의료 문헌에서의 절차적 지식 추출을 위한 단위 절차 추출 연구

송사광[O+], 오흥선[*], 최윤정[*], 장혜주[*], 맹성현[*], 최성필[+], 최윤수[+]
[+]한국과학기술정보연구원 정보기술연구실, {esmallj, spchoi, armian}@kisti.re.kr
[*]한국과학기술원 전산학과, ohs@kaist.ac.kr, {choiyj35, hyejujang}@gmail.com, myaeng@kaist.ac.kr

Procedural Entity Extraction for Procedural Knowledge on Medline Abstracts

Sa-kwang Song [O+], Heung-Seon Oh[*], Yoonjung Choi[*], Heju Jang[*], Sung-Hyon Myaeng[*], Sung-pil Choi[+], Yunsoo Choi[+]
[+]Korea Institute of Science and Technology Information, {esmallj, spchoi, armian}@kisti.re.kr
[*]Korea Advanced Institute of Science and Technology, ohs@kaist.ac.kr, {choiyj35, hyejujang}@gmail.com, myaeng@kaist.ac.kr

요 약

본 연구는 2인의 전문의와 함께 의료 문헌의 초록을 분석하여 의료문서에서의 절차적 지식을 모델링하고 텍스트 마이닝 기법을 적용하여 절차적 지식을 추출하는 방법론에 대해 기술한다. 절차적 지식은 목적과 해법의 묶음으로, 해법은 다시 단위 절차 지식의 네트워크로 정의 하였고, 목적과 해법 정보 추출과 단위 절차 지식의 구성요소인 대상/행위/방법 개체를 인식하기 위해, 품사태깅, 구문분석, 술어-논항 구조(Predicate-Argument Structure), 온톨로지 용어 매핑 정보 등에 기반한 기계학습 방법을 사용하였다. 실험을 위해 전문의와 함께 위암과 척추질환에 대한 1309 문서에 절차적 지식 태깅을 수행하였고, 이 문서 집합을 기반으로 목적/해법 추출 작업과 단위 절차 지식(대상질병/행위/적용방법) 추출 실험을 수행하여, 각각 82%와 63%의 F-measure 값을 얻을 수 있었다.

1. 서 론

절차적 지식(Procedural Knowledge)은 일반적으로 무언가를 수행하는 방법(how to do something)에 대한 지식 또는 기술에 대한 지식(knowledge of skills) 등으로 정의되어 왔고[1,3], 인공지능/언어처리 분야에서는 절차적 지식을 서술적 지식(Declarative Knowledge)과 구분하여 사용하고 있다. [1]는 이러한 절차적 지식에 대해 표 1과 같이 분야별로 두 용어를 비교 정리하였다.

표 1 서술적 지식 vs. 절차적 지식

	서술적 지식	절차적 지식
철학	Knowledge-that, propositional knowledge	Knowledge-how, procedural knowledge, abilities
심리학	Explicit knowledge, declarative knowledge	Implicit knowledge, tacit abilities, skills
인공지능	Declarative knowledge	Procedural knowledge

절차적 지식은 일반적으로 특정 목적을 달성하기 위한 순차적인 또는 구조적인 행위들의 모음 또는 단위 절차 지식의 모음으로 표현된다[3]. 즉, 개별 단위 절차 지식들이 서로간의 관계를 형성하여 '목적(Purpose)'을 달성하기 위한 '해법(Solution)'을 구성하게 된다. 위암 절제술에 대한 절차를 예로 들어보자.

"위암 절제술은 일반적으로 피검사, FOBT(Fecal occult blood test), CT Scan 등의 검사를 진행한 후, 마취 처치를 진행한다. 마취가 완료된 후 피부절개와 복막 절개 등의 절차를 순차적으로 수행한 후, 대상 부위를 절제하게 된다."

이 예에서 위암 절제술은 다수의 단위작업(사전검사, 수술절차 등)들이 순차적 관계로 구성되어 있다. 즉, 단위 절차들과 각 단위 절차들간의 관계로 표현된다. 이러한 절차 정보의 추출은, 다양한 응용을 가능하게 하는데, 특정 목적에 대한 방법론(절차)을 추출하고 이를 활용한다는 점에서 매우 중요한 시도라 할 수 있다. 예를 들어, 척추 전문의가 환자를 치료하기 위해 최신 치료 방법론 및 치료 절차에 대한 최신 경향을 습득하고자 할 경우, 절차적 지식의 제공은 의료 품질 향상으로 효과를 가져올 수 있다. 이와 같이 의료 문서 집합에서의 절차적 지식 추출은 질병에 대한 다양한 해결 방법 및 절차를 의료전문가에게 제공하여 의료 품질의 개선 및 선진화에 기여하고, 나아가 기존 연구 내용의 분석 및 정책 결정 등에 중요한 정보를 제공할 수 있다.

2. 관련 연구

사전, 시소러스, 온토롤지 등을 활용한 다양한 전문용어, 개체, 개념 등의 정보를 추출하는 연구는 최근까지 꾸준히 진행되어 오고 있고, 단백질 개체간의 관계 추출 연구 등 개체간의 연관 관계 또는 이벤트 등의 정보를 추출하는

연구들이 근래에 진행되어 오고 있다[5,6,8,9]. 하지만 이러한 단편적인 지식에서 벗어나, 구조화되고 절차화된 지식 추출에 대한 연구는 극히 찾아보기 어렵고, 다만 eHow, wikiHow 등의 웹 문서에서 절차적인 정보를 추출하는 연구[7]가 있으나, 이는 이미 사람들에 의해 구조화/순차화된 문장 집합에서 온톨로지를 추출하는 연구이다. 따라서, 본 논문에서는 의료문서를 대상으로 절차적 지식을 모델링하고, 절차적 지식의 각 요소를 추출하는 방법론에 대해 기술한다.

3. 본 론

3.1 절차적 지식 모델링

앞서 절차적 지식을 특정 목적을 달성하기 위한 순차적인 또는 구조적인 행위들의 모음 또는 단위 절차 지식의 모음이라고 언급하였다. 즉, 분석 대상이 되는 문서의 구조를, 문서를 작성한 목적 문장과 그 목적(Purpose)을 이루기 위한 해법(Solution) 문장의 집합으로 표현할 수 있고, 해법 문장은 다시 구체적인 해결 절차들을 포함한 문장의 모음으로 표현할 수 있다. 해법 내의 단위 절차들은 각 절차들간에 순차적, 병렬적, 또는 독립적 관계가 있다고 가정하였고, 각 단위 절차는 다시 절차 수행의 대상(Target), 방법(Method), 행위(Action)의 트리플(Triple)로 구성하였다. 이는 단위 절차 내에서의 실험 수행 목적 또는 대상이 되는 부분을 대상(Target)으로, 이 대상에 적용하는 구체적인 실험방법을 방법(Method)로, 그리고 실험대상에 실험방법을 어떻게 적용했는지를 행위(Action)으로 정의하였다. 그림 1은 이러한 절차적 지식을 모식적으로 표현한 것이다.

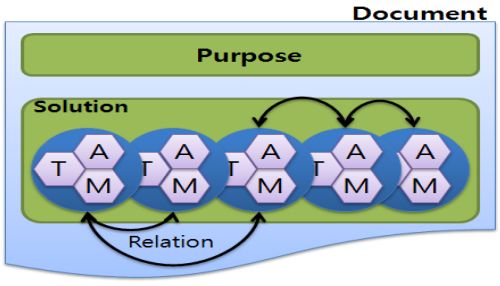

그림 1. 의료문서 요약에서의 절차적 지식 모델링: T(대상), A(행위), M(방법)

대상(Target)은 병명, 질병명, 증상, 효과, 수치 등으로 한정하였고, 방법(Method)는 치료 방법, 수술 방법, 복약 방법 등으로 정의하였다. 또한 행위(Action)은 대상과 방법을 연결해 주는 술어부분을 의미하며, "'대상'에 '방법'을 적용하여 '행위'를 하였다"와 같이 해석이 가능하다. 이때 "have", "be" 등과 같은 보편적인 술어는 고려하지 않았다. 이러한 절차적 지식 모델링은 의료문서 분석 결과의 가장 큰 수혜자라고 할 수 있는 전문 의료인들과 함께 문서의 내용/특성 등을 분석하며 수행하였다.

3.2 절차적 지식 추출 방법

위와 같이 모델링된 대상 문서는 4단계의 과정을 거쳐 절차적 지식이 추출된다. 먼저 대상 문서는 품사태깅, 구문분석, 술어-논항 구조 분석, 전문용어 추출 등 다양한 언어처리 기법을 적용하고, 이를 기반으로 논문의 목적을 기술하고 있는 섹션과 그 목적을 해결하기 위한 방법론을 포함하고 있는 섹션을 구분해 낸다. 다음단계는 추출된 방법론 섹션에서 단위 절차를 추출하는 과정인데, 이는 해법 섹션이 실험 대상과 실험방법에 대한 구체적인 내용을 포함하고 있기 때문이다. 단위 절차는 3.1절에서 언급한 것처럼 대상, 행위, 방법의 트리플로 구성된다. 마지막으로, 이렇게 인식된 단위 절차들 간의 관계 정보를 추출하는 과정을 거쳐 대상 문서에 대한 절차적 지식이 완성된다.

3.3 대상 문서

절차지식 추출 대상 문서는 Medline 초록 중에서 위암(Gastric Cancer)과 척추질환(Spinal Disease)의 두 도메인 문서로 한정하였다. 특히, 두 도메인 중에서 8개(위암 2개, 척추질환 6개) 질병에 해당하는 문서들만을 대상으로 하였는데, 이는 전문의사들과 Medline 초록을 분석을 통해, 이들 질병에 관한 문서가 절차적인 지식 추출에 적합하고, 동시에 일반적으로 많은 관심을 갖는 질병 분야라는 결론을 얻었기 때문이다. 대상 문서인 Medline 초록은 일부 최근 문서의 경우 섹션 정보를 포함하여 제공하고 있는데, 섹션 정보라 함은 저자에 의해 구분된 의미적인 블록 정보로, 일반적으로 OBJECTIVE, BACKGROUND, METHODS, RESULTS, CONCLUSIONS 등의 섹션으로 구분되어 있다. 아래에 이러한 섹션 정보를 포함하고 있는 예를 볼 수 있다.

> **OBJECTIVE**: To examine the impact of malignancy and location of the cerebellar tumor on motor, cognitive, and psychologic outcome. **BACKGROUND**: Although many …생략…. **METHODS**: Children, aged from 6 to 13 years, with a cerebellar malignant tumor (MT; MT group, n=20) or a cerebellar benign tumor (BT; BT group, n=19) were … 생략… **RESULTS**: Parents and teachers reported high rate of learning and academic difficulties, …생략… **CONCLUSIONS**: Dentate nuclei lesions are major risk factors of motor and cognitive impairments in both cerebellar BT and MT.

초록 내의 섹션은 대개 아래와 같이 5개로 분류 가능하고, 그 일반적인 의미는 다음과 같다: OBJECTIVE(논문의 목적), BACKGROUND(논문의 배경 정보), METHODS(논문의 목적을 이루기 위한 해법), RESULTS(적용된 방법에 대한 결과), CONCLUSIONS(논문의 결론). 하지만, 본 연구에서는 구체적인 목적 및 해법을 기술한 구절(문장)을 찾고, 찾은 구절에서 단위 절차 지식을 추출하는 것을 목표로 하고 있기 때문에, 각 문장을 각 섹션으로 분류하는 연구와는 차별화된다. 즉, 초록내의 각 문장이 목적을 기술하고 있거나, 해법을 기술하고 있는 문장을 찾는 연구이므로 복수의 문장이 동일한 섹션에 포함되어 있다 할지라도 단위절차 정보가 포함되어 있지 않은 문장은 목적 또는 해법 분류에서 제외된다. 그러나 초록의 기본 구조가 의미상 섹션으로 구조화되어 있기 때문에 이러한 특성은 목적/해법 문장 분류에 매우 중요한 단서가 된다. 이에 대한 상세한 설명은 4.2절에서 제공된다. 해법 섹션은 일반적으로 하나 이상의 방법론 구절(문장)을 포함하고 있고, 각 방법론들은 서로간의

절차적인 관계를 내포하고 있는데, 위의 예에서 해법 섹션을 간략화하여 살펴보도록 한다.

Children with a cerebellar malignant tumor or a cerebellar benign tumor were examined at least 6 months after the end of treatment using the ① international cooperative ataxia rating scale, the ② Purdue pegboard for manual skill assessment and the ③ age-adapted Weschler scale.

여기서, 대상자 또는 대상질병은 cerebellar malignant tumor나 cerebellar benign tumor를 갖고 있는 아이들이고, 세가지 실험 방법(①, ②, ③)을 적용하였다. 또한 'after'라는 단어를 기준으로 2개의 절차가 구분되어 있고 이들의 관계는 '순차적'이라는 정보를 담고 있으며, 위 세가지 실험 방법은 병렬 관계로 구분할 수 있다. 따라서, 대상 문서에서 절차적 지식 추출하는 작업은 먼저 초록으로부터 목적 문장, 해법 문장을 인식하고, 각 문장에서 단위 절차 지식인 대상질병/행위/적용방법의 트리플을 추출한 후 단위 절자 지식들 간의 관계를 인식하는 것으로 정리할 수 있다. 본 논문에서는 분량상 단위 절차의 요소인 대상질병/행위/적용방법 정보 추출하는 단계까지의 내용을 표현한다.

3.4 학습 코퍼스

기존에 절차적 지식 추출을 목적으로 구성된 테스트 컬렉션이 부재하기 때문에, 본 연구에서는 2명의 전문의를 활용하여 총 1,309개의 논문 초록에서 목적/해법 문장과, 대상질병/행위/적용방법의 트리플 및 그들 간의 관계 정보를 태깅 하였다. 1,309개의 초록을 두 사람이 반으로 나눠서 각각 태깅 하였으며, 태깅을 모두 마친 후, 상대방이 태깅한 문서를 서로 검증하는 단계를 거쳤다.

4. 실험

실험은 크게 두 부분으로 구분되는데, 초록 문서의 구조를 인식하여 초록의 목적 섹션과 해법 섹션을 구분해 내는 실험과, 추출된 세션에서 구체적인 단위 절차로써 대상질병(Target)/행위(Action)/적용방법(Method) 트리플 정보를 인식해 내는 실험이다. 다음에 이 두 실험에 대한 상세한 설명을 하였고, 이에 앞서 대상 문서 전처리 작업에 대한 기본적인 설명을 추가하였다.

4.1 대상 문서 전처리

대상 문서는 다양한 언어처리 작업을 거쳐, 품사태깅, 구문분석, 술어-논항 구조(Predicate-Argument Structure), 전문 용어 추출 등을 포함한다. 술어-논항 구조는 술어(predicate)와 논항(argument) 관계를 이용하여 문장 내에 존재하는 각 단어 간의 유의미한 연관관계를 표현하는 구조이다. 전문 용어 추출이라 함은 잘 알려진 UMLS(Unified Medical Language System), UniProt, GO(Gene Ontology) 등과 같은 의생명 분야 온톨로지를 기반으로 문서 내의 단어절 또는 다어절 용어를 태깅한 정보인데, 이는 대상 문서에 포함된 용어들이 해당 분야의 전문용어임을 고려할 때, 심층 지식 추출의 효율성을 위해 필수요소라 할 수 있다.

4.2 목적/해법 문장 추출

논문 초록에 포함된 각 문장이 목적 또는 해법을 기술하는지를 판별하는 문제는, 각 문장을 목적, 해법, 기타의 세가지 분류 중 하나에 할당하는 문제로 간주할 수 있다. 이는 초록에 있는 문장들을 미리 정의한 섹션에 분류하는 섹션 분류 문제와 비슷하다. 본 연구에서는 이를 해결하기 위해서 섹션 분류에서 사용되는 기계 학습 알고리듬들(SVMs[2]과 CRFs[4])을 이용하여 목적/해법 문장을 추출하였다. 실험에 사용된 자질은 크게 내용 자질, 위치 자질, 이웃 문장 자질, 전문용어 자질 등 4가지로 구분된다. 각각에 대한 설명은 다음과 같다.

● 내용 자질: 분류하고자 하는 문장으로부터 unigram과 bigram을 추출하여 이용함. stemming 및 stopword 제거.

● 위치 자질: 목적 문장은 대부분 초록의 앞 쪽에서 많이 나오고 해법 문장은 목적 문장의 뒤 쪽에 많이 나옴. 이를 바탕으로 초록에서 문장의 상대적인 위치를 자질로 활용함.

● 이전 이후 k 문장: 현재 문장의 앞 뒤 k 번째 문장의 내용 자질을 문맥 자질로 활용함.

● 전문용어 자질: UMLS/Uniprot/GO Ontology 온톨로지 태깅 정보

4.2 단위 절차 정보 추출

단위 절차 정보는 Target(대상질병)/ Action(행위)/Method(적용방법)의 트리플로 구성됨으로 단위 절차 정보 추출 과정은 이러한 트리플을 문장 내에서 추출하는 것을 의미한다. 따라서 목적/해법 문장에서 Target, Action, Method에 해당하는 의미적 개체 추출에 대해 아래에 기술한다.

단위 절차 추출의 기본 요소인 Target, Action, Method의 추출을 위해 사용된 자질은 어절 자체의 자질, 문맥 자질, 술어-논항 구조 자질, 전문용어 자질 등, 크게 4가지로 분류된다.

● 어절 자체의 자질: 어절 자체, 어절의 기본형, 품사 태그, 품사 분류 정보(즉 동사, 명사, 기호 등), 어절의 대문자로 시작 여부 또는 전체 대문자 여부

● 문맥 자질: 이전/이후 N개의 어절 및 품사 태그 정보

● 술어-논항 구조:술어 및 논항(argument) 해당 구절

● 전문용어 자질: UMLS/Uniprot/GO Ontology 온톨로지 태깅정보

실험을 위해 대상 문서에서 앞에서 언급한 관련 자질을 추출하였고, 추출된 정보를 기반으로 CRFs 모델 학습을 수행하였다. CRFs에서 입력은 위 자질 정보 집합이고, 출력은 각 단위 절차 요소(Target, Action, Method)의 BIO 표기법(예를 들어, B-Target, I-Target, B-Action, I-Action, B-Method, I-Method 등)을 이용하였다. 실험은 Target, Action, Method 각 객체의 경계인식과 객체 분류를 동시에 수행하였는데, 다어절 후보 객체의 경우 전체 일치의 경우에만 제대로 인식된 것으로 판단하였다. 따라서, 경계인식과 객체 분류 모두가 옳게 인식된 결과에 대해

각각의 정확도, 재현도, F-measure를 측정하였다.

5. 실험 결과

5.1 목적/해법 문장 추출

학습 데이터와 테스트 데이터를 8:2의 비율로 구성하여 실험을 수행하였고, 그 결과는 표 2에 나타나 있다. CRFs를 이용한 목적 문장에 대한 추출 정확도는 F-1 수치로 0.84의 비교적 높은 성능을 보였지만, 해법을 기술하고 있는 문장을 추출하는 성능은 0.69정도로 예상보다 낮게 나타나고 있다. 이는 해법 문장이 초록 내 문장 순서에 어느 정도 종속적이긴 하지만, 실제로는 해법 문장 사이 사이에 단위 절차정보(Target, Action, Method) 정보가 포함되지 않은 문장이 삽입되어 있는 경우가 많기 때문으로 분석된다. 표 3는 동일한 데이터 및 자질을 이용해서 지지벡터기계(SVMs) 알고리듬에 적용한 결과이다. SVMs는 목적과 해법 문장 분류에서 각각 0.87과 0.79의 높은 성능을 나타냈다. 목적 문장 분류가 해법 문장 분류보다 상대적으로 높은 성능을 보이고 있는데, 이는 목적 문장의 기술 방식이 일관성 있고 단서가 되는 어절이 한정(예를 들어, the aim of this study, the goal is 등등)되는 반면, 해법 문장은 이러한 표현의 일관성을 찾기가 어렵고, METHODS섹션에 포함되는 문장임에도 구체적인 대상질병이나 적용방법 등이 명시되지 않아 해법 문장으로 분류가 어려운 경우가 있기 때문이다.

표 2 CRFs를 이용한 목적/해법 문장 분류 결과

	정확도	재현도	F-1
Purpose(목적)	0.8326	0.8578	0.8450
Solution(해법)	0.6923	0.6913	0.6918
전체	0.7279	0.7326	0.7303

표 3 SVMs를 이용한 목적/해법 문장 분류 결과

	정확도	재현도	F-1
Purpose(목적)	0.8462	0.9009	0.8727
Solution(해법)	0.8333	0.7610	0.7955
전체	0.8369	0.7957	0.8158

5.2 단위 절차 정보 추출

실험은 8:2의 비율로 학습과 테스트를 수행하여 표 4와 같이 전체적으로 62%의 F-1값을 보였다. 단위 절차 지식 중 Action 객체 인식의 경우, 상대적으로 높은 인식 성능을 나타내고 있는데, 이는 Action의 경우 대부분이 동사파생 어절이고, 그 길이가 길지 않기 때문이다. 반면 Target이나 Method 객체 인식의 경우는 0.5 중반의 정확도를 나타내고 있는데, 이는 의미적으로 문장 내에서 Target과 Method를 구분하는 것이 어려움을 나타낸다고 할 수 있다. 그리고, 그 길이가 대부분 복수의 어절들로 구성되어 있고, 단순히 명사어구들의 모음이 아니라, 부사/형용사를 포함하거나 전치사구나 동명사구 등 다양한 경우가 많아 상대적으로 인식 정확도가 높지 않다.

표 4 CRFs를 이용한 단위 절차 정보 추출 결과

	정확도	재현도	F-1
Target(대상)	0.5212	0.5696	0.5443
Action(행위)	0.7878	0.7753	0.7815
Method(방법)	0.6014	0.5078	0.5507
전체	0.6401	0.6102	0.6248

6. 결론

본 연구에서는 전문의들과의 의료 문서 분석과정을 통해, 의료 도메인에서의 절차적인 지식을 모델링 하고, 1309개 문서로 구성된 절차 지식 코퍼스를 구축하였다. 모델의 각 요소 추출 방법론으로써, 다양한 언어처리 정보를 자질로 활용한 기계학습 기반 방법론을 제시하였고, 실험결과 목적/해법 추출의 경우는 82%로 높은 성능을 얻을 수 있었고, 단위절차 추출의 경우도 63%로 의미있는 결과를 얻을 수 있었다.

참고문헌

[1] Baljinder Sahdra, Paul Thagard, Procedural knowledge in molecular biology, Philosophical Psychology, vol. 16, No. 4, 2003

[2] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[3] Georgeff, M.P., Lansky, A.L., Procedural knowledge, Proceedings of the IEEE, vol. 74, No. 10, pp. 1383-1398, 1986

[4] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.

[5] Miwa, Makoto, Rune Sætre, Jin-Dong Kim, and Jun'ichi Tsujii. Event Extraction with Complex Event Classification Using Rich Features. Journal of Bioinformatics and Computational Biology (JBCB). 8(1). pp. 131-146, February 2010.

[6] Pyysalo, Sampo. Entities, relations, events: representing biomolecular semantics. BMC Bioinformatics. 11(Suppl 5). pp. O6, 2010.

[7] Yuchul Jung, Jihee Ryu, Kyung-min Kim and Sung-Hyon Myaeng. "Automatic Construction of a Large-Scale Situation Ontology by Mining How-to Instructions from the Web", Journal of Web Semantics, Vol. 8, Issues 2-3, pp. 110-124, 2010

[8] 김재훈, 김형철, 최윤수, 기계학습 기반 개체명 인식을 위한 사전 자질 생성, 정보관리연구, vol 41, no.2 2010, pp.31-46, 2010

[9] 최성필, 최윤수, 정창후, 맹성현, 구문 트리 가지치기 및 소멸 인자 조정을 통한 트리 커널 기반 단백질간 상호작용 추출 성능 향상, 한국정보과학회논문지, vol 37, No.2, pp. 85-

94, 2010