Predicate-Argument Structure 기반의 어휘적 패턴을 이용한 관계 추출

정창후*, 전홍우*, 최윤수*, 최성필*
*한국과학기술정보연구원
e-mail:chjeong@kisti.re.kr

Relation Extraction using Lexical Patterns based on Predicate-Argument Structure

Chang-Hoo Jeong*, Hong-Woo Jhun*, Yun-Soo Choi*, Sung-Pil Choi*
*Korea Institute of Science and Technology Information

요 약

문서 내에 존재하는 개체들 간의 관계를 자동으로 추출할 때 다양한 형태의 문서 분석 결과를 활용할 수 있는데, 본 논문에서는 문장 내에 존재하는 각 단어의 predicate-argument 관계를 분석하여 자질로 활용하는 PAS 패턴 기반 관계 추출 시스템을 제안한다. 관계 종류별로 구축된 PAS 패턴 집합을 활용하여 관계 식별기를 개발하였고, 실험을 통하여 개발된 관계 식별기의 성능을 측정하였다. 실험 결과 개체 간의 유의미한 관계를 표현해주는 PAS 패턴이 관계 추출 작업에 유용한 정보임을 알 수 있었다.

1. 서론

문서 내에 존재하는 중요한 개체 간의 관계를 자동으로 추출하는 작업은 정보 추출 작업 중에서 핵심적인 작업으로 꼽히면서도 가장 어려운 작업으로 알려져 있다. 이러한 개체 간의 관계를 추출할 때 문서 내에 존재하는 다양한 자질을 활용할 수 있는데, 본 논문에서는 이러한 다양한 자질 중에서 PAS(Predicate-Argument Structure) 기반의 어휘적 패턴을 이용한 관계 추출 방법을 제안한다.

2. 기존 연구

PAS는 문장을 구성하는 각 단어에 대한 predicate-argument 관계를 이용하여 문장 내에 존재하는 각 단어 간의 유의미한 관계를 표현하고 있는 구조이다[1, 2]. CFG[3]를 사용하는 전통적인 파서와 달리 HPSG[4]를 사용하는 Enju[1] 파서는 효과적으로 문장의 구문적/의미적 구조를 분석하여 predicate-argument 관계를 제공한다. 따라서 사용자는 파싱 결과로부터 직접적으로 문장에 있는 단어들 사이의 의미적 연관관계를 파악할 수 있다. 결과적으로 문장의 의미가 핵심 역할을 수행하는 정보추출, 자동요약, 질의응답과 같은 고수준 자연어 처리 애플리케이션에서 PAS 패턴은 유용하게 사용될 수 있다.

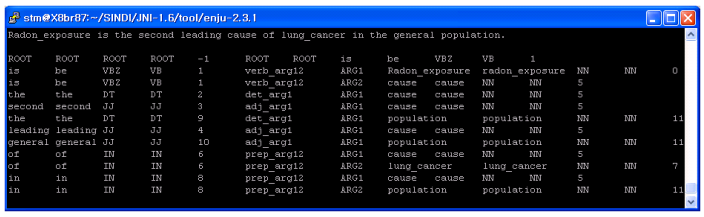
본 논문에서는 문장 내에 존재하는 개체들 사이의 관계를 유의미하게 표현해주는 PAS 패턴을 활용하여 관계를 예측하는 방법을 제안한다.

3. PAS 패턴의 분류를 이용한 관계 예측

PAS는 predicate-argument 관계를 이용하여 문장 내에 존재하는 각 단어 간의 유의미한 연관관계를 표현하는 구조이다. 그리고 PAS 패턴은 문장을 구성하는 모든 단어에 대한 predicate-argument 관계 그래프에서, 중요하게 지정된 개체와 개체를 연결하는 최소 집합의 predicate-argument로 구성된 순서열을 의미한다. 이러한 특성 때문에 PAS 패턴은 문장 내에서 상호작용하는 두 개체 간의 연관관계를 표현해주는 중요한 자질 정보가 된다. 따라서 한 개체로부터 시작해서 다른 개체로까지의 의미적 연결고리를 제공해주는 PAS 패턴을 이용하여 관계 식별을 수행할 수 있다.

본 논문에서는 PAS 패턴을 추출하기 위해서 Enju 파서를 이용하였다. Enju 파서를 이용한 문장 분석 결과의 예는 그림 1과 같다.

1) http://www-tsujii.is.s.u-tokyo.ac.jp/enju



(그림 1) Enju 파서의 문장 분석 결과

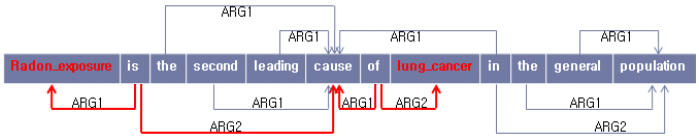
그림 1에서 보는 바와 같이 Enju 파서는 문장을 입력으로 받아서 문장을 구성하는 각 단어의 predicate-argument 관계를 분석하여 제공한다. 행으로 나열된 predicate-argument 분석 결과의 각 필드에 대한 설명은 표 1과 같다.

<표 1> Enju 파서 분석 결과의 각 열에 대한 설명

열 번호	상세 설명
1	predicate 단어
2	predicate 단어의 기본형
3	predicate 단어의 품사
4	predicate 단어의 기본형의 품사
5	문장에서 predicate 단어의 위치
6	predicate 종류
7	predicate와 argument 사이의 관계 레이블
8	argument 단어
9	argument 단어의 기본형
10	argument 단어의 품사
11	argument 단어의 기본형의 품사
12	문장에서 argument 단어의 위치

표 1에서 설명한 내용을 바탕으로 “Radon_exposure is the second leading cause of lung_cancer in the general population.” 문장에 대한 분석 결과인 그림 1의 2번째 행과 3번째 행을 설명하면, 우선 단어 ‘is’는 동사로서 argument 1과 2를 갖는데 그 중 argument 1은 명사인 단어 ‘radon_exposure’를 지칭하고 argument 2는 또 다른 명사인 단어 ‘cause’를 지칭한다는 사실을 나타낸다. 분석 결과의 1번째 행은 단순히 문장의 root predicate를 표현하는 것이고, 4번째 행부터는 2번째와 3번째 행을 해석한 것과 같은 방식으로 해석하면 된다.

Enju 파서에서 제공된 결과를 이용하여 각 단어의 predicate-argument 관계 그래프를 그리면 그림 2와 같이 표현된다.



(그림 2) Predicate-Argument 관계 그래프

그림 2에서 실제적으로 문장 내에 존재하는 두 개체 간의 유의미한 관계를 표현하는 PAS만을 추출하여 패턴을 구성하면 그림 3과 같다. 화살표의 연결은 한 개체로부터 상호작용하는 다른 개체로까지의 predicate-argument 관계를 추적할 수 있다는 것을 의미한다. 따라서 ‘radon_exposure’와 ‘lung_cancer’ 사이의 관계를 추적해보면 ‘is cause of’와 같은 중요한 어휘적 패턴을 기반으로 관계가 형성되어 있음을 알 수 있다. 다시 한 번 말하지만, 이러한 패턴은 두 개체 간의 상호작용을 식별하는데 중요한 자질로 사용될 수 있다.

(그림 3) PAS 패턴

결과적으로 개체 1과 개체 2의 관계는 두 개체를 유의미한 관계로 연결해주는 PAS 패턴에 의하여 식별될 수 있다. 따라서 관계 별로 나타나는 패턴의 집합을 구축하여 이 패턴 집합을 관계 예측의 근거 자질로 활용하면 개체 간의 관계를 추출하는 관계 추출 시스템에 활용할 수 있다.

PAS 패턴을 어휘 자질로 활용하기 위해서 predicate-argument 구조, 즉 그림 3에서 화살표로 연결되는 predicate 단어와 argument 단어, predicate의 종류, 그리고 predicate와 argument 사이의 관계 레이블을 이용하여 벡터 값을 생성하였다. 그리고 이 벡터 값을 SVM의 내장 커널 중 하나인 RBF 커널을 이용하여 관계 식별을 수행하였다.

4. 실험 및 분석

PAS 패턴이 관계 예측의 유용한 자질로 활용될 수 있는지의 가능성을 검사해보기 위해서 AIMed[2] 컬렉션을 대상으로 다음과 같이 실험을 수행하였다.

첫 번째로 개체 간의 유의미한 관계를 표현해주는 모든 PAS 패턴을 대상으로 관계 식별 실험을 수행

2) ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/

하였다. 여기서 대상으로 삼은 패턴은 어휘적 패턴 뿐만 아니라 괄호나 쉼표와 같은 기호적 패턴까지도 포함하였다.

두 번째로 개체 간의 유의미한 관계가 가장 잘 표현되는 동사구가 포함된 패턴만을 대상으로 실험을 수행하였다. 동사구를 일정 수준 이상으로 추상화하면 관계 종류가 될 정도로 동사구는 관계 식별에 중요한 단서가 될 수 있다.

libsvm[3]을 이용하여 위에서 제시한 두 가지 실험을 수행하였는데, 실험 결과는 표 2와 같다.

표 2 RBF 커널을 이용한 실험 결과

	옵션 값		결과 값
	cost	gamma	accuracy
모든패턴	512.0	0.0078125	86.2478
동사구포함패턴	8.0	0.5	87.1499

본 실험을 통해서, 개체 간의 관계를 추출할 때 두 개체를 연결하는 PAS 패턴이 아주 유용한 단서가 될 수 있음을 파악하였고, 더불어서 동사구와 같은 핵심 단어가 포함된 패턴이 성능 향상에 좀 더 기여할 수 있음을 확인하였다. 따라서 관계 추출 시스템을 개발할 때 본 연구에서 얻어진 결과들을 활용하면 좀 더 성능 좋은 관계 추출 시스템을 개발할 수 있다.

5. 결론 및 향후 연구

본 논문에서는 문장 내에 존재하는 개체 간의 유의미한 관계를 표현해주는 PAS 패턴을 활용하여 관계 추출 시스템을 개발하였고, 실험을 통하여 PAS 패턴이 관계 추출을 위한 중요한 자질로 활용될 수 있음을 증명하였다.

향후 연구로는 본 논문에서 증명한 PAS 패턴의 관계 식별력을 기존의 다른 자질 활용 방법, 예를 들면 구문트리의 유사성을 이용하는 방법과 결합하여 혼합 커널을 구성하는 방법에 대한 연구가 필요하다. 구문적 유사성을 활용하는 구문트리 기법과 의미적 유사성을 활용하는 PAS 패턴 기법을 결합하면 보다 성능 좋은 관계 추출 시스템을 개발 할 수 있을 것으로 사료된다.

참고문헌

[1] Akane Yakushiji, Yusuke Miyao, Yuka Tateisi, and Jun'ichi Tsujii, "Biomedical Information Extraction with Predicate-Argument Structure Patterns", SMBM2005.

[2] Akane Yakushiji, Yusuke Miyao, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii, "Automatic Construction of Predicate-argument Structure Patterns for Biomedical Information Extraction", EMNLP 2006.

[3] Context Free Grammar, http://en.wikipedia.org/wiki/Context-free_grammar

[4] Head-driven Phrase Structure Grammar, http://en.wikipedia.org/wiki/HPSG

3) http://www.csie.ntu.edu.tw/~cjlin/libsvm/