# 경험 규칙에 의한 대명사의 Coreference Resolution

안영훈°, 강승식, 우종우, *윤보현
국민대학교 컴퓨터학부
*한국전자통신연구원 언어공학부
{xfilecom, sskang, cwwoo}@cs.kookmin.ac.kr, ybh@etri.re.kr

# Coreference Resolution of Pronouns by Heuristic Rules

Young-Hoon Ahn°, Seung-Shik Kang, Chong-Woo Woo, Bo-Hyun Yun
School of Computer Science, Kookmin University
Linguistic Engineering Department, ETRI

**요 약**

정보추출과 정보검색 시스템에서 문서의 내용을 보다 정확히 분석하기 위해 3인칭 대명사 "그/그녀/그들/그녀들"의 선행사를 결정하는 방법을 제안한다. 일반적으로 3인칭 대명사의 선행사는 현재문장 또는 이전문장의 주어인 경우가 많고, 또한 3인칭 대명사가 2회 이상 반복되는 경우가 자주 발생한다. 이러한 특성을 이용하여 현재 문장과 이전 문장에 출현한 인칭명사들 중에서 선행사로 사용되는 경우를 조사하여 경험적인 방법으로 선행사 결정 규칙을 발견하였다. 이 경험 규칙은 3인칭 대명사의 격에 따라 조금씩 달라지기 때문에 대명사의 격에 따라 "주격/목적격/소유격"으로 구분하여 기술하였다. 실험 결과, 3인칭 대명사의 선행사 결정 정확도는 주격, 소유격, 목적격에 대해 각각 88.6%, 90.3%, 81.5%로 나타났다.

## 1. 서 론

정보검색 시스템이나 문서 분류, 문서 요약, 클러스터링 시스템은 문서 분석 기법에 의해 문서에 출현한 용어들을 추출한다. 이 때 각 용어가 문서 내용을 대표하는 정도를 계산하는 방법으로 단순빈도와 상대빈도가 사용되고 있다. 그런데 인칭명사를 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 '그'가 여러 번 반복되었을 때 단순빈도에 의해 용어의 가중치를 계산한다면 '김대중'은 저빈도어가 되어 가중치가 낮아지는 문제가 있다.

정보추출 시스템에서도 인명을 추출할 때 대명사의 선행사를 결정해야 하며, 가중치를 부여하기 위한 빈도 계산이나 정보추출 과정에서 발생하는 대용어 해소 문제는 대명사로 국한되지 않는다. 즉, 회사명을 '그 회사'로 지칭하거나 제품이름을 '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 문제이고 이러한 coreference resolution에 대한 연구가 외국에서는 MUC, TREC, IREX 등 정보검색 및 정보추출 학술회의를 중심으로 활발히 진행 중이다.

대용어 해소는 다양한 유형의 명칭에 대해 해결되어야 할 문제이지만 모든 경우를 포괄하는 규칙을 발견하기는 쉽지 않으며, 본 논문에서는 3인칭 대명사의 선행사를 결정하는 경험규칙을 제안한다.

## 2. 관련 연구

coreference에 관한 연구는 다음과 같이 3가지로 구분된다[1]. 첫째, 전통적인 언어지식과 도메인 지식을 이용한 방법으로 언어의 품사정보, 구문정보, 의미정보 등을 이용하여 상호 조응관계를 조사하는 방식으로 많은 사람의 노력, 처리속도와 도메인 독립적이지 못한 단점을 갖는다[2,3,4]. 둘째, knowledge-poor approach방법은 최근의 추세이고 첫째 방법과 대등한 결과를 보인다[5]. 셋째, cooccurence와 선택제약 패턴을 이용하여 선행사 후보를 치환했을 때 충분히 높은 빈도의 공기패턴일 경우에 선행사로 인정하는 방법이다[6].

Baldwin(1997)은 CogNIAC에서 영어 대명사의 선행사를 결정하는 방법으로 6개의 경험규칙을 이용하여 영어 대명사의 coreference resolution을 시도하였다. Baldwin이 제시한 6개의 경험 규칙은 다음과 같다[7].

[CogNIAC의 경험규칙]

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 이전문장에서 선행사라고 판단된 단어가 유일할 경우 이를 선택한다.
4. 현재 분석중인 단어가 소유대명사이면 exact string match인 단어를 선택한다.
5. 현재 문장에서 선행사라고 판단된 단어가 유일하면 이를 선택한다.
6. 이전문장에서 선행사라고 찾은 단어와 현재 분석중인 단어가 격이 같은 경우는 선택한다.

CogNIAC의 경험 규칙에 의한 영어 대명사의 선행사 인식 방법은 매우 높은 정확률을 보이고 있다. 그러나 CogNIAC은 정확률에 중점을 두고 있기 때문에 재현율이 높지 않다.

### 3. 3인칭 대명사의 Coreference Resolution

3인칭 대명사의 선행사를 찾기 위해서 '그/그녀/그들'에 대해서 세 가지 유형으로 구분하여 처리하였다. 첫째, 3인칭 대명사에 주격조사 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 경험규칙을 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[8]. 3인칭 대명사가 주격 또는 보조사 '은/는'과 함께 쓰인 경우에 대한 검색방식과 경험규칙은 다음과 같다.

[검색방식] 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 인칭명사인지를 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 문장에서 동시에 나타난 경우는 주격을 우선으로 하고, 주격이 없는 경우는 소유격을 선행사로 선택한다. 3인칭 대명사 소유격의 검색방식과 경험규칙은 다음과 같다.

[검색방식] 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 named entity가 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 검색방식과 경험규칙은 다음과 같다.

[검색방식] 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] 인칭명사 뒤에 바로 3인칭 대명사가 출현하고 이 인칭명사가 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 주격조사 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 coreference resolution 실험을 위해 신문기사에서 정치관련 기사들을 수집하였고, 각 문서에서 3인칭 대명사의 문장 표현 형태가 주격, 목적격, 소유격에 해당하는 문장을 각각 100개씩을 실험 대상으로 하였다. 입력 문서에 대한 형태소 분석결과로부터 실험 대상이 되는 대명사를 인식하였다.
3인칭 대명사의 선행사 후보는 그 대명사의 앞 부분에 출현한 인칭명사 혹은 대명사로 제한된다. 그런데 현재 형태소 분석기의 분석결과는 어떤 명사가 인칭명사인지, 아닌지를 알 수 없으므로 인칭명사를 수동으로 표시해 주는 방법으로 실험하였다. 이전문장의 검색 범위를 5개 문장으로 제한하여 실험한 결과는 표 1과 같다.

표 1. 3인칭 대명사의 실험결과

| | 재현율 | 정확률 |
|---|---|---|
| 3인칭 대명사의 주격 | 78% | 88.6% |
| 3인칭 대명사의 목적격 | 75% | 81.5% |
| 3인칭 대명사의 소유격 | 84% | 90.3% |

실험에서 발생한 오류의 내용을 살펴보면 다음과 같다. 주격 실험의 경우 22개의 오류 중 10개는 정답을 제시하기는 했지만 잘못된 결과를 제시한 경우이고, 12개는 이전 5개 문장에서 규칙을 만족하는 답이 없어서 선행사를 발견하지 못한 경우이다. 이 경우 이전 문장을 5개에서 10개로 확장하면 정확도가 향상될 것으로 기대된다. 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 25개 중 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.

### 5. 결론

본 논문은 3인칭 대명사에 대하여 선행사를 결정하는 방법으로 7개의 경험 규칙을 제안하였으며, 이 경험적인 규칙들을 이용하여 선행사 후보의 탐색범위를 이전 문장 5개로 제한하여 실험한 결과 재현율은 78%~84%이고, 정확률은 81%~90%였다. 실험결과에서 선행사가 결정되지 않은 경우가 7%~12%로 이는 탐색 범위를 이전 5 문장으로 제한했기 때문일 것으로 예상된다. 따라서 탐색범위를 이전 문장 10개 또는 문서 처음까지로 확장한다면 재현율이 높아질 것이다. 또한, '-께서/-에게'와 같은 조사를 선행사 요건에 추가하여 경험 규칙을 개선하면 성능이 향상될 것으로 기대된다. 본 논문에서 제안한 경험규칙을 적용하려면 문서내에 출현한 명사가 인칭명사인지를 판단해야 하므로 named entity의 유형을 결정하는 방법이 선행되어야 한다.

### 6. 참고 문헌


[1] 강병주 외 4인, "지시대명사 '이것'의 선행사 추정", http://kibs.kaist.ac.kr/~bjkang/anaphora/report.html.

[2] 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.

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

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

[5] 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.

[6] 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, 1-3, 1990.

[7] Breck 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.

[8] Sanda M. 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.

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

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

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

[12] Ruslan Mitkov, "Robust Pronoun Resolution with Limite Knowledge", COLING-98, pp.869-875, 1998.

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