실용영어에서 고유명사 일치를 위한 자료구조

김 종 선
해군사관학교 전산과학과

Efficient Data Structures of Coreference Resolution for Proper Names

Kim, Jong-Sun
Department of Computer Science, R.O.K. Naval Academy

요 약

고유명사가 문장속에서 다시 언급될때는 여러 가지 변형된 형태로 나타난다. 즉 같은 의미의 이름으로 사용되면서 서로 다른 이름 형태를 갖게된다. 이러한 경향은 coreference 처리를 어렵게 만든다 본 논문에서는 고유명사의 coreference와 의미상으로 인식되지 않은 고유명사의 식별에 이용될 수 있는 효율적인 자료구조를 제시한다.

1. 서론

고유명사의 coreference 모델은 같은 이름을 나타내면서도 서로 다른 형태를 갖는 패턴에 기본을 두고 있다. 축약된 형태의 이름으로 원래 형태의 이름을 찾기 위해서 효율적인 자료구조를 이용한다. 제시된 모델은 두가지 점에서 효율적이다. 새로운 이름이 주어지면 지금까지 언급된 이름 중에서 형태는 다르면서 같은 이름이 있는지 찾아야 한다 이때 후보 이름들의 탐색 공간을 줄임으로써 탐색 시간을 줄일 수 있다 제시된 모델은 주어진 이름의 구성 단어들중 하나의 단어로도 탐색이 가능하다. 여기에 추가하여 언급된 고유명사의 의미 분류(semantic category)가 주어지면, 의미 분류를 사용함으로써 후보이름의 탐색공간을 더욱 줄일 수 있다.

2. 고유명사의 패턴

신문이나 잡지의 실용적인 영어 문장에서 하나의 이름을 지칭하는 고유명사는 여러 가지 축약된 형태로 나타난다[1-3].

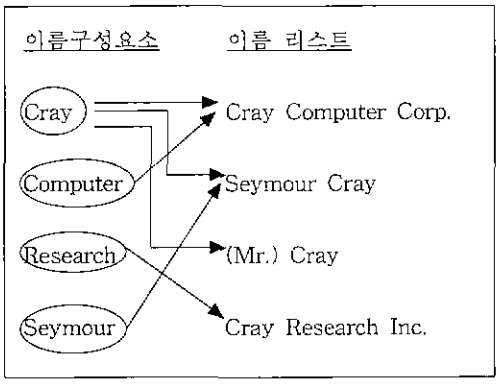
축약어는 공식적 혹은 비공식적 작문에서 사용하지 않고 전 세계적으로 통용되는 경우에만 사용한다고 정의하고 있지만[4]. 오늘날 발행하고 있는 신문기사나 정기 간행물에서 사용되고 있는 축약어들이 전체 단어의 25 퍼센트를 넘고 있다[5].

서로 다른 형태를 갖는 축약어는 고유명사의 인식을 어렵게 만든다 변형된 이름의 공통된 특색을 살펴보면, 처음으로 언급되는 고유명사는 생략 없이 완전한 형태로 주어지고 다시 언급될 때는 축약된 형태로 나타난다. 예를 들면, 사람 이름이 한 번 언급된 이후에 다시 나타날 때는 타이틀과 성(Dr. Cray)만 표시된다. 회사이름의 경우에는 이름의 일부 단어만 사용하여 축약어로 사용된다. 이와 같이 이름의 종류에 따라 축약되는 형태도 다르게 나타날 수 있다. 축약된 이름에서 이름의 의미 분류를 식별하는데 사용되는 주요 단어가 (예. Co, Corp., Inc., Mr. 등) 생략된 경우에는 고유명사의 식별을 어렵게 만든다.

3. 이름 일치를 위한 자료구조

이름 일치를 위한 자료구조는 고유명사의 coreference를 찾는데 이용될 뿐만 아니라 축약어의 원어를 인식하는데도 사용된다.

이를 위한 기본적인 개념은 기사에서 이름이 나타날 때마다 서로 다른 형태의 이름들을 이름 리스트에 유지한 다음 새로운 이름이 언급될 때 이름 리스트에서 주어진 이름과 공통의 단어를 포함하고 있는 이름들만 후보자로 선택한다. 예를 들면 그림1에서 Cray노드는 Cray를 포함하는 모든 이름(Cray Computer Corp, Seymour Cray, Cray, Cray Research Inc.)과 연결하는 링크를 갖고 있어서 'Cray' 단어로 구성된 모든 이름을 연결하게 된다.



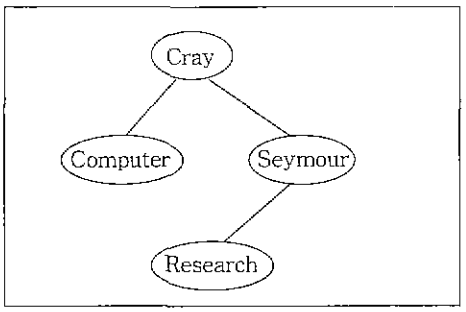
<그림 1> 이름 구성 요소의 노드구조

이러한 구조는 효율적인 이름 매치 구조를 제공한다. 왜냐하면 이름 리스트에 나열된 이름들은 그 이름의 어떠한 구성요소에 의해서도 이름을 찾을 수 있기 때문이다. 예를 들면 'Seymour Cray'는 Seymour나 Cray노드중 어느 노드에서나 접근이 가능하다. 이름이 축약된 형태로 사용될 경우에는 생략되는 단어에 대한 일정한 규칙이 있는 것은 아니다. 따라서 이름의 일부분을 가지고서 원래 이름의 탐색이 가능한 구조가 필요하다.

이러한 방법은 문자(character) 단위로도 원래의 이름을 찾을 수 있다. 약어는 이름을 구성하는 주요단어의 첫 문자로 구성된다. 예를 들면 NIH (National Institute of Health)의 full name을 찾기 위해서는 이름 구성요소 중에서 'N'으로 시작하는 노드를 찾으면 쉽게 접근할 수 있다.

3.1 이름 구성 리스트

이름 구성 리스트 (C-list)는 이름을 구성하는 서로 다른 단어들의 리스트를 말한다 예를 들면 처음으로 언급되는 이름 Cray Computer Corp.는 두 개의 요소 Cray와 Computer 노드가 만들어진다. 이름이 언급될 때마다 이름을 구성하는 새로운 단어들은 C-list의 노드가 생성된다 이때 노드의 탐색을 효율적으로 하기 위하여 노드는 이진탐색트리(binary search tree)에 저장된다 즉 C-list를 구성하기 위하여 이진탐색트리 구조를 사용한다. 그림2는 그림1의 이름리스트에 보이는 이름의 C-list를 보여준다



<그림 2> C-list를 위한 이진탐색트리

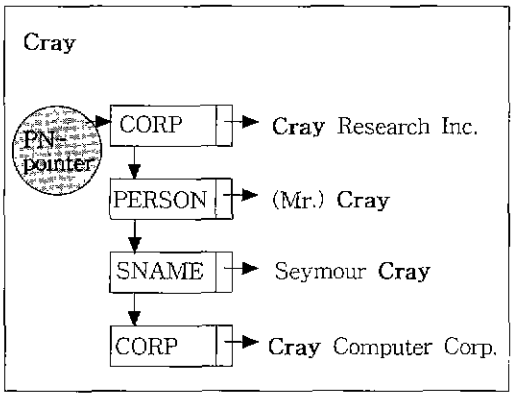
3.2 이름 리스트

트리의 각 노드는 서로 다른 이름 구성요소를 포함하는데, 이것은 고유한 단어를 나타낸다. 각 노드는 고유한 단어와 그 단어를 포함하는 이름과 연결하기 위하여 이름리스트(PN-list)라 불리는 동적 리스트 구조를 갖는다 리스트의 처음은 PN-pointer가 가르키고 있다(그림3 참조).

PN-list는 새로운 형태의 이름이 인식될 때마다 노드가 자동적으로 생성되는데, 노드는 이름의 의미형태 (semantic type)와 포인터를 갖게된다.

포인터는 이름을 연결하는데 사용되고, 의미 형태

는 이름의 의미적인 값(사람, 회사, 도시 등)을 저장하게 된다. 이름의 분류가 식별되지 않을 경우에는 미식별(unknown)로 분류된다. 그림3은 Cray 노드 내부의 이름 구성 리스트의 구조를 나타낸 것이다.



<그림 3> Cray 노드의 이름리스트 구조

PN-list 구조는 이름을 일치시키기 위한 탐색 공간을 줄여준다. 예를 들면 새로 언급된 이름의 의미형을 알게되면 (Mr. Cray: Cray는 사람 성을 나타냄) 해당 노드 (Cray) 안에서 PN-list의 의미형이 SNAME (surname을 의미함)을 찾아서 링크를 따라가면 쉽게 찾을 수 있다.

그리고 PN-list 에 생성된 링크들은 생성된 시간에 따라 자동으로 정렬되어 있다. 즉 새로운 노드가 생성될 때마다 리스트의 앞에 삽입함으로써 PN-pointer에 가까운 노드는 멀리 있는 노드 보다 최근에 생성되어진 노드이다 리스트를 순차적으로 방문하면 자동적으로 최근에 언급된 이름을 먼저 찾게되는 구조를 가지고 있다. 예를 들면, 'Cray Computer Corp.'와 일치하는 이름을 찾기 위해서는 Cray 노드를 방문한다. Cray 노드의 PN-list를 차례대로 방문하면 Cray 단어를 포함하는 이름 중에서 최근에 언급된 순서대로 방문하게 된다 이와 같은 구조는 recency constraint[6]를 구현한 것이다.

4. 결론

본 연구에서는 실용적인 영어 기사에서 고유명사(proper names)의 coreference 해결을 위한 효율적인 데이터 구조의 모델을 제시하였다. 이 모델은 C-list와 PN-list의 두 가지 기본적인 자료구조를 갖는다.

이름의 구성요소들은 C-list에 저장함으로써 같은 의미의 이름으로 변형된 형태의 이름을 찾을 수 있다. 뿐만 아니라 해당 구성 요소 노드를 직접 방문함으로써 후보자 이름의 대상을 줄일 수 있다.

제시된 coreference 모델은 월스트리트 저널에서 고유명사의 의미를 추출하고, 문맥상에서 서로 관련된 이름을 찾아서 추출된 의미를 결합하는데 이용되었다.

[참고문헌]

[1] Carroll, John M., 1985, *What's in a Name*, W.H. Freeman and Company, New York

[2] Amsler, Robert, 1987, Words and Worlds. *Proceedings of the Third Workshop on Theoretical Issues in Natural Language Processing (TINLAP3)*, New Mexico State University at Las Cruces, NM, January 7-9, 1987, 11-15.

[3] Coates-Stephens, Sam, 1993, The Analysis and Acquisition of Proper Names for the Understanding of Free Text, *Computer and Humanities*, Vol. 26, 1993, 441-456.

[4] Guralnik, David B. (ed.), *Webster's New World College Dictionary*, Merriam Webster, Springfield. MA, 1996

[5] De Sola, Ralph, 1986, *Abbreviation Dictionary*: Augmented International Seventh edition, Elservier Science Publishing Co. Inc., New York.

[6] Allen, James, 1995, *Natural Language Understanding*, The Benjamin/Cummings Publishing Company, Redwood City, CA.