한국어 백과사전에 등장하는 영대명사(Zero Pronoun)의 복원에 관한 전산학적 연구

신효식[0] 강영수[0] 최기선[0] 송만석*
한국과학기술원 전산학과 전문용어언어공학연구센터
{gerling, yskang, kschoi}@world.kaist.ac.kr
mssong@december.yonsei.ac.kr

Computational Approach to Zero Pronoun Resolution in Korean Encyclopedia

Hyo-Shik Shin[0] Young-Soo Kang[0] Key-Sun Choi[0] Mansuk Song*
KORTERM KAIST[0]
Dept. of Computer Science, Yonsei University[00]

요 약

이 논문은 한국어 백과사전에 등장하는 질병에 대한 요약문 생성의 일환으로 내용을 비교하고 중복성을 제거하기 위해 논리표현으로의 변환과정에서 중요한 영대명사의 복원을 다룬다.백과사전의 요약적인 기술 특성상 자주 등장하는 영대명사의 복원을 위해 통사의미적 혹은 담화적 언어지식에 의존하기보다는 질병에 관한 개념지도를 토대로 복원할 수 있다는 지식기반 방식을 제안한다.

1. 서론

영대명사(Zero Pronoun)란 대명사가 등장할 자리가 형태론적으로 비어있는 경우를 말한다. 영대명사 현상은 풍부한 형태체계를 가진 로만어 계통 언어에서는 물론 한국어 및 일본어에서도 자주 발견되는 현상이다. 전자의 언어군에서는 영대명사의 빈자리를 복원하는데 단서가 될만한 형태적 표지가 있지만, 후자의 언어군에서는 오로지 담화상의 정보구조의 흐름 속에서 파악된다는 점에서 차이가 있다. 이러한 차이는 영대명사의 문법화와 비문법화의 대비로 요약된다. 전자의 언어군에서는 주로 주어 위치에만 영대명사를 허용하는 반면에, 후자의 언어군에서는 주어 자리는 물론 목적어 및 여타의 문장성분에 대해서도 영대명사를 허용한다. 즉, 후자의 언어군에서는 형태론적 혹은 통사론적 제약이 없이 영대명사 현상이 등장한다는 점이 특징이다.

영대명사의 복원이란 대명사의 선행사를 찾는 인덱싱의 문제를 일컫는다. 본 논문에서는 대상영역을 한정하여, 백과사전의 인간관련 병명의 기술에 등장하는 영대명사를 설명하는데 초점을 둔다. 따라서, 해결책 또한 지식기반 설명으로 언어일반적이기 보다는 영역한정적인 설명을 시도한다. 그러나, 영역한정적이라 하더라도 지식기반의 영대명사의 복원이 여타의 영역, 언어에로 확장될 수 있다는 점에서 영역 독립적, 언어독립적이라는 일반성을 획득한다.

2. 관련연구

영대명사의 연구는 인덱싱의 문제인 대용사(anaphora)의 연구와 관련하여 기계번역 및 질의응답, 문서요약 등의 분야에서 많은 관심을 끌어 왔다. 통사론 혹은 구문해석에 입각한 분석[Hobbs 78]으로부터 시작하여, 담화의 절편으로서 문장을 대상으로 담화정보상 가장 두드러진 문장성분을 "중심(center)"이라 칭하여 대용사의 선행사(인덱싱의 결정요소)로 삼는 중심화 이론이 있었다.[Grosz et al. 95] 한편, 구문해석의 언어학적 정보를 이용한 지식기반(knowledge-based) 접근 방식에 반대하여, 언어학적 정보를 이용하지 않는 지식배제(knowledge-poor/independent) 접근방식도 대안으로서 제시되었다.[Nasukawa 94][Dagan etc. 90] 전자의 방식은 높은 정확률을 보이는 반면에 고비용 부담이라는 단점이 있으며, 후자의 방식은 저비용으로 다량의 처리가 가능하다는 장점이 있지만, 상대적으로 낮은 정확률을 보인다. 이러한 양단의 단점을 극복하려는 시도로는 부분구문분석 정보만을 이용한 [Ferrandez 2000]의 접근방식을 들

수 있다.
한국어 대용과 생략현상에 관련하여 [차건희 외 97]에서는 중심화(Centering) 이론의 틀 안에서 베이시안 확률을 적용하는 기법을 제안했다. 또한 한국어 담화의 영대명사 현상에 관련하여 [이민행, 이익환 99]에서는 통제된 중심화 이론(Controlled Centering Theory) 하에서 프레임과 슬롯구조를 이용하여 다루었다. 이러한 설명의 장점은 대명사의 선행사가 앞 문장에 통사적으로 등장하지 않는 경우에도 인덱싱해 줄 수 있다는 점에 있다.

3. 백과사전의 병명기술에 토대한 개념지도

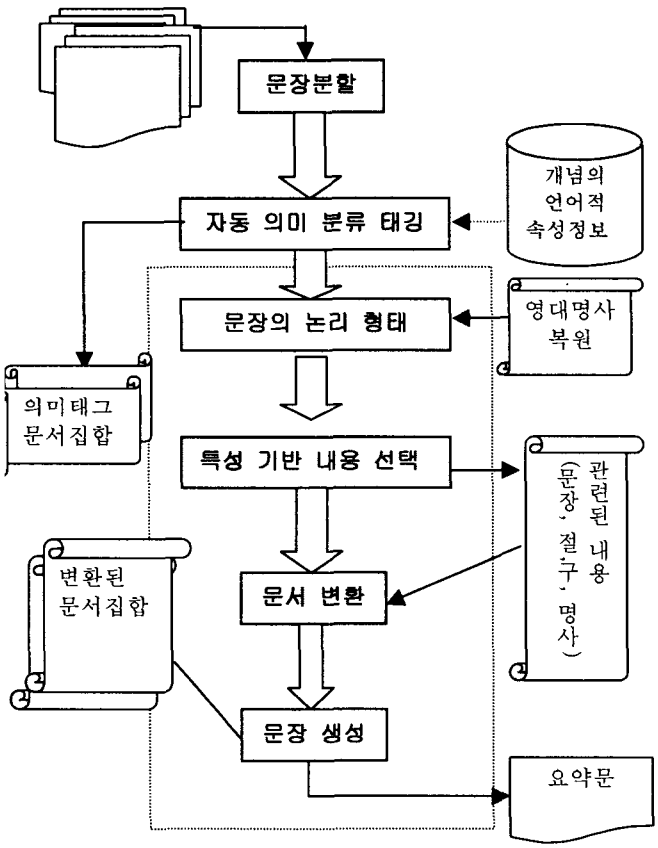

그림 1: 요약시스템 흐름도

본 논문은 [그림 1]과 같은 다중문서의 요약시스템 흐름도에서 문장분할을 통해 단문화된 문장에 의미범주가 태깅된 후 논리형식으로의 변환과정에 초점이 맞춰져 있다. 다중문서로부터 추출된 단문장들은 해당 의미범주별로 수집된다. 수집된 문장들의 의미비교를 위해서는 문장들이 구문분석된 후 논리형식으로 변환되어야 한다.
문장의 논리형식을 위해서는 영대명사는 선행사를 찾아 인덱싱되어야 하는데, 본 논문에서는 이 절차를 영대명사의 복원이라고 칭한다.

본 연구의 기초자료로 사용된 ETRI와 계몽사가 1997년 공개한 백과사전은 23,000 항목의 표제어를 담고 있다. 인간 병명 관련 표제어는 [그림2]에서처럼 구성상의 체계적 특성을 보여준다. <id>는 각 표제항목에 부여되는 고유번호이며, <title>에는 표제어로서 병명이 등장한다. <contents>에는 표제어에 대한 설명이 등장하는데, 병명에 관련한 일정한 특징적 의미성분을 포함한다. 예를 들어, "뇌빈혈"에 대한 설명부분은 [1]정의, [2]증상, [3]/[4]원인, [5]치료라는 의미적 범주화가 가능하다.

```
<id>      04270
<title>   뇌빈혈     腦貧血
<contents>
[1]뇌혈관의 혈액 순환이 나빠져서 혈액이 부족하여 일어나는 병
[2]갑자기 얼굴빛이 창백해지고 현기증이 일어나서 쓰러지며, 심하면 의식을 잃는다.
[3]뇌빈혈은 대부분 놀라거나 공포 따위가 원인이 되어, 뇌동맥이 수축함으로써 일어난다.
[4]그 밖에 많은 출혈이나 심장 쇠약, 다른 부위로 혈액이 지나치게 흘렀을 경우에도 일어난다.
[5]응급 처치로는 옷을 느슨하게 하고, 머리를 낮게 하여 휴식을 취하게 한다
```

그림 2: 백과사전 병명 기술의 예

더 나아가 각 의미범주에 등장하는 어휘는 특정한 의미관계를 형성한다. 이 관계는 다음과 같이 도식화되며 일종의 질병에 대한 지식지도로 간주될 수 있다. [Paik, H.-S., etc. 2001]에서는 [그림3]과 유사한 질병에 관한 개념지도를 설정하고 있다.

이 개념지도에 의하면 질병을 크게 4가지 의미범주로 구분하여, 각 의미범주별로 사용되는 명사(구) 및 술어들의 어휘군을 수립하였다[1]. 특히 '환자'는 '증상', '원인', '치료' 의미군에 관여적이며, '치료사'는 '치료' 의미군에 관여적임을 보여준다. 이 개념지도는 단편적인 것으로 많은 데이터를 통해서 보완 통합되어야 한다[2].

[1] 위 4가지 의미범주에 해당되지 않는 내용은 기타 의미군으로 분류할 수 있다.
[2] '치료사'는 의사나 간호인일 수도 있고, 경우에 따라서는 환자 자신일 수도 있다.

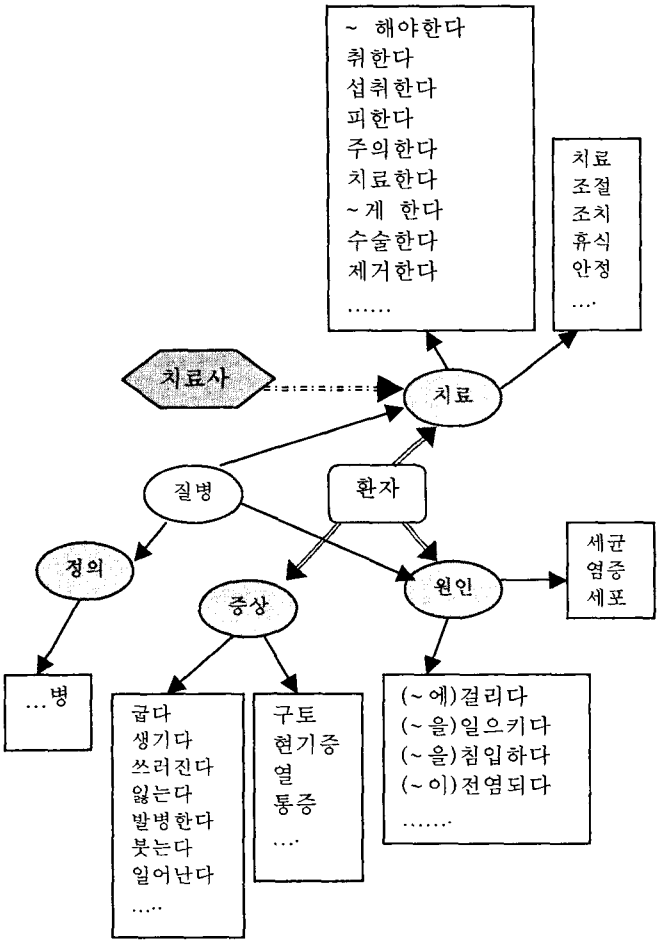

그림 3: 질병에 대한 개념지도

4. 영대명사의 분포

[그림2]의 텍스트에서 생략된 문장성분 즉, 영대명사를 복원하면 [그림4]과 같다.

<id> 04270
<title> 뇌빈혈　　腦貧血
<contents>
[1] 뇌빈혈은 뇌혈관의 혈액 순환이 나빠져서 혈액이 부족하여 일어나는 병이다./정의
[2] 환자는 갑자기 얼굴빛이 창백해지고 현기증이 일어나서 쓰러지며, 심하면 의식을 잃는다./증상
[3] 뇌빈혈은 대부분 놀라거나 공포 따위가 원인이 되어, 뇌동맥이 수축함으로써 일어난다./원인
[4] (뇌빈혈은) 그 밖에 많은 출혈이나 심장 쇠약, 다른 부위로 혈액이 지나치게 흘렀을 경우에도 일어난다./원인
[5] 뇌빈혈의(/는) 응급 처치로는 (치료사가) 환자를 옷을 느슨하게 하고, 머리를 낮게 하여 휴식을 취하게 한다/치료

그림4: 영대명사를 복원한 예

백과사전의 질병 기술에 나타난 영대명사 복원은 백과사전의 언어적 특징을 분석하므로써 가능하다. 한국어의 영대명사는 대명사가 등장할 자리가 형태론적으로 비어있는 일종의 생략현상으로 간주될 수 있다. 생략된 성분의 복원은 문맥 혹은 상황으로부터 가능하며, 따라서 문어 보다는 구어가 두드러진 생략 현상을 통해서 표현의 간결성을 보인다. 마찬가지 이유에서 문어 텍스트 중에서도 백과사전은 표제어와 설명부로 이루어진 구성상의 특징 때문에 설명부는 가능하면 표제어가 생략된 채 기술된다. 생략된 표제어는 대부분의 설명부 문장에서 일종의 주제어(Topic)로서 복원될 수 있다 ([1], [4], [5] 참조)[3]. 표제어 병명은 대부분의 경우에 주제어 형태로 복원될 수 있다. 주제어는 주어나 목적어와 같은 논항성분일 수도 있지만 ([1], [4]) 부가어 혹은 소유격 명사로부터 변환될 수도 있다([5]). 논항 성분으로서의 복원은 문법적인 측면에서 후자와 차이난다. 동사의 하위범주화 정보를 만족하기위해서 전자만이 필수적이기 때문이다. 본 논문에서 관심은 하위범주화 정보와 관련된 전자의 경우에 있다.

문제는 표제어 이외에도 전형적으로 생략되는 성분이 있다. 즉, "환자"가 생략된 경우([2], [5])와 "치료사"가 생략된 경우([5])가 그렇다.

이제 문제는 위 세가지 종류의 어휘가 생략될 수 있다면, 그들 영대명사의 복원 규칙을 세우는 것이다.

5. 영대명사의 복원 규칙

본 논문에서 다루고자 하는 영대명사의 선행사는 앞 문장/문맥 속에 등장하지 않는다는 점에서 중심화 이론의 "중심(Center)" 개념이 적용될 수 없다. 또한 통제중심화 이론을 받아 들여 "병명"과 "환자", "치료사" 등을 프레임과 슬롯 분석에 포함시킨다 하더라도 어떻게 이들 영대명사를 복원시킬지에 관해서는 추가적으로 많은 제약을 부과해야하는 난점이 있다.

본 논문에서는 [그림 3]에서 제시한 질병에 관한

[3] 한국어에 있어서 기저구조적 주제어와 통사적 변형에 따른 주제어에 대한 자세한 논의는 서정수(1991)를 참조.

개념지도에서 영대명사 복원의 해결책을 찾고자 한다. 다음은 영대명사 복원을 위한 규칙의 일부이다.

1. '/치료'로 의미범주가 태깅된 곳에서는
   ㄱ. 다음과 같은 동사와 관련된 통사적 환경에서 "치료사"를 주어위치에 복원한다: "치료하다", "~(하)게/도록 하다", "제거하다", "주사하다" 등
   ㄴ. "~(하)게 하다"의 환경에서 "환자"를 목적어 위치(내포절 주어)에 복원한다.
   ㄷ. 기타 통사적 환경에서는 "환자"를 필요에 따라 해당 위치에 복원한다.
2. '/정의'라고 의미범주가 태깅된 곳에서는 표제어 병명을 복원한다.
3. 위 두 의미범주를 제외한 여타 의미범주에서는
   ㄱ. 술어가 주어 위치에 [+animate]를 선택제약할 경우 "환자" 를 복원한다.
   ㄴ. 여타의 경우에 해당 위치에 표제어 병명을 복원한다.

그림5: 영대명사의 복원규칙

[그림5]에서 수립한 영대명사의 복원규칙이 적용된 예를 보면 다음과 같다.

1.
   ㄱ. (치료사는) 환자를 따로 떼어 놓고 치료해야 한다.
   ㄴ. (치료사는) (환자를/가) 안정을 취하도록 해야 한다.
   ㄷ. (환자는) 자극이 심한 음식물을 피해야 한다.
2. (백혈병은) 혈액 속의 백혈구 수가 무제한으로 늘어나는 병(이다)
3.
   ㄱ. (환자가) 갑자기 얼굴빛이 창백해지고 현기증이 일어나서 쓰러지며, 심하면 의식을 잃는다
   ㄴ. (당뇨병은) 만성이 되기 쉽다.

그림6: 영대명사의 복원규칙이 적용된 예

[그림5]의 영대명사의 복원규칙은 다음과 같은 특징을 갖는다.

첫째, 영역 한정적인 전문지식을 토대로 한다. 즉, 질병에 관한 지식지도를 가정하여 일정하게 태깅된 의미범주를 정보를 이용했다 (1, 2).
둘째, 한정적으로 통사정보를 이용했다. 동사의 하위범주화 정보를 토대로 주어 혹은 목적어 위치에 해당 영대명사를 복원했다 (1ㄴ).
셋째, 한정적으로 의미적 선택제약 정보를 이용했다. 주어 위치에 [+animate] 자질이 관여적이거나 동사의 어휘적 정보에 따라 주어 혹은 목적어가 결정되는 경우가 그렇다 (1ㄱ, 3ㄱ).

위에서 가정된 영대명사 복원규칙의 장점은 기왕에 문서 요약을 위해 가정되는 지식지도를 이용했다는 점과 구분분석에 토대한 사전의 어휘정보를 이용했다는 점에 있다. 즉, 영대명사 복원을 위한 추가적인 별도의 어떠한 장치도 도입하지 않았다는 것이다.

이제 위 규칙이 일정한 적용의 순서를 갖는다는 점에 주목하자. [그림5]의 항목 1에서 "치료사"의 복원을 우선적으로 적용한 다음 "환자" 의 복원을 검토했다. 항목 3에서는 "환자" 의 복원을 검토한 다음 표제어 병명을 복원하도록 했다. 따라서, 다음과 같은 복원의 순위를 보인다.

치료사 > 환자 > 병명

그림7: 영대명사의 복원 순위

이러한 복원 적용은 [그림 3]의 개념지도상 특수 의미범주에 속할수록 우선 순위를 갖는다고 일반화될 수 있다.

6. 결론 및 남은 문제

지금까지 논의는 다음과 같이 정리된다.
백과사전의 질병에 대한 기술은 영역특수적인 개념지도를 토대로 조직화되어 있다고 전제하였다. 영대명사의 복원을 위해 일반영역에서 필요한 언어지식 기반 방식에 대한 논의를 배제하고, 오로지 영역 한정적인 처리를 위해서는 개념지도 기반 방식이 효율적이라는 입장을 취했다. 이를테면, 질병 기술에는 표제어 병명, '환자' '치료사' 가 영대명사의 위치에 복원되는데, 의미범주 정보에 토대하여 복원의 우선순위가 결정된다는 것이었다.
본 논문에서 생략한 개념지도의 구축에 대한 논의는 더 많은 연구를 필요로 한다. 구조적인 면에서 그리고 양적인 면에서 개념지도의 완성을 위해서는 보다 많은 데이터의 보충이 요구된다. 본 논문에서

제시한 영대명사의 복원 규칙에 관한 실험도 이루어지지 않은 점은 본 모델의 검증을 위해서도 향후 과제로 남는다.

5. 참고 문헌

박철우. 한국어 정보구조에서의 화제와 초점. 서울대 박사학위 논문, 1998.

서정수. 현대 한국어 문법연구의 개관, 한국문화사, 1991.

이민행, 이익환. "한국어 대화에서의 대명사의 선행사 탐색 - 통제된 중심화이론적 접근," 제 11회 한글 및 한국어정보처리 학술대회 발표논문집, 382-388, 1999.

차건희, 송도규, 박재득. "한국어 대용과 생략 해결을 위한 센터링 이론의 적용" . ' 97 한글과 한국어 정보처리 발표논문집, 1997.

ETRIKEMONG Set. Electronics and Telecommunications Research Institute & (C) Kemong, 1997.

Ferrandez, A., J. Peral. "A Computational Approach to Zero-pronouns in Spanish." ACL, 166-172, Hongkong, 2000.

Grosz, B.J., A.K. Joshi, S. Weinstein. "Centering: a framework for modeling the local coherence of discourse." Computational Linguistics, 21, 1995.

Hobbs, J. "Resolving pronoun references." Lingua, Vol. 44, 1978.

Dagan, I., A. Itai. "Automatic processing of large corpora for the resolution of anaphora references" . COLING ' 90, Helsinki, 1990.

Nakaiwa, H. "Atutomatic Extraction of Rules for Anaphora Resolution of Japanese Zero Pronouns from Aligned Sentence Pairs" , NTT , 1997.

Nasakawa, T. "Robust method of pronoun resolution unsing full-text information." COLING ' 94, Kyoto, 1994.

Paik, H.-S., Y.-S. Kang, K.-S. Choi. "" Analysis of Linguistic Features for Identifying Information Constituents of a Concept. The 6th NLPRS, Nov. 2001 (forthcoming).