패턴 학습을 이용한 고유명사 추출

김현준*0, 김정화*, 강승식*, 우종우*, 윤보현**
*국민대학교 컴퓨터학부
**한국전자통신연구원 언어공학부
{hjkim, jhkim, sskang, cwwoo}@kookmin.ac.kr, ybh@etri.re.kr

Proper Noun Extraction Using Pattern Learning

Hyun-Joon Kim*0, Jeong-Hwa Kim*, Seung-Shik Kang*, Chong-Woo Woo*,
Bo-Hyun Yun**
*School of Computer Science, Kookmin University
**Linguistic Engineering Department, ETRI

요 약

본 논문은 고유명사를 활용하여 특정 정보를 좀더 효율적으로 추출하기 위한 연구이며, Named Entity 의 한 범주인 사람 이름에 대하여 어휘 사전이나 실마리 사전의 사용 없이 초기에 주어지는 몇 개의 인칭 명사들을 태그가 부착되지 않은 코퍼스에 적용시켜 고유명사 추출을 위한 패턴을 학습하고, 그 패턴을 적용하여 새로운 고유명사를 생성해 내는 작업을 통해 인칭 명사들을 효율적으로 추출할 수 있는 방법을 제안한다.

1. 서론

정보의 양이 증가함에 따라 기존의 정보 검색 시스템들은 중복되는 많은 검색 결과를 추출하고 있다. 이러한 검색 결과에서 원하는 정보를 추출하기란 매우 어렵고 많은 시간과 노력이 든다. 정보 추출은 이러한 수많은 정보들로부터 미리 정의된 주제나 관심분야의 정보만을 효율적으로 추출하는 연구분야 이다.

정보 추출을 위한 한 단계로써 Named Entity들을 추출하는 작업이 선행되어야 하는데, 이때, Named Entity가 될 수 있는 정보로는 일반적으로 문서를 대표할 수 있는 명사이다. 이중에서도 고유명사는 문서 내에서 다른 어휘에 비해 문서의 내용을 효과적으로 대표할 수 있다. 이러한 정보추출에 관한 연구는 미국 정부 주도의 MUC(Message Understanding Conference) 학술회의로 대표되며, Named Entity와 동일지시어(Coreference)에 대한 연구가 활발히 진행되고 있다[1]. 국내에서의 고유명사 추출에 대한 연구는 주로, 미리 기재된 어휘사전이나 실마리 사전에 의존적이다[5][6]. 본 논문에서는 Named Entity 의 한 범주인 인칭 명사에 대하여, 어휘 사전이나 실마리 사전의 사용 없이 초기에 주어지는 몇 개의 인칭 명사를 태그가 부착되지 않은 코퍼스에 적용시켜 인칭명사 추출을 위한 패턴을 학습하고, 그 패턴을 적용하여 새로운 인칭명사를 생성해 내는 작업을 통해, 문서 내의 모든 인칭 명사들을 추출할 수 있는 방법을 제안하고자 한다.

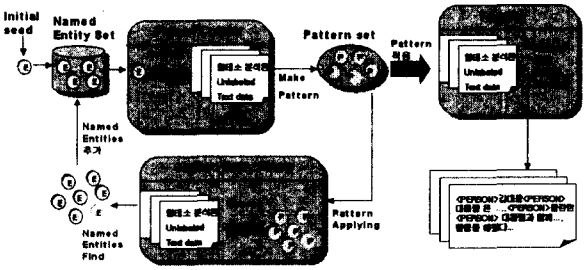
2. 관련 연구

고유명사 추출에 관한 연구는 MUC -6과 MUC -7의 Named Entity 컨테스트를 통해 활성화 되기 시작하였고, 일본에서도 IREX(Information Retrieval and Extraction Exercise) Workshop을 통해 Named Entity 를 추출하려는 연구가 활발히 진행되고 있다. 기존의 통계적인 방법이나 사람에 의해 직접 만들어진 패턴 적용 규칙을 통해 이루어졌던 고유명사 추출작업은 새로운 도메인의 적응성에 대해 한계에 이르렀다. 이러한 문제를 해결하기 위한 한 방법으로서, 일본에서는 형태소들 사이의 문맥적 관계를 학습하여 일본어내의 Named Entity를 추출하려는 연구가 있었으며, 태깅된 코퍼스를 이용한 실마리 학습을 통해 문서 내의 Named Entity를 추출하려는 연구가 있었다[2][3]. 이와는 달리 태깅되지 않은 코퍼스에서 문맥적 규칙(contextual rule)과 철자법 규칙(spelling rule)을 사용한 Named Entity 분류에 관한 연구가 있다[4].

국내에서는 고유명사 추출에 관한 연구로서, 형태소 분석 결과로부터 고유명사 출현 패턴과 패턴 부가 정보를 사용하여 인명, 기관 명, 회사명 등의 고유명사 추출의 정확도를 높이려는 연구가 있었으나, 이는 어휘사전에 의존적이다[5]. 그리고 정보 추출이나 정보 검색, 문서 요약과 같은 분야에 사용하기 위하여 실마리 학습을 통한 고유명사의 범주 결정을 통하여 고유명사를 추출하려는 연구가 있었으나 이는 태깅된 코퍼스를 사용한 실마리 학습이 선행되어져야 했다[6]. 그 외에도, 데이터 수집기를 통해 이름집합, 접사집합, 단서집합을 이용하여 고유명사를 추출하는 연구가 있으나, 외래어로 표기되는 고유 명사들에 대해서는 낮은 추출 확률을 나타내고 있다[7].

3. 고유명사 학습 시스템

3.1 고유명사 학습 시스템 구조



[그림1] 고유명사 학습 시스템의 구조

본 논문의 고유명사 학습 시스템은, Named Entity들의 집합을 사용하여 패턴을 생성하는 '패턴 생성기(Pattern Creator)' 와 패턴 생성기에 의해 생성되는 패턴 집합을 사용하여 문장 속에서 인칭명사를 찾는 '고유명사 추출기 (Named Entity Finder) ' 의 2개의 모듈로 구성된다[그림 1].

고유명사 추출기는 초기에 미리 주어지는 몇 개의 인칭 명사들을 사용하여 형태소 분석된 문서 내에서 패턴 생성기를 통해 패턴들을 생성한다. 생성된 패턴들은 다시 같은 문서 내에서 인칭 명사를 추출한다. 패턴에 의해 추출된 인칭명사가 Named Entity 집합에 속하지 않으면 새로운 인칭명사로써 추가되고, Named Entity 집합은 확장된다. 또한 이렇게 확장된 Named Entity 집합은 이전에는 없던 다른 패턴들을 생성함으로써 패턴 집합을 확장 시킨다. 따라서 이 두 가지 모듈이 반복적으로 수행됨으로써 패턴 집합과 Named Entity 집합이 학습되며, 이때, 이 두 모듈의 반복 수행은 더 이상 패턴이 증가하지 않을 때까지나, 임의로 정해진 반복 횟수만큼 수행된다. 패턴 학습이 끝나면 최종 패턴 집합을 가지고 문서 내에 존재하는 모든 인칭 명사들을 찾아 태깅하게 된다.

3.2 패턴 생성기

본 논문에서의 패턴 생성기는 기본적으로 Named Entity chunking/tagging에 사용되고 문맥적 단서로써 앞뒤의 형태소들 (preceding /subsequent morphemes) 을 고려하는 모델을 사용한다[2]. 문장 내에서 Named Entity 집합에 속하는 단어에 일치하는 어근을 가진 어절(W_0)이 찾아지면, 그 어절(W_0)의 앞에 오는 어절(W_{-1})과 뒤에 오는 어절(W_1)의 어근이 나타내는 문맥 정보를 이용한다. 이때, 앞뒤 어절의 어근이 명사나 동사 이외의 품사 (감탄사, 부사, 형용사 등)일 경우는 문장에 있어서 중요 정보가 될 수 없으므로 생성할 패턴의 앞뒤 문맥정보로 사용하지 않는다.

패턴을 생성하는 규칙은 아래와 같은 과정으로 나타낼 수 있다.

1) 문장 내에서, Named Entity 집합에 있는 Named Entity와 일치하는 어근을 가진 어절(W_0)을 찾는다.
2) 일치되는 어절(W_0)의 왼쪽 어절(W_{-1})의 어근이 명사인지 검사한다.
   - 명사일 경우 : 왼쪽 어절(W_{-1})의 어근이 패턴의 왼쪽 문맥 정보로 사용됨
   - 명사 이외의 경우 : 패턴의 왼쪽 문맥 정보는 null 값을 가짐
3) 일치되는 어절(W_0)의 오른쪽 어절(W_1)의 어근이 명사나 동사인지를 검사한다.
   - 명사나 동사일 경우 : 오른쪽 어절(W_1)의 어근이 패턴의 오른쪽 문맥 정보로 사용됨
   - 명사나 동사 이외의 경우 : 패턴의 오른쪽 문맥 정보는 null 값을 가짐
4) 일치되는 어절(W_0)의 어근이 미등록어일 경우에만 패턴을 생성한다.
5) 생성된 패턴을 패턴 집합에 추가 시킨다. (만약 생성된 패턴이 패턴 집합에 존재할 경우에는 추가하지 않는다.)

패턴 생성의 예는 다음과 같다.

예) 일치하는 인칭명사가 '홍길동'이고 Named Entity 타입은 'PERSON' (사람 이름)일 경우

(W_{-1})	(W_0)	(W_1)		
맛진	홍길동	대통령은	⇨	null : PERSON : 대통령
동사	미등록어	명사		생성된 패턴

3.3 고유명사 추출기

고유명사 추출기는 아래와 같이 인칭명사를 추출한다.

1) 문장 내의 한 어절이, 패턴의 왼쪽&오른쪽 문맥 정보에 일치하는 왼쪽&오른쪽 어절의 어근을 가지고 있는지를 검사한다.
2) 패턴이 일치되었을 때, 그 어절의 어근이 미등록어일 경우에는 인칭명사로 추출하고, Named Entity 집합에 추가한다.

3.4 인칭명사 태깅

인칭 명사 태깅에서는 최종 학습된 패턴 집합에 있는 패턴들을 사용하여, 문서 내의 인칭 명사들을 찾아 태깅한다. 이때, 현재 구현된 고유명사 추출기는 한 어절로 이루어진 인칭명사만을 찾기 때문에 두 개나 세 개의 어절로 이루어진 인칭명사를 하나로 묶어 주는 작업이 필요하다. 본 시스템에서는 문장에서 인칭명사가 연속으로 나올 경우에, 연속으로 이어지는 이 인칭 명사들을 하나의 인칭명사로 묶어준다. 이러한 과정에는 이전에 나오는 인칭명사가 하나의 형태소로만 이루어져 있어야 한다는 전제 조건이 따른다. 두&세 어절로 이루어지는 인칭명사를 찾는 예제는 다음과 같다.

(NE)	(NE)	(NE)	(NE)	
미하일	고르바초프는	부시와	클린턴을	만나
하나의 NE로 태깅		조사로 인해 묶이지 않음		

=> <PERSON>미하일 고르바초프</PERSON>

4. 실험 및 결과

4.1 실험 환경

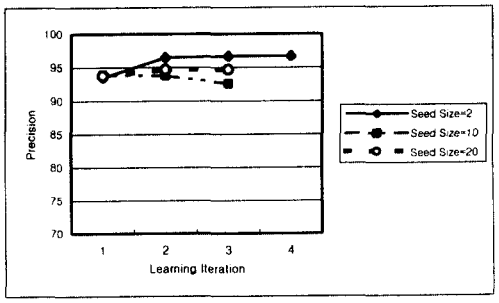
패턴을 생성하고 새로운 인칭명사를 추출하기 위해서는 모든 문서 내의 문장이 형태소 분석을 거쳐야 한다. 본 논문에 사용되는 고유명사 학습 시스템은 형태소 분석기 HAM version 5.0.0a를 사용하여 구현하였다[8]. 사용되는 문서는 '세종 말뭉치 98'에서 발췌한 것으로써, 문서의 크기가 520K인 원시 데이터이다. 현재 이 논문에서는 Named Entity 타입을 'PERSON'이라고 정한 인칭명사에 대해서만 실험 결과를 도출하였다.

4.2 초기에 사용될 인칭명사 선정

초기에 사용할 인칭 명사들은 문서 내에서 자주 출현하는 인칭 명사 몇 개를 선정한다. 본 실험에서는 'PERSON'이라는 Named Entity 타입을 가지는 2개, 10개, 20개의 초기 인칭 명사들이 사용되었다. 초기 인칭 명사들이 선정되면 이것들을 이용하여 Named Entity 추출기를 수행시킨다.

4.3 인칭명사 추출 실험 결과

이 실험은 반복 학습을 이용해 더 이상 새로운 인칭명사가 생성되지 않을 때까지 수행하였다. 수행과정 중에 패턴에 의해 추출된 새로운 인칭 명사들 중 인칭명사가 아닌 경우에는 제거 한 후 패턴을 추출했고, 찾아진 패턴 중 양쪽에 아무런 정보도 포함하지 않은 패턴(null:PERSON:null)은 자동 제거했다. [그림 2]는 이 실험 결과로 생성된 패턴이 어느 정도 정확하게 인칭 명사들을 찾아내는지에 관한 반복 수행마다의 정확률을 보여준다.



[그림2] 반복 학습에 따른 정확률

대체로 생성된 패턴은 인칭명사 추출에 93%이상의 정확률을 보인다. 또한 초기에 사용된 인칭명사의 개수가 20개인 경우가 10개인 경우 보다 높은 정확률을 보이는 이유는 반복 할 때마다 찾아지는 인칭명사의 개수가 잘못된 인칭명사의 개수 보다 증가율이 더 크기 때문이다. 이렇게 높은 정확률을 보이는 반면 재현율은 초기에 사용된 인칭 명사들을 어떻게 주느냐에 따라 다른 성능을 보인다. 이 실험에서 사용된 문서에 있는 사람 이름은 약 500개이며, 문서 내에서 출현되는 개수는 약 1273개이다. 초기에 사용된 인칭 명사들이 2~20개의 경우 3~4번의 반복 수행 후 약 84~100개(약 20%) 정도가 찾아진다. 그 이유는 초기에 주어지는 인칭 명사들이 특정 분야의 사람인 것과 실험에 사용된 문서가 '세종 종 말뭉치 98'에서 임의로 발췌한 다양한 분야의 문서이기 때문이다. 따라서 패턴들은 앞뒤의 문맥으로 특정 분야에 관련된 것들로 많이 생성되며, 이러한 특정 패턴들에 의해 찾아지는 인칭 명사들은 대부분이 특정 분야에 관련된 것들이다. 예를 들면, 실험 초기에 정치에 관련된 인칭명사(김대중, 클린턴 등)를 사용한 결과, 정치에 관련된 인칭명사(김영삼, 박정희, 엘친, 고르바초프 등)들은 대부분 찾아냈지만, 다른 분야의 인칭명사(세익스피어, 해밍웨이 등)들은 대부분 찾을 수 없었다. 따라서 이렇게 찾아진 인칭명사가 문서 내에서 출현 빈도수가 높으면 재현율이 높아지지만, 빈도수가 낮을 경우에는 낮은 재현율을 보이게 된다.

5. 결론

본 논문은 정보추출을 위한 기반으로서 Named Entity를 태깅하기 위해, 우선적으로 문서내의 인칭명사에 대한 Named Entity를 추출하는 방법을 제안했다. 이 방법은 사람 이름이 고유명사로 미등록어라는 점과 바로 앞뒤 단어들로 이루어진 패턴들의 반복 학습을 이용하여 인칭 명사들을 추출하는 것으로, 반복 학습 동안 93%이상의 정확률을 보였으며, 학습된 패턴만을 사용하여 국내외의 인칭명사까지 추출할 수 있었다. 그러나 아직은 초기에 사용되는 인칭 명사들에 의해 학습된 분야의 인칭명사만을 추출하는 단점이 있다. 향후 과제로는 분야에 상관없이 보편적인 인칭 명사들을 찾아내는 연구가 필요할 것이며, 이는 재현율을 높이는 방법이 될 수 있을 것이다.

참고 문헌

[1] MUC, Proc. of 7 th Message Understanding Conference(MUC-7). MUC. 1998.

[2] Utsuro, T. and M. Sassano, "Minimally Supervised Japanese Named Entity Recognition Resources and Evaluation", In *Proc. Of the 2 nd International Conference on Language Resources and Evaluation*, pp 1229-1236, 2000.

[3] Collins, M. and Y. Singer, "Unsupervised models of named entity classification ", In *Proc. 1999 Joint SIGDAT Conference on EMNLP and VLC*, 1999.

[4] Stevenson, M. and R. Gaizauskas, "Improving Named Entity Recognition using Annotated Corpora", *LREC Workshop on "Information Extraction meets Corpus Linguistics"*, 2000.

[5] 정래준, 김준태, "고유 명사 출현 패턴을 이용한 색인의 성능 향상에 관한 연구", 한글 및 한국어 정보처리 학술발표논문집, pp 68-72, 1996.

[6] 노태길, 이상조, "규칙 기반의 기계학습을 통한 고유명사의 추출과 분류", 정보과학회 추계 학술발표회,pp 170-172, 2000.

[7] 김태현, 이현숙, 하유선, 이만호, 맹성현, "데이터 집합을 이용한 고유명사 추출", 한글 및 한국어 정보처리 학술발표논문집, pp 11-18, 2000.

[8] 강승식, "한국어 형태소 분석기와 한국어 분석 모듈", http://nlp.kookmin.ac.kr/