텍스트로부터 용어 정의문의 자동 추출 방법

신효식[0] 김재호 이해윤 최기선
한국과학기술원 전산학과 전문용어언어공학연구센터
{gerling, jjaeh, haeyun, kschoi}@world.kaist.ac.kr

A Method for Automatic Extraction of Term Definition from Text

Hyo-Shik Shin[0] Jae-Ho Kim Hae-Yun Lee Key-Sun Choi
KORTERM / KAIST
Dept. of Computer Science

요 약

본 연구는 텍스트 코퍼스로부터 용어의 정의를 자동으로 추출하여 용어의 자동 추출기술과 통합하여 다목적의 용어뱅크를 구축하기 위한 목적으로부터 출발하였다. 지식정보의 확산에 따라 기존 전문분야 용어집에 수록되지 않은 용어의 수는 폭발적으로 증가하고 있다. 기존의 용어집 혹은 용어사전의 디지털화만으로는 새로운 전문용어의 포괄성에서 한계가 있는 것이다. 정보의 획득이라는 면에서 보면 이러한 한계를 극복하고 모든 용어에 대해서 즉시적으로 용어의 정의를 제공받는 것이 바람직하다. 자동으로 구축된 용어집의 응용은 여러 가지로 기대된다. 새로운 용어에 대한 의미 파악을 위해서는 물론, 확장된 전문용어집의 작성이나 전문분야 온톨로지의 구축 등에도 이용될 수 있다.

1. 서론

본 연구는 텍스트 코퍼스로부터 용어의 정의에 해당하는 정보를 자동적으로 추출하기위한 시스템을 구현하는데 목표를 둔다. 따라서 지식구축의 자동화를 위한 연구의 일환이다.

지식 정보의 확산에 따라 기존 전문분야 용어집에 수록되지 않은 용어의 수는 폭발적으로 증가하고 있다. 일반적으로 전문용어 사전 혹은 용어집은 전문가 집단의 합의에 따라 장기간의 계획하에 편찬되기 때문에 수록된 용어는 안정화된 것들로 대상이 한정된다. 즉 지면의 한정, 시간상의 제약이라는 현실적인 한계를 안고 있다. 정보의 획득이라는 면에서 보면 이러한 한계를 극복하고 모든 용어에 대해서 즉시적으로 용어의 정의를 제공받는 것이 바람직하다.

기존에 용어의 자동 인식 혹은 추출에 대한 연구는 많이 이루어져 왔다. 이에 비해 용어 정의문의 자동 추출에 대한 연구는 미약한 편이다.

자동으로 구축된 용어집의 응용은 여러 가지로 기대된다. 새로운 용어에 대한 의미 파악을 위해서는 물론, 새로운 전문용어집의 작성이나 전문분야 온톨로지의 구축 등에도 이용될 수 있다.

정의문 추출의 연구는 한편으로 개별언어적 언어현상을 반영해야 하며, 다른 한편으로는 분야특수적 텍스트 현상을 반영해야 한다는 점에서 미시적 연구가 필요하지만, 기술개발의 방법론적인 면에서는 개별언어 및 제반 분야를 포괄하는 거시적 연구가 가능하다. 본 연구에서는 한국어 뉴스 기사, 특히 정보과학 섹션의 텍스트로부터 급증적으로 늘고 있는 이 분야 용어의 정의문 정보를 추출하고자 한다. 이 연구는 텍스트 스타일과 같은 몇 가지 특징만 고려한다면 학술분야, 여타 기술 분야 텍스트에까지 확장될 수 있다는 기대에서 비롯되었다.

본 연구의 구성은 다음과 같다.

2장에서는 정의문의 자동 추출에 관한 기본 연구를 개관하며, 새로운 방향의 접근방법론을 동기화한다. 3장에서는 용어 정의에 대한 여러 정의 모형을 비교하고, 한국어에서 정의문은 어떻게 실현되는지를 살펴본다. 4장에서는 3장에서 살펴본 한국어의 정의문 형식을 패턴화하여 정의문 추출 알고리즘을 수립한다. 5장에서는 실험 결과를 보여주고 그 결과를 분석함으로써 성능향상을 위한 방안을 모색한다. 마지막으로 결론 및 향후 연구과제를 정

리한다.

2. 관련연구

언어별로 다른 목적이긴 하지만 텍스트 코퍼스로부터 용어의 정의 정보를 추출하려는 일련의 연구들이 진행되고 있다.

[5]는 일본어의 방송용 뉴스원고를 이용하여 용어집을 자동구축하기 위한 연구이다. 뉴스 원고 중에서 용어를 설명하기 위한 표현을 3가지 패턴으로 분류하여 패턴매칭을 통해 용어 정의문을 추출한다. 추출대상을 강조를 의도하는 꺽쇠괄호로 묶여진 명사구에 한정하여 정의문을 추출한다는 점에서 이러한 용어 표지가 없는 무표지 텍스트에 대해서 어떻게 적용될 수 있는지는 미지수다. 그렇지만, "Term은 NP이다" 패턴이 아닌, "Term은 VP-다" 패턴과 같이 소위 비정형 정의문 형식으로부터도 정의문 추출을 시도했다는 점에서는 정의문 추출의 심층적 연구로 간주할 만 하다.

[6]는 중국어의 코퍼스 기반 용어은행(term bank)의 구축에 관한 설계를 소개하고 있다. 정보과학 기술분야에 한정된 코퍼스를 기반으로 훈련코퍼스와 실제 정보를 추출하기 위한 대용량 확장코퍼스를 구축하였다. 훈련코퍼스에 대해서는 용어정의를 위한 참조 주석화를 통해서 일정한 패턴화를 시도한다. 문장, 문단 혹은 완전 텍스트가 용어정의의 참조용으로 주석화 될 수 있다. 이러한 주석화는 용어 정의문 템플릿을 찾기 위한 중간과정에 해당한다. 이러한 주석화 코퍼스를 통한 기계적 학습 기제는 타 영역, 타 언어에로 확장적용 될 수 있다는 점에서 바람직하기는 하지만, 고비용과 많은 시간의 투자가 요구된다는 점에서 한계를 안고 있다. 또한 이 시스템은 아직 구현되지 않아 실험결과가 전혀 없는 실정이다.

[11][12][14]은 의학분야에 대해서 규칙기반 텍스트 마이닝 기법을 이용하여 비전문가용 디지털 영어 용어사전을 구축하려는 DEFINDER 개발에 관한 연구이다. 비전문가인 사용자를 위한 의학분야 텍스트로부터 한편으로는 약식 텍스트처리 모듈을 통해 용어 정의문 패턴, 혹은 텍스트 표지를 통해 정의문을 추출하고, 다른 한편으로는 의존적 어휘문법에 기반한 문법분석 모듈을 통해 복합적인 언어현상인 동격, 조응사 문제까지 포괄하여 용어 정의문을 추출 한다. 이 연구는 텍스트에 등장하는 보다 많은 유형의 정의문 정보를 추출할 수 있는 다양한 장치를 마련했다는 점에서 가장 발전적인 연구라고 평가할 수 있다.

이상에서 개관했듯이, 여러 언어에서 용어의 정의문 추출에 관한 연구가 지식구축의 일환으로 추진되고 있음을 알 수 있다. 그러나, 아직 한국어에 관해서는 용어 추출에 관한 연구는 되고 있지만 [15], 정의문의 자동 추출에 관한 연구가 거의 없다는 점에서 본 연구는 그 시발점이 될 것이다.

3. 용어 정의의 정의 및 언어적 표현

3.1. 용어의 정의

언어기호의 정의는 의미에 관련된 것으로 전통적으로는 철학과 언어학의 주된 연구 대상이었다.[1] 용어는 단어에 상응하는 전문분야의 언어기호로 정의되며, 사전학(lexicon), 사전편찬학(lexicography)에서 혹은 학문적 글쓰기에서 중요하게 다루어져 왔다. 예를 들어, 용어담당 국제표준화기구 기술위원회인 ISO/TC 37의 국제규격 ISO 704 [8]가 사전학에 관한 것이라면, ISO 1087-1 [9], ISO 10241 [10]은 사전편찬학의 관점에서 용어의 정의 문제를 다룬다.[2] 학문적 글쓰기에서는 사전에서 언급하는 용어의 정의 방식을 따르기 보다는 논리적인 내용 전개나 내용의 명확성을 위해서 용어의 정의를 어떻게 제시할 것인가에 관한 실질적인 문제가 다루어 진다. 본 연구에서 목표로 삼고 있는 코퍼스로부터 용어의 정의 정보 추출을 위해서는 학문적 글쓰기에 대한 이해가 보다 중요하다. 코퍼스란 글쓰기의 결과물이기 때문이다.

용어 정의문의 형식 및 내용에 관해서는 여러 견해들이 있지만, 정의 표현의 추상적 모형에 대해서는 대략의 합의가 도출되었다. 본 연구에서는 ISO 1087-1, ISO 704 에서 규격화하여 권장하고 있는 정의문 형식과 부합되는 정의 정보만을 추출대상으로 삼는다.[3] 전통적인 정의 방식에 기초한 ISO 704 규격의 정의문 형식은 일종의 템플릿으로써 다음과 같다.

(1)
X = Y + 차별적 의미특질소(distinguishing characteristics)

여기서 X는 정의될 용어를 말하며, Y는 X에 대한 상위어이며 '차별적 의미특질소'란 동위어(Cohyphonyme)들로부터 그 용어를 구별해주는 특징

---

[1] 정의도 Austin의 화행이론에 따르면 일종의 화행행위이다. « defining exercitive »와 « defining expositive »가 있다. 전자는 개념의 정의를 처음 내릴 때, 후자는 이미 내린 정의를 설명목적으로 되풀이 할 때 사용된다.

[2] 일반 사전에서의 정의문제는 [16](3장)를 참조.

[3] Cobuild 사전방식의 설명적 정의에 대해서는 [16](3장)를 참조.

적인 의미속성을 말한다.[4]

위의 정의문 형식은 세 개의 슬롯을 가지고 있다[16]. 'X', 'Y', '=' 가 그것인데, 이 슬롯들의 'filler' 즉, 언어적 실현체는 일정한 조건을 만족해야 한다. 이 조건은 언어 보편적인 측면도 있지만, 개별 언어적 특성도 갖는다. 이를테면, 'X' 의 슬롯 필러는 언어보편적으로 용어여야 한다. 이 용어 명사의 앞에 부정관사가 나와야 되는지 여부는 언어에 따라 다르다. 'Y' 의 슬롯 필러는 용어이거나 특정한 분류어(class word)이어야 한다. 분류어로는 '과정(process)', '방법(method)', '기능(function)', '속성(property)', '체계(system)' 과 같은 일반적 용어가 해당된다. 이들 일반적 용어는 관사류를 제외한 어떠한 수식어도 있어서는 안 된다는 영어의 특성도 있다. 동치관계를 나타내는 '=' 의 슬롯 필러가 될 수 있는 동사(구)는 '연결동사(hinge/connective verb)' 라고 불리는 동사들이다. 영어의 경우 'be', 'mean', 'consist of' 등의 동사가 해당된다.

더 나아가 연결동사는 현재시제, 직설법 형태를 가져야 한다. 조동사는 can만 허용된다. 이상의 언어적 특징은 정의의 일반성, 항구성과 관련 있다.

다음은 위의 정의문 형식이 구체적으로 실현되는 예들을 영어를 대상으로 제시한다.

(2)

a. A knife is an instrument which is used for cutting
b. Aluminium is a metal produced from bauxite
c. A dentist is a person who takes care of people's teeth. [16]

위 예에서 용어('knife', 'aluminium' 등)는 문장의 주어로서, 연결어(=)는 be 동사로 실현되고 있다. 그리고, 정의부분은 관계절의 수식을 받는 명사구로 일정하게 나타나고 있다. 이처럼 정의문 형식에 대한 모든 슬롯 필러가 한 문장 안에서 실현되기도 하지만, 문장 경계를 벗어나 실현될 수도 있다.

(3)

During sexual reproduction, a male sperm joins with a female egg. This is called fertilization. [16]

위 예문에서는 정의문 형식이 지시대명사를 매개로 두 문장에 걸쳐 실현된 경우이다.[5]

[4] 어휘 의미론적으로 말하자면, X의 내포적 의미자질 집합(set of intensional semantic features) 중에서 Y의 내포적 의미자질 집합을 제거하고 남은 유일적인 의미자질을 말한다.

3.2 한국어의 용어 정의의 언어적 표현

앞 절에서 살펴본 정의문 형식이 한국어 텍스트에서 어떤 언어표현 형식으로 실현되는지 훈련코퍼스 분석을 통해서 살펴보았다. 연합뉴스 정보과학 섹션의 36개의 기사를 훈련코퍼스로 선정하였다. 이 코퍼스는 6,087어절로 구성되어 있으며, 53개의 전문용어 정의문이 포함되어 있다. 시사적인 내용을 간결하게 표현하는 뉴스기사의 글쓰기 방식이 학술 텍스트와는 다르다는 점에서 장르 한정적인 특징도 반영되어 있다.

우선적으로 ISO 규격에서 규정한 형식 (1)에 부합하는 경우를 살펴보면 다음과 같다. 정의문 구성의 요소들이 모두 나타나는 경우는 다음의 통사적 패턴으로 제시될 수 있다.

(4) Term-은 $[\text{Rel N}]_{NP}$-다

a. 모바일 그룹웨어 ASP는 기업들이 그룹웨어를 무선인터넷 환경에서도 이용할 수 있도록 시스템을 구축, 관리, 운영해 주는 서비스다.
b. 벼 입선충은 볍씨나 왕겨에 달라붙어 월동한 뒤 종자가 발아하면 밖으로 나와 잎과 이삭 등에 기생, 벼잎을 말라 비틀어지게 하거나 쌀알을 검게 변색시켜 미질을 떨어뜨리는 등 큰 피해를 주는 해충이다.
c. ㈜지씨티세미컨덕터는 국내 기술과 인력으로 지난 98년 미국 실리콘밸리 산타클라라에 설립한 무선통신용 반도체 개발 전문회사이다.

(4) 예문에서는 관형절을 통해 차별적 의미특질소가 표현되고, 상위어가 등장한다. 아래 (5) 예문도 마찬가지 구조를 갖지만 Y가 '방법', '것', '제품' 과 같은 일반적 분류어를 취한다.

(5)

a. 사전 냉각은 혈액이 근육에서 비켜가는 것을 지연시켜, 핵심온도를 낮추는 방법이다.
b. M-커머스는 이동통신 단말기를 통해 각종 서비스를 이용하는 것을 말한다.
c. GSM 단말기는 16화음 멜로디 기능을 채용한 듀얼폴더형으로 4그레이 LCD(액정화면)와 7가지 색상의 발광다이오드(LED)를 채택, 고급화를 추구한 제품이다.

[5] 정의문 형식은 많은 경우에는 불충분하게, 혹은 변형된 형태로 실현되기도 한다.

i. Digital transfer links are used to interconnect interface adaptors to form signaling data links.
ii. Graphite has a very high melting point.

위 예문에서는 상위어가 없이 특질소만으로 정의정보가 표현된 경우이다.

또한, 정의문 정보를 문장성분의 일부분에서도 추출할 수 있다. 문장의 일부분에 포함되어 나타나는 경우로서 상위어와 차별적 의미특질소가 관계절 형태로 나타난다.

(6)

a. [유방암 치료에 효과가 있는 것으로 알려진 신약인 카페시타바인(capecitavine)이] 영국에서 처음으로 시판에 들어갔다고 BBC 방송이 14일 보도했다.

b. [EMS는 전자제품 생산을 위탁받아 자사 상표없이 제조를 전문으로 하는 기업으로] 주문자상표부착생산(OEM)과는 달리 전세계를 상대로 한 다품종 다량생산체계이며 일부 설계 등까지 수행하면서 대량구매 등을 통해 가격경쟁력을 제고하는 방식으로 운영된다.

(6)a에서는 용어의 정의가 관형절의 형태로 표현되고 있다. 즉, "카페시타바인"은 "유방암 치료에 효과가 있는 것으로 알려진 신약이다"라고 정의된다. (6)b에서는 자격격조사 "-으로" 부가구를 통해 일종의 정의가 표현된다. 즉, "EMS"는 "전자제품 생산을 위탁받아 자사 상표없이 제조를 전문으로 하는 기업이다"라고 정의될 수 있다.

그러나, 코퍼스상에서 등장하는 용어의 정의문 패턴은 앞서 제시된 표준적 정의문 형식에서 벗어나는 경우가 많다. 그 첫 예로써 다음 (7)에서 보는 바와 같이, 정의문 내에 상위어가 나타나지 않는 경우이다.

(7)

a. [림프절 페스트는 쥐 등 갉아먹는 설치류 동물의 벼룩에 의해 전염된다].

b. 기업은행은 [해외로 송금하거나 해외로부터 송금을 받아야 하는 고객이 10분 이내에 대금을 찾을 수 있는 '초고속 해외송금 서비스'를] 제공하기로 했다고 16일 밝혔다.

(17)에서는 정의 정보가 "Term은 VP-다" 형식으로 표현되고 있다. 즉, 정의 부분이 (17)a에서는 "전염되다"라는 동사로 끝나며, (17)b에서는 "찾을 수 있다"라는 복합 동사구조로 끝나고 있다.

또한 다음 예에서 보는 바와 같이, 차별적 의미특질소가 나타나지 않고 상위어만 등장하는 경우가 있다.

(8)

충북 제천 지역의 한 축사에서 최근 [2종 가축 전염병인 돼지 오제스키 병이] 발생했으나 돼지가격 하락을 우려한 시가 이 사실을 은폐해 비난을 받고 있다.

위 예문에서는 "돼지 오제스키 병"이 용어이며, 관형절의 "2종 가축병"은 상위어에 해당한다. 사실 한국어에서는 연결동사로서 계사(copular) "이다"가 등장하는 경우에 모호성이 발생한다. 한국어에서는 관사의 문법적 기능이 미약하므로 주어와 술어명사구의 관계가 불명료하다[6]. 다음 예문은 동치관계를 보여준다.

(9)

… 유럽에서는 [세계 최대 이동통신 네트워크 정비업체인 에릭슨의] 신용등급이 정크수준으로 하향조정되는 등 세계통신업계가 위기를 맞고 있다고 전했다.

위 예문에서 "에릭슨"이 용어이며, "세계 최대 이동통신 네트워크 정비업체"와는 동치관계에 있다. 동치관계는 정의문으로 간주될 수 있지만, 상하관계는 정의문으로 간주될 수 없다.[7] 그러나, 통사적 표지가 아닌 의미적 판단을 동원해야 한다는 점에서 동치관계 여부를 판정하기란 쉽지 않다.[8]

이상으로 한국어에서 정의문 형식이 표현될 수 있는 여러 가지 표현 가능성을 형태 통사적인 측면에서 살펴 보았다. 정의 정보의 추출 관점에서 본다면, 정의 정보를 패턴화하여 단순히 형태 통사적인 형식을 통해서만 특징짓는 것으로는 충분하지 않다.

예를 들어, (7)의 상위어가 생략된 구조에서 처럼 정의 정보가 "VP"로 표현되는 경우에는 비정의적인 절대 다수의 문장과 구별될 수 없다. 이 경우에는 어휘 의미적인 기준이 추가되어야 한다. 혹

---

[6] 영어나 독일어와 같이 관사의 사용이 문법화된 언어에서는 계사의 술어명사구에 관사의 종류에 따라 동치관계인지 포함관계(혹은 상하관계)인지가 결정된다. [1] 참조

[7] 상하관계는 동격구문에서도 발견된다.
'온라인게임 업체 (주)엔씨소프트는 상반기 실적 집계결과 매출액이 증가했다고 밝혔다.'에서 '(주)엔씨소프트'가 용어라면 동격의 '온라인게임 업체'와는 포함관계에 있으며 따라서 상하관계에 있다고 이해된다.

[8] 관형절로부터 정의문을 파악하는 것도 한국어에서는 쉽지 않다. 한정절 혹은 제한절 등 구분을 통해 선행명사에 대한 관계절의 수식범위가 결정되지만, 한국어에서는 의미특질소에 해당하는 관형절의 의미가 선행사(즉 상위어) 모두에 해당하는지 여부를 문법적인 표지로 알 수 없기 때문이다.

은 문맥을 분석하여서 비로소 용어의 정의 정보를 획득할 수 있는 경우가 있다.[9]

본 논문에서는 어떤 의미적 정보가 정의와 관련되는 것인지 다루지 않았으며[10] 문맥 분석을 통해서 비로소 용어의 정의 정보를 획득 할 수 있는 경우도 다루지 않았다.

4. 시스템의 구현

본 장에서는 용어 정의문 추출을 위해 구현한 시스템에 대하여 기술한다. 먼저 전체 시스템 구조도를 살펴 보고 시스템을 이루는 세부 모듈을 자세히 살펴본다.

4.1 시스템 구조도

본 연구에서 제안하는 시스템의 구조도는 그림 1과 같다.

시스템은 크게 패턴 생성기, 전처리기, 정의문 추출기 세 부분으로 이루어져 있다. 먼저 패턴 추출기에서는 용어 정의문이 부착된 코퍼스를 입력으로 받아 정의문을 이루는 일반적인 패턴을 빈도수를 고려하여 뽑아낸다.

전처리기에서는 정의문을 추출할 패턴을 적용하기 위한 준비 작업을 한다. 정의문을 추출할 대상 문서에 대해 형태소 분석 및 태깅을 한 후 전문용어 추출기와 명사구 인식기를 이용하여 대상문서에서 전문용어와 명사구를 인식하여 용어 정의문 추출기에서 패턴을 적용하여 용어 정의문을 추출한다.

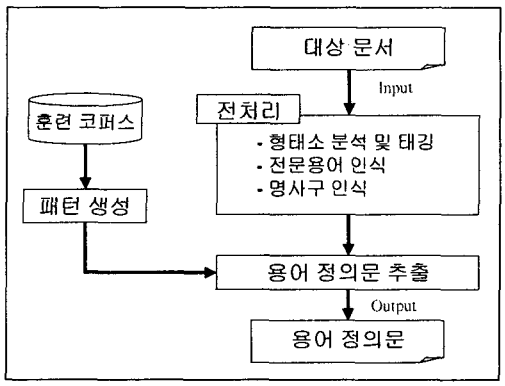

그림 1 용어 정의문 추출을 위한 시스템 구조도

4.2 정의문 표현의 패턴화

앞 장에서 형태 통사적 측면에서 다양한 방식의 언어적 표현으로 정의문의 정보가 표현되고 있음을 훈련코퍼스에 기초하여 살펴보았다. 본 실험에서는 앞서 (1)에 제시된 형식의 슬롯 필러가 모두 등장하는 경우를 주요 연구대상으로 삼고자 한다. 약간의 어휘적 제약만 가한다면 패턴화 할 수 있다는 점에서 정의 관련 정보의 자동추출화를 위한 초기 단계 연구로 적절하기 때문이다.

즉, 정의 관련 정보의 자동 추출을 위한 1차적인 패턴은 다음과 같다.

(11)
a. Term-은/는/이/가 [Rel N]$_{NP}$ -이다
b. [[Rel N]$_{NP}$인 Term]$_{NP}$

이 때, 미래나 추정을 나타내는 단어인 '방침', '예정', '계획' 등은 일반적인 분류어가 될 수 없으므로 NP의 핵어인 N이 될 수 없다. 또한 "이다" 대신에 "~라 알려져 있다" 또는 "불린다" 등의 상당어가 쓰일 수 있다. 이러한 어휘적 제한 혹은 확장은 보다 큰 코퍼스의 학습을 통해서 추가적으로 작업되어야 한다.

더 나아가 의미적 특질소는 등장하지 않지만, 상하관계 혹은 동의어를 뜻하는 다음 패턴도 포함하도록 한다.

(12)
a. Term-은/는/이/가 NP-이다

---

[9] [웹사이트의 유저 프라이버시 보호가 어느 정도 이뤄지는 지를 쉽게 체크할 수 있는 국제 표준]이 지난 5년여의 줄다리기 끝에 16일 확정돼 온라인 프라이버시 보호에 새로운 발판을 마련했다. 월드 와이드웹 컨소시엄은 이날 '플랫폼 퍼 프라이버시 프리퍼런시스' (Platform for Privacy Preferences; 일명 P3P)를 확정했다고 발표했다.

위 예에서는 용어 "플랫폼 퍼 프라이버시 프리퍼런시스"의 정의 정보가 앞 문장의 [] 안에 제시되고 있다. 조응소를 통해 정의문이 표현되는 경우도 문맥분석을 통해 탐지가 가능하다.

[10] 실제로 전문용어 사전에서는 동일한 분야의 정의문에 등장하는 차별적 의미 특질소가 서로 다른 경우가 많다. ETRI와 계몽사가 공개한 백과사전(1997)에 따르면, 인간질병명의 정의에 원인, 증상 등의 상이한 의미특질소가 사용되고 있다.

i) 습진: [피부의 염증으로 생기는]$_{원인}$ 피부질환
ii) 야뇨증: [밤에 잠을 자다가 무의식 중에 오줌을 자주 싸는]$_{증상}$ 병증

i)에서는 '원인', ii)에서는 '증상'에 해당하는 정보가 정의에 이용되고 있다. 생성사전론의 틀 안에서 행한 정의에 관련된 다양한 의미정보의 구분에 대해서는 [4]를 참조하라.

b. NP-인 Term

단순한 이 패턴은 비록 충분한 정의 관련 정보를 제공하지는 않지만 용어의 온톨로지 구성에도 중요하고, 정의문 생성에도 중요한 상위어 혹은 동의어를 제공해 줄 수 있기 때문이다.

본 논문에서는 위 4개의 패턴으로 대상 문서에서 정의 관련 정보를 추출하고, 슬롯 필러로 상위어 혹은 일반 분류어가 등장하지 않는 패턴을 제외하였다.

4.3 전처리기

대상문서에 정의문 추출 패턴을 적용하기 위해서는 각 문장에 대해서 전문 용어와 명사구 단위가 인식되어 있어야 하며, 각 단어들도 형태소 태깅이 되어 있어야 한다.

이에 전처리기에서는 먼저 대상 문서를 한나눔[2]을 사용하여 형태소 분석 태깅을 한다. 전문용어 추출기[15]로 형태소 태깅된 결과에서 전문용어를 인식하여 정의문 추출기에서 그 전문용어들에 대해 정의문을 추출하도록 한다.

명사구 단위는 연속된 명사들을 묶어 주는 간단한 규칙으로 구현한 명사구 인식기를 사용하여 인식한다.

4.4 용어 정의문 추출

정의문 추출기에서는 전처리기를 거친 문서에 대해 앞에서 생성한 패턴을 적용하여 전문 용어에 대한 정의문을 추출한다. 같은 문장에서 여러 개의 정의문이 추출될 경우 각 정의문에 대한 점수를 계산하여 가장 높은 점수의 정의문을 선택하여야 하지만 본 논문에서는 전문용어를 포함한 문장이 정의문인지 아닌지를 판단하는 분류의 문제로 보고 정의문을 포함한 문장을 모두 추출한다.

5. 실험 및 결과

이 장에서는 본 연구에서 제시한 정의문 추출 시스템을 평가한다.

평가는 두가지 코퍼스에 대하여 진행하였다. 하나는 패턴을 추출하기 위해 사용한 앞에서 설명한 훈련코퍼스이고 다른 하나는 성능평가를 따로 제작한 평가코퍼스이다. 평가코퍼스의 모법답안(gold standard)는 두명의 언어학자가 참여하여 만들었다. 형태 통사론적 혹은 의미적 기준을 동원하여 각자 정의 관련 정보를 주석화하였으며, 차이나는 부분에 대해서는 합의를 통해서 수정안을 만들었다.

훈련코퍼스는 패턴이 얼마나 잘 만들었는지를 살피고, 시스템 성능 향상을 위해서 사용되었다. 그리고 평가코퍼스를 통해 제작된 패턴으로 정의문이 잘 추출되는지를 살핀다.

연합뉴스 정보과학 섹션의 158개의 기사를 정의문 추출을 위한 평가코퍼스로 선택하였다. 이 문서는 17,354어절로 구성되어 있고 63개의 전문용어 정의문을 포함하고 있다.

실험 결과는 표 1과 같다.

표 1 용어 정의문 추출 실험 결과

	훈련코퍼스	평가코퍼스
전체 정의문 수	53개	63개
시스템이 제시한 정의문 수	24개	39개
시스템이 정확히 제시한 정의문 수	22개	36개
정확률	91.67 %	92.31 %
재현율	41.51 %	57.14 %

패턴을 몇 개 밖에 적용하지 않아 재현율이 낮음을 알 수 있다. 그러나 단 4개의 적은 수의 패턴으로도 훈련코퍼스에서 41%, 평가코퍼스에서 57%의 결과를 얻었다는 것은 뉴스 기사에서 정의문은 비교적 일정한 패턴을 가지고 있다는 것을 알 수 있다. 정확률은 두 코퍼스에서 모두 90%가 넘는 좋은 결과를 보였다. 사용한 패턴이 일반적이고 정의문 추출에 적합하다는 것을 알 수 있다.

다음은 뽑혀진 정의문의 예를 보여준다.

> * ITS는 *카메라, 감지기 등을 통해 수집된 차량 속도, 교통량 등 실시간 교통정보를 휴대폰과 인터넷, 전광판, ARS 등을 통해 운전자에게 전달하는 시스템이다.*
>
> * *벼바이러스병을 옮기는 주요 매개충인* 애멸구가 2천548마리로 지난해보다 20%, 끝동매미충은 2천278마리로 50% 증가했다.

어떠한 패턴으로 이루어진 정의문이 추출되었는지를 살펴보기 위해 표 2와 같이 각 패턴에 의해 뽑아진 정의문의 수를 조사하였다. 괄호안의 숫자는 틀린 정의문의 수이다.

표 2 각 패턴에 의해 뽑아진 정의문의 수

패턴	훈련코퍼스	평가코퍼스
Term은 $[\text{Rel N}]_{NP}$-이다	5(0)개	6(0)개
$[[\text{Rel N}]_{NP}$인 Term$]_{NP}$	4(1)개	1(0)개
Term은 NP-이다	3(0)개	0(2)개
NP-인 Term	12(1)개	29(1)개

짧은 시간에 많은 정보를 전달해야 하는 뉴스기사의 특성상 정의문은 "NP인 Term"인 형태로 가장 많이 쓰이는 것을 볼 수 있다.

시스템이 제시한 정의문 중 틀린 정의문 하나를 살펴보자.

> * 재정경제부 관계자는 8일 '연초 계획대로 정보통신부가 [*시장지배적 사업자인*] SK텔레콤으로부터 상반기 경영실적을 9월말까지 제출받아 1~2개월의 검토를 거쳐 요금을 인하할 방침'이라고 말했다.

위 문장은 "NP인 Term" 패턴에 의해 정의문으로 추출되었지만, "시장지배적 사업자"는 정의문이 되지 않는다. 정의문이 될 수 있는 형식이나 어휘를 좀 더 학습하여 구체화시킬 필요가 있다.

다음은 패턴에 없어서 시스템이 뽑지 못한 정의문들의 예이다.

> * *성인 당뇨병 환자의 포도당을 측정하는* 글루코시계 측정기는
>
> * 애벌구는 *국내에서 월동하는 해충으로 직접적인 피해는 없지만 벼줄무늬잎마름병과 벼검은줄오갈병, 보리복지모자익병 등 바이러스에 의한 병을 옮기는 매개곤충 역할을 한다.*
>
> * *어항을 개발하고 남은 부지*(배후부지)
>
> * 이 병은 *구제역, 콜레라 등과 함께 전염성이 강한 2종 가축 전염병으로 임신한 돼지의 경우 유산이나 사산을 일으키고 새끼돼지는 치사율이 매우 높은 것으로 알려져 있다.*

"Rel Term", "Term은 VP" 등의 패턴으로 정의되는 정의문을 많이 볼 수 있는데 이것은 정의문외에도 많이 서술되는 너무 일반적인 형태이어서 아무런 제약없이 쉽게 사용할 수 없다.

3번째 예와 같이 용어를 설명하고 괄호로 용어를 쓰는 형태도 좋은 패턴이 될 수 있을 것이다. 4번째 예와 같이 용어를 앞 문장에서 언급하고 대명사를 써서 다음 문장에서 정의를 하는 경우도 많이 있었다. 이를 처리하기 위해선 조응해소가 미리 수행되어야 하지만 조응해소가 그리 쉬운 문제는 아니다. 앞으로 위와 같은 예들을 처리할 수 있는 패턴을 더 생성하고 정교화한다면 더 좋은 결과를 기대할 수 있을 것이다.

6. 결론 및 향후 계획

본 연구에서는 용어의 정의문을 자동으로 생성하기 위해서 텍스트 코퍼스로부터 용어의 정의 관련 정보를 자동으로 추출하는 방법을 제시하였다.

훈련 코퍼스를 분석하여 만들어낸 정형적인 정의문 패턴 4개만을 통해 정보과학 분야 뉴스 기사에서 비교적 높은 정확률의 정의문 정보를 추출할 수 있었다. 추출된 정의문 정보를 토대로 정의문 생성하는 것도 중요한 모듈이지만 본 연구에서는 다루지 못했다.

문맥 분석이 필요하거나 혹은 'Term은 VP-다"와 같은 비정형 정의 형태의 정보까지 자동 추출할 수 있도록 패턴을 더 추가하는 문제는 향후 과제로 남아 있으며, 이 경우 문장내의 의존관계, 시소러스의 이용도 함께 고려해 보아야 할 것이다. 더 나아가 조응사 해결 모듈을 적용하여, 복합적인 정의문 정보도 추출할 수 있도록 이 시스템은 확장되어야 한다. 추출된 정보가이 단순 설명문인지 정의문인지를 판별할 수 있는 모델을 통해 패턴을 더 정교화한다면 정의문 추출 시스템의 성능은 향상될 것이다.

감사의 글

본 연구는 전문용어언어공학연구센터에서 수행한 문화부의 21세기 세종계획 "전문용어 정비", 과학재단의 국제공동연구사업 "자연언어처리 기반 이동통신 시스템에 관한 기초 연구", 과학기술부의 뇌신경정보학사업 "인간의 지식처리 모델링을 위한 전문분야 지식 베이스 원형 구축 및 활용" 과제의 일환으로 수행되었습니다.

7. 참고 문헌

[1] 신효식. 최소주의 문법에 입각한 소위 계사 'sein'의 통사적 접근, 독일문학 56집, 1995.

[2] 이운재, 김선배, 김길연, 최기선. 모듈화된 형태소 분석기의 구현. 제 11회 한글 및 한국어 정보처리 제 1회 한국어 형태소 분석기 및 품사태거 평가 워크숍 논문집. P.123-136, 1999.

[3] 이익섭, 임홍빈. 국어문법론, 학연사, [8]1988.

[4] 이해윤 외. 전문용어 정의의 정의. 세미나 초록, KAIST/KORTERM, 2002.

[5] 山田一郎, 田正啓, 金淵培. 뉴스원고를 이용한 용어집 작성 검토. NHK방송기술연구소, ms, 2001.

[6] Bai Xiaojing, Hu Junfeng, Zan Hongying, Chen Yuzhong and Yu Shiwen. A Corpus-based Approach to Term Bank Construction. In: LREC 2002 Workshop Proceedings of

International Standards of Terminology and Language Resources Management, 80-83, 2002.
[7] Barrière, C. and Hermet, M. Causality taking root in Terminology, ms., 2002.
[8] (-) ISO 704: Terminology work - Principles and methods, 2000.
[9] (-) ISO 1087-1: Terminology work - Vocabulary - Part 1: Theory and application, 2000.
[10] (-) ISO 10241: International terminology standards - Preparation and layout, 1992.
[11] Klavans, J. and Muresan, S. DEFINDER: Rule-based methods for the extraction of medical terminology and their associated definitions from on-line text. In: Proceedings of AMIA Symposium 2000.
[12] Klavans, J. and Muresan, S. Evalution of DEFINDER: A system to mine definitions from consumer-oriented medical text. In: Proceedings of The First ACM+IEEE JCDL 2001.
[13] Landau, S. I. Dictionaries: The Art and Craft of Lexicography, Cambridge: Cambridge University Press, 1984.
[14] Muresan, S. and Klavans, J. A Method for Automatically Building and Evaluation Dictionary Resources. In: LREC 2002, 231-234, 2002.
[15] Oh, Jong-Hoon, Key-Sun Choi. Automatic Terminology Recognition: Using the Lexical Resource of the Specific Fields. Second International Conference on Language Resources and Evaluation, Terminology Resource and Computation Workshop, (WTRC), 2000.
[16] Pearson, J. Terms in Context, Amsterdam/ Philadelphia: John Bejamins, 1998.
[17] Trimble, L. English for Science and Technology: A Discourse Approach. Cambridge: Cambridge University Press, 1985.
[18] Zweigenbaum, P., Bouaud, J., Bachimont, B., Charlet, J., Seroussi, B. and Boisviex, J.F. From Text to Knowledge: a Unifying Document-Oriented View of Analyzed Medical Language. Proceedings of IMIA WG6. 1997.