미]
성상형용사	VAB	있[성상형용사]+는[관형사형 전성어미]				
성상관형사	ANA	주도적[성상관형사]				
지시관형사	ANB	A	A+	이런[지시관형사]		
수관형사	ANC			여러[수관형사]		
능동전성사	VFA	C	C+			
수동전성사	VFB					
사동전성사	VFC					
긍정지정사	VFD			것[기타의존명사]+이[긍정지정사]+라고[종속연결어미]		
성상정도부사	ADA	D	D+	대단히[성상정도부사]		
성상상태부사	ADB			열심히[성상상태부사]		
성상의성부사	ADC			부르르[성상의태부사]		
성상의태부사	ADD			텅텅[성상의태부사]		
지시처소부사	ADE			쿵쿵[성상의성부사]		
지시시간부사	ADF			이제는[지시시간부사]		
부정부사	ADG			아니[부정부사]		
문장양태부사	ADH			오히려[문장양태부사]		
문장접속부사	ADI			하지만[문장접속부사]		
종속연결어미	EEG			E	E+	고통스럽[성상형용사]+지[종속연결어미]
관형사형전성어미	EEI	갈[성상형용사]+는[관형사형 전성어미]				

한글태그	영문태그	축약태그	형태소경계	사용 예제
부사형전성어미	EEJ	S	S+	찾[성상형용사]+게[부사형전성어미]
명사형전성어미	EEK			매[일반동사]+기도[명사형전성어미]
의문형종결어미	EEC			되[일반동사]+는가[의문형종결어미]
명령형종결어미	EED			생각하[일반동사]+오[명령형종결어미]
청유형종결어미	EEE			하[일반동사]+자[청유형종결어미]+.[문미기호]
대등연결어미	EEF			언[일반동사]+고[대등연결어미]
높임선어말어미	ERA			주[일반동사]+시[높임선어말어미]+아[종속연격어미]
공손선어말어미	ERB			받[일반동사]+자오[공손선어말어미]+아[종속연격어미]
현재시제선어말어미	ERC			
과거시제선어말어미	ERD			하[일반동사]+었[과거시제선어말어미]+다[평서형종결어미]
미래시제선어말어미	ERE			살[일반동사]+겠[미래시제선어말어미]+다면[종속연격어미]
사동선어말어미	ERH			
피동선어말어미	ERI			
평서형종결어미	EEA			가[일반동사]+니다[평서형종결어미]
감탄형종결어미	EEB			없[성상형용사]+구려[감탄형종결어미]+![문미기호]
문미기호	SYA			없[성상형용사]+구려[감탄형종결어미]+![문미기호]
원열림기호	SYB			'[원열림기호]+세계관적[성상관형사]+'[오른열림기호]
오른열림기호	SYC			'[원열림기호]+세계관적[성상관형사]+'[오른열림기호]
컴마기호	SYD			권력[용언불가능보통명사]+,[컴마기호]
기타기호	SYE			한[지명고유명사]+.[기타기호]+미[지명고유명사]
단위기호	SYH	15[숫자수사]+m[단위기호]+로[부사격조사]		
빈칸	SYI			
어절구분자	SYJ	한[지명고유명사]+-[어절구분자]+미[지명고유명사]		
감탄사	INA	I	I+	얼씨구나[감탄사]