문서요약을 위한 조응 대용 해결

김상수[0] 김계성 노태길 이상조
경북대학교 컴퓨터공학과
{atri01[0], kskim, nayas}@sejong.knu.ac.kr, sjlee@bh.knu.ac.kr

Resolution of Context Anaphora for Text Summarization

Sang-Soo Kim[0] Kye-Sung Kim, Tae-Gil Noh, Sang-Jo Lee
Dept. of Computer Engineering, Kyungpook Nat'l University

요 약

한 문서에서 동일한 개체(Entity)를 지칭하는 고유명사가 다른 형태로 출현하는 현상은 문서요약의 품질을 떨어지게 만드는 요소이다. 이런 문제를 해결하기 위해서는 각각의 고유명사 및 지칭어를 인식하고 이들간의 상관 관계를 밝혀야 한다.

본 논문에서는 이런 문제를 개체명 조응 대용 관계로 정의하고 출현 특성에 따라 분류한 후 특성에 맞는 처리 방법을 보인다. 이를 위하여 고유명사의 조응 출현 양상에 따른 휴리스틱을 만들고, 고유명사를 지칭하는 명사들의 시소러스를 구축한 후 이들을 처리하는 방법을 제안한다.

1.서론

인터넷의 급속한 발달로 인하여, 많은 정보들이 생겨나고 있다. 이들 정보를 담고있는 문서, 특히 웹 문서의 증가로 인하여 정보 검색 시스템의 필요성이 증가되었고, 이에 따라 자동 요약 시스템의 중요성도 함께 부각되고 있다.

문서 요약은 생성 요약과 추출 요약으로 나눌 수 있다. 생성 요약은 문서의 담화 구조를 파악한 후에 요약문을 생성하고, 추출요약은 중요 문장의 추출을 통하여 요약문을 생성한다. 그리고, 두 종류의 문서요약에서 중요한 요소중의 하나가 응집성이다. 응집성은 추출요약에서는 요약 결과의 일관성 및 가독성을 높이기 위해서, 생성요약에서는 담화 구조를 파악하기 위해 담화의 연결관계를 의미하는 응집성에 많은 연구가 있어왔다.

응집성의 여러 유형 중에서 지시(reference)와 반복(repetition)에 해당되는 동일한 고유명사가 다른 어휘로 반복적으로 표현되는 현상과 대용어 처리를 통하여 응집성을 높일 수 있다.[1]

본 논문에서는 동일한 개체의 고유명사가 다른 어휘로 자주 표현되는 현상과 대용어를 정의 및 분류하고 그 분류에 따른 처리 방법을 보인다.

2장에서는 대용에 관한 관련 연구를 살펴보고 3장에서는 대용의 분류 및 접근 방법을 알아보기로 한다. 4장에서는 실험 및 결과에 관하여 평가해보고, 5장에서 결론을 맺는다.

2.관련연구

문서 요약은 국내외에서 다양한 방법들로 많이 진행되고 있나, 국내에서는 주로 생성 요약보다는 문서를 대표하는 문장을 추출하는 추출 요약 방법이 주를 이루고있다[2]. 한편 고유명사(Named Entity)의 추출 및 참조(대용)에 관한 연구는 MUC 컨테스트를 통해 활성화되기 시작되었고[3,4], 일본은 IREX (Information Retrieval and Extraction Exercise) Workshop을 통해 개체명을 추출하는 연구가 활발히 진행되고 있다.

국내에서는 고유명사의 인식에 관한 연구[6]는 있었으나, 인식된 고유명사들 간의 상관관계에 관한 연구는 아직 발표되지 않았다. 그리고, 일반적인 대용어 처리의 한 부분으로 명사구, 지시·인칭 대명사 등의 대용 현상을 처리하는 연구는 활발하게 진행 되고있다.[7]

3. 개체명 조응 대용 해결

일반적인 대용 현상은 말의 경제성에 따라 문장의 구성요소인 선행어가 먼저 등장하고 이후에 대용어가 대치되어 사용되는 순행 대용을 뜻한다.

본 논문에서는 이런 대용 현상들 중에서 선행어로 특정 개체명이 등장하고 그 개체명의 축약형이 대용어로 사용되는 경우와 '지시사+대용사(명사,그,녀,..)'의 형태로 개체명을 대용하는 경우를 개체명 조응 대용이라 하고 이들을 처리한다. 개체명이란 다양한 고유명사들 중에서 사람(PSN), 단체(ORG), 장소(LOC)로 나누고 이 범주에 속하는 고유명사로 삼는다.

개체명 조응 대용을 선행어의 부분 음절과 주변의 명사가 결합되거나 선행어의 부분 음절들로만 구성되는 직접 조응 대용어와 지시어나 인칭어를 사용해서 선행어를 간접 조응 대용하는 것으로 나누어 처리했다. 간접과 직접 조응 대용으로 나누어 처리하는 이유는 이들 대용어들의 출현 특성이 다르기 때문이다. 직접 조응 대용어(이하 직접 대용)의 경우에는 선행어가 나타난 다음에 나타나는 대용어의 형태를 명확하게 알 수 있는 형태론적인 현상이 존재하지 않는다. 따라서 처음 등장하는 선행어(개체명)에서 주변 정보를 수집한 다음 축약되어 사용되는 형태를 추론하고, 그 결과로서 선행어 다음에 그 축약형이 나타나는지 검사하는 방법이 가장 좋은 방법이다.

간접 조응 대용(이하 간접 대용)은 직접 대용과 달리 인칭 대명사와 '지시 관형사+명사'의 형태로 그 출현을 확인 할 수 있다. 여기서 인칭 대명사는 사람(PSN)을 지칭하는 것을 알 수 있으나 '지시 관형사+명사'인 경우는 명사의 성격에 따라 문장 전체, 명사구, 날짜, 시간 등 다양하게 대용된다.

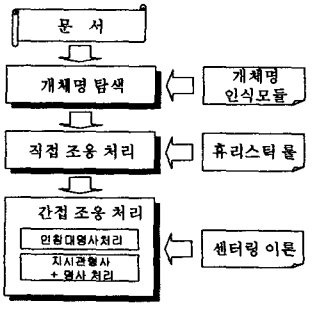

그림1. 시스템 구성도

전체적인 시스템은 먼저 직접 대용을 처리하고 간접 대용을 처리하는 순서로 처리된다. 그림1은 전체적인 시스템의 구성도를 보여준다.

3.1 직접 조응 대용어

직접 대용 처리는 개체명의 부분 음절들로만, 또는 부분 음절을 가지고 동일한 명사를 가지는 개체명은 동일한 개체를 의미한다고 보고 처리한다.

그림 2는 직접 대용어의 처리 과정을 보여준다. 처리 과정은 먼저 선행어인 개체명을 탐색한 다음 휴리스틱을 적용하여 직접 대용어 후보를 만들고, 만들어진 후보들을 선행어가 나타난 지점 이후에 존재 유무를 검사하는 순서로 처리하였다.

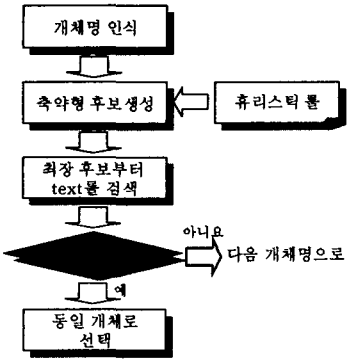

그림2 직접 조응 대용어 처리

휴리스틱 룰은 개체명을 사람(PSN), 단체(ORG), 장소(LOC)로 분류하고, 각 개체명의 분류에 따라 작성하였다.,

3.1.1 사람(PSN)에 관련된 휴리스틱

개체명을 인식하고 주변의 명사(실마리어)를 탐색한 후 직접 조응 대용어의 후보를 생성한다. 이때 주변 실마리어를 수집할 때 다른 종류의 개체명은 무시된다.

- 문장 S = $\{E_1,E_2,E_3,..,E_{n-1},E_n\}$
- 어절 E_i= $\{Sub_1, Sub_2, .., Sub_n\}$, Sub_i:1음절

표1. 사람에 관련된 휴리스틱

선 행 어	휴리스틱	예
<PSN : $E_1,E_2,..,E_n$> + Noun	E_nSub_1 + Noun	김 대통령
	E_nSub_1 + 인칭접미사	김씨
	E_nSub_{n-1}+ $EnSub_n$	대중
	E_n	히딩크

3.1.2 단체(ORG)에 관련된 휴리스틱

단체를 직접 조응하는 경우에는 개체명이 단체를 의미하는 접미사(~회,~련,~소,..)의 존재 여부에 따라 처리가 달라진다. 그 이유는 경우에 따라 직접 대용되는 현상이 다르기 때문이다.

표2. 단체에 관련된 휴리스틱

선 행 어	휴 리 스 틱	예
<ORG : $E_1,E_2,..,E_n$>	E_1Sub_1 + E_2Sub_1 + E_nSub_n	전경련, 한통
	E1 + E_nSub_n	기아차,현대차
	E_1+E_2Sub_1+E_3Sub_1+ E_nSub_n	한국상의
	E_{n-1}+E_n	제약협회
	E1	삼성
$E_1,E_n(E_k)$	E_k	SK 텔레콤(skt)

접미사가 존재하는 경우에는 개체명의 첫 어절이 주로 일반 명사 또는 지명과 같은 명사가 나타난다. 따라서 이런 어절들은 후보들은 후보에서 제외시켜야 한다.

1.3 장소(LOC)에 관련된 휴리스틱

장소에 관한 직접 대용의 경우에는 주로 접미사가 생략되는 경우 또는 이름이 생략되고 접미사만 등장하는 경우 두 가지로 볼 수 있다. 그 이외로는 국가명에서 첫 음절만 사용되는 경우이다. 이 경우는 그 수가 많지 않아 테이블로 처리한다.

표3. 장소에 관련된 휴리스틱

선 행 어	휴 리 스 틱	예
<LOC:$E_1,E_2,..,E_n$>	E1 + E_nSub_n	서울시
	E_nSub_n	시,도
	E_nSub_{n-1}	서울
	E_1Sub_1	한,미,일,중

3.2 간접 조응 대용어 처리

간접 조응 대용어의 처리 대상으로는 지시 대명사를 제외한 '인칭 대명사'와 '지시관형사+명사'의 경우로 한정한다. 지시 대명사는 개체명 보다 명사구·절,문장 전체를 받는 경우가 많았다.

간접 대용은 생략, 대용 및 지시사(reference) 해결에 자주 사용되는 센터링 이론을 적용하여 처리하였다. 센터링 이론은 이전 발화의 담화요소인 Cf들의 가장 우선순위가 높은 Cp가 현 발화의 담화 중심요소인 Cb가 될 확률이 비교적 높다는 가정에서 출발한다[8]. 따라서 Cf의 우선순위가 어떻게 적용되느냐가 가장 중요한 문제가 된다.

처리 과정은 먼저 간접 조응 대용어가 발생한 문장으로부터 앞 3문장까지의 개체명을 수집하여 Cf를 작성하고, 가중치를 부가하여 Cp를 구했다. 가중치를 부가하는 것은 기존의 Cf 목록의 순위를 변경하여 처리하였다. Cf 목록의 순위는 다음과 같다.

① 후보 개체명과 대용어와의 관계
② 후보 개체명의 수식 명사와 대용어와의 관계
③ 문장간 거리 ④ 주어 ⑤ 목적어 ⑥ 보어

여기서 이들 명사와 대용어와의 관계를 정의하기 위하여 이들 명사들의 성격과 상관관계를 파악하고, 시소러스를 구축하였다. 그 예는 다음과 같다

표4. 개체명 참조 명사 시소러스

사람(PSN)		단체(ORG)		장소(LOC)	
그	1000	회사	2000	장소	3000
사람	1010	신문	2100	곳	3100
대변인	1020	일간	2110	...	
당국자	1030	경제지	2120		
장관	1040	잡지	2200		
...		...			

그림3은 인칭 대명사 처리의 예를 보여준다. 인칭 대명사는 개체명의 종류가 사람(PSN)인 경우와 단체 및 장소 종류의 개체명+사람을 뜻하는 명사가 붙은 개체명으로 한정한다.

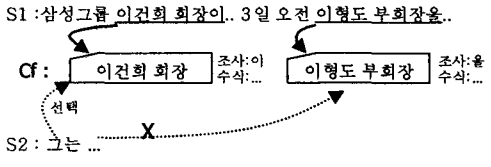

그림3. 인칭대명사 처리 예

'지시 관형사+명사'의 경우는 명사의 성격에 따라 개체명을 조응 대용할 수 도 있고, 문장 전체나 앞 문장의 명사구를 대용하는 경우가 많다. 따라서 명사를 어떻게 처리하는 것이 중요한 문제이다. 그림4는 지시 관형사 + 명사의 처리 예를 보여준다.

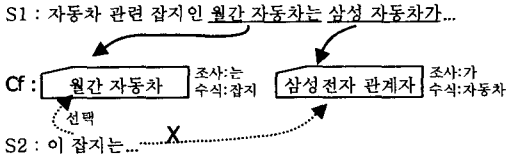

그림4. 지시관형사 + 명사 처리 예

4. 실험 및 평가

실험 코퍼스는 문장의 길이가 5문장 이상인 중앙일보 경제면 기사 91개로 구성하였다. 입력 데이터로는 수동으로 개체명의 원형만 태깅을 하였고, 처리를 위한 형태소 분석기는 HAM[9]을 사용했다. 표1은 코퍼스에서 직접 대용 처리의 빈도를 보여준다.

표5 직접 조응 대용 빈도

	PSN	ORG	LOC	전 체
전체 빈도수	202	751	422	1375
초기 원형	96	404	223	723
처리대상(%)	106(52.4)	347(0.46)	119(28.1)	572(41.6)

한 문서에서 문장의 수는 평균 8.02 문장이고, 한 문서에 개체명의 출연 빈도는 약 15.9개로 한 문장에 2개정도의 개체명이 출연하였다. 처리 대상은 전체 개체명들 중에서 약 41%로 나타났다. 간접 조응 대용은 전체 코퍼스에서 인칭대명사가 31개가 나타났고, 지시관형사+명사는 총 58개 출연하였으나 개체명을 지칭하는 경우는 모두 18개였다. 따라서 전체 처리 대상은 총 49개 였다.

평가는 정보 검색에 사용되는 정확율과 재현율을 사용하여 평가를 수행하였다.

표6. 처리 결과

	직 접				간접
	사람	단체	지역	평균	
재현율 (%)	90.5 (96/106)	97.4 (338/347)	94.9 (189/199)	95.5 (623/652)	85.7 (42/49)
정확율 (%)	94.1 (96/102)	76.2 (338/443)	98.4 (189/192)	84.5 (623/737)	93.3 42/45)

단체인 경우 정확율이 낮게 나타났고, 사람 및 지역은 상대적으로 높게 나타났다. 이는 단체의 축약 후보로 생성된 어휘들이 일반 명사로 사용되는지, 아니면 개체명의 대용으로 사용되는지 판단하는 기준을 더 마련해야 할 것으로 판단된다.

간접 대용은 개체명을 지칭하는 단어이지만 한 분류를 설명하는 명사구를 대용하는 경우에 처리가 불가능했다.

전체 처리대상이 전체 개체명의 41%가 처리 대상이 된다. 따라서 이들이 원래의 개체명으로 환원되면 추출 요약에서 사용되는 문장간 유사도 계산에 영향을 줄 뿐만 아니라 추출 후 요약의 질을 높일 수 있다.

아래 표 5는 문서를 입력한 후 추출요약으로 2문장을 추출한 결과물을 보여준다. 표 5의 첫 번째 문장의 '이 중국 본사회장'과 두 번째 문장의 '그'가 누구인지는 추출 문장만을 읽어서는 알 수 없다. 그러나, 본 논문의 시스템을 거치면 이들이 '이형도 부회장' 하나의 개체임을 알 수 있다.

표5. 개체명 대용 처리 이전 요약 생성물

이 중국본사 회장의 취임은 작년 10월 이건희(李健熙) 회장이 윤종용(尹鍾龍)삼성전자 부회장, 이형도 부회장 등 그룹 핵심인사들과 함께 중국 주요 인사들을 면담하고, 중국내 주요 사업장들을 시찰하는 등 중국시장의 중요성을 강조한 뒤 이루어진 것이어서 삼성의 중국 전략과 관계가 있는 의미 있는 인사 발령이다.

그는 또 삼성전자 종합연구소장이 된 후HDTV 기술, 프린터 기술의 연구 등을 시작하여 오늘의 디지털 제품들을 탄생시키는 초석도 마련했으며, 그후 삼성전기 대표로 부임하여 8년간 회사 규모를 6배나 키웠고 장기간 지속되던 적자구조를 흑자구조로 탈발꿈시켰다.

5. 결론 및 향후 연구

본 논문은 휴리스틱 룰 및 센터링 이론을 사용한 개체명 참조 해결 시스템을 제안했다. 먼저 직접적으로 대용되는 직접 대용어와 간접적으로 적용되는 대용어(인칭, 지시관형사+명사)를 해결했다. 이를 위해 개체명을 지칭하는 명사들의 상관관계를 밝히는 명사들의 시소러스를 구축 방안을 제공한다.

앞으로 처리 대상을 개체명 뿐만 아니라 명사구, 명사절, 문장 전체로의 확장을 통하여 참조(coreference)처리 방향으로 나아가야 할 것이다.

참 고 문 헌

[1] Jan Renkema, "담화연구의 기초", 한국문화사, pp.63-71, 1997

[2] 김계성, 이현주, 이상조, "단락 자동구분을 통한 중요 문장 추출", 한글 및 한국어 정보 처리 학술 발표 논문집,200년 10월,pp.233-237

[3] Yeal Ravin, Nina Wacholder, "Extracting Names from Natural-Language Text", IBM Research T.J. Watson Research Center

[4] Baldwin, B., "CogNiac: A discourse processing engine". Ph.D. Thesis, University of Pennsylvania, Department of Computer and Information Sciences. 1995

[5] 노태길, 이상조, "규칙 기반의 기계학습을 통한 고유명사의 추출과 분류" ,정보과학회 2000년 추계학술대회, VOL.27 NO.02 pp.0170-0172, 200년 10월

[6] 김정해, "HPSG 파서에 기반한 한국어 문맥 조응대용어의 해결", 박사학위 논문, 1996,경북대학교

[7] 김학수, 서정연, "다중모드 대화 시스템에서 이중 캐시 모델의 센터링 알고리즘을 이용한 명사 대용어구 처리." 2000년 11월 정보과학회 논문지B pp 1133-1140

[8] 차건회, 송도규, 박재득, "한국어 대용과 생략 해결을 위한 센터링 이론의 적용", 한글 및 한국어 정보처리 학술 발표 논문집. 1997년 10월, pp.347-352

[9] 강승식, "HAM;한국어 형태소 분석기와 한국어 분석 모듈"