특허 개체명 인식에 대한 기계학습 사례

이태석○, 강승식
한국과학기술정보연구원, 국민대학교 컴퓨터공학부
tsyi@kisti.re.kr, sskang@kookmin.ac.kr

Named Entity Recognition for Patent Data by Machine Learning

Tae-Seok Lee○, Seung-Shik Kang
KISTI, Kookmin University

요 약

특허 분석에서 관심 있는 기술명, 서비스명, 제품명을 인식하도록 기계학습 기법을 사용해 개체명 인식기의 성능을 평가해 보았다. 개체인식을 위한 엔진은 스탠포드 대학의 NER과 CRF++을 사용하였다. 그 결과 F1값인 0.5612로 나타났다. 이것은 인명, 지역명, 조직명 개체를 인식하는 다른 연구에서 나타난 0.7857보다 0.2245 떨어지는 결과이다. 특허 개체명 인식에 대한 자질값 선정과 사전처리에 대한 더 많은 연구가 필요하다.

주제어: 특허 데이터, 개체명 인식, 기계학습, CRF++

1. 서론

기업은 새로운 수요를 창출하는 제품개발요구에 직면함에 따라 지금까지 이어온 대량생산과 규모의 경제를 통한 경쟁력으로는 더 이상 장기간 생존이 힘들어지게 되었다. 이러한 환경에서 다양한 지식과 고도화된 정보화 기술의 활용이 가속화되고 있다. 또한, 빠르고 정확한 지식공유 기반이 구축되고 우수지식 콘텐트 보유량이 증가되어 빅데이터의 이용이라는 새로운 트렌드로 변화하는 현 시점은 지식재산권의 보유와 분석을 통한 전략적 활용에 대한 관심과 투자가 필요하게 되었다.

특허의 경제적 가치는 특허권자가 특허발명에 관한 라이선스를 행사하여 이를 사용하고자 하는 제3자로부터 기술사용료를 받는 것이다. 라이선스 관계는 국내뿐 아니라 국제적 차원에서 기술의 공여 및 기술도입계약을 통해서 이루어진다. 그리고 특허권자는 권리침해로 인한 손해배상을 청구할 수 있다.

기업은 자사의 제품과 관련기술에 대한 특허를 분석함으로써 적절한 기술로드맵을 만들고 기업의 미래 경쟁력을 준비하는데 적극적인 투자를 하고 있다. 전통적인 특허 계량분석뿐만 아니라 비정형화 된 텍스트 마이닝 기술과 기계학습을 통한 특허분석으로 그 영역이 넓혀지고 있다[1]. 본 논문은 자연어처리에서 연구되어온 개체명 인식기를 특허 자료에 접목하여 제한된 특허맵 분석을 자동화하고 다양화 할 수 있는 특허 기술 및 제품에 대한 개체명 인식을 시도하고 그 성능을 평가하고자 한다.

2. 관련 연구

언어처리 및 개체명 인식은 경험적 지식에 의존하는 규칙기반 기법과 대량의 수집자료를 기반으로하는 기계학습 방법을 사용하고 있다[2]. 규칙기반 기법은 전문가의 통찰력을 가지고 규칙을 찾아 경험에 의존하여 만드는 방법으로 자동화 할 수 없는 단점이 있지만, 경험적인 규칙의 반영으로 그만큼 좋은 결과를 얻을 수 있다. 반면, 기계학습은 양질의 학습 데이터가 필요하며, 통계적 기반의 학습 처리 기법을 사용한다. 대량으로 빠르게 처리할 수 있는 반면, 양질의 학습 데이터를 만들기가 어려우며 규칙기반에 비해 정확도가 떨어진다. 최근의 언어처리는 인터넷 데이터를 기반으로 많은 양의 자료를 기반으로 통계적인 처리 방법인 기계학습의 정확도가 매우 높은 수준까지 와 있다[3].

기계학습은 학습과정과 성능 관리하는 방법에 따라 Supervised machine Learning, Semi-Supervised Learning, Unsupervised Learning으로 나누어진다. Supervised machine Learning은 학습과정에서 작업자가 계속 학습 데이터를 만들어 성능 관리를 지속적으로 해주어야 한다. 기계학습 알고리즘으로는 Hidden Markov Models, Decision Trees, Maximum Entropy models, Support Vector Machines, Conditional Random Fields 등이 있다[3]. 최근에는 보다 많은 자질을 사용하는 Conditional Random Fields 가 많이 쓰이고 있다.

Semi-Supervised Learning은 학습 과정과 성능 관리가 부분적으로 자동화되는 기계학습 기법이다. 초기에는 Supervised machine Learning과 마찬가지로 작업자가 학습데이터 및 학습 작업에 관여하지만, 이후 추가되는 학습 작업은 결과를 재사용하는 방식으로 학습된 지식을 확장해 가는 bootstrapping 과 같은 방식으로 2개 시스템에서 한쪽 결과를 다른 쪽 모델로 사용하여 교대로 성능을 높여나가는 방식이다[4]. 하지만, 어느 정도 횟수가 증가할수록 한계가 있다.

Unsupervised Learning은 대량의 자료를 자동적으로 군집화 시켜주는 방식으로 작업자는 그 결과를 해석하고 정리하는 작업을 주로 한다. 사용되는 알고리즘은 클러스터링 기법으로 LSA(Latent semantic analysis), LDA(Latent Dirichlet Allocation)를 사용한다[5,6,7].

실제 개체명 인식을 위한 과정은 개체명 추출을 위한 정보 구축 단계와 개체명 추출 처리 과정으로 나누어진

다. 개체명 추출을 위한 정보 구축 단계에서는 좌우 문맥에 대한 품사정보, 문자 유형 정보를 토대로 자질값을 선정하고 개체명 사전과 결합 단어 사전을 구축한다. 개체명 추출 처리 과정은 우선 후보군을 추출하고 후보 개체명 주변정보를 이용하여 보통명사와 분리한다. 문장 구조에 따라 개체명을 제한하는 용언을 고려하여 대상 집합을 줄이고 종합하여 어절간의 관계를 통해 개체명의 범주와 범위를 결정한다.

본 논문에서는 Conditional Random Fields 알고리즘을 사용하는 스탠포드 NER을 사용하여 특허 데이터를 학습하고 학습 모델을 테스트하고 평가하였다[8]. 스탠포드 NER은 오픈소스로서 자질값 확장으로 간단하고 다양하게 처리 가능하며, 텍스트에 있는 단어의 개체 명 인식(NER) 레이블 시퀀스를 이용한다. 특징으로 개체 명 인식을 위해 잘 설계된 특징 추출과 결합 선형 체인 조건부 랜덤 필드 (CRF) 시퀀스 모델 사용한다. 적용 예는 사람, 조직, 위치, 시간, 돈, 비율, 날짜, 기타 등 개체명 인식에 사용된다. 배포되는 자료에는 CoNLL 2003 영어 학습 데이터에 대한 교육 모델 포함되어 있으면 자바 언어로 구현되어 있다.

3. 개체명 인식을 위한 자료 구축

한국과학기술정보연구원에서 수행한 미국등록 특허 2,400건을 대상으로 수작업 처리한 결과물인 Gold Standard 총 919,254건을 사용하였다. 미국등록 특허 2,400건은 IPC A-H 분류별 2-3건씩 고루 포함된 자료이다. 작업은 7명이 작업하였으며 최종 작업결과에 대한 의견 일치도 (Kapa Score)는 0.6 정도로 보통 수준이다. 문장은 청구항, 디스크립션, 초록, 제목에 대해서 추출하였다. Gold Standard 데이터를 가지고 그림 1과 같이 작업을 하였다.

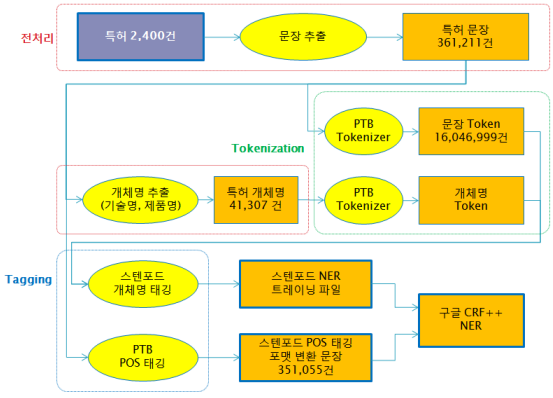

그림 1. 데이터 작업 흐름도

작업자는 그림 2의 태깅 작업도구를 사용하여 작업을 수행하고 그 결과 데이터를 변환하여 NER 입력 데이터로 사용하였다. 특허 2,400건의 작업 결과는 특허번호, IPC 분류, 문장의 추출 섹션, 문장번호 순으로 개체명 태그와 품사 태그, 관계 태그로 이루어져 있다. 개체명과 개체명 사이의 관계 패턴과 개체명 앞과 뒤의 단어에 대한 정보도 함께 저장하고 있다. 상세한 내용은 그림 3과 같다.

그림 2. 태깅 작업도구(DQtagger)

1열 Instance_id 총 919,254 개
2열 Entity1 (예, lamp box, power source 등) 식별된 제품명
3열 Begin(En1) 시작 숫자 값
4열 End(En1) 끝 숫자 값
5열 PoS of Entity1 품사 태그 스페이스 분리로 여러 개 표시 (예, NN, NNS)
6열 Entity2 (예, lamp box, power source 등) 식별된 제품명
7열 Begin(En2)
8열 End(En2)
9열 PoS of Entity2 품사 태그 2개가 옴 NN, NN
10열 Entity1_type [Similar_Product | Product] 식별 제품개체의 수준
11열 Entity2_type [Similar_Product | Product] 식별 제품개체의 수준
12열 Relation_type 두개의 개체사이의 관계, [NONE | partof]
13열 Patent_id 미국특허 등록번호 6355283 ??? 발행연도가 없어 식별이 어려울 수 있음.
14열 Patent_section (A,B,C,D,...) // 특허 IPC 분류 코드 : A21D/013-00
15열 Patent_class
16열 Patent_subclass
17열 Patent_maingroup
18열 Patent_subgroup
19열 Section (title, abstract, claim, description) 특허 섹션 제목, 초록, 청구항, 디스크립션 구분
20열 Sentence_id 문장 번호
21열 Sentence 문장 내용
22열 Sen_PoS (문장내 모든 단어의 품사열)
23열 Pattern (두 개체 사이의 패턴) [Similar_Product CONTAINING A NP OF Similar_Product]
24열 Pattern_PoS (두 개체 사이의 단어 품사열) [Similar_Product VBG DT NN IN Similar_Product]
25열 개체1의 바로 앞 단어 Lexical
26열 개체1의 바로 앞 단어 PoS
27열 개체1의 바로 뒤 단어 Lexical
28열 개체1의 바로 뒤 단어 PoS
29열 개체2의 바로 앞 단어 Lexical
30열 개체2의 바로 앞 단어 PoS
31열 개체2의 바로 뒤 단어 Lexical
32열 개체2의 바로 뒤 단어 PoS
33열 두 개체 사이의 관계 패턴, 대부분 ~ing, ~ed 형태 모두 나열 구분자 스페이스

그림 3. 데이터 작업 결과물 형식

```
The             O
method          O
for             O
making          O
such            O
improved        O
liquid          Product
binder          Product
for             O
pre-application         O
combination     Similar_Product
with            O
cosmetic        Similar_Product
powders         Similar_Product
,               O
eye             Product
shadows         Product
and             O
eyebrow         Similar_Product
makeup          Similar_Product
```

그림 4. Stanford NER 학습 데이터 형식

스탠포드 NER에 필요한 데이터 처리는 펜실베이니아 대학의 PTB 토큰처리기를 사용하였다[8,9]. 스탠포드 NER의 학습데이터 형식은 그림 4와 같이 PTB 토큰 워드와 NER 태그로 이루어져있다. 작업자는 Product(제품), Similar Product(제품후보), Technology(기술), Service(서비스), unknown(확인필요)와 같이 5개의 개체명을 태깅하였다.

표 1. 개체명 토큰수

개체명	토큰 수	비율
Service	2,648	0.02%
unknown	34,379	0.21%
Technology	411,184	2.56%
Product	942,744	5.87%
Similar_Product	1,801,056	11.22%
O	12,854,988	80.11%
총합	16,046,999	100%

개체명 인식의 자질값은 한 어절부터 연속 세 개의 어절 까지의 자질을 각각 테스트하여 F1 점수가 높은 것들만 선택하여 표 1 과 같이 최종 우수 자질 집합을 얻었다.

표 2. 최종 우수 자질 집합

상태 자질(y_i)	전이 자질(y_{i-1})
x_i x_{i-1}/x_i, x_i/x_{i+1} $x_{i-2}/x_{i-1}/x_i$, $x_{i-1}/x_i/x_{i+1}$	x_{i-2}, x_{i-1}, x_i, x_{i+1} x_{i-3}/x_{i-2}, x_{i-2}/x_{i-1}, x_{i-1}/x_i, x_i/x_{i+1}, x_{i+1}/x_{i+2} $x_{i-3}/x_{i-2}/x_{i-1}$, $x_{i-2}/x_{i-1}/x_i$, $x_{i-1}/x_i/x_{i+1}$, $x_i/x_{i+1}/x_{i+2}$

4. 실험 및 평가

특허 문서에 대해 작업자가 실시한 태깅 결과를 가지고 개체인식 학습 데이터와 실험데이터로 나누어 실험하였다. 이때 그림 5에서와 같이 고빈도 개체명 길이는 주로 5-23 문자로 이루어진 것이 81.58%에 해당한다. 따라서 실험에 사용한 대상은 길이가 5-23 문자로 이루어진 개체명을 대상으로 하였다. 길이 제한으로 짧은 약어와

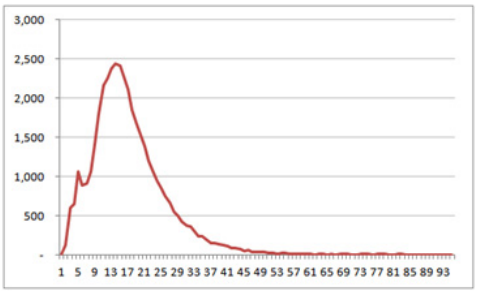

그림 5. 개체명 길이와 빈도수 그래프

긴 화학식은 대부분 제외 되었다. 학습용 데이터와 테스트용 데이터를 각 3천개씩 추출하여 스탠포드 NER로 학습하였다. 학습하는데 걸린 시간은 4.5분이 걸렸고 테스트하는 데는 17초가 걸렸다. F1 값은 0.5612로 나타났다.

CRFClassifier tagged 100000 words in 1 documents at 12229.42 words per second.

Entity	P	R	F1	TP	FP	FN
Product	0.5346	0.4059	0.4614	1539	1340	2253
Similar_Product	0.6847	0.5757	0.6255	3814	1756	2811
Technology	0.6690	0.4763	0.5565	1510	747	1660
unknown	0.3254	0.2989	0.3116	55	114	129
Totals	0.6361	0.5020	0.5612	6918	3957	6863

이것은 의료 분야의 사례와 비교하였을 때 많이 떨어지는 결과이다. 의료 분야의 임상기록에대한 개체명 인식 연구 결과는 본 논문에서 사용한 CRF 기법을 동일하게 사용하여 F1 점수가 81.48로 높게 나타났다.[10]

5. 결론

특허 문서의 개체명(기술명, 제품명, 서비스명) 인식 결과가 좋지 않은 것은 인식에 적합한 문형 자질, 단어 수준 자질, 사전조사 자질을 활용해야 할 것으로 생각된다. 본 연구에서 사용한 어절 중심의 자질만으로는 개체명을 인식하는 데 한계가 있음을 알게 되었다.

개선 방안으로 특허의 IPC 분야 (A~H)를 구분하여 각 분야에 적합한 성장형 학습 모델을 활용할 수 있으며, 특허 문서에 출현하는 패턴중심의 개체 속성 사전을 구축하여 화학식이나 약어에 대한 중의성을 해소하는 것도 좋은 방법이다. 특허 객체명 인식에서도 품사정보를 추가 자질로 사용하면 좋은 결과를 얻을 수 있을 것으로 기대한다.

특허 문서에 대한 분석을 통해 좀 더 좋은 자질을 발굴하여 정확률을 많이 떨어뜨리지 않으면서 재현율을 높이는 방법으로 핵심 feature의 적절한 선택과 좋은 트레이닝 데이터 생산을 통해 제품명, 서비스명, 기술명 인식 성능을 높여나갈 수 있으리라 기대한다.

참고문헌

[1] Gurulingappa Harsha, Mueller Bernd, Klinger Roman, Mevissen Heinz-Theodor, Hofmann-Apitius Martin, Fluck Juliane, and Friedrich Christoph M, Patent Retrieval in Chemistry based on Semantically Tagged Named Entities. FRAUNHOFER INST SANKT AUGUSTIN (GERMANY) ALGORITHMS AND SCIENTIFIC COMPUTING SCAI, pp.1-9, 2009.

[2] Silviu Cucerzan, and David Yarowsky, Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence, Proceedings of the 1999 Joint SIGDAT, pp.90-99, 1999.

[3] David Nadeau, and Satoshi Sekine, A survey of named entity recognition and classification, Lingvisticae Investigationes, Volume 30, Number 1, pp.3-26, 2007.

[4] Joseph Polifroni, Imre Kiss, and Mark Adler, Bootstrapping Named Entity Extraction for the Creation of Mobile Services,IREC 2010, pp.

1515-1520, 2010.
[5] Dr. Edel Garcia, Latent Semantic Indexing (LSI) A Fast Track Tutorial, 2006.
[6] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman, Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science. Volume 41, Issue 6, pp.391-407, 1990.
[7] Latent Dirichlet Allocation, Journal of Machine Learnig Research 3, pp.993-1022, 2003.
[8] http://nlp.stanford.edu/software/crf-faq.shtml
[9] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini, Building a large annotated corpus of English the Penn Treebank, Journal Computational Linguistics - Special issue on using large corpora: II, Volume 19 Issue 2, pp.313-330, 1993.
[10] Yefeng Wang, Annotating and Recognising Named Entities in Clinical Notes, ACLstudent '09 Proceedings of the ACL-IJCNLP 2009 Student Research Workshop, pp.18-26, 2009