Coreference Resolution을 위한 3인칭 대명사의 선행사 결정 규칙

강 승 식† · 윤 보 현†† · 우 종 우†

요 약

정보 검색 시스템에서 문서의 내용을 대표하는 용어를 추출하거나 정보 추출 및 텍스트 마이닝에서 특정 정보만을 추출하려면 고유명사에 대한 대용어 문제가 해결되어야 한다. 대용어 해소 문제는 인칭 명사에 대한 대명사의 선행사 결정 문제가 대표적이다. 본 논문에서는 한국어에서 문서의 내용을 보다 정확히 분석하기 위해 3인칭 대명사 "그/그녀/그들/그녀들"의 선행사를 결정하는 방법을 제안한다. 일반적으로 3인칭 대명사의 선행사는 현재 문장 또는 이전 문장의 주어인 경우가 많고, 또한 3인칭 대명사가 2회 이상 반복되는 경우가 자주 발생한다. 이러한 특성을 이용하여 현재 문장과 이전 문장에 출현한 인칭 명사들 중에서 선행사로 사용되는 경우를 조사하여 선행사 결정 규칙을 발견하였다. 이 경험 규칙은 3인칭 대명사의 격에 따라 조금씩 달라지기 때문에 대명사의 격에 따라 주격, 목적격, 소유격으로 구분하여 기술하였다. 제안한 방법의 타당성을 검증하기 위하여 신문 기사의 정치 관련 문서에서 대명사의 격에 따라 100개씩 총 300개의 실험 대상을 선정하였으며, 실험 결과로 3인칭 대명사의 선행사 결정 정확도는 재현율이 79.0%, 정확률이 86.8%로 나타났다.

Antecedent Decision Rules of Personal Pronouns for Coreference Resolution

Seung-Shik Kang† · Bohyun Yun†† · Chong-Woo Woo†

ABSTRACT

When we extract a representative term from text for information retrieval system or a special information for information retrieval and text mining system, we often need to solve the anaphora resolution problem. The antecedent decision problem of a pronoun is one of the major issues for anaphora resolution. In this paper, we are suggesting a method of deciding an antecedent of the third personal pronouns, such as "he/she/they" to analyze the contents of documents precisely. Generally, the antecedent of the third personal pronouns seem to be the subject of the current statement or previous statement, and also it occasionally happens more than twice. Based on these characteristics, we have found rules for deciding an antecedent, by investigating a case of being an antecedent from the personal pronouns, which appears in the current statement and the previous statements. Since the heuristic rule differs on the case of the third personal pronouns, we described it as subjective case, objective case, and possessive case based on the case of the pronouns. We collected 300 sentences that include a pronoun from the newspaper articles on political issues. The result of our experiment shows that the recall and precision ratio on deciding the antecedent of the third personal pronouns are 79.0% and 86.8%, respectively.

키워드 : 정보 추출(Information Extraction), 대용어 해소(Coreference Resolution), 선행사 결정 규칙(Antecedent Decision Rules)

1. 서 론

정보 검색 시스템이나 문서 분류, 문서 요약, 클러스터링 시스템은 문서 분석 기법에 의해 문서에 출현한 용어들을 추출한다. 이때 각 용어가 문서 내용을 대표하는 정도를 계산하는 방법으로 단순 빈도와 상대 빈도가 사용되고 있다. 그런데 고유 명사를 반복할 때는 용어 자체를 반복하는 대신에 대명사를 사용하기 때문에 특정인이 주제어인 문서에서 인명이 한 번만 출현하는 경우가 발생하게 된다. 예를 들어, 김대중 대통령에 관한 문서에서 '김대중'이라는 인명이 한 번 사용되고 대명사 '그'가 여러 번 반복되었을 때 단순 빈도에 의해 용어의 가중치를 계산하면 '김대중'은 출현 빈도가 낮기 때문에 용어 가중치가 낮아지는 문제가 발생한다.

텍스트 마이닝과 정보 추출 시스템에서도 인명을 추출할 때 대명사의 선행사를 결정해야 하며, 가중치를 부여하기 위한 빈도 계산이나 정보 추출 과정에서 발생하는 대용어

※ 본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.
† 정 회 원 : 국민대학교 컴퓨터학부 교수
†† 정 회 원 : 목원대학교 컴퓨터교육학과 교수
논문접수 : 2003년 8월 4일, 심사완료 : 2004년 3월 30일

해소 문제는 대명사를 비롯하여 약어 등 고유명사를 지칭하는 표현들이 있다. 즉, 회사명을 '그 회사'로 지칭하거나 제품 이름을 '그 제품'으로 지칭하는 등 대용어 문제는 모든 명칭(named entity)에 대해 발생하는 문제이다. 이처럼 문서의 내용을 보다 정확히 파악하는데 있어서 대용어의 선행사를 어떻게 결정하는지가 매우 중요한 문제이다. 이러한 대용어 해소(coreference resolution)에 대한 연구가 외국에서는 MUC, TREC, IREX 등 정보 검색 및 정보 추출 학술회의를 중심으로 활발히 진행 중이다.

대용어 참조 해소는 "빌 게이츠"가 "William H. Gates," "Mr. Gates," "William Gates," "Bill Gates," "Mr. Bill H. Gates"와 같이 다양한 형태로 표현되었을 때 동일 개념에 대한 용어들의 관계를 인식하는 문제이다[1-4]. 예를 들어, "Bill Gates is the richest man in the world… Many people in the software industry fear and respect the guy."에서 "the richest man in the world"와 "the guy"는 모두 "Bill Gates"를 가리킨다. 그런데 "the guy"는 앞에 출현한 대용어에 의존적인 반면에 "the richest man in the world"는 이와 무관하다는 점에 차이가 있다.

대용어 참조 해소는 MUC-6에서 대용어 참조 태스크가 추가되었으나 그 전에도 정보 추출 시스템이 정보를 추출하기 위해 그 기능이 구현되어 왔다. 대용어 참조 해소의 대상이 되는 용어로는 고유 명사(names), 별칭(aliases), 동격명사(appositives), 특정 명사를 가리키는 명사구(definite noun phrases), 대명사(pronouns), 서술 명사(predicate nominals) 등이 있다. 예를 들어, '마이크로소프트'와 관련된 대용어 해소 대상이 되는 용어들은 'Microsoft Corporation,' 'MS,' '세계 최대의 소프트웨어 업체', '그 회사' 등이다.

대용어 해소에 관한 연구는 다음과 같이 3가지로 요약된다. 첫째, 전통적인 언어 지식(linguistic knowledge)과 도메인 지식(domain knowledge)을 이용하는 방법이다. 이 접근 방법은 형태소 분석과 구문 분석, 그리고 의미 분석 및 담화 분석을 통해 언어의 품사 정보, 구문 정보, 의미 정보 등을 파악하고 이를 기반으로 용어간의 상호 조응 관계를 결정하는 방식이다[5, 6]. 그런데 언어 분석 기법에 의존하는 방법은 많은 사람의 노력이 필요하고, 처리 속도가 느리며 도메인 독립적이지 못한 단점을 갖는다. 둘째, 언어 지식과 도메인 지식을 이용하는 방법의 단점을 극복하기 위한 knowledge-poor approach 방법으로서 형태소 분석과 품사 태깅 결과를 이용하거나, 부분 구문 분석(partial parsing)을 기반으로 하는 연구이다[7-9]. 셋째, 말뭉치에 기반한 통계적 기법으로서 대용어 관계가 설정되어 있는 말뭉치로부터 대용어 관계를 파악하는 패턴 정보를 학습하는 방법이다. 말뭉치로부터 패턴 정보를 학습하기 위하여 공기관계(cooccurrence relation)와 선택 제약 패턴을 이용하여 선행사 후보를 치환했을 때 충분히 높은 빈도의 공기 패턴들을 자동으로 추출하여 대용어 해소 규칙으로 사용한다[10-12].

대용어 해소는 다양한 유형의 고유명사 및 고유명사를 지칭하는 표현들에 대해 대용어를 결정하는 문제이지만 모든 대용어 표현을 포괄하는 규칙을 발견하기는 쉽지 않다. 따라서 본 논문에서는 대용어 중에서 빈도가 높은 유형인 인칭 명사로 그 범위를 제한하여 3인칭 대명사에 대한 대용어를 결정하는 방법을 제안한다.

2. 관련 연구

대용어 참조 해소 기법으로는 수동으로 대용어 해소 규칙을 작성하는 방법과 대용어 태깅된 말뭉치로부터 자동으로 규칙을 학습하는 통계적인 방법이 있다. 수동으로 규칙을 작성하는 방법은 referent가 앞 부분에 출현한다는 전제 조건에 의해 앞 부분에 출현하는 대용어 후보들을 수집하여 성(gender), 수(number) 등의 제약 조건을 만족하는 후보들 중에서 경험적으로 가능성이 높은 후보를 선택하는 규칙을 적용하여 referent를 선택한다. 수동으로 작성된 경험 규칙에 의한 방법은 다양한 유형의 대용어에 대해 일반적인 규칙을 발견하기가 쉽지 않으나 전문가에 의해 정교한 규칙을 작성하기 때문에 3인칭 대명사 혹은 회사명 등 특정 유형에 대한 대용어 해소에 적합하다.

이에 비해, 통계적 기법은 학습 말뭉치로부터 대용어 관계를 결정하는 규칙을 학습한다. 그 예로서, Aone(1995)의 MLR(machine-learning-based resolver)과 McCarthy(1995)의 RESOLVE는 통계적 학습 알고리즘에 의해 학습 말뭉치에 표시된 대용어 관계 정보로부터 임의의 두 용어들에 대한 대용어 관계 여부를 좌우 문맥 정보와 함께 추출하여 결정 트리(decision tree)로 구현하는 결정 트리 추론 시스템이다[11, 12]. 학습 말뭉치의 대용어 참조 정보는 학습 알고리즘에 따라 다르지만, 시스템이 대용어 관계를 구별할 수 있는 <속성, 값>쌍 형태의 특성 정보와 대용어가 출현한 문맥으로 구성된다. MLR의 경우에 학습 예제들은 66개의 특성 정보들로 기술되어 있다.

두 시스템을 각각 50개 문서 집합과 250개 문서 집합에 대한 실험 결과에 의하면, RESOLVE는 재현율과 정확율이 각각 80%~85%, 87%~92%로 나타났으며 MLR은 67%~70%, 83%~88%였다. 그런데 이 시스템들을 MUC-6 대용어 태스크의 25개 문서 집합에 적용했을 때 RESOLVE는 재현율이 41%~44%, 정확률이 51%~59%이다. 이 결과는 대용어 해결 알고리즘을 수동으로 작성한 상위 5개 시스템의 재현율이 51%~63%, 정확률이 62%~72%인데 비해 매우 낮다. 즉, 대용어 해소 알고리즘은 통계적인 방법보다는

수작업으로 규칙을 정교하게 작성했을 때 재현율과 정확률이 높아진다.

Hobbs(1976)의 단순한 알고리즘에 의해 대명사의 선행사를 결정하는 방법을 제안하였다[13]. Hobbs의 알고리즘은 대명사가 발견된 앞 부분의 명사구들에 대해 순서대로 성(gender), 수(number)와 같은 선택 제약 조건을 적용하여 최초로 제약 조건을 만족하는 명사구를 선행사로 결정하였다. Baldwin(1997)은 Hobbs의 알고리즘을 발전시켜 영어 대명사의 선행사를 결정하는 시스템으로 CogNIAC이라는 대명사 참조 해소 엔진을 개발하였다[14]. CogNIAC은 3인칭 대명사의 선행사를 결정하는데 6개의 경험 규칙을 이용하고 있는데, CogNIAC의 경험 규칙에 의한 영어 대명사의 선행사 인식 방법은 매우 높은 정확률을 보이고 있다. Baldwin이 제시한 6개의 경험 규칙은 다음과 같다.

[CogNIAC의 경험 규칙]

① 현재 분석중인 단어의 앞 부분에서 선행사라고 판단되는 단어가 유일하면 이를 선택한다.
② 현재 분석중인 단어가 재귀 대명사이면 현재 문장에서 가장 가까운 선행사 후보를 선택한다.
③ 현재의 문장과 바로 이전 문장에서 선행사라고 판단된 단어가 유일할 경우 이를 선택한다.
④ 현재 분석중인 단어가 소유 대명사이면 문자열이 정확히 일치(exact string match)되는 단어를 선택한다.
⑤ 현재 문장에서 선행사라고 판단된 단어가 유일하면 이를 선택한다.
⑥ 이전 문장에서 선행사라고 찾은 단어와 현재 분석중인 단어가 격이 같은 것을 선택한다.

이외에도 대용어 해소를 위한 구문적 제약 범위에 관한 연구와 일본어에서 지시대명사의 대용어 연구, 포르투갈어의 소유 대명사의 대용어 해소에 관한 연구가 있다[15-17]. 한국어에서는 인칭 명사를 인식하는 연구로서 '대통령', '대표이사' 등 명사 뒤에 오는 직함과 관련된 표현을 이용하여 고유명사를 인식하는 방법이 있다[18, 19].

3. 3인칭 대명사의 선행사 결정 규칙

한국어에서 인칭 명사를 지시하는 3인칭 대명사는 '그'이며 '그'의 품사로는 인칭 대명사, 지시 대명사, 그리고 관형사이다. 입력 문서로부터 '그'가 포함된 어절을 조사하여 그 중에서 주격, 목적격, 소유격으로 사용된 인칭 대명사 용례들을 수집하였다. 3인칭 대명사 '그'에 대한 선행사를 결정하는 규칙을 발견하기 위하여 '그/그녀/그들'에 대해서 세 가지 유형으로 구분하여 처리하였다. 첫째, 3인칭 대명사에 주격 조사 또는 보조사 '은/는'이 결합된 어절(주어)과, 둘째는 소유격 조사 '의'가 결합된 어절, 마지막으로 목적격 조사 '를'이 결합된 어절이다. 세 가지 유형에 대해 각각 3인칭 대명사의 선행사를 결정하는 경험 규칙을 발견하였다. 3인칭 대명사의 선행사를 결정하는데 보편적으로 적용되는 선행사 후보 제약은 다음과 같으며, 선행사 후보는 세 가지 요건을 모두 만족해야 한다.

[선행사 요건 1] 주격, 소유격, 목적격 어절
[선행사 요건 2] 인칭명사, 3인칭 대명사
[선행사 요건 3] 대명사와 수(number)가 일치하는 명사

3.1 3인칭 대명사의 주격 및 보조사 '은/는'

3인칭 대명사 주격(보조사 '은/는' 포함)의 선행사는 현재 문장 또는 이전 문장의 주어인 경우가 많다. 따라서 현재 문장과 이전 문장들에서 선행사 요건을 만족하는 주어가 발견되면 이를 선행사로 선택한다. 이 때, 선행사 후보가 2개 이상 발견되면 3인칭 대명사에 가까운 선행사를 선택한다[20]. 3인칭 대명사가 주격 또는 보조사 '은/는'과 함께 쓰인 경우에 대한 탐색 방식과 경험 규칙은 다음과 같다.

[탐색방식] 주격 또는 보조사 '는'이 포함된 3인칭 대명사를 찾는다. 해당 3인칭 대명사로부터 좌측으로 주격(보조사 포함) 선행사 후보를 검색한다.

[규칙 1] 현재 문장 또는 이전 문장에서 주격 또는 보조사 '은/는'에 해당하는 선행사가 발견된 경우 선행사 요건을 만족하면 선행사로 선택한다.

[규칙 2] 현재 문장 또는 이전 문장에서 주격 후보가 2개 이상 발견되는 경우는 3인칭 대명사에 가장 가까운 주격을 선택한다.

예) ***노태우씨는*** *한마디로 비교적 안정된 기조 위에 점진적 개혁을 추구하는 합리적 정치인으로 평가될 수 있을 것이다. 그러나* ***그는*** *자신의 건곤일에도 불구하고 해방 이후 우리나라 정치사에 중대한 변수로 치부되어 왔던 군 출신임에 틀림없다.*

위의 예제에서 '그는'의 선행사는 '노태우씨는'이다. [규칙 1]을 적용하여 이전 문장의 주어인 '노태우씨는'을 찾고, 어근 '노태우씨'가 선행사 요건에 만족하기 때문에 선행사로 선택된 것이다. 그리고 '노태우씨'는 접미사 '씨'를 단서로 하여 인칭 명사인지를 판단한다.

3.2 3인칭 대명사의 소유격

3인칭 대명사 소유격의 선행사는 주격 및 소유격 후보를 검색하여 선행사 요건을 만족하는지 확인한다. 현재 문장에서 주격과 소유격이 동시에 존재할 때는 3인칭 대명사에 가까운 것을 선택한다. 그러나 현재 문장이 아닌 이전 문장

에서 동시에 나타난 경우는 주격을 우선으로 하고, 주격이 없는 경우는 소유격을 선행사로 선택한다. 3인칭 대명사 소유격의 탐색 방식과 경험 규칙은 다음과 같다.

[탐색방식] 소유격 3인칭 대명사를 찾는다. 해당 3인칭 대명사로부터 좌측으로 주격(보조사 포함), 소유격 선행사 후보를 검색한다.

[규칙 3] 문장 내에서 주격이 발견되고, 선행사의 요건을 만족하면 선행사로 선택한다.

[규칙 4] 문장 내에서 선행사 요건을 만족하는 소유격을 찾은 경우 주격 후보가 있는지를 확인한다. 주격 후보가 없으면 소유격을 선행사로 선택하고, 주격 후보가 있으면 주격을 우선으로 선택한다.

[규칙 5] 문장 내에서 발견된 선행사 후보들이 소유격만 존재할 경우 경우는 3인칭 대명사에 가장 가까운 것을 선택한다.

예) ***김당선자는*** *그런 뜻에서 작은 의리와 구연을 버려야 한다.* ***그의*** *당선을 위해 불철주야로 일한 수많은 동지가 있다.* ***그의*** *오늘이 있기까지 모든 불리한 여건에도 불구하고 그의 곁을 지켜온 많은 측근이 있다.*

마지막 문장 "**그의** 오늘이 …"에서 '그의'의 선행사는 현재 문장에 선행사 후보가 없으므로 이전 문장에서 후보를 찾는다. 이 때, 앞 문장에는 주격 후보 '동지가'가 발견되지만 명칭(named entity)이 아니므로 "***그의*** 당선을 …"의 '그의'가 선행사로 선택된다. 두 번째 문장에서 '그의'에 대한 선행사는 첫 번째 문장의 '김당선자'가 선행사로 선택된다. 또한, 주격이 두 번 이상 발견되는 경우는 [규칙 2]로 선행사를 결정한다.

3.3 3인칭 대명사의 목적격

목적격의 선행사는 현재 또는 이전 문장에서 주격, 소유격뿐만 아니라 목적격도 선행사 후보가 된다. 주격, 소유격과 마찬가지로 한 문장내 주격, 소유격, 목적격이 2개 이상 발견된 경우는 3인칭 대명사에 가장 가까운 것을 우선으로 선택한다. 3인칭 대명사 목적격에 대한 탐색 방식과 경험 규칙은 다음과 같다.

[탐색방식] 목적격 3인칭 대명사를 찾고, 해당 3인칭 대명사로부터 좌측으로 검색하여 주격(보조사 포함), 소유격, 목적격 선행사 후보를 검색한다.

[규칙 6] 문장 내에 주격 선행사 후보가 존재하면 선택하고, 존재하지 않으면 3인칭 대명사로부터 가장 가까운 선행사 요건을 만족하는 후보를 선택한다.

[규칙 7] 인칭 명사 뒤에 바로 3인칭 대명사가 출현하고 이 인칭 명사가 선행사 요건을 만족하면 선행사로 선택한다.

예) *대통령제하의 극한 대립양상 운운은 그럴듯한 명분일 뿐이다. 그들은* ***김영삼씨가*** *아무리 민자당 대표위원이라고 해도* ***그를*** *여권의 사람으로 보지 않고 있으며 여권의 맥을 이을 사람으로는 더더욱 보지 않고 있음이 점차 확실해지고 있다.*

위 예에서 '그를'과 같은 문장 내에서 주격 조사 및 보조사 '은/는'이 있는 어절은 '그들은'과 '김영삼씨가'이다. 그러나 '그들은'은 '그를'과 수가 다르기 때문에 선행사 요건을 만족하지 못하므로 '김영삼씨가'를 선행사로 선택한다.

4. 실험 및 평가

3인칭 대명사의 대용어 해소 실험을 위해 신문 기사에서 정치 관련 기사들을 수집하였고, 각 문서에서 3인칭 대명사의 문장 표현 형태가 주격, 목적격, 소유격에 해당하는 문장을 각각 100개씩을 실험 대상으로 하였다. 입력 문서에 대한 형태소 분석 결과로부터 실험 대상이 되는 대명사를 인식하였다. 3인칭 대명사의 선행사 후보는 그 대명사의 앞부분에 출현한 인칭명사 혹은 대명사로 제한된다. 그런데 현재 형태소 분석기의 분석 결과는 어떤 명사가 인칭 명사인지, 아닌지를 알 수 없으므로 인칭 명사를 수동으로 표시해 주는 방법으로 실험하였다. 이전 문장의 검색 범위를 5개 문장으로 제한하여 실험한 결과는 <표 1>과 같다.

〈표 1〉 3인칭 대명사의 대용어 실험 결과

	재 현 율	정 확 률
3인칭 대명사의 주격	78.0%	88.6%
3인칭 대명사의 목적격	75.0%	81.5%
3인칭 대명사의 소유격	84.0%	90.3%
평 균	79.0%	86.8%

Balwin(1997)은 298개의 대명사에 대한 실험 결과에서 77.9%의 정확도를 보이고 있으며, MUC-6의 30문서에 대한 실험에서는 재현율 75%, 정확률 73%를 보이고 있다. 영어와 한국어라는 언어의 특성 차이가 있을 수 있지만, 단순히 재현율과 정확률을 비교했을 때 본 연구의 실험 결과는 재현율 79.0%, 정확률 86.8%로서 CogNIAC에 비해 좀더 나은 성능을 보이고 있다.

실험에서 발생한 오류의 내용을 살펴보면 다음과 같다. 주격 실험의 경우 22개의 오류중 10개는 정답을 제시하기는 했지만 잘못된 결과를 제시한 경우이고, 12개는 이전 5개 문장에서 규칙을 만족하는 답이 없어서 선행사를 발견

하지 못한 경우이다. 소유격 실험에서는 16개의 오류 중 9개가 잘못된 결과를 제시한 경우이고, 7개는 규칙을 만족하는 결과가 없는 경우이다. 목적격은 위 두 가지 경우보다 다양한 오류를 보인다. 전체 오류 25개중 잘못된 결과 8개, 현재 또는 이전 문장에서 선행사를 찾지 못한 경우 9개, '그를'이 '그것을'의 준말로 쓰인 경우가 8개였다. 목적격 '그들'과 '그것을'의 준말 '그들'을 구분할 수 있다면 좀 더 향상된 결과를 얻을 수 있을 것이다.

선행사의 탐색 범위를 이전 문장 10개 또는 문서 처음까지로 확장한다면 재현율이 높아질 수 있으나, 이 경우에 선행사 후보 대상이 많아지므로 오류율 또한 증가할 것으로 예상된다. 또한, 본 실험에서는 편의상 빈도가 높은 주격, 소유격, 목적격 조사를 대상으로 하였는데, 실용적으로 활용하려면 부사격 '-께서/-에게'와 같은 조사를 추가하여 성능을 개선할 필요가 있다.

5. 결 론

대용어 문제는 인명 뿐만 아니라 명칭과 관련된 용어로 인명, 지명, 기관명, 제품명 등에서 발생한다. 본 논문에서는 인명에 대한 지시 관계 중에서 대명사의 선행사를 결정하는 문제를 해결하는 방법을 제안하였다. 3인칭 대명사의 선행사를 결정하는 방법으로 대명사의 격에 따라 주격, 소유격, 목적격에 대한 7개의 경험 규칙을 제안하였다. 이 규칙들을 이용한 대용어 해소 실험의 결과로 재현율은 79%이고, 정확률은 86.8%로 나타났다. 실험 결과에서 선행사가 결정되지 않은 경우가 격 유형에 따라 7%~12%였는데 그 원인을 분석한 결과에 의하면 선행사 후보의 탐색 범위인 이전 문장 5개에 대용어가 없는 경우가 다수 발견되었다.

본 연구에서는 실험 문서에서 발견된 인칭 명사들을 수동으로 표시하여 실험을 하였는데, 정보 추출 시스템 등 실용적인 시스템에 적용하기 위해서는 문서 내에 출현한 명사의 유형이 인칭 명사인지를 자동으로 인식하는 기능이 선행되어야 한다. 또한, 실용적인 시스템에 적용되는 대용어 해소는 인칭 명사의 대명사를 비롯하여 약어와 동의어 등 고유명사를 지칭하는 표현들로 확대되어야 한다. 일반적인 명칭에 대한 대용어 표현으로 대용어 해소 기능을 확장하려면 인명, 지명, 기관명, 제품명 등 고유명사의 유형을 인식하는 연구와 대용어 표현 기법에 관한 연구가 필수적이다. 본 연구에서는 대용어 해소 문제를 해결하기 위한 첫 단계로서 인칭 명사에 대한 대명사의 선행사 결정 방법을 제안하였으며, 향후 연구로서 명칭 인식 문제와 대용어 표현 기법에 관한 연구를 수행할 예정이다.

참 고 문 헌

[1] C. Cardie, "Corpus-Based Acquisition of Relative Pronoun Disambiguation Heuristics," Proceedings of the 30th Annual Meeting of the ACL, Association for Computational Linguistics, pp.216-233, 1992.

[2] C. Cardie, "Learning to Disambiguate Relative Pronouns," Proceedings of the Tenth National Conference on Artificial Intelligence, American Association for Artificial Intelligence, pp.38-43, 1992.

[3] A. Kehler, "Probabilistic Coreference in Information Extraction," Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp.163-173, 1997.

[4] R. Weischedel, M. Meteer, R. Schwartz, L. Ramshaw, and J. Palmucci, "Coping with Ambiguity and Unknown Words through Probabilistic Models," Computational Linguistics Vol.19, No.2, pp.359-382, 1993.

[5] J. Carbonell and R. Brown "Anaphora resolution : a Multi-strategy Approach," Proceedings of the 12th International Conference on Computational Linguistics COLING '88, pp.96-101, 1988.

[6] D. M. Carter, "Interpreting Anaphora in Natural Language Texts," Chichester : Ellis Horwood, 1987

[7] R. Mitkov, "Robust Pronoun Resolution with Limited Knowledge," COLING '98, pp.869-875, 1998.

[8] R. Mitkov, "An Integrated Model for Anaphora Resolution," Proceedings of the 15th International Conference on Computational Linguistics COLING '94, pp.1170-1176, 1994.

[9] C. Kennedy and B. Boguraev, "Anaphora for Everyone : Pronominal Anaphora Resolution without a Parser," Proceedings of the 16th International Conference on Computational Linguistics COLING '96, pp.113-118, 1996.

[10] I. Dagan, and A. Itai, "Automatic Processing of Large Corpora for the Resolution of Anaphora References," Proceedings of the 13th International Conference on Computational Linguistics, COLING '90, Vol.III, pp.1-3, 1990.

[11] C. Aone and W. Bennett, "Evaluation Automated and Manual Acquisition of Anaphora Resolution Strategies," Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp.122-129, 1995.

[12] J. F. McCarthy and W. G. Lehnert, "Using Decision trees

for Coreference Resolution", Proceedings of the Fourteenth International Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence, pp.1050-1055, 1995.

[13] J. Hobbs, "Pronoun Resolution", Research Report #76-1, City College, City University of New York, 1976.

[14] B. Baldwin, "CogNIAC : High Precision Co-Reference with Limited Knowledge and Linguistic Resources," ACL '97/EACL '97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution, pp.38-45, 1997.

[15] R. Stuckardt, "Anaphora Resolution and the Scope of Syntactic Constraints," COLING-96, pp.937-943, 1996.

[16] H. Nakaiwa and S. Shirai, "Anaphora Resolution of Jpanese Zero Pronouns with Deictic Reference," COLING-96, pp.812-817, 1996.

[17] I. Paraboni and V. L. S. Lima, "Possessive Pronominal Anaphor Resolution in Portuguese Written Texts", COLING-98, pp.1010-1014, 1998.

[18] 정래정, 김준태, "고유명사 출현 패턴을 이용한 색인의 성능 향상에 관한 연구", 제8회 한글 및 한국어 정보처리 학술발표논문집, pp.68-72, 1996.

[19] 황이규, 윤보현, "HMM에 기반한 한국어 개체명 인식", 정보처리학회논문지B, 제10-B권 제2호, pp.229-237, 2003.

[20] M. Sanda Harabagiu and Steven J. Maiorano, "Knowledge-Lean Coreference Resolution and its Relation to Textual Cohesion and Coherence," Proceedings of the ACL-99 Workshop on the Relation of Discourse/Dialogue Structure and Reference, pp.29-38, 1999.

강 승 식

e-mail : sskang@kookmin.ac.kr

1986년 서울대학교 컴퓨터공학과(학사)

1988년 서울대학교 컴퓨터공학과(석사)

1993년 서울대학교 컴퓨터공학과(박사)

1994년~2001년 한성대학교 정보전산학부 부교수

2001년~현재 국민대학교 컴퓨터학부 부교수

관심분야 : 한국어 정보처리, 정보검색, 텍스트마이닝 등

윤 보 현

e-mail : ybh@mokwon.ac.kr

1992년 목포대학교 전산통계학과(학사)

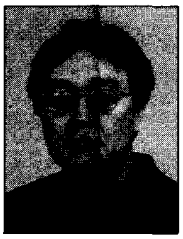
1995년 고려대학교 컴퓨터학과(석사)

1999년 고려대학교 컴퓨터공학과(박사)

2001년~2002년 한국전자통신연구원 선임연구원 언어이해연구팀 팀장

2003년~현재 목원대학교 컴퓨터교육학과 조교수

관심분야 : 정보검색, 자연언어 처리, XML/SGML, 지식정보처리 등



우 종 우

e-mail : cwwoo@kookmin.ac.kr

1978년 서울대학교 농생물학과(학사)

1983년 미국 Minnesota State University at Mankato 전산학과(석사)

1991년 미국 Illinois Institute of Technology 전산학과(박사)

1992년~1993 국방정보체계연구소 선임연구원

1994년~현재 국민대학교 컴퓨터학부 부교수

관심분야 : ITS, 에이전트, 전문가 시스템 등