규칙 기반의 기계학습을 통한 고유명사의 추출과 분류[1)]

노태길 이상조
경 북 대 학 교 컴 퓨 터 공 학 과
nayas@sejong.knu.ac.kr sjlee@bh.knu.ac.kr

Extraction and Classification of Proper Nouns by Rule-based Machine Learning

Tae-Gil Noh Sang-Jo Lee
Dept. of Computer Engineering, Kyungpook National University

요 약

고유명사를 추출하고 그 범주를 파악할 수 있다면, 이는 정보 추출이나 정보 검색, 문서 요약과 같은 분야에 도움을 줄 수 있다. 본 논문에서는 고유명사를 추출하고 그 범주를 찾는 방법을 제시한다. 고유명사가 태깅된 코퍼스로부터, 고유명사의 내부와 주변에 반복적으로 나타나는 실마리들을 규칙 기반으로 학습한다. 이를 통하여 고유명사를 찾고 그 범주를 정한다. 구현한 시스템은 경제기사 코퍼스에서 4가지 범주로 고유명사를 추출하고 분류함에 있어 79.8%의 재현율과 92.9%의 정확률, 그리고 F 평가치에서 85.8의 성능을 보인다.

1. 서론

고유명사는 특정한 사물이나 사람의 이름을 나타내는 명사이다. 보통명사와 달리 고유명사의 수는 유한하지 않다. '사람'이라든가 '회사'와 같은 개념의 종류는 유한하다고 볼 수 있지만 이들 각 '개념'의 구체적인 '개체'인 고유명사는 늘 새로이 생성되기 때문이다. 이를테면 사람 이름이나 회사 이름의 집합은 늘 새로운 이름에 대하여 열려 있다. 고유명사는 일반적인 개념과 비교하면 더 구체적인 대상을 지칭하고 있다. 이런 특징으로 인해 고유명사는 검색어휘로 더 높은 가치를 지님이 알려져 있다.[1] 또한, 정보 추출(Information Extraction) 분야나 자동요약분야에서도 고유명사는 다른 어휘에 비해 텍스트가 지칭하는 사건이나 사실에 필수적인 요소가 될 확률이 더 높다. 뉴스와 같은 도메인의 텍스트에서, 핵심 내용인 '어디에서', '누가', '무엇을'에 해당하는 요소가 일반적으로 고유명사라는 것이다.[2] 명사나 명사구가 고유명사인지를 파악하고, 그 고유명사가 어떤 것인지를(사람인지, 장소인지, 조직인지 등을) 파악할 수 있다면 정보검색이나 정보추출에 크게 도움이 될 수 있다. 처리 대상이 되는 도메인에 따라 정도의 차이는 있으나, 고유명사를 모두 사전에 기재하여 처리하는 것은 불가능하다. [3] 따라서 사전에 기재되어 있지 않은 고유명사를 처리하고 인식하는 방법이 필수적이다.

본 논문에서는 고유명사에 태그가 붙여진 학습 코퍼스로부터 고유명사의 어휘 내부 실마리와 문맥 실마리를 학습하고, 이렇게 학습된 규칙을 통해서 일반 텍스트에서 고유명사를 추출하고 사람, 조직, 장소, 기타의 4가지 범주로 분류하는 방법을 제시한다. 본 논문에서 제시하는 방법은 규칙에 기반하지만 학습을 통해 범주의 변화나 도메인의 변화에 적응할 수 있는 방법이다.

1) 본 연구는 정보통신연구진흥원의 99'대학기초연구지원사업의 지원과제 "*Web 상에서 적확한 검색을 위한 문서의 대표 개념어 생성 및 요약 시스템*"의 일부로 수행되었다.

2. 관련 연구

고유명사에 대한 영어권의 추출 연구가 활성화 된 것은 MUC(Message Understanding Conference)의 NE(Named Entity) 컨테스트[4]의 영향이 크다. MUC-6와 MUC-7의 NE 컨테스트는 정보추출(IE , Information Extraction)을 위한 한 단계로 명명 개체(Named Entity)를 이름(NAMEX - 조직, 사람, 장소), 시간(TIMEX, - 날짜와 시간), 수치(NUMEX, - 금액, 퍼센트, 등)요소로 나누어서 추출하고 분류하는 대회를 개최하였다. MUC의 NE 컨테스트에 참여한 시스템들은 크게 나누면 규칙 기반의 시스템[5]과 통계적인 방법의 시스템[6,7], 그리고 두 방법의 혼합 시스템으로 나누어 볼 수 있다. 규칙 기반 방법은 연구자가 수작업으로 작성한 추출 규칙이나 언어학적 규칙에 기반하고 있으며, 통계적인 시스템은 HMM이나 최대 엔트로피(ME)방법 등을 사용하여 고유명사 문제를 접근하고 있다. 규칙 기반 방법은, 일반적으로 텍스트의 도메인이나 범주의 항목이 변화되면 적응하기 힘든 단점이 있다고 지적되고 있다.[7]

국내에서는 고유명사 관련 연구는 다른 자연어 처리 문제의 일부로 연구되었다. 상당수의 고유명사가 미등록어이거나 미등록어를 포함하기 때문에 미등록어 처리를 위해 용례 분석을 통해 고유명사를 처리하는 연구[8], 문서 분류를 위해 문서에 등장하는 개체(entity)를 찾는 연구[9] 등이 있었지만, 이때 관심의 초점은 정확한 품사를 부착하는 것을 돕거나 문서 분류를 돕는 도구로서 되어 있기 때문에 고유명사 추출 및 분류와는 거리가 있다. 색인어로 사용하기 위해 고유명사중 인명과 기관명을 자동으로 추출하는 연구[10]가 있으나, 그 외에 정보 추출의 관점에서 고유명사를 대상으로 한 연구는 드물다. 본 논문에서는, MUC의 NE 문제 중에서 수치 요소나 시간 요소를 제외하고 고유명사만을 추출한 후 사람, 조직, 장소, 기타의 4가지 범주로 자동 분류한다.

3. 규칙 기반 학습

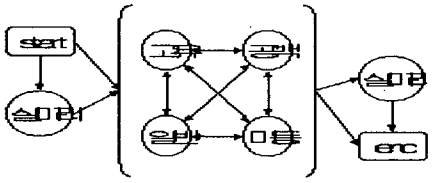
텍스트에 등장하는 고유명사를 인식하는 경우를 세 가지 경우로 나누어 볼 수 있다. 첫번째 경우는 고유명사를 이루는 어휘 내부에 실마리가 있어, 이를 통해 고유명사임과 그 대상의 성질을 알 수 있는 경우이다.(예; 세느강, 국립과학수사연구소, 성미전자, 김 위원장) 둘째 경우는 어휘 내부에는 실마리가 없고 주변 문맥에 그 미등록 고유명사의 성질을 알려주는 실마리가 있는 경우이다. (예; 러시아의 핵잠수함 '크루스크호'는... , 인터넷 중권사인 이트레이드는 ...) 세 번째 경우는 어휘 내부에도, 주변 문맥에도 고유명사에 대한 정보를 알려주는 실마리가 없는 경우로 따로 준비된 지식, 즉 사전을 통하지 않고서는 해결 할 수 없는 고유명사이다.(예; 경제기사에서 코스닥, 나스닥) 시스템이 학습을 위해 사용하는 코퍼스는 표2처럼 일반 문서에 고유명사의 앞뒤로 태그를 달아 고유명사임을 밝히고 그 범주를 기록한 문서들이다.

<표 2> 고유명사가 태깅된 코퍼스의 예

(<PN LOC>대전</PN>-<PN ORG>연합뉴스</PN>) <PN PSN>윤석이</PN> 기자 - <PN ORG>한국수자원공사 대청댐관리사무소</PN>는 집중호우로 인한 <PN LOC>금강</PN>유역의 홍수조절을 위해 지난 27일부터 실시되던 <PN LOC>대청댐</PN>의 수문방류를 31일 오전 10시를 기해 종료했다고 밝혔다. <PN ORG>대청댐관리소</PN>의 이 같은 조치는 ...
ORG=조직, PSN=사람, LOC=장소 및 지명, ETC=이에 속하지 않는 나머지

3.1 어휘내 실마리어 학습

학습 코퍼스에서 발견된 같은 범주에 속하는 고유명사들 중에 왼쪽 끝이나 오른쪽 끝의 동일한 위치에 반복적으로 나타나는 부분 문자열이 있다면, 이 문자열을 추출하여 실마리어로 삼는다. 이 실마리어가 나타날 때 결합된 명사 및 명사열이 그 범주의 고유명사인 확률을 학습 코퍼스로부터 통계적으로 계산하고 학습한다. 부분문자열이 실마리어로 지적되면 이를 포함하는 고유명사에서 실마리어를 떼어내고 나머지 부분들에 대해 양방향으로 사전 검색을 통해 최장 일치되는 명사로 분리한다. 실마리어가 결합한 복합명사/명사열을 그림1과 같이 결합한 오토마톤으로 생각하고 각 결합이 고유명사가 되는 통계치를 얻는다.



<그림 1> 실마리어와 결합하는 고유명사의 모델

이 모델은 코퍼스에서 동일 범주의 고유명사들 내부에 반복적으로 증권, 전자, 연구소, 경기장 같은 어휘가 부분 어휘로 자주 등장했다면, 이들이 나타날 때의 각 조합이 해당 범주의 고유명사일 확률이 얼마나 되는지 통계치를 축적하는 것을 의미한다. 특히, 학습코퍼스에서 이 실마리어가 결합했음에도 고유명사가 아니었던 경우는 부정어 학습으로 따로 기록한다. 이때 부정어란 실마리어가 높은 확률의 조합과 위치에 등장했음에도 불구하고 명사/명사구가 고유명사가 되지 않았을 때 그 원인을 실마리어와 결합한 특정어휘로 보고 정의한 것이다. 부정어와 통계치, 그리고 그 범주가 각 실마리어별로 따로 기록된다.

각 실마리어마다 부정어와 통계치를 따로 기록하는 이유는, 실마리어마다 고유명사의 구성에 기여할 수 있는 정도가 다르기 때문이다. 경제기사의 기업 이름에 자주 등장하는 실마리어 '산업'은 역시 기업에 자주 등장하는 실마리어 '투자신탁'에 비해 고유명사를 이루는 비율이 낮다. 특히 일반 명사와 결합했을 때에 그러하다. (표3) 이는 사람에게도 '미래 산업'과 같은 예는 전통 산업에 대한 반대 의미의 개념인지 특정 기업의 이름인지 문맥없이 판단하기 어렵다는 사실과 일치한다.

<표 3> 학습된 실마리어의 예

실마리어 37 산업$ (ORG)		실마리어 38 투자신탁$ (ORG)	
결합	고유명사 ORG인 비율	결합	고유명사 ORG인 비율
미등록* + [실마리]	0.9	미등록* + [실마리]	1.0
일반* + [실마리]	0.42	일반* + [실마리]	0.9
고유* + [실마리]	1.0	고유* + [실마리]	1.0
미등록*일반*+[실마리]	0.7	...	...
...	....		
부정어 : 전통, 기존, 디지털, ...		부정어 : (없음)	

*(*는 한 개 혹은 그 이상의 반복. $는 단위의 오른쪽 끝)*

3.2 문맥 실마리 학습

문맥 실마리 학습 단계는 학습코퍼스의 고유명사 중 실마리어를 포함하지 않는 고유명사를 대상으로 한다. 이런 고유명사는 문맥 없이 고유명사만을 기록해 놓으면 어느 범주에 속하는지 알 수 없는 미등록어/기등록어의 나열로, 문맥에서 수식언이나 용언의 결합을 통해 그 성격을 파악할 수 있는 어휘들이다. 이들 중에서, 모호성 없이 나타나는 수식어구나 술부의 패턴을 연구자가 미리 지정해 두고, 학습 코퍼스의 고유명사를 수식하거나 서술하는 문장이 이 패턴에 일치하면 수식어구/술부가 어떤 카테고리의 고유명사를 설명해 주는지를 학습하는 단계이다. 표4는 신문기사 도메인에서 자주 나타나는 패턴인 'A인 B'의 패턴으로부터 학습하는 예이다. 그림과 같이 이 패턴에 일치하는 B위치에 고유명사가 나타나면, A가 구체화된 개체는 B의 카테고리에 속한다는 것을 학습하게 된다.

위와 같이 학습할 문맥의 패턴은 정해져 있으며, 여기에 등장하는 학습 코퍼스의 고유명사로부터, 일반적인 개념(명사)의 한 구체적인 개체(고유명사)는 무엇(범주)에 속한다는 것을 학습하게 된다. 본 논문에서는 학습을 위하여 [A인B], [A B], [A는 B로서..], [A는 ... B이다]. 등의 14개의 패턴을 사용하였다.

<표 4> 'A인 B' 패턴의 문맥 실마리 학습의 예

... 주식형 *신탁상품*인 '<PN ETC>하이테크주가금전신탁</PN>'과... ... 금전*신탁상품*인 '<PN ETC>머니텍 1호</PN>'를 ...	신탁상품의 instance는 ETC에 속한다.
... 종합여신금융*회사*인 <PN ORG>현대캐피탈</PN>은 ... ... <PN ORG>이트레이드코리아</PN>의 모*회사*인 <PN ORG>이트레이드인터네셔널</PN>은 ...	회사의 instance는 ORG에 속한다.

3.3 사전 기제

어휘내 실마리어 학습과, 문맥 실마리 학습에서도 걸러지지 않은 고유명사에 대해서는 실마리 없이 사용된 고유명사로 파악, 사전 지식을 가정하고 쓴 고유명사로 보고 사전에 기록하여 관리한다. 문맥 실마리 학습이 의미적인 요소를 파악하지 못하기 때문에, 실지로 사람이 사전 지식으로 고려하지 않을 어휘들도 상당수 사전으로 기제된다

4. 학습 데이터의 적용 ; 자동 추출과 분류

고유명사 추출과 분류는, 일반 텍스트에 대해서 가능한 지식을 기반으로 차례로 태그를 부착하는 순서로 이루어진다. 입력으로 받은 일반 텍스트에 대해서 먼저, 고유명사의 후보열을 구하기 위해서 명사와 명사열을 찾는다. 형태소 분석 결과를 통해 각 어절이 용언 어절인지 체언

어절인지를 구분하고 체언 어절이 아닌 것을 후보에서 모두 제외한다. 체언 어절에서 조사를 떼어내고 남는 명사나 명사열들을 고유명사가 될 수 있는 후보로 삼고 각 단계를 적용한다.

- 사전 적용 단계
학습단계에서 실마리어나 문맥 실마리에 등장하지 않아 따로 사전에 기재된 내용과, 미리 유지되고 있는 고유명사 사전에 기재된 내용에 따라 먼저 태그를 붙인다.
- 어휘 내부 실마리 적용 단계
실마리어가 나타나면 학습한 통계치를 적용하여 그 실마리어를 포함한 어휘가 고유명사일 확률을 구한다. 높은 확률의 결합이어도 부정어일 경우에는 태그를 부착하지 않는다. 고유명사로 판별되면 그 범주는 실마리어에 부여된 범주를 따르게 된다.
- 문맥 실마리 적용 단계
각 명사/명사열을 대상으로 미리 지정된 문맥 패턴이 나타나면 패턴에 따라 고유명사로 태깅한다. 이때 그 범주는 학습된 개념 - 인스탄스 관계를 따른다.
- 정리 규칙 적용 단계
이 단계에서는 최종 결과를 출력하기에 앞서 마지막으로 학습단계와 관계없이, 연구자가 시스템에 미리 지정한 규칙을 적용하는 단계이다. 직접 고유명사를 추출하는 규칙은 없으며, 추출된 내용에 대한 정리 규칙들이다. 하나의 명사가 단계별로 서로 다른 범주로 분석되었을 때 어느 단계의 결과를 최종 결과로 삼을 것인지 결정하는 해소 규칙과, 같은 범주의 고유명사들이 이어지면 이를 하나의 고유명사로 연결하는 정리 규칙 등을 이 단계에서 적용한다. (예, *<LOC 대구시> <LOC 수성구> <LOC 사월동> ---> <LOC 대구시 수성구 사월동>*)

5. 실험

실험에 사용한 학습 코퍼스는 경제 부문 기사 100개에 고유명사 태그를 부착한 코퍼스로 총 어절 수는 19000여 어절이다. 이 코퍼스에 고유명사는 1479번 나타났으며 699개의 고유명사였다. 추출된 실마리어는 51개이고 이들을 통해 학습대상에 오른 어휘는 290개이다. 남은 409개의 고유명사 중에 259개가 문맥 실마리 학습에 기록되었으며, 나머지 고유명사는 문맥 실마리도, 어휘실마리도 등장하지 않아 따로 기록되었다. 실험은 시스템이 이전에 본 적이 없는 3100여 어절의 20개의 경제 부문 신문기사를 대상으로 하였다. 시스템은 Linux상에서 Perl을 사용하여 구현하였다. 체언 어절을 구분해내기 위한 형태소 분석기는 HAM[11]을 사용했으며, 복합명사를 분리하고 미등록 명사와 고유명사, 일반 명사를 구분하기 위해 필요한 2만여 단어의 일반명사와 고유명사 사전은 따로 구축하였다.

고유명사 추출 시스템 혹은 NE 시스템의 평가는 일반적으로 시스템의 출력과 사람의 수작업 태깅을 비교하여 정확률과 재현율, 그리고 이 두 수치를 하나로 표현하는 F_β 평가값(F_β measure)으로 평가한다. 특히 MUC의 NE대회는 $\beta=1$값으로 시스템의 등위를 매겼다. F값의 계산은 다음 식과 같다.

(1) $$F_\beta = \frac{(\beta^2+1)\times 정확률 \times 재현율}{\beta^2 \times 정확률 + 재현율}$$

영어권의 연구에서 사람이 수작업으로 태깅할 때 $F_{\beta=1}$값은 96정도인 것으로 보고되고 있다.[3] MUC-7의 F 평가값은 상위 시스템의 92~93에서 12위의 69사이의 다양한 값이 보고되었다.

실험대상에는 221번 고유명사가 나타났으며, 시스템이 추출한 185번 중 172개는 고유명사의 경계와 카테고리가 정확하게 추출되었다. 추출하지 못한 것이 40개, 경계와 카테고리 중 하나는 틀리는 부분일치가 9개, 완전히 잘못 추출한 것이 4개였다. 재현율은 79.8%, 정확률은 92.9%, F 값은 85.8을 보였다.

표 5에 고유명사 추출에 실패한 예를 들었다.

<표 5 고유명사 추출에 실패한 예들>

... 최근 반등 타이밍에서 선두에 섰던 로커스가 상한가를 넘나들며 상승폭이 컸고 다음과 <PN ORG>엔씨소프트</PN>도 상승폭이 컸다.
(찾지못함) 문맥의 '의미'로 결정되는 예(로커스). 기업이름이 일반명사 단일어절인 경우, 고유명사사전에 있어도 모호(다음)
<PN ORG>대우 구조조정협의회</PN>는.... . 또, 구조협은 ...
(찾지못함) 약어나 부분으로 줄여쓰는 경우. 실마리어 상실.
만기 1년인 <PN ORG>통안증권</PN>은 12.04%로 ...
(오류) 통안증권은 증권의 한 종류. 부정어 추가 학습으로 극복가능.

6. 결론 및 향후과제

본 논문에서는 학습 코퍼스로부터 규칙 기반으로 학습하여, 텍스트에서 고유명사를 추출하고 사람, 조직, 장소, 기타의 범주로 분류하는 방법을 제시하였다. 본 논문에서는 고유명사의 출현을 세 종류로 나누어 접근하였다. 각 경우마다 실마리어나 문맥의 어휘가 고유명사의 어떤 범주를 지정하는지 학습함으로서 고유명사를 찾고 그 범주를 결정하는 방법을 제시하였다.

문맥을 통해서 뜻이 드러나는 고유명사들이 가장 큰 문제가 된다. 이들에 대하여 더 다양한 실마리나 패턴을 찾을 필요가 있다.

실마리나 패턴이 어느 범주에 속하는지 연구자가 지정하지 않고, 고유명사가 태깅된 코퍼스로부터 학습하기 때문에, 본 논문의 방법은 도메인이나 분류의 범주가 변화해도 적용할 수 있으리라 본다. 이를 증명하기 위하여 도메인과 범주를 변화해가며 학습하고 실험할 필요가 있다.

7. 참고문헌

[1] Paul Thompson and Christopher C. Dozier. "Name searching and information retrieval" , Computation and Language, June 1997

[2] Min-Yen Kan and Kathleen R.McKeown, Columbia University Computer Science Technical Report, CUCS-030-99, 1999

[3] David D. Palmer and David S. Day , "A statistical profile of the Named-Entity Task" , Proceedings of the Fifth Conference on Applied Natural Language Processing. 190-193, 1997, ACL

[4] Chinchor, N. and Marsh, E. , "MUC-7 named entity task definition" , Proceeding of the Seventh Message Understanding Conference (MUC-7) , 1998

[5] Krupka, G.R. and Hausman, K. "IsoQuest : Description of the NetOwl(tm) extractor system as used in MUC-7", Proceeding of the Seventh Message Understanding Conference(MUC-7), 1998

[6] D.Bikel, S. Miller, R. Schwartz and R. Weischedel. "NYMBLE : a high-performance learning name-finder" Proceedings of the Fifth Conference on Applied Natural Language Processing. 1997 ACL

[7] Andrew Borthwick, "A Maximum Entropy Approach to Named Entity Recognition", Ph.D dissertation, New York University, 1999

[8] 박봉래, 황영숙, 임해창, "용례 분석에 기반한 미등록어의 인식" 정보과학회 논문지 제 25권, 제 2호, pp397-407, 1998

[9] 이경순, 최기선, "전문용어 및 정보추출에 기반한 문서분류시스템", 한글 및 한국어 정보처리 학술발표논문집, pp79-84, 1999

[10] 정래정, 김준태, "통계 정보와 어의 정보를 이용한 미등록 고유 명사의 자동 색인", 정보과학회 96년 가을 학술발표논문집 vol.23 no.2 1996

[11] 강승식, HAM v4.70c "한국어 형태소 분석기와 한국어 분석 모듈", http://ham.hansung.ac.kr/ham/ham.html