Suffix Tree와 Distant Supervision을 이용한 관계 추출

이현구○, 최맹식, 김학수
강원대학교 컴퓨터정보통신공학과
nlphglee@kangwon.ac.kr, nlpmschoi@kangwon.ac.kr, nlpdrkim@kangwon.ac.kr

Relation Extraction Using Suffix Tree and Distant Supervision

HyunGoo Lee○, Maengsik Choi, Harksoo Kim
Kangwon National University Computer and Communication Engineering

요 약

자연어처리 분야에서 관계 추출은 중요한 연구 분야이다. 많은 관계 추출 연구는 지도 학습 방법을 사용하지만 정답을 구축하는 비용이 큰 문제가 있다. 본 논문에서는 distant supervision을 이용하여 데이터를 구축하고, suffix tree를 이용한 규칙기반 관계 추출 모델을 제안한다. Suffix tree를 이용한 관계추출의 Macro F1-measure는 84.05%로 관계 추출에서 사용이 가능함을 보였다.

주제어: 관계 추출, Suffix Tree, Distant supervision

1. 서론

관계 추출(Relation Extraction)은 문장 내에 두 개체명 사이의 관계를 찾는 것이다. 요즘같이 웹이 발달한 시대에서는 위키피디아(Wikipedia), 트위터(Twitter), 전자신문 등 자연어 문서에서 많은 양의 관계 정보를 추출할 수 있다. 이러한 관계 정보를 자동으로 추출한다면 질의응답 시스템(Question Answering System) 등 여러 자연어처리 분야에서 효율적으로 사용될 수 있다[1]. 하지만 관계 추출은 주로 제한된 말뭉치를 이용한 트리 커널 기반의 지도 학습 방법(Supervised Learning)으로 연구되어 학습 데이터가 적고 말뭉치 도메인에 종속적인 문제가 있다. 이러한 문제를 완화하기 위해 자동으로 말뭉치를 생성하는 distant supervision[2]을 이용한 연구가 진행되었다. Distant supervision은 많은 데이터를 자동으로 생성하여 말뭉치 제한을 해소한다. 따라서 도메인 종속성 문제를 해결하고 데이터 구축에 드는 비용을 절감할 수 있다.

본 논문에서는 디비피디아 온톨로지(DBPedia ontology)를 이용한 distant supervision으로 레이블 데이터를 생성하고, 생성된 레이블 데이터와 suffix tree를 이용한 규칙기반 관계 추출 모델을 제안한다. 본 논문의 구성은 다음과 같다. 먼저, 2장에서 관련 연구에 대해 알아보고, 3장에서 suffix tree를 이용한 규칙기반 관계 추출 모델을 제안한다. 4장에서 실험 및 평가를 하고, 5장에서 결론을 맺는다.

2. 관련 연구

기존의 많은 관계 추출 연구는 ACE(Automatic Content Extraction) 말뭉치[3]를 이용한 트리 커널 기반의 지도 학습 방법이 사용되었다[4,5,6,7]. 하지만 적은 학습 데이터와 도메인 종속성 문제로 인해 최근에는 distant supervision을 이용하여 많은 레이블 데이터를 생성하고, 생성된 데이터들을 기반으로 지도 학습 방법을 사용하여 복잡한 알고리즘이 아닌 빠르고 간단한 규칙 및 자질 기반의 방법들이 활발히 연구되고 있다[8,9]. 또한 suffix tree로 생성된 규칙을 이용하는 질의응답 시스템 연구도 진행이 되었다[10].

본 연구에서는 디비피디아 온톨로지를 이용한 distant supervision으로 많은 양의 레이블 데이터를 생성하고, suffix tree를 이용하는 규칙 기반 관계 추출 모델을 제안한다.

3. Suffix Tree를 이용한 관계 추출

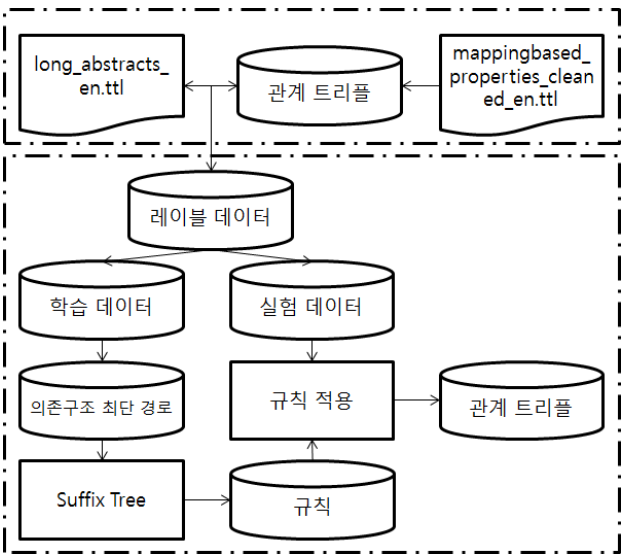

그림 1. 시스템 구성도

제안 시스템은 [그림 1]과 같다. 첫째 지식베이스에서 distant supervision으로 레이블 데이터를 추출하는 부분, 둘째 추출된 데이터로부터 규칙을 생성하는 부분, 마지막으로 추출된 규칙을 적용하여 관계 트리플을 추출하는 세 가지 부분으로 나누어진다.

3.1 Distant Supervision으로 레이블 데이터 생성

본 논문에서는 디비피디아 온톨로지를 이용하여 데이터를 생성한다. 디비피디아 온톨로지는 위키피디아에 의미적 주석을 달아두어 제공해주는 온톨로지이다. 위키피디아 인포박스의 내용이 가공되어 있는 'mappingbased_properties_cleaned_en.ttl'을 이용하여 주어, 술어, 목적어 구조(Subject, Predicate, Object)를 가진 관계 트리플을 추출하고 위키피디아의 문서내용 'long_abstracts_en.ttl'에서 Subject와 Object가 모두 포함된 문장을 추출한다.

3.2. Suffix Tree를 이용한 규칙 추출

규칙 추출을 위해 의존구조 최단 경로(Shortest Path)[11]를 이용한다. 문장에 비해 최단 경로는 술어에 연관된 단어들로 구성되어 있다. 즉 술어와 관계되지 않은 단어들을 제거하여 질이 좋은 규칙을 생성하게 된다. Suffix tree[12]는 문자열의 부분 문자열을 저장하는 트리이다. [그림 2]와 같이 "I like apples."은 "apples.", "like apples.", "I like apples."의 임시 문자열을 생성하여 suffix Tree를 생성한다.

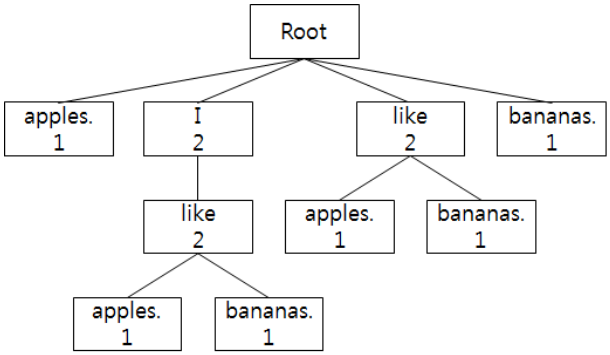

그림 2. Suffix tree의 예

규칙을 추출하기 전 suffix tree를 생성하는 방법은 다음과 같다. 본 연구에서는 최단 경로를 구하기 위해 Apache OpenNLP[13]를 사용한다.

1) OpenNLP의 문장 분리 툴인 SentenceDetectorME를 사용하여 문장을 분리한다.
2) OpenNLP의 구문 분석 툴인 Parser의 결과로부터 head word를 통해 의존 구조로 변경한다.
3) 관계 트리플의 주어와 목적어 사이의 최단 경로를 구한다.
4) 최단 경로에서 주어와 목적어를 클래스 정보로 바꾸어준다. 이 때 클래스 정보는 디비피디아 온톨로지 'instance_types_en.ttl'를 이용한다.
5) 최단 경로에 주어와 목적어의 클래스를 제외한 문자열만 술어 suffix tree에 저장하고, 주어와 목적어 클래스 사이에서 나타난다고 표시를 한다.
6) 만약 중복된 내용이 저장되면 suffix tree의 노드에 빈도를 1증가시킨다.

표 1. 생성된 레이블 데이터의 예

	값	
개체명	Lawrence Frankopan	London
클래스	Agent|Person	Place|PopulatedPlace|Settlement|City
레이블 데이터	Lawrence Frankopan, born 29 December 1977, is a sports agent and businessman in London.	

[표 1]은 distant supervsion으로 생성된 레이블 데이터의 예이다. [표 1]로부터 [그림 3]의 최단 경로를 생성한다. [그림 3]에서 생성된 최단경로 "Lawrence Fankopan is businessman in London"에서 "is businessman in"만 술어 suffix tree에 저장한다.

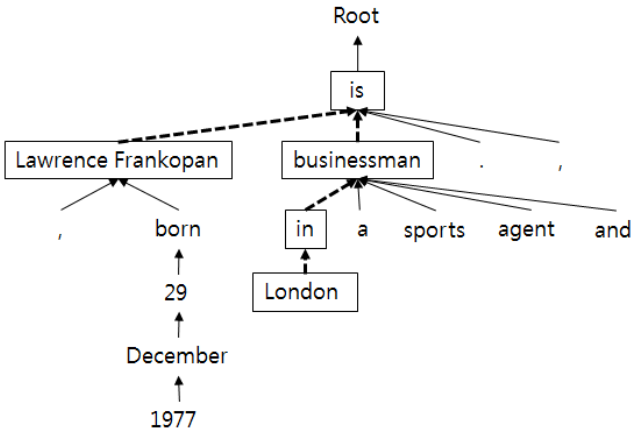

그림 3. 의존구조 트리와 최단 경로 추출의 예

생성된 suffix tree에서 특정 빈도이상인 것을 규칙으로 뽑는다. 예로 [그림 2]의 suffix tree에서 빈도 2를 기준으로 추출되는 규칙은 "I like", "like"가 된다. 그리고 뽑힌 규칙 중 불용어[14]로만 이루어진 규칙은 제거한다.

3.3 규칙을 이용한 관계 트리플 추출

추출된 규칙을 적용하기 위해 각 규칙에 점수를 부여한다. 점수는 어떤 규칙이 각 술어에서 얼마나 더 중요한 규칙인지를 확인하는 용도로 쓰인다. 즉 다른 술어에서 적게 나타날수록 점수가 높게 부여된다. 각 술어에 대한 규칙의 점수는 수식 (1)로 계산된다.

$$Score_P(pred) = \frac{\frac{Count_{pred}(P)}{Count_{total}(P)} \times \log_2(Count_{pred}(P)+1)}{MaxScore} \quad (1)$$

수식 (1)에서 $Count_{total}(P)$은 규칙 P가 모든 술어에서 나타난 수, $Count_{pred}(P)$은 해당 술어 pred에서 규칙 P가 나타난 수이다. 마지막 $MaxScore$는 모든 술어의 모든 규칙 점수 중 가장 큰 값으로 모든 점수를 나누어주어 0 ~ 1사이의 값으로 정규화를 한다.

규칙적용을 통한 관계 트리플 추출과정은 다음과 같다.

1) 입력 데이터의 주어와 목적어가 포함된 레이블 데이터를 찾는다.
2) 주어와 목적어 사이의 최단 경로를 구한다.
3) 주어와 목적어의 클래스 정보를 찾는다.
4) 각 술어별 규칙 목록 중 주어와 목적어 클래스 사이에서 나타날 수 있는 규칙만 최단 경로문장에 포함되는지 검사한다.
5) 수식 (2)를 이용하여 가장 높은 누적점수를 가지고 있는 술어를 반환한다. $N(D_{pred})$는 학습 데이터에서 술어 pred의 수이다. 술어 데이터 수의 차이를 정규화 해주기 위해 $N(D_{pred})$로 나눠준다.

$$Pred(E_1,E_2) = \arg max_{pred_i \in Pred} \frac{\sum_P Score_P(Pred_i)}{N(D_{pred})} \quad (2)$$

6) 만약 포함되는 규칙이 하나도 없다면 클래스 계층을 한 단계 올려 4번부터 다시 진행한다. 예를 들어 클래스 'Agent|Person'에서 포함되는 규칙이 없으면 상위 클래스 'Agent'로 다시 진행한다. 즉 'Agent'의 하위 클래스 'Agent|Person', 'Agent|Family' 등을 모두 포함하여 다시 진행한다.

4. 실험

4.1 실험 준비

본 논문에서는 레이블 데이터를 생성하기 위해 디비피디아 온톨로지 3.9[15]를 사용하였다. 술어는 distant supervision으로 질이 좋은 문장이 추출되는 디비피디아 템플릿 5개를 사용하였다. 'ActiveYearsStartYear'과 'ActiveYearsEndYear'는 활동의 시작년도, 활동의 마지막년도, 'BirthPlace'는 태어난 장소, 'Nationality'는 국적, 'Award'는 수상을 나타내는 술어이다. [표 2]는 술어의 분포이다.

표 2. 데이터 별 술어 분포도

술어	학습 데이터	실험 데이터	골드 데이터
ActiveYearsStartYear	8,913	853	41
ActiveYearsEndYear	9,045	914	47
Award	1,627	125	4
BirthPlace	53,100	5,065	198
Nationality	8,754	845	126
Total	81,439	7,802	416

[표 2]에서 학습 데이터와 실험 데이터는 distant supervision으로 생성된 데이터 집합, 골드 데이터는 실험 데이터에서 임의의 문장을 뽑아 사람이 직접 정답을 부착한 것이다. 예를 들어 distant supervision으로 추출된 'Ruth Vollmer'와 'Munich'는 'Nationality'의 관계가 있지만 레이블 데이터의 문장은 "Ruth Vollmer (1903 - 1982 New York), was a German artist born in Munich."로 'Nationality'가 아닌 'BirthPlace'의 정보가 담겨져 있다. 이러한 오류를 수정한 것이 골드 데이터이다.

4.2. 실험 결과

표 3. 실험 결과

	실험 데이터	골드 데이터
Accuracy	0.8334	0.7644
Macro precision	0.7922	0.8816
Macro recall	0.7931	0.8031
Micro precision	0.8432	0.7737
Micro recall	0.8334	0.7644
Macro F1-measure	0.7927	0.8405

[표 3]은 실험 결과이다. 골드 데이터의 accuracy는 실험 데이터보다 낮지만 macro precision과 macro recall이 더 높게 나왔다. Macro precision과 macro recall은 각 술어별 precision과 recall의 평균을 나타난 것이다. 즉 골드 데이터의 일부 술어가 성능이 매우 높게 나와 macro precision과 macro recall이 높게 측정된 것이다.

5. 결론

본 논문에서는 suffix tree를 이용한 관계 추출을 제안하였다. Distant supervision으로 레이블 데이터를 생성하여 학습한 실험의 accuracy는 76.44%, Macro F1-measure는 84.05%로 괜찮은 성능을 보여 suffix tree가 관계 추출에서 사용될 수 있음을 확인하였다. 향후 distant supervision에서 발생되는 오류를 줄이는 방법을 적용하여 성능을 향상시킬 계획이다. 또한 '관계 외 술어'를 인식할 수 있도록 전처리 모델을 적용하여 발전시킬 계획이다.

감사의 글

본 연구는 LG전자 산학연구용역 과제의 지원을 받아 수행되었음. 또한 2014년도 강원대학교 학술연구조성비로 연구하였음(과제번호-C1010876-01-01)

참고문헌

[1] J. Cowie, W. Lehnert, "Information extraction", communication of the ACM, vol.39, no.1, pp.80-91, 1996.
[2] Mintz, Mike, et al, "Distant supervision for relation extraction without labeled data", In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, vol.2, pp.1003-1011, 2009.
[3] The NIST ACE evaluation website, http://www.nist.gov/speech/tests/ace .

[4] Culotta, Aron, and Jeffrey Sorensen, "Dependency tree kernels for relation extraction", In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Article No. 423, 2004.
[5] Bunescu, Razvan C., and Raymond J. Mooney, "A shortest path dependency kernel for relation extraction", In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp.724-731, 2005.
[6] Zhang, Min, Jie Zhang, and Jian Su, "Exploring syntactic features for relation extraction using a convolution tree kernel", In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pp.288-295, 2006.
[7] Zhou, G., Zhang, M., Ji, D. H., and Zhu, Q, "Tree kernel-based relation extraction with context-sensitive structured parse tree information", In Proceedings of EMNLP-CoNLL, pp.728-736, 2007.
[8] Tseng, Yuen-Hsien, et al, "Chinese Open Relation Extraction for Knowledge Acquisition", In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp.12-16, 2014.
[9] Chen, Yanping, Qinghua Zheng, and Wei Zhang, "Omni-word Feature and Soft Constraint for Chinese Relation Extraction", In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp.572-581, 2014.
[10] Deepak Ravichandran and Eduard Hovy, "Learning Surface Text Patterns for a Question Answering System", In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp.41-47, 2002.
[11] Razvan C. Bunescu and Raymond J. Mooney, "A Shortest Path Dependency Kernel for Relation Extraction", In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp.724-731, 2005.
[12] Gusfield, "Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology. Chapter 6: Linear Time construction of Suffix trees", pp.94-121, 1997.
[13] Apache OpenNLP, https://opennlp.apache.org/ .
[14] Ingo Feinerer,"A text mining framework in R and its applications", pp.156-157. 2008.
[15] DBPedia ontology, http://wiki.dbpedia.org/Downloads39/ .