의학 전문용어의 정의문 자동 추출

김재호[O] 배선미 신효식 최기선
한국과학기술원 전산학과 / 전문용어언어공학센터 / 언어자원은행
{jjaeh[O], sbae, gerling, kschoi}@world.kaist.ac.kr

Automatic Extraction of Medical Term Definition from Texts

Jae-Ho Kim[O], Sun-Mee Bae, Hyo-Shik Shin, Key-Sun Choi
KAIST CS Dept. / KORTERM / BOLA

요 약

지식 정보의 확산에 따라 기존 전문분야 용어집에 수록되지 않은 용어의 수가 폭발적으로 증가하고 있다. 이에 따라 용어집을 자동으로 구축하는 작업이 필요하게 되었다. 본 논문에서는 의학분야 코퍼스에서 주어진 전문용어에 대한 정의문을 자동으로 추출하는 방법을 제안한다. 우선, 정의문의 구문적 패턴과 용어의 어휘구성 패턴을 이용하여 용어의 상위개념을 추정한다. 상위개념별로 구축된 특성 어휘 목록을 이용하여 구문적 패턴으로 뽑힌 문장에 등장하는 어휘의 적합성 여부를 판단하여 정의문을 추출한다. 실험 결과 코퍼스에 정의 정보가 있는 48개의 용어에 대하여 71.43%의 정확률을 보인다.

1. 서론

전문용어 사전이나 백과사전은 언어처리에 매우 유용한 언어 지식이다. 그러나 많은 사전구축 비용이 들며, 신조어는 사전에 바로 반영되지 못하는 문제가 있다. 이러한 한계를 극복하기 위하여 코퍼스나 웹문서로부터 용어집을 자동으로 구축하려는 연구들이 진행되고 있다 [1, 2, 3]. 이들 연구들은 수동 혹은 반자동으로 구축한 정의문의 구문적 패턴을 이용하고 있다. 그러나 언어표현 방식은 매우 다양하기 때문에 이러한 구문적 패턴만을 이용하는 방법은 한계를 가질 수 밖에 없다. 예 1-1 과 같이 동일한 용어라도 분야마다 사전마다 다른 형식과 내용으로 기술되기 때문에 모든 경우를 고려한 정의문의 구문적 패턴을 만들기는 어렵다.

(예 1-1) ' 바이러스' 에 대한 사전 정의문의 예
[생물학] DNA 나 RNA 중 하나를 게놈으로서 갖는 감염세포 내에서만 증식하는 감염성의 미소구조체. (생물학사전 [8])
[의학] 세균보다 작아서 세균여과기로도 분리할 수 없고, 전자현미경을 사용하지 않으면 볼 수 없는 작은 입자. (두산세계대백과 EnCyber [9])
[의학] 미세한 감염성 인자의 1군으로 폭스바이러스(poxvirus)를 제외하고는 일반광학현미경으로는 관찰할 수 없다. (표준의학사전 [10])

본 논문에서는 정의문을 자동으로 추출하기 위하여 구문적 패턴뿐만 아니라 용어의 어휘구성 패턴, 정의문의 의미적 패턴까지 고려한 전문용어의 정의문 추출 방법론

2. 관련연구

[1]은 규칙 기반 정의문 추출기를 제안하였다. 쉽고 잘 기술된 소규모의 의학 분야 코퍼스에서 중요어구 (" is called" , " is the term used to describe" , " is defined as" 등)와 마크(" ()" , " --")를 근거로 하여 구축한 정의문의 구문적 패턴을 이용하였다. [2]는 웹문서로부터 "What is X?" 에 대한 답을 제시하기 위한 용어 설명기를 개발하였다. 용어, 설명문, 출처 URL, 상위개념으로 이루어진 템플릿을 사용하는데, 설명문과 상위개념을 추출하기 위하여 수동으로 구축한 패턴을 이용하였다. 그러나 [1]과 [2]는 패턴이 너무 일반적이고 그 수가 작아서 패턴의 적용범위가 작은 단점을 가지고 있다.

[3]은 용어를 설명하는 설명문을 추출하여 백과사전적 지식을 만들어내는 방법을 제시하였다. 먼저 구글 검색엔진을 이용하여 용어에 관련된 문서를 모은 후, 사전에 기술된 용어 설명문에서 반자동으로 추출한 패턴과 웹문서의 HTML 구조를 이용하여 용어를 설명하는 부분을 추출하였다. 추출된 문장에 등장하는 어휘가 분야 특성 어휘인지, 이 문장이 실제 백과사전 설명문과 비슷한 구조인지를 확률적으로 계산하여 최종적으로 한 문장을 설명문으로 선택하였다. 반자동으로 구축한 패턴과 언어적 통계 모델을 잘 조합하였으나, 용어의 분류에 따른 설명문의 특징을 분석하지 않고 단지 분야 특성 어휘에 의해 설명문을 선택하기 때문에 정의문이나 더 중요한 적절한 설명문을 추출하지 못할 수도 있는 문제를 가지고 있다.

3. 정의문 기술 형식

을 제안한다. 여기서 대상 분야는 의학 분야로 한정한다.

추출하려는 정의문은 어떻게 구성되어야 하는 지 알아보자. 국제표준화기구 ISO 에서는 정의문의 표준 형식에 대하여 기술하였다 [4]. ISO 704 규격의 정의문 형식은 다음과 같이 표현할 수 있다 [5].

X = Y + 차별적 의미특질소

여기서 X는 정의될 용어를 말하며, Y는 X에 대한 상위개념이다. '차별적 의미특질소 (distinguishing characteristics)'란 동위어들로부터 그 용어를 구별해주는 특징적인 의미속성을 말한다.

자연언어처리와 정보검색의 여러 분야에서 널리 이용되고 있는 영어 어휘 데이터베이스 워드넷 (WordNet [6])을 살펴 보자. 워드넷에서는 synset (Set of synonymys)단위로 용어풀이 (gloss)가 존재하는데 이 용어풀이는 ISO 에서 말하는 정의문에 해당된다고 볼 수 있다. 워드넷에서 명사에 대한 정의문은 동위어들로부터 그 용어를 구별해 주는 형용사 또는 관계절로 수식을 받는 상위개념으로 구성되어 있다[7]. 워드넷에서의 정의문 기술 방식은 ISO 정의 방법과 유사하다고 볼 수 있다.

본 논문에서는 일관성과 보편성을 고려하여 " 용어 $=_{def}$ **의미특질소 + 상위개념**" 을 정의문 기술 방식으로 채택하고, 이에 입각하여 정의문 추출 시스템을 설계하고자 한다.

4. 정의문 추출 시스템

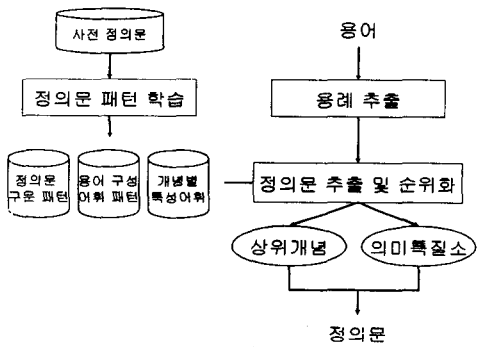

그림 1 전문용어 정의문 추출 과정

본 논문에서 제안하는 전문용어 정의문 추출의 전체 과정은 그림 1 과 같다. 주어진 용어에 대하여 전문용어 코퍼스에서 용례를 추출한 후, 사전 정의문을 학습하여 반자동으로 구축한 자원들을 이용하여 정의문을 추출한다.

4.1 정의문 기술 패턴 학습

표준의학사전의 용어 정의문 중 55,791 개를 훈련 정의문으로 사용하여 정의문 기술 패턴을 반자동으로 학습한다.

4.1.1 상위개념 추출 규칙의 학습

정의문의 가장 일반적인 형태는 예 4-1 과 같이 상위개념 명사 (이하 N_{sup} 라 칭한다)로 끝나는 것이다.

(예 4-1) Term = 의미특질소 + 상위개념 N_{sup}
극상골반 = 대단히 예리하게 돌출한 치골륭을 가진 (의미특질소) + 골반 (N_{sup})

위 예에서처럼 복합명사로 이루어진 전문용어의 경우 마지막 명사가 상위개념을 나타내기도 한다. 그래서 명사로 끝나는 정의문에서 그 명사가 용어를 구성하는 마지막 명사일 경우 그 명사를 용어의 상위개념으로 간주한다는 씨앗 정보 (기본규칙)를 만들어 낸다. 씨앗 정보로 뽑힌 상위개념 N_{sup} 은 N_{sup} 으로 끝나는 다른 용어 T'에 대해서도 상위개념일 가능성이 높다. 그래서 T'의 정의문에서 N_{sup} 이 나타나는 패턴을 모두 추출한 후 수작업으로 상위개념이 될 수 있는 규칙을 선별한다. 뽑힌 규칙은 표 1 과 같다.

표 1 훈련 정의문에서 추출한 상위개념의 구문적 패턴

규칙	패턴
0	~ N_{sup} (이다)$
1	~ N_{sup} 로(,) / ~ N_{sup} 으로(,)
	~ N_{sup} 로서(,)/ ~ N_{sup} 으로서(,)
2	~ N_{sup} 의 PART+J
3	~ N_{sup} 을 [말하-|총칭하|지칭하-|가르키-]
4	~ N_{sup} 나, / ~ N_{sup} 이나,
	~ N_{sup} 며(,) / ~ N_{sup} 이며(,)

· PART 는 휴리스틱으로 추출한 '한형', '하나', '분야', '총칭', '일종'등의 30 개의 어휘
· J 는 규칙 0, 1, 3, 4 에서 N_{sup} 다음에 기술된 조사 혹은 어미

상위개념 기술 패턴을 이용하여 정의문 기술 패턴을 생성한다. 사전 정의문에서는 문장의 앞에 용어가 보통 생략이 되기 때문에 이를 복원하여 다음과 같은 정의문 기술 패턴을 만든다.

· **패턴: 용어-조사 + 수식언 + 상위개념 + (부연설명)**

4.1.2 어휘구성정보를 이용한 용어의 상위개념 추정

앞 절에서 살펴본 바와 같이 복합명사로 구성된 용어는 마지막 명사가 상위개념을 추정하는 단서가 될 수 있다. 마지막 접미사를 통해서도 상위개념을 추정할 수 있다.

앞 절에서 만든 패턴으로 용어와 상위개념의 쌍을 추출한 후, 구성어휘에 대한 상위개념의 빈도를 구한다. 명사나 접미사로 어휘 L 을 포함한 용어가 상위개념 c 를 가지는 빈도를 Freq(L, c)라고 하자. Freq(L, c) 값을 이용하여 어휘 L 이 포함된 용어가 상위개념으로 c 를 가질 확률은 식 (1)과 같이 계산할 수 있다. 용어 T 가 주어졌을 때 상위개념을 추정하는 함수 f(T)는 식 (2)와 같다.

$$P(c \mid L) = \frac{P(L, c)}{P(L)} = \frac{1}{P(L)} \cdot \frac{Freq(L, c)}{\sum_{c'} Freq(L, c')} \quad \text{식 (1)}$$

$$f(T) = \arg\max_{c} P(c \mid T) = \arg\max_{c} \sum_{l \in T, \forall c} P(c \mid l) \quad \text{식 (2)}$$

4.1.3 용어의 상위개념에 따른 의미특질소의 분류

정의문에서 의미특질소를 이루는 요소는 다양하나 각 개념에 따라 의미특질소로 선호되는 요소가 있다. 예를 들면, 신체기관을 나타내는 용어는 그것이 어디에 위치해 있고 어떤 기능을 하는 지가 중요하고, 질병 이름을 나타내는 용어는 그것이 발생한 원인이 무엇이고 어떤 증상이 나타나는 지가 중요하다. 이러한 점에 착안하여, 훈련 사전정의문에서 용어에 대한 상위개념을 추출하고 각 상위개념에 대하여 정의문에서 공기하는 어휘를 수집한다. 이렇게 만든 개념별 특성 어휘 목록은 추출된 문장이 정의문의 의미특질소로 적합한지를 계산하는 데 쓰인다.

4.2 정의문 추출 및 순위화

용어에 대한 정의문을 추출하기 위하여 우선 코퍼스에서 용어가 포함된 용례를 모두 추출한 후, 정의문의 구문적 패턴을 이용하여 상위개념 후보를 추출한다. 상위개념 후보가 없을 경우에는 4.1.2 절에서 제안한 어휘구성정보를 이용하여 상위개념을 추정한다. 여러 상위개념이 추출된 경우에는 식 (1)의 값이 가장 높은 개념을 선택한다.

용어에 대한 상위개념이 추정되고 나면 정의문으로 가장 적절한 의미특질소를 선택한다. 용례의 수식언 M 에서 명사, 동사, 형용사를 키워드 K로 추출한 후, 상위개념 c에 대하여 식 (3)에 의해 P(K|c) 의 값을 계산한다. 수식언 M 에 대한 점수는 식 (4)에 의하여 구하고 그 값이 가장 높은 수식언 M 이 차별적 의미특질소로 선택된다. n_k 는 M에 등장하는 키워드 수로 수식언의 길이를 고려한 일반화를 위하여 삽입한다. w_k 는 키워드에 대한 가중치 값으로, 정의문에서 서술어가 중요하기 때문에 K가 명사일 경우보다 K가 서술어일 경우에 가중치를 높게 준다.

$$P(c \mid K) = \frac{P(c \mid K)}{P(c)} \times P(K) \approx P(c \mid K) \times P(K) \quad \text{식 (3)}$$

$$f(M \mid c) = \frac{1}{n_k} \sum_{K \in M} w_k \times P(c \mid K) \times P(K) \quad \text{식 (4)}$$

이와 같이 추출한 상위개념과 의미특질소를 결합하여 최종적으로 정의문을 제시한다.

5. 실험 및 결과

본 논문에서 제시한 정의문 추출 시스템을 평가해 보자. 우선 성능 평가를 위하여 평가용 용어를 선정하고 정답 코퍼스를 만든다. 대상 코퍼스는 Korterm 전문용어 의학 코퍼스 중 140 만 어절의 내과학 [11] 코퍼스를 사용하였다. 48 개의 용어에 대하여 정의문이 태깅되었다.

표 2 코퍼스에서 정의문 추출 결과

Set	정확률
Set I (36 개)	24/33 = .7273
Set II (12 개)	3.5/5.5 = .6363
전체 (48 개)	27.5/38.5 = .7143

- Set I: 정의문이 코퍼스에 등장하는 용어
- Set II: 상위개념이나 의미특질소 중 하나만 코퍼스에 등장하는 용어

표 2 는 48 개의 용어에 대한 코퍼스에서 정의문을 추출한 결과로, 상위개념이나 의미특질소가 맞을 경우에는 0.5 점, 둘 다 맞을 경우에는 1 점으로 계산하였다. 정의문이 존재하는 Set I 에 대하여 정확률이 약 73%로 어느 정도의 성능을 보였지만, 정확한 정의문을 뽑는 것이 중요하므로 앞으로 성능을 더 개선해야 할 것이다.

코퍼스에서 추출한 몇 개의 정의문은 사전의 정의문보다 좋았는데 그 예는 다음과 같다.

(예 5-1) 봉소염
- 코퍼스: 봉소염은 국소적 통증, 홍반, 종창, 열감 등을 특징으로 하는 피부의 급성염증이다
- 사전: 심부피하조직이나 가끔 근육의 급성 광범성, 확산성, 수종성, 화농성의 염증으로 농양형성을 동반하기도 한다.

코퍼스에서 정의문을 추출하는 작업은 신조어에 대한 정의문의 자동 생성이나 기존 용어 정의문의 재정비에 있어 매우 가치 있는 일이 될 수 있다는 점을 알 수 있다.

6. 결론 및 향후 연구

본 연구에서는 용어의 정의문을 자동으로 추출하기 위하여 텍스트 코퍼스로부터 용어 정의문에 관련된 정보를 자동으로 추출하는 방법을 제시하였다. 기존의 대부분의 정의문 추출 방법이 수동 또는 반자동으로 구축한 구문적 패턴만을 이용하는 것과는 달리 본 논문에서는 정의문의 구문적 패턴뿐만 아니라 용어의 어휘 구성 패턴, 정의문의 의미적 패턴까지 고려하여 정의문을 추출하였다.

앞으로 개념간의 상하위 관계, 유사 관계를 알 수 있는 시소러스를 잘 구축하여 이용한다면 상위개념과 의미특질소로 쓰이는 어휘를 제대로 얻을 수 있을 것이다.

감사의 글

본 연구는 과학기술부, 과학재단, 한국과학기술원 BK21 정보기술사업단의 지원을 받았습니다.

참고 문헌

[1] Muresan, S. and Klavans, J., "A Method for Automatically Building and Evaluation Dictionary Resources", In International Conference on Language Resources and Evaluation 2002, 2002.

[2] Sato, Satoshi, "Automated Editing of Hypertext Resume from the World Wide Web", Symposium on Applications and the Internet, 2001.

[3] Fujii, Atsushi and Tetsuya Ishikawa, "Organizing Encyclopedic Knowledge based on the Web and its Application to Question Answering", ACL-EACL 2001, 2001.

[4] ISO 704: Terminology work Principles and methods, 2000.

[5] Pearson, Jennifer, "Terms in Context", Amsterdam/ Philadelphia: John Bejamins, 67-88, 1998.

[6] WordNet http://www.cogsci.princeton.edu/~wn/

[7] Fellbaum, Christiane," WordNet - an Electronic Lexical Database" , The MIT Press Cambridge, Massachusetts London, England, 1-46, 1998.

[8] 생물학사전, 아카데미 서적 (한국생물과학협회), 1998.

[9] 두산세계대백과 EnCyber, http://kr.encycl.yahoo.com/

[10] 표준의학사전, 아카데미 서적, 1993.

[11] 내과학 I, II, 해리슨, 내과학 편찬위원회 편, 정담, 1997.