한국어에 적합한 자동 온톨로지 생성을 위한 모델 제안 및 구현

경북대학교 언어 정보 연구실
정영규† · 박미성 · 최재혁 · 이상조

Implementation and Model to Automatically Generate an Ontology for Korean

YoungGiu Jung, MiSung Park, JaeHyuk Choi, Sang Jo Lee
Department of Computer Engineering, Kyungpook National University, Deagu, Korea

요 약

본 논문은 언어학적 데이터로부터 자동으로 온톨로지를 생성하기 위한 모델을 제안하고 이를 구현한다. 모델 제안을 위해 온톨로지의 기본 구성 요소인 개념과 관계를 정의하고 이러한 개념과 관계 객체를 자동으로 추출하는 알고리즘을 제안한다. WordNet을 이용하여 개념을 자동으로 추출하고, 추출된 개념들간의 관계는 한국어의 구문적 특성을 이용하여 관계의 기본 형태를 정의하고 이를 기반으로 관계를 추출한다. 본 논문은 특허문서에서 전기통신기술문서를 대상으로 구현했으며, 제안된 알고리즘을 다른 영역으로 확장하여 이를 검증할 것이다.

서 론

정보의 보다 효율적인 검색을 위해 많은 사람들이 웹 기반의 온톨로지를 구축하기 위해 노력중이다. 이러한 노력의 결과로 몇 개의 정보 검색사이트가 온톨로지를 기반으로 한 정보 검색 서비스를 하고 있다. 오늘날 온톨로지에 대한 대부분의 연구는 유럽과 미국에서 이루어지고 있으며 대표적인 시스템으로 FRODO(a Framework for Distributed Organizational Memories)[1]와 KAON,[2] OKMS(Ontology-based knowledge management sy-stem)[3] 등이 있다.

온톨로지를 생성함에 있어서 가장 기본이 되는 처리는 개념 추출과 개념들을 연결하는 관계를 정확하게 정의하는 것이다. 이러한 개념 추출과 개념간을 연결하는 관계의 추출은 언어학적 데이터 내에 존재하는 어휘의 추상적인 의미를 추출하는 것으로 볼 수 있다. 그러나 어휘가 가지는 의미적 모호성 때문에 자동으로 개념과 개념들을 연결하는 관계를 추출하는 것에 많은 어려움이 따른다. 따라서 대부분의 온톨로지 개발자들은 개념과 개념을 연결시키는 관계를 추출함에 있어서 수동이나 반자동 방법으로 구축하고 있다.

본 논문은 언어학적 데이터에 나타나는 어휘를 분석하여 개념과 관계를 자동으로 추출하는 방법을 제시하고 이를 기반으로 도메인에 적합한 온톨로지의 자동 생성과정을 보여준다. 2장에서는 개념과 관계의 기본 모델을 정의하고 3장에서는 이를 자동으로 생성하기 위한 시스템 구조와 세부 알고리즘을 논의 할 것이다. 그리고 4장에서 이를 기반으로 해서 생성한 온톨로지의 예를 보이고 5장에서 결론을 맺는다.

자동 온톨로지 생성을 위한 모델

온톨로지란 특정 도메인에 적합한 개념 객체와 관계 객체들로 이루어진 방향 그래프로 정의되어진다. 각각의 객체 생성에 있어서 주요한 문제는 언어학적 데이터에서 개념 객체와 관계 객체를 추출하는 것이다.[4] 이러한 문제를 해결하기 위해서 개념 객체와 관계 객체의 모델을 정의하고 이를 기반으로 온톨로지를 자동 생성한다.

1 개념 객체

개념 객체란 다음 속성들의 집합으로 정의한다. 여기에서 선택한 속성의 종류는 Category, Label, Kor-Instance,

†E-mail : rerajO@sejong.knu.ac.kr,
E-mail : mspark@knu.ac.kr,
E-mail : jhchoi@silla.ac.kr,
E-mail : sjlee@knu.ac.kr

Eng-Instance, DependenceConcept으로 한다. DependenceConcept의 경우에는 한국어의 구문적인 특성을 반영한 관계 속성의 하나이다. Table 1은 각 속성에 대한 설명이다.

Fig. 1은 언어학적 데이터에서 추출된 "단말기"라는 어휘에서 생성된 개념객체의 예이다.

2. 관계 객체

관계 객체란 다음 속성들의 집합으로 정의한다. 여기서 선택한 속성의 종류는 Category, Label, Kor-Instance, Eng-Instance, LinkedConceptSet, DependenceConcept으로 한다. LinkedConcept--Set 속성은 여러 개의 개념들이 선택될 수 있으며 Table 2는 각 속성에 대한 설명이다.

Fig. 2는 언어학적 데이터에서 추출된 "공급"이라는 어휘에서 생성된 연결 객체의 추출의 예이다.

자동 온톨로지 생성을 위한 시스템 구조

자동 온톨로지 구축을 위한 시스템은 세개의 컴포넌트인 문장성분 추출기, 개념 추출기, 그리고 관계 추출기로 구성된다. Fig. 3은 개략적인 시스템 구조이다.

입력된 언어학적 문서들은 tagger를 이용해서 태깅한 후 문장성분 추출기를 이용하여 적절한 문장성분을 갖는 어휘를 추출한다. 그리고 이를 입력으로 하여 개념개체를 추출하고 추출된 개념객체와 태깅 문서들을 입력으로 하여 관계객체를 추출한다. 마지막으로 추출된 객체들을 이용하여 온톨로지를 만들게 된다. 다음 절에서 개념객체 추출기와 관계객체 추출기의 구조와 알고리즘을 설명한다.

Table 1. 개념 객체의 속성

속성	설명
Category	객체의 종류
Label	선택된 상위 개념
Kor-Instance	한글 명
Eng-Instance	영어 명
DependenceConcept	형용사에의 한 수식

Category	Concept Object
Label	Electronic equipment
Kor_Instance	단말기
Eng_Instance	terminal
DependenceConcepts	NULL

Fig. 1. "단말기"에서 생성된 개념객체의 예.

Table 2. 개념 객체의 속성

속성	설명
Category	객체의 종류
Label	선택된 상위 개념
Kor-Instance	한글 명
Eng-Instance	영어 명
LinkedConceptSet	연결된 개념들
DependenceConcept	부사에의 한 수식

1. 개념객체 추출기

개념이란 낱낱의 사물로부터 공통의 성질이나 일반적 성질을 추출하여 된 표상을 말한다.[6] 따라서 언어학적 데이터로부터 개념을 추출하기위해서는 일반적인 뜻을 표현할 수 있는 시소러스나, WordNet[5] 같은 도구가 필요하다. 본 논문에서는 WordNet을 이용하여 개념객체를 추출하는 알고리즘을 제안한다. 아래의 Fig. 4는 개념객체 추출기의 구조이다.

개념객체의 추출은 두 단계로 구성되어진다. 첫번째 단계는 문장성분 추출기에서 추출된 명사들을 역어사전을 이용하여 영어로 변환한다. 여기에서 사용된 역어사전은 도메인에서 추출된 명사에 대해서 수동으로 구축한 것으로서 일반명사를 변환하는 사전과 복합명사를 변환하는 사전으로 나누어 구성한다. Fig. 5는 복합명사 사전의 예이다.

본 논문에서 구성한 복합명사 사전은 개념추출에서 발생하는 문제점인 개체명 인식문제와 WordNet에 나타나지 않는 어휘를 처리하기 위하여 사용한다. 두 번째 단계인 개념

Category	Relation Object
Label	supply
Kor_Instance	공급
Eng_Instance	supply
LinkedConceptSet	device, natural phenomenon
DependenceConcepts	electronic equipment

Fig. 2. "공급"에서 생성된 관계객체의 예.

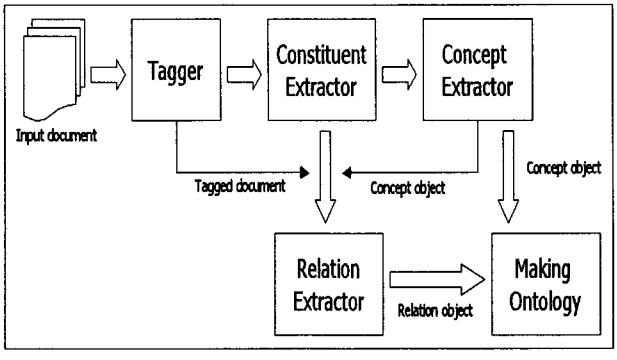

Fig. 3. 시스템 구성.

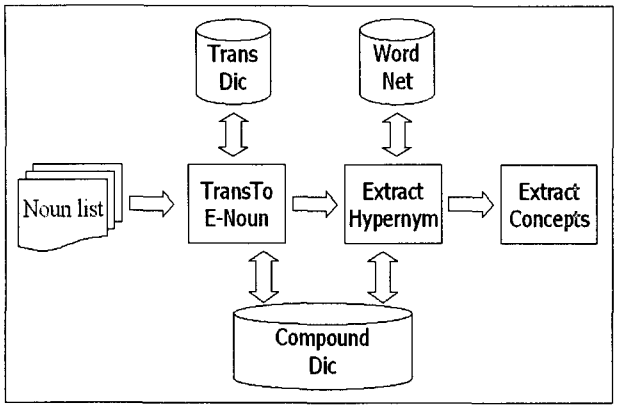

Fig. 4. 개념객체 추출기의 구조.

Meanning of Korean	Compound Term	concept
음성신호	aural signal	communication signal
전원	power source	physical phenomenon
아날로그신호	analog signal	communication signal
이진신호	binary signal	communication signal
병렬/직렬 변환기	parallel-serial converter	electronic device
카오디오	car audio	electronic equipment
오디오잭	audio jack	electronic device
이동통신단말기	mobile communication terminal	electronic equipent
데이터심볼	data symbol	basic cognitive process
패이딩성분	padding element	artifact
송수신기	transceiver	electronic equipent
레일리	rayleigh	name
……	……	……

Fig. 5. 복합명사사전의 예.

추출단계에서는 WordNet의 Hypernym과 복합명사사전을 이용하여 언어학적 데이터에서 개념을 추출한다. 알고리즘 1은 WordNet과 복합명사사전을 이용해서 개념객체를 추출하는 알고리즘이다.

WordNet에서 추출한 Hypernym에서 개념을 선택할 때에 문제점은 상위어의 높이를 얼마로 결정하는가이다. 실험 데이타에서 추출한 어휘들의 Hypernyms들을 관찰한 결과, 대부분의 어휘가 2개 이상의 상위어를 가지고 있으며 2번째 상위어부터는 의미가 추상화 되어서 객체를 분별할 수 없게 됨을 확인하였다. 따라서 본 논문에서는 Hypernym의 높이를 1로 결정하여 개념 객체 추출알고리즘을 수행한다.

2 관계객체 추출기

자동화된 온톨로지 구축에서 가장 어려운 부분중의 하나는 개념간의 관계 추출이다. 하지만 현재 대부분의 개념간의 관계설정은 수동으로 하고있다. 본 논문에서는 개념간의 관계를 다음 두 가지 문제로 정의하고 이 문제들을 해결하기 위한 방법론을 제기한다.

- 무엇을 관계로 정의 할 것인가?

알고리즘1. 개념객체 추출 알고리즘

```
입력 : 어휘리스트
출력 : 개념객체리스트
1. while(!WordList.IsEmpty())
2. {
3.       while(복합명사가 있는 동안){
4.          개념객체=WordList.복합명사를찾음
5.          If(형용사 수식이 있는가?)
6.             개념객체에서 DependenceConcept
               을 채움
        }
7.      WordNet검색
8.      검색된 어휘의 Hypernyms을 추출.
9.      적절한 Hypernyms선정.
10.     개념객체의 속성을 채움
11.     If(형용사 수식이 있는가?)
12.        개념객체에서 DependenceConcept을
           채움
  }
```

- 정의된 관계를 언어학적 데이터로부터 어떻게 추출할 것인가?

개념간의 관계객체를 추출하는 문제에서 중심이 되는 것은 관계를 무엇으로 정의하느냐이다. 본 논문은 구문적인 특성에 기반한 관계를 정의하고 이를 이용하여 관계객체를 추출한다. 다음은 본 논문에서 사용하고 있는 관계객체의 정의이다.

정의 1 [직접연결관계] :

만약 언어학적 문서에서 개념객체들이 관계객체에 의해서 직접 연결된다면 이를 직접연결관계라고 정의한다.

정의 2 [수식연결관계] :

1) 만약 문장 내에서 추출한 관계객체가 직접연결관계를 만족하고 이것에 개념객체의 구문적인 수식관계가 이루어진다면 이를 수식연결관계라고 정의한다.

2) 만약 문장 내에서 추출한 개념객체와 그 객체 앞에 존재하는 개념객체 사이에 수식 관계가 이루어진다면 이를 수식연결관계라고 정의한다.

정의 3 [상하위어관계] :

만약 개념객체들이 직접적인 상하위어관계에 있을 때 이를 상하위어관계라고 한다.

언어학적 데이터로부터 관계객체를 추출하기 위한 정의는 다음에 나열하는 규칙과 연결되어 관계객체 추출 알고리즘에 사용되어진다. 그리고 정의 3의 경우는 개념객체들

```
알고리즘 2. 관계객체 추출 알고리즘.
입력 : 태깅된문서, 원문, 개념객체리스트
출력 : 관계객체 리스트
1. while(모든 문서){
2.     문서내에서 찾은 개념객체들로 문서를 대치.
3.     while(한 개의 문서){
4.         한 개의 문장을 읽음.
5.         if(규칙1이 존재하는가?){
6.             관계객체를 추출
7.             if(규칙2가 존재하는가)
8.                 관계객체의 DependenceConcept을
                   채움
               }
           }
       }
   }
```

을 추출한 후 WordNet에 나타나는 Hypernym관계를 비교하여 연결관계를 형성한다.

규 칙 1 :

용언화동사(관형어말어미(동사형명사의 형태를 갖는 어휘인 경우 정의 1에 의거하여 용언화동사를 관계객체로 추출.

규 칙 2 :

문장 성분이 부사인 경우 정의 2-1에 의하여 관계객체의 의존 개념객체로 추출.

규 칙 3 :

개념객체를 수식하는 개념객체가 형용사일 때 정의 2-2에 의거하여 개념객체의 의존 개념객체로 추출.

알고리즘 2는 위에서 사용한 정의와 규칙을 기반으로 언어학적 데이터에서 관계객체를 추출하는 알고리즘이다.

구 현

본 논문은 특허 문서의 의미검색을 위한 자동 온톨로지 구축 논문으로써 현재 대상으로 사용하고 있는 영역은 전기통신 기술부분에서 추출된 문서들이다. 전체 문서의 개수는 50이고 특허 문서의 청구범위에 기술된 언어학적 데이터를 대상으로 한다. Fig. 6은 알고리즘 1과 2를 수행한 뒤 생성된 온톨로지의 부분적인 예이다.

Fig. 6에서 생성된 온톨로지는 내부적으로는 본 논문에서 제안한 개념객체와 관계객체의 구조를 가지고 있다

Fig. 6에서 표기의 설명은 다음과 같다. 사각형은 개념객체를 나타내고, 육각형은 관계객체를 표시한다. 그리고 화살표는 개념간의 관계 방향을 나타낸다.

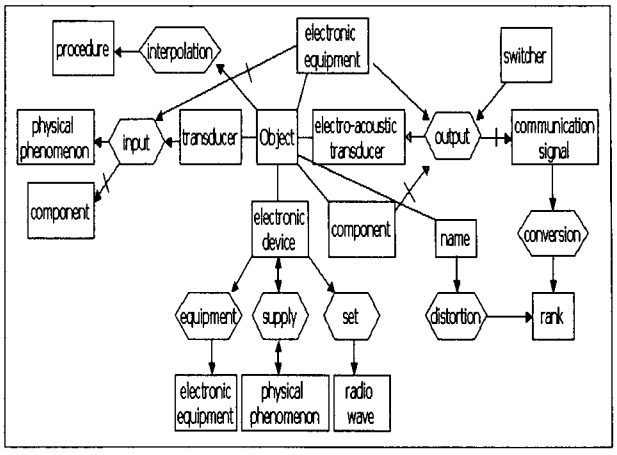

Fig. 6. 생성된 온톨로지의 예.

결 론

본 논문은 언어학적 데이터로부터 온톨로지를 자동으로 구축하기위한 모델과 구현을 목적으로 한다. 온톨로지를 개념객체와 관계객체로 이루어진 방향 그래프로 정의하고 개념객체와 관계객체를 자동으로 추출할 때 발생하는 문제점을 제기한 후 효과적인 객체 자동 추출 알고리즘을 제안하였다.

개념객체의 추출은 WordNet의 Hypernym을 이용하였고 생성된 개념객체는 Hypornym을 이용하면 확장 가능할 것이다. 그리고 관계객체의 추출은 한국어의 구문적인 특성을 반영한 관계를 정의하고 이를 반영한 규칙을 제안한 후 이를 이용하여 관계객체를 추출하였다.

본 논문에서 제시된 알고리즘을 특허 문서의 다른 도메인에 적용한 후 알고리즘의 문제점을 파악하여 이을 개선하고, 생성된 온톨로지를 기반으로 하는 특허문서 검색 시스템을 구성하여 특허문서에 대한 의미적 검색이 가능하도록 구현할 것이다.

REFERENCES

1) A. Abecker et al, "Toward a Technology for Organizational Memories," IEEE Intelligent Systems, vol.13, no. 3, May/June 1998, pp40-48
2) A. Maedche and B. Motik and L. Stojanovic and R. Studer and R. Volz, An Infrastructure for Searching, Reusing and Evlving Distributed Ontologies, WWW2003, May 20-24, 2003, Budapest, Hungary
3) A. Maedche, B. Motik, L. Stojanovic, R. Studer, and R. Volz, Ontologies for Enterprise Knowledge Management, April, 2003, IEEE Computer Society
4) Scott Farrar, William D. Lewis, and D. Terence Langendoen, An Ontology for Linguistic Annotation, Argust, 2002, EMELD Language Digitization Project Conference 2002
5) Fellbaum C(1998) : *WordNet-An electronic lexical database. MIT Press.*
6) 이희승. 국어 대서전, 민중서림