DOI: 10.3745/KIPSTB.2008.15-B.4.365

자질별 관계 패턴의 다변화를 통한 온톨로지 확장

이 신 목[†] · 장 두 성[††] · 신 지 애[†††]

요 약

본 논문에서는 패턴의 다변화를 통하여 관계를 점진적으로 추출함으로써 온톨로지를 확장하는 모델을 제안한다. 패턴 다변화 과정에서 위키피디아로부터 추출한 관계 패턴 후보를 자질별로 다변화시킨다. 다변화된 패턴 후보로부터 말뭉치 빈도수에 따른 신뢰도를 이용하여 패턴을 선별한다. 선별된 패턴은 위키피디아로부터 관계를 추출하는 데 사용되며, 추출된 관계는 다시 관계 패턴 확장에 사용된다. 본 논문에서는 점진적 학습 과정에서의 패턴 다변화를 통하여 패턴 선택의 범위를 확장함으로써, 선택되는 패턴이 점진적으로 정제되는 모델을 제시한다. 이를 통하여, 관계의 확장성과 정확도를 향상시키고자 하였다. 단일 자질 패턴 모델에 대한 실험을 통하여, 어휘, 중심어, 상위어 정보는 신뢰도에, 품사, 구문 정보는 확장성에 유리하며, 구문 단위 유형별로 필요한 자질 유형이 다름을 관찰하였다. 이와 같은 특성에 기반하여 현재 연구 진행 중인 복합 자질 패턴 모델을 제안한다.

키워드 : 온톨로지 확장, 관계 추출, 패턴 다변화, 온톨로지 진화, 점진적 관계학습

Incremental Enrichment of Ontologies through Feature-based Pattern Variations

Sheen-Mok Lee[†] · Du-Seong Chang[††] · Ji-Ae Shin[†††]

ABSTRACT

In this paper, we propose a model to enrich an ontology by incrementally extending the relations through variations of patterns. In order to generalize initial patterns, combinations of features are considered as candidate patterns. The candidate patterns are used to extract relations from Wikipedia, which are sorted out according to reliability based on corpus frequency. Selected patterns then are used to extract relations, while extracted relations are again used to extend the patterns of the relation. Through making variations of patterns in incremental enrichment process, the range of pattern selection is broaden and refined, which can increase coverage and accuracy of relations extracted. In the experiments with single-feature based pattern models, we observe that the features of lexical, headword, and hypernym provide reliable information, while POS and syntactic features provide general information that is useful for enrichment of relations. Based on observations on the feature types that are appropriate for each syntactic unit type, we propose a pattern model based on the composition of features as our ongoing work.

Key Words : ontology enrichment, relation extraction, pattern variation, ontology evolution, incremental relational learning

1. 서 론

온톨로지 분야에서 온톨로지의 진화는 중요한 문제이다. 온톨로지가 대상으로 하는 분야나 그 분야의 데이터가 시간에 따라 변화하거나, 응용 온톨로지에서의 사용자 요구사항이 변화할 경우, 적응적으로 변화시킬 수 있어야 한다. 특히, 새로운 지식이 빠르게 유입되는 정보통신 분야의 경우, 온톨로지에서 대상으로 하는 지식의 증가량이 많으므로, 온톨로지를 연속적으로 확장하여야 하며, 분야지식의 증가속도에 발맞추기 위하여, 온톨로지 확장 과정을 효율적으로 자동화할 필요가 있다.

본 논문에서는, 온톨로지 확장 과정의 일환으로 점진적인 온톨로지 관계 지식의 확장 모델을 제안하고자 한다. 관계 확장을 위한 자원으로 온톨로지와 위키피디아 백과사전 말뭉치를 활용한다. 온톨로지는 초기 관계 지식의 제공원으로서 활용하며, 기구축된 정보통신분야 온톨로지를 사용한다. 말뭉치는 위키피디아 정보통신분야 용어에 대한 위키피디아 항목 중 여러 개의 문장으로 구성된 텍스트 부분으로 구성된다. 이는 관계 및 관계 패턴을 추출하기 위한 자원으로 사용한다. 위키피디아 백과사전에서 제공하는 다양한 정보

† 준 회 원 : 한국과학기술원 전자전산학과 박사과정
†† 정 회 원 : KT 미래기술연구소 수석연구원
††† 정 회 원 : 한국정보통신대학교 공학부 교수

논문접수 : 2008년 6월 6일
수 정 일 : 2008년 7월 18일
심사완료 : 2008년 7월 22일

를 사용하지 않고, 텍스트 부분만을 추출하여 이용한 이유는 다음과 같다. (i) 본 연구에서 사용하는 기구축 온톨로지가 텍스트에 기반하여 추출되었기 때문에, 문장 내에서 사용되는 구와 절 형태의 객체가 자주 나타난다는 점이다. (ii) 관계 추출을 위한 패턴을 한 문장 내에서의 구문에 기반하여 단순하게 정의하기 위하여서이다. 위키피디아 백과사전의 다양한 자질을 사용할수록 다양한 패턴과 관계를 추출할 수 있지만, 패턴의 형태가 복잡해지기 때문에, 패턴 내 각 자질의 특성을 파악하기 어려워진다. 따라서, 본 연구에서는, 설명항 텍스트의 한 문장 내에서 구문으로 연결되는 객체간의 관계로 추출의 범위를 한정시킨다.

본 논문에서는 확장 과정에서의 신뢰성을 유지하기 위하여 문장 내에서 관계를 나타낼 수 있는 관계 패턴을 이용한다. 구문 정보를 이용하여 추출한 주요 구문 단위의 다양한 정보들을 자질로 사용한다. 주요 구문 단위만을 사용하는 이유는, 정의역과 치역을 나타내는 문장 상의 구나 절 사이의 거리가 먼 경우에도 패턴의 적용률을 유지할 수 있기 때문이다.

또한, 기존의 패턴 기반 관계 추출 연구와는 달리, 패턴 내의 자질들의 중요도가 서로 다르다는 점에 착안하여, 문서로부터 추출한 초기 패턴 후보를 자질에 따라 다양하게 일반화시켜서 확장하는 패턴의 다변화 과정을 도입한다. 이와 같이 다변화된 패턴 후보들은 말뭉치에서의 관계 및 패턴 빈도수에 따른 신뢰도를 이용하여 선별함으로써, 보다 중요한 자질을 많이 가진 패턴이 선택될 수 있는 모델을 설계한다. 선별된 패턴은 위키피디아로부터 관계를 추출하는데 사용하고, 추출된 관계는 다시 관계 패턴 확장에 사용됨으로써 관계와 패턴의 점진적 확장이 이루어진다.

본 논문에서는 어휘, 품사, 구문, 중심어 및 기존 온톨로지에서 추출한 상위어를 패턴의 기본 자질로 이용한다. 이들 기본 자질을 이용하여 다섯 가지의 단일 자질 기반 패턴 모델을 구성하고, 이를 통하여 각 자질의 확장성과 신뢰도, 구문 단위 유형별 적합성 등의 특성을 분석하여, 복합 자질 패턴 모델에서의 자질 구성을 정의하여 제안한다.

본 논문의 구성은 다음과 같다. 2장에서, 기존의 관계 추출 연구에 대하여 살펴보고, 3장에서 전체 시스템의 구성을 설명한 후, 4장에서 관계와 패턴의 점진적 확장 과정을 상술한다. 5장에서, 관계와 패턴의 확장을 위한 실험을 보이며, 6장에서 결론과 본 논문에서 제시한 모델의 확장과 관련해서 진행중인 연구에 관해 언급한다.

2. 관련연구

본 절에서는 현재까지 제안된 다양한 종류의 관계 추출 모델을 알아본다.

2.1 규칙 기반 추출

[3]에서는 구문과 의미 정보에 의한 관계 추정 규칙을 이용하여 두 개체 사이의 관계를 추출한다. [2]에서는 UMLS에서 정의된 특정 관계의 정의역을 포함하는 개체와 치역을 포함하는 개체가 한 문장 내에서 관계를 표현하는 동사의 주어와 목적어로 사용된 경우, 해당 관계를 추출한다. [2]는 구문 구조에 기반한 규칙을 사용하므로, 본 논문에서 사용하는 구문 구조 기반의 패턴 표현이 가능하다. [4]에서는 관계를 나타내는 동사 간의 의미체계와 사전을 이용하여 자연언어 파싱 과정에서 관계를 추출하기 위한 규칙을 적용한다. 규칙 기반 방법은 미리 정하여진 규칙을 이용하여 관계를 추출하므로 추출을 위한 템플릿을 학습할 수 없다는 단점이 있다.

2.2 패턴 기반 추출

[1]에서는 말뭉치에 있는 개체들 사이의 구문구조로부터 추출한 패턴을 이용하여 관계를 학습하고, 학습한 관계들을 GENIA 온톨로지의 개념체계를 이용하여 일반화한다.

[6]에서는 관계와 패턴, 패턴과 패턴 간의 유사도를 이용하여 말뭉치로부터 추출한 삼진관계를 관계에 매핑한다.

관계를 표현하는 동사구와 주어/목적어 등의 변수사이의 삼진관계를 이용하여 관계를 추출하는 방법에는 [9]가 있다. 여기서는 동사를 중심으로 한 수식 관계를 이용하여 관계를 추출하였다. 관계 추출에 있어서 동사가 중요한 역할을 하는 것은 사실이지만, 다양한 형태의 패턴을 보지 못하므로, 적용률의 향상을 위하여 패턴의 확장이 필요하다.

[12]에서는 개념들 간의 인과관계를 추출하기 위해 어휘와 구문구조를 이용한 패턴을 코퍼스로부터 반복학습하는 방법론을 제안하였다. 특히 [12]는 본 논문에서 사용하는 패턴의 기본 형태를 제공한다. 하지만, 패턴에서 사용하는 자질이 어휘 정보로 한정되어 있고 자질의 특성이나 중요도를 고려할 수 있는 방법이 없다는 한계점을 지닌다.

일반적인 구문구조에 기반하여 패턴을 구성하는 방법인 [10]에서는 술어논항구조를 이용하여, 어휘와 품사 정보로 구성된 패턴의 기본 단위를 구한 후, 각 기본 단위를 연결시켜서 하나의 패턴을 구성한다. 이 방법은 다양한 형태의 패턴을 구성할 수 있다는 장점이 있다. 하지만 패턴 기반 모델에서는 매칭에 있어서 서로 다른 역할을 수행하는 다양한 자질이 사용되는데 현재까지 다양한 형태의 자질의 특성을 관찰하기 위한 연구는 아직까지 보지 못하였다. 본 연구의 모델에서는 패턴에 포함된 어휘, 품사, 구문, 의미 정보에 대한 자질 각각에 대한 비교를 통하여 자질의 속성 및 결합 특성을 관찰한다.

또한, 본 연구에서는 패턴의 모든 일반화된 형태를 고려하는 자질 다변화 방법을 적용함으로써 보다 폭넓은 형태의 패턴을 고려하는 것이 가능하다.

2.3 자질 기반 추출

자질 기반 방법은 ACE 등의 성능 평가에서 최근 비교적 높은 성능을 보이고 있다. 이들 방법에서는 관계 추출 대상이 되는 객체에 대한 자질을 추출하고, 자질에 의하여 관계를 추출한다. [11]에서는 위키피디아로부터 위키피디아의 엔트리나 문장 상에 나타나는 객체쌍에 대한 자질을 뽑은 후,

SVM에 의하여 관계를 추출한다. 자질에 의한 관계추출의 경우 역시, 온톨로지 확장에 적용하기 위하여서는 자질의 신뢰성 문제를 해결하여야 한다. 본 논문의 패턴에서 사용하는 자질의 중요도와 역할을 관찰함으로써, 자질 기반 모델의 성능 향상에도 도움을 줄 수 있다.

3. 시스템 구성

본 장에서는 온톨로지의 관계 확장을 위한 시스템의 각 모듈이 각각 어떻게 구성되어 있고 상호 간에 어떻게 결합되어 있는지를 살펴보고자 한다. (그림 1)은 온톨로지의 점진적 확장 과정을 보여준다. 시스템에서 이루어지는 점진적인 과정을 5 단계로 나누어 설명하면 다음과 같다.

패턴 후보 추출 단계: 온톨로지 관계를 이용하여 위키피디아 말뭉치로부터 패턴 후보를 추출한다. 위키피디아 말뭉치에서 온톨로지 관계의 정의역과 치역이 함께 있는 문장으로부터 정의역과 치역을 구문적으로 연결하는 패턴 후보를 추출한다. 예를 들어, “node”와 “tree” 사이의 partOf 관계를 나타내는 패턴은 문장상의 “the nodes of a tree”로부터 추출할 수 있다. 어휘 자질만을 고려할 경우, (1)의 패턴을 얻을 수 있다. (DOM과 RNG는 각각 정의역과 치역을 의미한다.)

node:DOM of: tree:RNG (1)

패턴 후보 다변화 단계: 추출한 패턴 후보를 모든 가능한 일반화된 형태로 확장한다. 패턴 후보를 패턴에서 사용하는 자질별로 일반화시키면서, 모든 가능한 일반화된 형태의 패턴 후보를 생성한다. 패턴 (1)을 다변화시키면, (2)의 패턴을 얻을 수 있다.

:DOM of: :RNG
node:DOM of: :RNG
:DOM of: tree:RNG (2)

패턴 선정 단계: 관계 추출을 위한 패턴 후보와 그에 해당하는 관계 간의 상호 연관성을 나타내는 신뢰도를 계산하여, 신뢰도가 높은 패턴 후보를 관계 추출을 위한 패턴으로 선정한다.

관계 추출 단계: 선정된 패턴을 이용하여 위키피디아 말뭉치로부터 각 패턴에 해당하는 관계를 추출한다. 추출한 관계들은 온톨로지에 추가되며, 다음 단계의 패턴 후보를 추출하기 위한 단서가 된다.

패턴과 관계의 점진적 확장 단계: 추출한 관계는 다시 위키피디아 말뭉치로부터 패턴 후보를 확장하는 데 사용된다. 온톨로지 관계뿐 아니라, 추출된 관계들에 대하여서도 위키피디아 말뭉치로부터의 패턴 추출 과정을 적용한다. (그림 1)에서와 같이, 이와 같은 과정을 매회 반복함으로써, 관계와 패턴의 점진적 확장이 이루어진다.

4. 패턴과 관계의 점진적 추출

본 장에서는 기존의 관계를 이용하여 위키피디아 말뭉치로부터 관계 추출을 위한 패턴을 추출하고, 추출한 패턴으로부터 관계를 추출함으로써, 패턴과 관계를 점진적으로 확장하는 본 논문의 모델을 각 단계별로 상세히 설명한다.

4.1 패턴 후보 추출

본 절에서는 본 논문에서 사용하는 패턴의 자질과 구조, 표현방식에 대하여 설명하고, 자질별 특성을 관찰하기 위한 자질별 모델에 대하여 알아본다. 또한, 온톨로지와 말뭉치로부터 추출한 기존의 관계를 이용하여, 위키피디아 말뭉치로부터 패턴 후보를 추출하는 과정을 보인다.

4.1.1. 패턴의 형태

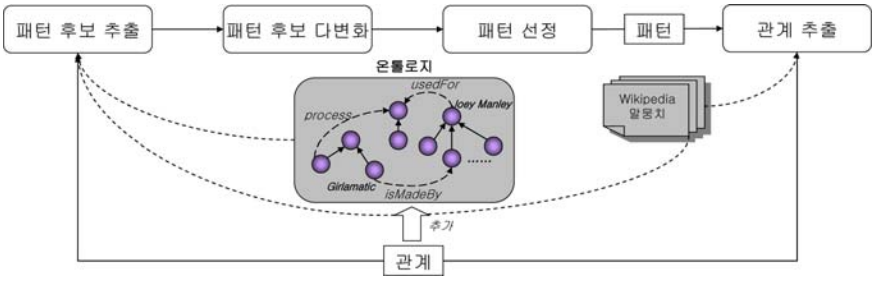
패턴이 어떠한 자질들로 이루어지고, 어떠한 구조적 특성을 지니며, 어떻게 표현되는지에 대하여 살펴본다.

패턴의 자질

본 논문에서 사용하는 패턴은 5 가지의 자질을 사용한다. 각각의 자질을 표로 표현하면 <표 1>과 같다.

<표 1>에서는 각 자질들을 본 논문의 나머지 부분에서 표현하기 위한 기호와 그 의미를 설명한다.

5가지 자질은 위키피디아 말뭉치를 커넥서 파서에 의하여 자연언어 분석한 결과와 온톨로지로부터 추출한다. WORD는 어휘 정보, POS는 품사 정보, SYN은 구문 정보, HW는 중심어 정보, HYPER는 상의어 정보를 각각 의미한다. 앞의 네 가지 자질은 구문 분석 말뭉치로부터, 상위어 정보는 구문 분석 말뭉치와 온톨로지로부터 추출한다. 상위어 정보의



(그림 1) 관계 패턴의 다변화를 통한 온톨로지 확장

〈표 1〉 패턴에서 사용하는 자질

자질 기호	자질의 의미
WORD	구문단위에 포함된 단어의 원형태
POS	구문단위의 중심어의 품사
SR	구문단위의 중심어의 구문적 역할
HW	구문단위의 중심어
HYPER	구문단위 혹은 구문단위 중심어의 온톨로지 상위어

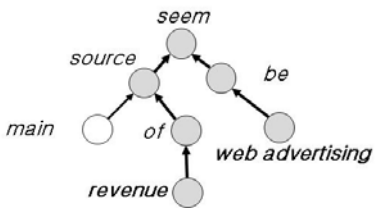
경우, “isa”, “instanceOf” 등의 상하위 구조를 이루는 관계를 이용한다. 구문 단위의 중심어가 정의역의 중심어와 같은 상하위 관계의 치역에서 중심어를 취함으로써 패턴을 형성한다. 이를 위하여, 상하위 관계 정의역의 중심어와 치역의 중심어 쌍 목록을 미리 구축한다. 예를 들어, “CAD”를 상하위 관계 정의역의 중심어로 가지는 경우가 <표 2>의 네 가지가 있는 경우, “CAD”를 중심어로 가지는 구문 단위의 상위어 정보는 “CAD”, “design”, “automation”, “layout”이 된다.

〈표 2〉 “CAD”를 정의역의 중심어로 가지는 isa관계의 정의역과 치역

정의역	치역
architectural CAD	CAD
CAD	design
circuit CAD	electronic design automation
circuit layout CAD	circuit layout

패턴의 구조

본 논문의 패턴은 구문분석 말뭉치의 의존구조에 기반하여 구성된다. 예를 들어, “Main source of revenue seems to be web advertising.”이라는 구문을 의존구조에 의하여 분석한 (그림 2)와 같은 결과로부터, “MadeBy” 관계의 정의역에 해당하는 revenue와 치역에 해당하는 “web advertising” 사이의 최단 경로인 회색 표시의 노드로 이루어진 경로를 패턴으로 표현한다. 이와 같이 정의역과 치역을 연결하는 의존구조 상에서 불필요한 수식어를 제외한 최소 경로를 구문구조의 최소경로라 정의한다. 구문구조 최소 경로는 본 논문에서 사용하는 패턴의 기본 구조를 이룬다. 따라서, (그림 2) 예제의 패턴에서는 seem, source, of, revenue, be, web advertising의 구문 단위에 대한 자질별 기술이 필요하다.



(그림 2) 구문 구조 트리: “Main source of revenue seems to be web advertising.”

패턴의 표현

패턴의 구문단위별 표현은 각 구문단위별 자질들의 나열로 이루어진다. 즉, 구문단위에 대한 어휘 정보, 품사 정보, 구문 정보, 중심어 정보, 상위어 정보 등의 나열로 나타나고 각 구문단위가 정의역인지 치역인지의 여부를 함께 표현한다. 구문 단위 u에 대한 패턴 단위의 표현 R_u를 정규식으로 나타내면 다음과 같다.

$$R_u=(\text{WORD}(;\text{WORD})*)?:\text{POS}?:\text{SR}?:\text{HW}?:(\text{HYPER}(\#\text{HYPER})*)?:\text{ROLE} \quad (3)$$

(3)에 나타난 각각의 자질을 뜻하는 기호는 <표 1>에 정의되어 있다. ROLE은 해당하는 구문 단위가 패턴이 나타내고자 하는 온톨로지 관계에서 정의역 혹은 치역의 역할을 하는지의 여부를 나타낸다. 예를 들어, (그림 1)의 “web advertising”이라는 구문 단위에 대하여 노드에 대한 패턴 단위를 만들기 위하여, 어휘 정보인 “web advertising”, 품사 정보인 “N”, 중심어 정보인 “advertising”, 온톨로지로부터의 상의어 정보인 “disseminating”을 (3)의 형태로 기술하면, (4)와 같은 형태가 된다. (4)에서 RNG는 구문 단위가 패턴이 나타내는 온톨로지 관계에서 치역을 표현함을 알려준다.

web advertising:N::advertising:disseminating:RNG (4)

이와 같은 각 구문단위 표현을 이용한 의존구조 최소경로는 [8]에서 제시한 선순위 문자열 표현 방법에 의하여 나타낸다. 선순위 문자열 표현을 통하여 패턴 매칭 시간을 선형시간으로 줄일 수 있다[12]. 이와 같은 표현 방식으로 (그림 2)의 구문구조 최소경로를 표현하면 (5)와 같다.(DOM은 정의역을, RNG는 치역을 나타낸다.)

seem:V::seem:: source:N:SUBJ:source:: of:::of:: revenue:N::revenue::DOM 0 0 0 be:V:::: web;adverti-sing:N::advertising:disseminating:RNG 0 0 (5)

4.1.2. 패턴 후보의 추출

본 논문에서는 패턴에 포함된 각 자질의 특성을 관찰하기 위하여 자질별 모델을 제시한다. 본 장에서는 각 자질별 모델의 구체적인 형태를 설명하고, 구문분석 말뭉치로부터 패턴 후보를 추출하는 과정을 보인다.

자질별 패턴 모델

자질별 패턴 모델에서는 각각의 자질별로 한 개씩의 모델을 구성한다. 즉, <표 3>과 같은 형태의 패턴 정규식 표현을 사용한다. <표 3>에 사용된 각 기호들의 의미는 <표 1>

〈표 3〉 자질별 패턴 모델

모델	정규식 표현
R-type	((WORD;)+:(DOM|RNG)?SPACE)+
P-type	(POS:(DOM|RNG)?SPACE)+
S-type	(SR:(DOM|RNG)?SPACE)+
H-type	(HW:(DOM|RNG)?SPACE)+
Y-type	((HYPER#)+:(DOM|RNG)?SPACE)+

에 정의되어 있다.

패턴 추출 과정

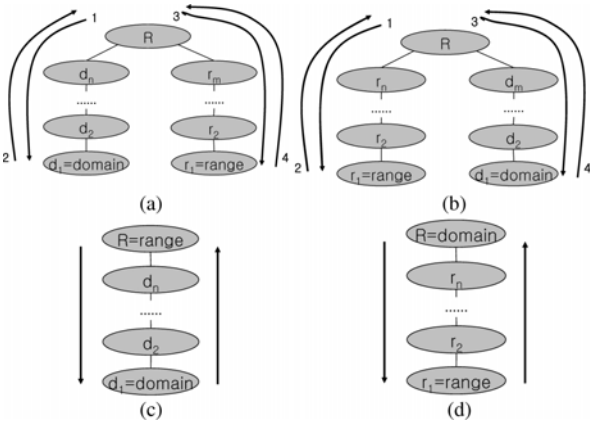
자질별 패턴 모델은 3장에서 소개한 관계 및 패턴의 점진적 추출 과정에 적용된다. 위키피디아 말뭉치로부터 추출한 관계와 기존 온톨로지의 관계는 패턴의 점진적 추출 과정을 위한 단서가 된다. 패턴의 추출은 관계의 정의역과 치역을 함유한 위키피디아 문장으로부터 이루어진다. 구체적인 과정을 (그림 3)에 표현한다.

단계1: 개체 및 사건 단위 인식: 본 논문에서는 개체 및 사건 간의 관계를 추출 대상으로 삼으므로, 패턴 추출 과정에서도 개체 및 사건 단위를 인식할 필요가 있다. 이를 위하여, 개체 및 사건 단위를 말뭉치로부터 인식할 필요가 있다. 본 논문에서는 명사구를 개체 및 사건의 기본 단위로 삼는다. 하지만, 기존 온톨로지에는 용어/구/문장 등 다양한 단위의 관계가 있으므로, 개체 및 사건 인식 단계에서는 기존 온톨로지 관계의 정의역/ 치역 단위들과 함께 규칙에 기반하여 추출한 명사구가 개체 및 사건의 단위가 된다.

단계 2: 문장 추출: 기존 온톨로지 및 현재까지 위키피디아 말뭉치로부터 추출한 관계의 정의역과 치역을 함께 가진 문장을 추출한다.

단계 3: 패턴 후보 추출: 문장 내에서의 정의역과 치역 사이의 구문 경로를 분석하여 유형별로 구문 패턴 후보를 산출한다. 정의역과 치역의 조상-후손 관계에 따라서 구문 경로의 형태를 구분하면 (그림 4)와 같은 네 가지 유형이 나타난다. (a), (b)는 정의역과 치역 사이에 조상-후손 관계가 존재하지 않는 경우이다. (c)는 치역이 조상 노드인 경우이고, (d)는 정의역이 조상 노드인 경우이다. 각 경우를 [8]의 선순위 문자열을 이용하여 표현하면 <표 4>와 같다.

(그림 3)에서 보듯이, 패턴 후보의 추출 단계에서는 관계 추출 단계에서 추출한 명사구 간 관계들이 확장되어 입력으로 들어온다. 따라서, 추출되는 패턴 후보는 매회 확장되는 관계를 반영하도록 확장된다. 또한, 패턴 후보는 다음 회의 관계 확장에 사용된다.



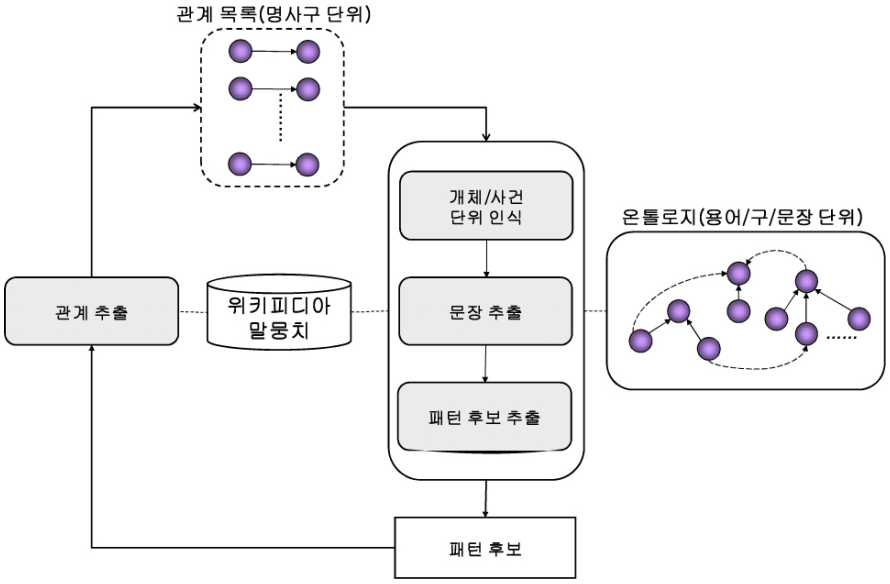
(그림 4) 정의역과 치역 사이의 구문 경로의 유형

〈표 4〉 구문 경로 유형별 패턴 후보의 형태

구문 경로 유형	패턴 후보 형태
(a)	R d_n d_{n-1} … d_1 0_n r_m r_{m-1} … r_1 0_m
(b)	R r_n r_{n-1} … r_1 0_n d_m d_{m-1} … d_1 0_m
(c)	R d_n d_{n-1} … d_1 0_n
(d)	R r_n r_{n-1} … r_1 0_n

4.1.3. 패턴 후보의 다변화

본 논문에서 제안하는 관계 패턴의 다변화를 통한 확장에 대하여 설명한다. 본 논문에서는 패턴에 정의된 모든 자질



(그림 3) 관계 패턴의 다변화를 통한 온톨로지 확장

조건을 만족할 경우에 한하여 관계를 추출한다. 하지만, 패턴 내의 각 자질별 중요도는 서로 다르다. 따라서, 모든 자질을 표현한 패턴보다 중요한 자질만을 표현한 패턴이 더 유용할 수 있다.

본 논문에서는 패턴별 중요도를 판단하기 위하여 패턴의 다변화와 신뢰도를 이용한 판별 모델을 제안한다. 즉, 4.1.2에서의 모든 가능한 패턴 자질의 조합에 의하여 패턴을 생성함으로써 패턴 자질의 모든 일반화된 형태를 구한 후, 패턴별 신뢰도를 계산하여 중요도를 판별한다. 본 절에서는 본 논문의 패턴 다변화 과정을 보이고자 한다. 패턴 후보의 다변화를 위한 알고리즘은 다음과 같다.

Algorithm: DiversifyPatt(P)
Input: P: 패턴의 집합
Output: updated set of P
foreach p in P:
    feat_num=패턴 p에 포함된 자질의 개수;
    num_expanded_patt=2^{feat_num};
    for i=0 to (num_expanded_patt-1):
        num=i의 2진 표현;
        패턴 p_{new}를 초기화한다.
        for j=1 to feat_num:
            if (num의 j번째 자리수가 1)
                패턴 p의 j번째 자질을 p_{new}에 추가한다.
            endif
        end
        패턴 new_patt을 패턴 집합 P에 추가한다.
    end
end

입력 패턴 p에 n개의 자질이 있다면, 모든 생성 가능한 n자리수의 이진 표현을 구하고, 각 이진 표현에 대응하는 확장 패턴을 정의한다. <표 5>는 이진 표현(*num*)과 확장 패턴(p_{new})을 대응시키는 과정이다.

예를 들어, 입력 패턴 "node:N::node:connection; shape;point:"에 대하여, 111000이라는 이진 표현과 대응되는 확장 패턴은 앞의 세 가지의 자질만 표현된 "node:N::node::"이다.

이와 같은 변환 과정을 통하여, 입력 패턴 p의 모든 일반화된 패턴을 구할 수 있다. 위의 방법은 문자열 형태의 패턴을 직접 다루기보다는 이진 표현 형태로 다룬 후 문자열에 대응시킴으로써, 패턴 간 상하위 관계 구축 등 다방면으로의 응용이 용이하다.

〈표 5〉 초기 패턴의 이진 표현과 확장 패턴의 대응 과정

num=d_1d_2 …d_n (d_i=0,1)
if (d_i=1) include the i-th feature of p into p_{new}
if (d_i=0) don't include the i-th feature of p into p_{new}

4.1.4. 패턴 후보의 선정

본 절에서는 4.1.3절에서 다변화시킨 각 패턴에 대하여 신뢰도를 계산하고, 신뢰도를 바탕으로 관계 추출에 사용할 패턴을 선정하는 과정을 설명한다.

특정한 패턴 p가 특정한 관계 r을 잘 반영하는지의 여부를 평가하기 위하여서는 r에 대하여 p가 추출된 빈도수의 r에 대하여 추출된 전체 패턴 후보 수에 대한 비율(=p_a)을 계산하여야 한다. 역으로 생각하면, 특정한 관계 r이 특정한 패턴 p를 잘 반영하는지의 여부를 평가하기 위하여서는 r에 대하여 p가 추출된 빈도수의 패턴 p의 전체 출현 빈도에 대한 비율(=p_b)을 계산하여야 한다.

이와같은 두 가지 척도는 각 패턴 후보별로 작성한 변형 규칙을 적용한 경우의 성능 평가 척도와 유사한 개념이다. 즉, 관계 r을 나타내는 패턴 후보 p를 이용한 관계-패턴 변형 규칙이 <표 6>과 같은 경우, p_a는 변형 규칙의 정확률과, p_b는 변형 규칙의 재현율과 일치한다.

두 가지 척도를 이용하여 하나의 신뢰도를 산출하기 위하여, 정확률과 재현율의 F-value를 사용한다.

〈표 6〉 관계-패턴의 변형 규칙

condition	말뭉치의 한 문장에서 패턴 후보 p를 발견
action	패턴 후보에 나타난 정의역과 치역에 해당하는 객체에 관계 r을 태깅

관계r과 패턴p에 의한 변형 규칙의 정확률(P)과 재현율(R), F-value(F)를 구하는 공식을 정리하면 (7)과 같다. 아래의 식에서, C는 위키피디아 말뭉치를 의미하며, 임의의 집합 S에 대하여, n(S)는 집합 S의 원소의 개수를 의미한다. β는 정확률과 재현율의 비율을 조정하는 상수이다.

$RC = \{(e_1, e_2) \mid e_1$과 e_2는 C의 한 문장에서
패턴 p로 연결된 개체나 사건.
e_1은 관계r의 정의역, e_2는 치역의 역할을 한다.$\}$

$related = \{(e_1, e_2) \mid e_1$과 e_2는 C의 한 문장에 나타난 개체나 사건.
e_1은 관계r의 정의역, e_2는 치역의 역할을 한다.$\}$

$connected = \{(e_1, e_2) \mid e_1$과 e_2는 C의 한 문장에서
패턴 p로 연결된 개체나 사건$\}$

$$P = \frac{n(RC)}{n(connected)}, R = \frac{n(RC)}{n(related)}$$
$$relib = F = \frac{(\beta^2+1)PR}{P+\beta^2 R} \quad (7)$$

5. 실 험

본 장에서는 본 논문에서 정의한 단일 자질 패턴 모델을 이용한 패턴 및 관계의 점진적 확장 과정을 실험을 통해 관찰한 결과를 서술한다.

〈표 7〉 각 자질별 모델에 대한 1회 실험결과

모델	초기패턴후보 개수	확장패턴후보 개수(확장비)	선택된 패턴(의미관계)	추출된 의미관계 인스턴스 수
R-type	79	1548(19.59)	pass: can: :DOM 0: 0: through: transformers:RNG 0: 0:(p_beTransfromedBy)	3
P-type	63	292(4.63)	N:RNG V: N:DOM 0: 0:(p_toTransmit)	73052
S-type	71	149(2.10)	OBJ:RNG :DOM 0:(p_toReceiveFrom)	17764
H-type	75	1110(14.80)	offer:shogun:DOM 0: :RNG 0:(capableOf)	1
Y-type	70	6537(93.39)	equipment:DNP equipment:RNP 0:(p_toSendTo)	577

5.1 실험 환경

본 논문의 실험에서 사용된 온톨로지와 위키피디아 말뭉치의 구축 환경을 설명한다.

설명의 편의를 위하여, 본 논문에서 사용하고자 하는 온톨로지의 관계에 해당하는 용어의 의미를 다음과 같이 정의하고자 한다.

- 의미관계: 온톨로지에서 동일한 역할을 하는 개념 간 관계의 집합
- 의미관계 인스턴스: 특정한 두 개념 사이에 존재하는 관계를 특정한 의미관계에 의하여 명명하고 분류한 결과

확장을 위한 온톨로지로 IT 분야의 온톨로지인 IT 코아 온톨로지[1]를 선정하였다. IT 코아 온톨로지는 IT 분야에서 일반적이며 다른 서비스를 위한 기반이 될 수 있는 지식을 대상으로 한다.

IT 코아 온톨로지는 IT 분야의 클래스/인스턴스/관계 지식을 포함하며, 149개의 의미관계와 27450개의 의미관계 인스턴스를 포함한다. 149개의 의미관계 중 18개[2]는 ConceptNet의 관계 유형을 사용하였으며, 나머지는 온톨로지 구축 과정에서 자체적으로 만들어서 사용하였다. 본 논문에서는 이 가운데 23개의 의미관계[2]에 대하여 확장을 수행하였다. 확장 대상 의미관계를 선정한 근거는 다음과 같다. (1) 향후 온톨로지 응용 추론에서 비교적 중요한 역할을 할 수 있는, 특정한 상태의 변화와 관련이 있는 관계를 위주로 선정하였다. (2) 본 논문에서 사용하는 구문 기반 관계 추출의 장점을 극대화하기 위하여, IsA, PartOf 등 비교적 단순한 관계 유형은 제외하였다. 선정한 의미관계의 IT 코아 온톨로지 내에서의 빈도수 분포를 살펴보면 <표 7>과 같다.

추출을 위한 대상 말뭉치로 IT 분야 용어 100,000개에 대한 위키피디아(http://en.wikipedia.org) 정의문 말뭉치를 구성하였다. 말뭉치는 2650만 토큰으로 구성되며, 'computer science'와 'information technolgoy', 그리고 그 하위부류에 속한 문서들로 이루어진다[7].

〈표 8〉 관계 유형별 관계 인스턴스의 분포

관계 유형	분포	관계 유형	분포
isMadeBy	26.4%	capableOf	4.1%
Provide	19.5%	effectOf	0.7%
Process	8.3%	isDesignBy	0.6%

5.2 실험 결과

<표 8>에 본 논문에서 각 자질별 패턴 모델을 1회씩 실험한 결과를 보인다. 두번째 열에서 보듯이, 어휘, 중심어, 상위어 정보와 같이 패턴별로 자질들이 <표 8>에 본 논문에서 각 자질별 패턴 모델을 1회씩 실험한 결과를 보인다. 두번째 열에서 보듯이, 어휘, 중심어, 상위어 정보와 같이 패턴별로 자질들이 구체적이고 다양한 형태로 나타나는 경우는 패턴의 확장성이 높은 반면, 품사, 구문 자질과 같이 자질의 종류의 수가 적어서 일반적인 형태의 표현만이 가능한 경우는 확장성이 낮은 편이다.

특히 상위어 패턴의 경우, 한 개의 정의역과 치역 개념에 대하여 매우 다양한 의미적 상위어가 나타날 수 있으므로, 확장률이 초기 패턴의 93.39배로 매우 높게 나타났다. 표의 네 번째 열에서는 신뢰도에 의한 패턴 선택 단계에서 각 유형별로 선택한 관계들을 보인다. 이 관계를 이용하여 말뭉치로부터 추출된 관계의 수는 다섯 번째 열에 나타난다. 이를 통하여, 어휘 패턴과 중심어 패턴의 경우, 하나의 패턴에 의하여 추출되는 관계의 수가 매우 적은 반면, 품사 패턴과 구문 패턴의 경우, 다량의 관계를 한번에 추출할 수 있음을 알 수 있다. 의미 패턴의 경우, 역시 하나의 패턴의 의해 비교적 많은 관계를 추출할 수 있었다. <표 9>에서는 각 모델별로 관계 추출을 위한 10개 이상의 패턴이 생성될 때까지 점진적 확장 과정을 반복한 결과를 보인다. 품사나 구문 정보 자질 패턴의 경우와 같이 1회 반복에서 생성된 패턴의 신뢰도가 너무 낮거나 평가가 불가능할 정도로 다량의 관계가 산출되는 경우 반복을 중단하였다. 신뢰도 산출을 위한 식(7)의 상수, β는 0.3으로 계산하였다. 두번째 열에서는 패턴 다변화 이전 단계에서 말뭉치로부터 직접 추출한 초기 패턴의 개수에 대하여, 세번째 열에서는 패턴 다변화 직후에 확장된 패턴의 개수에 대하여 마지막 회 반복에서 첫 회 반복보다 증가한 비율을 나타낸다.

1) http://coreonto.kaist.ac.kr/project_03.as

2) p_toDevelop, p_beTransmittedBy, desirousEffectOf, isMadeBy, p_toTransform, provide, p_toEnable, madeOf, p_toRecord, p_isProcessingBy, p_isRecordingOf, p_toSendTo, capableOf, p_toIncrease, p_isCodingBy, p_isModulationBy, p_isCommunicationBy, p_beTransformedBy, p_isDesignBy, process, p_toTransmit, effectOf, p_toReceiveFrom

어휘 패턴(R-type)과 중심어 패턴(H-type)을 비교해 보면, 어휘 패턴은 중심어 패턴에 비하여 패턴에서 제시하는 자질 조건이 비교적 다양하고 구체적이어서, 자질에 의한 제약이 보다 많으므로, 추출되는 관계 인스턴스의 수가 적으며, 따라서, 패턴 후보의 증가 비율 또한 비교적 낮다. 의미 패턴의 경우, 하나의 상의어에 연결되는 어휘가 다양하기 때문에 추출되는 관계의 수가 어휘 패턴이나 중심어 패턴에 비하여 매우 높은 편이다. 따라서, 패턴의 증가 비율 또한 매우 높다. 또한, 초기 패턴이 늘어날수록 한 개의 구문 단위에 대하여 매우 다양한 의미적 상의어가 나타나는 경우 또한 증가하므로, 고려하는 자질의 개수가 증가하여, 확장패턴 후보에 대한 증가비가 기하급수적으로 증가한다.

<표 8>에서 한 개의 패턴에 의하여 추출되는 의미관계 인스턴스의 수가 Y-type(577개)이 H-type(1개)보다 월등히 많은데도 불구하고, <표 9>에서 최종적으로 추출된 관계 인스턴스의 개수가 의미 패턴의 경우에 중심어 패턴의 경우보다 적은 이유는 다음과 같다. 첫째, Y-type에 의하여 선택된 패턴의 개수가 H-type에 비하여 적기 때문이다. 둘째, 의미 패턴의 경우, 한 개의 초기 패턴에서 다변화 과정을 통하여 매우 다양한 형태의 패턴으로 확장된다. 한 개의 초기 패턴에서 확장되는 패턴들은 신뢰도가 비슷한 경우가 많아서 한 번의 반복과정에서 함께 선택되는 경우가 많다. 이들 패턴은 서로 일반화-구체화의 관계에 있기 때문에 추출되는 관계들이 중복되어 나타나는 경우가 많다.

예를 들어, <표 8>에서 Y-type의 “equipment:DNP equipment:RNP 0:” 패턴은 본래 “signal#interference#conversion#object:RNP:transmitter#equipment#device:DNP 0: 0:”와 같이 7개의 상의어 자질을 지닌 패턴으로부터 2^7 개의 패턴으로 확장된 패턴의 하나이므로, 유사한 신뢰도의 패턴을 매우 다양하게 가지는 반면, “offer:shogun:DOM 0: :RNG 0:”와 같은 H-type 패턴은 “offer: :shogun:DOM 0: implementation:RNG 0:”와 같은 3개의 중심어 자질을 지닌 패턴으로부터 확장된 패턴이므로, 유사한 신뢰도의 패턴이 비교적 적다. 따라서, 추출되는 관계가 겹치지 않는 다양한 패턴을 취하게 되기 때문에, 한 개의 패턴에 의하여 추출되는 관계의 개수가 비교적 적음에도 불구하고, 총수는 오히려 많음을 알 수 있다.

<표 10>에서는 선택된 패턴의 예로서 어휘(R-type) 패턴에 의한 관계 추출 과정에서 최종적으로 선택된 10개의 패턴을 보인다. 이 중, “pass: can: :DNP 0: 0: through: transformers:RNP 0: 0:”와 같이 자질이 구체적이고 많은 패턴일수록 산출 되는 관계의 수가 적고, “:DNP receiver:RNP 0:”와 같은 자질의 수가 적은 패턴일수록 추출되는 관계가 많다. 또한, <표 10>에서 알 수 있듯이, 두 패턴이 서로에 대하여 일반화와 구체화의 관계가 존재할 경우[3], 신뢰도가 유사하게 나타나는 경우가 있어서, 같은 관계로부터 추출되는 패턴이 유사한 순위에 나타나는 경향이 있다. 이러한 특성 때문에 유사한 패턴이 서로 다른 반복과정에서 추출됨으로써, 불필요한 반복을 되풀이하는 경우가 발생한다. 이와 같은 문제의 해결을 위해서 패턴 간의 상하위 관계 구조의 구축 및 활용에 관한 연구를 진행 중이다.

품사 자질(P-type)에 대한 실험과 구문 자질(S-type)에 대한 실험에서는 첫번째 반복과정에서 “p_toTransmit” 관계에 대한 “N:RNG V: N:DOM 0: 0:”, “p_toReceiveFrom” 관계에 대한 “OBJ:RNG :DOM 0:”를 최고 신뢰도 패턴으로 추출할 수 있었다. 하지만, 위의 두 패턴에 의하여 추출되는 관계의 수는 73052개와 17764개로 너무 많고 대부분이 매우 부정확한 결과를 보이므로, 정확률 산출 및 이후 반복과정의 적용이 불가능하다. 즉, 품사와 구문 자질 패턴은 높은 확장성과 낮은 신뢰도를 보인다.

5.3 평가

관계의 평가는 전문가의 수동 판단에 의한 정확률을 이용한다. 정확률은 (8)에 의하여 계산한다.

$$ER = \{r \mid r\text{은 말뭉치로부터 추출된 관계}\}$$
$$CR = \{r \mid r \in ER, r\text{은 평가자가 정확하다고 판단한 관계}\}$$
$$P = \frac{|CR|}{|ER|} \tag{8}$$

각 패턴별 성능평가를 위한 패턴 정확률을 (9)에 의하여 계산한다.

$$ER_p = \{r \mid r\text{은 말뭉치로부터 패턴p에 의하여 추출된 관계}\}$$
$$CR_p = \{r \mid r \in ER_p, r\text{은 평가자가 정확하다고 판단한 관계}\}$$
$$P_p = \frac{|CR_p|}{|ER_p|} \tag{9}$$

정확도(P) 평가를 위하여 최종 추출 결과 중 빈도수를 기준으로 한 개의 의미관계를 평가 대상으로 정한다.

어휘 패턴에 의한 정확도 평가 결과, 63개의 결과가 맞는

〈표 9〉 각 자질별 모델에 대한 반복 실험결과

모델	반복회수	초기패턴 후보 증가비	확장패턴 후보 증가비	선택된 패턴 개수	최종 추출된 관계 인스턴스 개수	정확률
R-type	8회	1.08	1.87	10	133	48.1%
H-type	9회	1.12	1.29	23	3029	51.0%
Y-type	3회	15.3	1091	14	1471	37.0%

3) 예들, pass: can: signal:DNP 0: 0: through: :RNP 0: 0: “pass: can: :DNP 0: 0: through: :RNP 0: 0:

〈표 11〉 패턴별 정확도 평가

관계	패턴	추출관계수	정확도
p_toSendTo	:DOM receiver:RNG 0:	68	53.6%
p_beTransformedBy	pass: can: :DOM 0: 0: through: :RNG 0: 0:	14	35.7%
p_beTransformedBy	pass: can: bpl:DOM 0: 0: through: :RNG 0: 0:	1	100%
p_beTransformedBy	pass: can: signal:DOM 0: 0: through: :RNG 0: 0:	1	100%
p_beTransformedBy	pass: can: signal:DOM 0: 0: through: transformers:RNG 0: 0:	3	100%
p_toSendTo	transmitter:DOM :RNG 0:	38	26.3%
p_toSendTo	transmitter:DOM receiver:RNG 0:	8	87.5%

것으로 판단되었고, 48.1%의 결과를 얻었다. 패턴별 정확도는 <표 11>과 같다.

“pass: can: signal:DNP 0: 0: through: transformers:RNP 0: 0:”와 같이 패턴 내에 자질의 개수가 많아서 다양한 제약이 가하여지는 경우, 산출되는 패턴의 수는 많지 않지만 100%의 정확률을 보인다. “transmitter:DNP receiver:RNP 0:”와 같이 자질에 의한 제약조건의 개수는 많지 않지만 어휘 자질을 이용한 패턴 표현 중 가장 구체적인 형태의 경우 역시 87.5%의 비교적 높은 정확률을 보인다.

반면, “pass: can: :DNP 0: 0: through: :RNP 0: 0:”와 같이 자질의 개수는 많지만 일반화의 정도가 비교적 높은 패턴의 경우와 “DNP:receiver :RNP 0:”와 같이 자질의 개수가 많지 않으며 일반화 비율이 높은 형태의 패턴 표현은 다량의 관계를 산출하여, 관계 확장에 도움이 되었지만 정확률은 매우 낮음을 알 수 있다.

6. 결 론

본 논문에서는 위키피디아에서 추출한 패턴 후보를 사용된 자질 중 일부를 생략한 일반화된 형태의 패턴 집합으로 확장하고 신뢰도에 기반하여 선택한 패턴을 바탕으로 관계를 점진적으로 확장하는 온톨로지 관계와 패턴의 확장 모델을 제시하였다.

본 모델은 현재까지의 패턴 기반 관계 추출 모델에서 시도하지 않았던 패턴 자질별 특성 및 자질 간의 결합 특성을 관찰하였다. 또한 패턴의 모든 일반화된 형태를 고려하는 패턴의 다변화 과정을 통하여, 패턴 선택 및 확장되는 관계의 범위를 넓히고자 하였다. 이를 통하여 각 자질 모델별로 다양하게 일반화된 패턴들이 확장되었고, 그 결과로 인하여 추출되는 관계 또한 다양해졌다.

본 논문에서 다룬 자질별 패턴 모델을 통하여 각각의 자질들이 관계 추출에 있어서 수행하는 역할 및 특성에 대하여 알아보았다. 어휘 자질, 중심어 자질, 상위어 자질과 같이 패턴에 단독으로 사용되었을 경우 신뢰도 있는 정보를 제공하는 자질도 있지만, 품사 자질, 구문 자질과 같이 단독으로 사용되기에는 일반화 정도가 너무 강한 자질도 있었다. 반면, 후자의 경우 다량의 관계를 제공하는 관계 확장에 유리한 정보를 제공하였다.

이와 같이 본 논문에서는 각 자질별 패턴 모델을 이용하여 각 자질의 특성을 관찰하였으나, 패턴 상에서 다양한 자질의 특성을 동시에 활용하지는 못하였다. 따라서, 자질별 특성을 활용하여 구성한 복합 자질 패턴 모델에 관한 연구가 진행 중이다. 본 논문의 실험에서, 어휘 자질은 일반적으로 연결 구문 단위[4]에서, 상위어 자질은 개념 구문 단위[5]에서, 중심어 자질은 양쪽 모두에서 중요한 역할을 함을 관찰할 수 있었다. 이러한 특성에 기반하여 복합 자질 패턴 모델에서는 개념 구문 단위 자질로 중심어와 의미 자질을 주요 자질로, 구문 자질을 보조 자질로 사용하며, 연결 구문 단위 자질로 어휘와 중심어 자질을 주요 자질로, 품사 자질을 보조 자질로 사용한다. 이를 기반으로 <표 12>와 같은 16가지의 복합 자질 패턴 모델을 정의할 수 있다. 이들을 이용하여 하나의 패턴에 두 개 이상의 자질이 결합되었을 경우의 결합 특성 및 결합력에 대하여 관찰한다. 복합 패턴 모델은 또한 단일 자질 패턴 모델에 비하여 보다 다양한 형태로의 일반화가 가능하다는 장점이 있다.

또한 각 자질별 실험 결과, 패턴 유형별로 선택된 패턴에 의하여 추출된 관계들을 관찰하면 자질의 일반화 정도와 사용된 자질의 개수에 따라서 관계의 확장성과 정확률에 뚜렷한 경향성이 드러남을 알 수 있다. 이와 같은 차별성을 통하여, 패턴의 신뢰도에 의한 선택과정에서 패턴의 일반화에 따른 영향력을 고려하여야 함을 알 수 있다. 이를 위하여, 모든 패턴 후보에 대하여 자질의 일반화 정도에 따른 패턴 간 상하위 구조를 구축하고, 구조 내에서 패턴을 선정하는 모델을 연구 중에 있다.

〈표 12〉 복합 자질 패턴 모델

연결자질 \ *개념자질*	*중심어*	*중심어+구문*	*의미*	*의미+ 구문*
어휘	HR	HSR	YR	YSR
어휘+품사	HRP	HSRP	YRP	YSRP
중심어	HH	HSH	YH	YSH
중심어+품사	HHP	HSHP	YHP	YSHP

참 고 문 헌

[1] M. Ciaramita, A. Gangemi, E. Ratsch, J. Saric and I. Rojas, “*Unsupervised Learning of Semantic Relations between Concepts of a Molecular Biology Ontology*,” in 19th

4)패턴에서 정의역과 치역을 연결하는 연결부를 표현하는 구문 단
5) 패턴에서 정의역과 치역의 개념을 표현하는 구문 단

International Joint Conference on Artificial Intelligence, 2005.
[2] C. Ramakrishnan, K. J. Kochut and A. P. Sheth, "*A Framework for Schema-Driven Relationship Discovery from Unstructured Text*," International Semantic Web Conference, 2006.
[3] R. Gaizauskas, G. Demetriou, P. J. Artymiuk and P. Willett, "*Protein Structures and Information Extraction from Biological Texts:The PASTA System*," Bioinformatics, Vol.19, Issue1, pp.135-143, 2003.
[4] C. Friedman, P. kra, H. Yu, M. Krauthammer and A. Rzhetsky, "*GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles*," Bioinformatics, Vol.17, pp.1367-4803, 2001.
[5] D. Zelenko, C. Aone and A. Richardella, "*Kernel Methods for Relation Extraction*," Journal of Machine Learning Research, Vol.3, pp.1083-1106, 2003.
[6] L. Specia and E. Motta, "*A hybrid approach for extracting semantic relations from texts*," 2nd Workshop on Ontology Learning and Population (OLP2) at COLING/ACL 2006, pp.57-64. 2006.
[7] P. Ryu and K. Choi, "*Automatic Acquisition of Ranked IS-A Relation from Unstructured Text*," Proceedings of the Workshop on From Text to Knowledge: The Lexicon/Ontology Interface, the 6th ISWC and ASWC, pp67-77, 2007.
[8] F. Luccio, A. M. Enriquez, P. O. Rieumont and L. Pagli, "*Exact rooted subtree matching in sublinear time*," Technical Report TR-01-14, Universita Di Pisa, 2001.
[9] A. Schutz and P. Buitelaar, "*RelExt: A tool for relation extraction in ontology extension*," in the Proceedings of the Fourth International Semantic Web Conference, pp.593-606. 2005.
[10] A. Yakushiji, "*Relation Information Extraction Using Deep Syntactic Analysis*," Ph.D. Thesis. University of Tokyo, 2006.
[11] G. Wang, Y. Yu and H. Zhu, "*PORE: Positive-Only Relation Extraction from Wikipedia Text*," in the Proceedings of the Sixth International Semantic Web Conference, pp.580-594, 2007.
[12] D. Chang and K. Choi, "*Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities*," Information & Processing Management, Vol.42, Issue3, pp.662-678, 2006.
[13] R. Girju, Automatic "*Detection of Causal Relations for Question Answering*," Proceedings of the 41st ACL, Workshop on Multilingual Summarization and Question Answering, 2003.
[14] R. Girju and D. Moldovan, "*Mining Answers for Causation Question*," AAAI Symposium on Mining Answers from Texts and Knowledge Bases, 2002.
[15] S. Lee and H. Kim, "*Pattern-based Extraction of Causal Relations in Korean*," accepted for publication of Proceeding of 2008 International Conference on Artificial Intelligence and Pattern Recognition(AIPR-08), 2008.
[16] C. S. G. Khoo, J. Kornfilt, R. N. Oddy and S. H. Myaeng, "*Automatic Extraction of Cause-Effect Information from Newspaper Text without Knowledge-based Inferencing*," Literary and Linguistic Computing, Vol.13, Issue4, pp.177-186, 1998.
[17] P. Pantel and M. Pennacchiotti, "*Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations*," Joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics, 2006.
[18] J. Huang, J. Shin and K. Choi, "*Enriching Core Ontology with Domain Thesaurus through Concept and Relation Classification*," OntoLex Workshop, ISWC, 2007.
[19] 이신목, 신지애, "전자장비 고장진단 질의응답을 위한 인과관계 정의 및 추출", 한국정보과학회 논문지 소프트웨어 및 응용, 제35권 5호, pp.335-346, 2008.

이 신 목

e-mail : smlee@world.kaist.ac.kr
1999년 한국과학기술원 전자전산학과 (학사)
2001년 한국과학기술원 전자전산학과 (공학석사)
2001년~현 재 한국과학기술원 전자전산학과 박사과정.
관심분야 : 온톨로지, 관계추출, 질의응답

장 두 성

e-mail : dschang@kt.com
1990년 전남대학교 전산학과(학사)
1993년 KAIST 전산학과(석사)
2005년 KAIST 전산학과(박사)
1993년~현 재 KT 미래기술연구소 수석연구원
관심분야 : 한국어 정보처리, 음성언어처리, 대화시스템 등

신 지 애

e-mail : jiae@icu.ac.kr
1983년 부산대학교 계산통계학과 (학사)
1986년 한국과학기술원 전산학과 (석사)
1990년 Columbia University, Computer Science (석사)
2004년 New York University, Computer Science (박사)
2005년~현 재 한국정보통신대학교 공학부 교수
관심분야 : AI Planning & Scheduling, Semantic Technology