텍스트에서 IS-A 관계의 자동 추출 및 순위화

류법모 최기선
한국과학기술원 전산학전공/시멘틱웹첨단연구센터
pmryu@world.kaist.ac.kr kschoi@cs.kaist.ac.kr

Automatic Acquisition of Ranked IS-A Relation from Unstructured Text

Pum-Mo Ryu Key-Sun Choi
Computer Science Division/SWRC, KAIST

요약

본 논문에서는 의존 구조 매칭과 약한 지도식 학습 방법을 적용하여 텍스트에서 IS-A 관계를 자동으로 추출하고 순위화하는 방법을 제안한다. 텍스트에서 잠재적인 IS-A 관계를 표현하는 [관계 표현, 하위어, 상위어]의 삼진관계 리스트를 추출하고, 관계 표현과 IS-A 관계 인스턴스, IS-A 관계 후보, 사이의 상호 관련성을 이용하여 각각의 점수를 반복적으로 정제한다. 제안한 방법의 대표적인 특징은 다음과 같다. 1) 의존 구조에 기반한 패턴 매칭 방법을 적용하여 정규 표현에 기반한 방법보다 다양한 형태의 삼진관계를 추출할 수 있고, 2) 도메인 코퍼스에서 통계적으로 추출한 어휘 사이의 관련성 정보를 이용하여 도메인에 적합한 IS-A 관계 인스턴스의 순위를 높일 수 있으며, 3) 관계 표현과 관계 인스턴스의 점수를 상호 관련성에 기반한 방법으로 반복적으로 점수화하여 IS-A 관계 인스턴스 사이의 변별력을 높일 수 있다. 실험에서 순위화된 관계 인스턴스는 전문가의 판단과 66%이상 일치함을 보였고, 의존 구조를 이용한 유연한 패턴 매칭 방법은 정규표현을 이용한 방법보다 43.6%의 추가적인 삼진관계를 추출하였다.

1 서론

최근에 웹과 같은 방대한 양의 텍스트에서 유용한 정보를 자동으로 추출하는 방법의 필요성이 증가하고 있다. 특히 블로그나 위키와 같은 새로운 종류의 웹 데이터는 전문적인 시스템에서 필요한 고급 정보를 내재하고 있다 [1]. 웹에서 추출할 수 있는 다양한 정보 중에서, 개체 간의 의미 관계는 질의응답 시스템과 같은 응용 시스템에서 유용하게 활용할 수 있고, 온톨로지 구축 및 확장을 위한 자원으로 사용될 수 있다 [2]. 따라서 구조화되지 않은 대규모의 텍스트에서 개체 간 의미 관계를 학습할 수 있는 효율적인 알고리즘이 필요하다.

정보 추출 작업의 초기 단계에서는 정규 표현에 기반한 패턴 매칭 방법을 이용하는 것이 일반적이었다. 그러나 각 응용 시스템이 해결하여야 하는 문제가 점점 복잡해 지면서, 특히 질의 응답 시스템에서 질의와 응답 사이의 표층 어순이 서로 다른 경우, 단순 패턴 매칭 기반의 정보 추출 방법으로 처리하기 어렵다 [3]. 의미 관계 추출 문제는 자연언어 처리에서 흔히 적용하는 동사와 인자 사이의 선택 제약 문제와 밀접한 관련이 있기 때문에, 문장에 포함된 어휘 사이의 의존 구조와 같은 구문 관계를 파악하는 자연언어 처리 기술을 적용한 관계 추출 방법에 대한 연구가 진행되고 있다 [4,5,6].

그러나, 구문 패턴에 기반한 관계 추출 방법만으로는 도메인에 적합한 관계를 선별하기 어려운 단점이 있다. 추출한 관계를 점수화하기 위하여 도메인 코퍼스에서 추출한 분포 가정에 기반한 어휘 간 관련성 정보를 이용할 수 있다. 이와 같은 점수화는 오류가 있거나 도메인에 부적절한 관계를 제거할 수 있으며 [4,7,8], 신뢰할 수 있는 관계 표현을 확장하여 관계 추출 시스템의 재현율을 높일 수 있다. 어떤 패턴이 유효한 관계를 많이 추출할 수 있는 경우 신뢰도가 높다고 한다. 따라서 유효한 관계는 신뢰도 높은 패턴과 관련성 (association)이 높다 [9].

상하위 관계 또는 IS-A 관계는 온톨로지 구조에서 기본 프레임을 제공하고, 하위개념-상위개념 관계 또는 인스턴스-소속개념 관계를 표현한다. 어휘의 문맥정보 사이의 포함관계를 이용하는 방법과 비교하여, 패턴 매칭을 이용하는 방법은 정확률이

높지만 재현율이 낮은 단점이 있다. 따라서 패턴 기반 IS-A 관계 추출 방법에서 재현율을 높이면서 정확률의 손실을 최소화하는 연구가 필요하다. 앞으로의 설명을 위하여 다음과 같이 기본 개념을 정의한다.

정의 1. **IS-A 삼진관계 (IS-A triple)**: IS-A 관계를 표현하는 [관계표현, 하위어, 상위어]의 세 개 구성요소를 가진 삼진관계를 말한다. 텍스트에서 IS-A 관계를 표현하는 패턴을 이용하여 추출한다. 예를 들어, 문장 “Content-addressable memory is a special type of computer memory used in certain very high speed searching applications.”에서 IS-A 관계를 위한 삼진관계 [be a special type of, content-addressable memory, computer memory]를 추출한다.

정의 2. **관계 표현 (relational expression)**: IS-A 삼진관계에서 하위어와 상위어 사이의 의미적인 관계에 대한 언어적인 표현을 말한다. 위의 예에서 관계 표현 “be a special type of”는 “content-addressable memory”와 “computer memory” 사이의 관계를 언어적으로 표현하고 있다.

정의 3. **관계 인스턴스 (relation instance)**: IS-A 삼진관계에서 [하위어, 상위어] 쌍을 말한다. 위의 예에서 [content-addressable memory, computer memory]는 한 개의 관계 인스턴스이다. 관계 인스턴스는 IS-A 관계의 후보가 된다.

본 연구에서는 문장의 의존 구조에 기반한 패턴 매칭 방법을 이용하여 자동으로 텍스트에서 IS-A 관계를 표현하는 삼진관계 리스트를 추출하고, 삼진관계에 포함된 IS-A 관계 인스턴스에 점수를 부여하여 순위화하는 방법을 제안한다. 관계 표현의 신뢰도가 높은 경우 그 관계 표현으로 연결된 관계 인스턴스의 신뢰도도 높다고 가정하고, 반대 방향으로, 관계 인스턴스의 신뢰도가 높은 경우 그 관계 인스턴스를 연결하는 관계 표현의 신뢰도도 높다고 가정한다. 따라서 추출한 삼진관계 리스트에서 나타나는 모든 관계표현과 관계 인스턴스의 신뢰도를 상호 관련성을 이용하여 정제되는 방향으로 반복적으로 계산한다. 한편 구문 구조에 기반한 패턴 매칭 방법은 기존의 정규 표현에 기반한 방법보다 재현율을 높이는 특징이 있다. 또한 해당 도메인 코퍼스에서 추출한 어휘 사이의 통계적인 관련성은 도메인에 적합한 관계를 추출하도록 하는 특징이 있다. 제안하는 방법은 추출하려는 관계의 종류에 의존적이지 않기 때문에, 기본 패턴만 다시 정의하면 전체-부분 관계와 같은 다른 종류의 의미 관계 추출에 쉽게 적용할 수 있는 특징이 있다.

이 논문의 구성은 다음과 같다. 2장에서는 기존의 패턴 기반 관계 추출 방법들의 장점 및 유용성을 설명한다. 3장에서는 위키피디아를 이용한 도메인 코퍼스 구축 및 분석, IS-A 관계 추출을 위한 패턴 및 패턴 매칭 프로세스, 학습 알고리즘 등 제안한 방법을 자세히 설명한다. 4장에서는 실험 내용을 설명하고, 실험 결과를 다양한 측면에서 분석한다. 마지막으로 5장에서는 제안한 방법의 특징과 장단점을 정리하고, 향후 연구 방향을 제안한다.

2 관련연구

90년대 초반 Hearst [10] 의 연구 이후로 패턴 매칭 방법을 이용하여 관계를 추출하거나 발견하는 많은 연구들이 진행되고 있다 [1, 11, 12]. 그 이후로 자동으로 온톨로지를 구축하거나, 질의응답 시스템에서 정확한 정답을 찾기 위하여 자연언어 처리 기술을 이용하여 개체 사이의 관계를 추출하는 연구가 시작되었다. 단순한 패턴 매칭 기술로는 인식하기 어려운 잠재적인 관계 삼진관계를 의존 구조 분석을 통하여 인식하였다. 이 방법은 영어에서 동사와 간접 목적어 사이의 관계를 인식하거나 거리가 먼 단어 사이의 의존관계를 파악할 수 있게 해 준다. Litkowiski [13]은 RDFS 데이터베이스를 구축하기 위하여 자연언어 처리 기술을 적용하여 질의 응답 시스템을 위한 온톨로지 자동 구축 방법의 가능성을 보였다. Reinberger & Spyns [14], Sabou [15], Specia & Motta [2] 등은 텍스트에서 온톨로지를 자동으로 학습하는 과정에서 의존 구조 분석 방법을 적용하였다. Reinberger & Spyns [14]는 생물의료 분야의 코퍼스에서 개체 사이의 관계를 추출하기 위하여 (술어-목적어) 또는 (전치사-명사) 사이의 의존 구조에 빈도수 정보를 통합하는 통계적인 관계 추출 방법을 적용하였다. Sabou [15]는 웹서비스 관련 문서에서 분야 전문적인 온톨로지를 추출하는 방법을 제안하였다. 이 방법에서는 온톨로지 계층 구조를 구축하기 위하여 자연언어 처리 도구를 적용하였고, 생물의료분야 서비스에서 추출한 온톨로지를 전문가 관점에서의 적절성과 정답 온톨로지 (gold standard)와 비교를 통하여 평가하였다. Specia & Motta [2]도 단순한 패턴 매칭 방법의 문제점을 극복하기 위하여 구문 의존 파서를 이용하여 삼진관계를 생성하였다.

한편 부적절하거나 오류가 있는 관계를 제거하기 위하여 관계 추출 단계에서 새로운 의미적 제약조건을 추가하는 연구가 진행되었다. Gliozzo et al. [7]는 의미적으로 관련이 있는 어휘는 도메인 코퍼스에서도 공기할 가능성이 높다고 가정하고, 도메인 제한 가설 (domain restriction hypothesis)를 제안하였다. 이 연구에서는 패턴 기반 방법과 분포 가정에 기반한 제약 조건을 결합하여 의미 관계의 도메인 전문적인 특징과 발화의 통합적(syntagmatic) 특성을 적용하였다. 발화의 통합적 특성의 관점에서

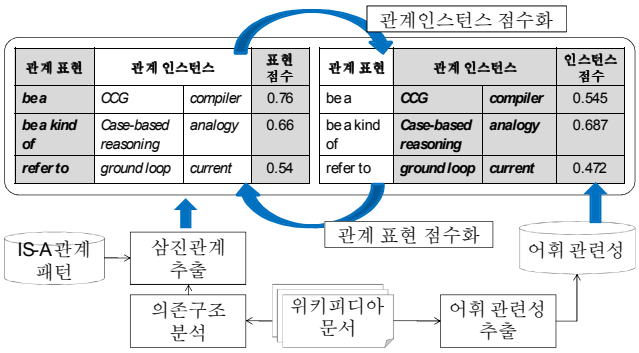

그림 1 시스템 구성 및 IS-A 관계 점수 계산

는 두 개의 개체 X와 Y가 어떤 의미 관계에 의하여 연결될 때, 코퍼스에서 공기할 가능성이 높고, 따라서 특정 어휘-구문 패턴에 의해 연결될 것이라는 가정을 한다. 도메인 전문적인 특징의 관점에서는 두 개체 X, Y 사이에 어떤 의미 관계가 존재할 때, X와 Y가 동일한 도메인에 포함되어야 한다고 가정한다. 분포 가정 (distributional hypothesis)에 기반하여 주어진 도메인 말뭉치에서 두 어휘의 공기 단어 집합이 유사한 경우 두 어휘가 동일한 도메인에 포함된다고 말한다. 이와 유사한 방법으로 Cimiano & Wenderoth [8]는 웹에서 추출한 통계정보를 사용하여 어휘의 특질 구조 (qualia structure)를 순위화하였다. 어휘의 특질 구조는 어휘의 네 가지 역할을 표현한다. 이 연구에서는 주어진 어휘와 역할을 표현하는 어휘 사이에 웹 기반의 통계정보를 이용한 상호 연관성이 높은 경우에 높은 점수를 부여하였다.

공동학습(co-training)은 자질들을 두 가지 자질 집합으로 분리한 후 서로 다른 관점으로 학습하고, 그 결과를 상호 지원하는 방향으로 통합하는 학습 알고리즘을 말한다. 패턴 기반의 관계 추출 과정을 공동학습의 관점에서 모델링하면, 패턴의 점수화 작업과 추출한 삼진관계의 점수화 작업이 서로 상대방의 결과를 이용하여 반복적으로 진행되도록 설계할 수 있다. Pantel & Pennacchiotti [9]은 이 방법을 적용하여 패턴의 신뢰도 점수와 관계 후보의 신뢰도 점수를 서로 상대방의 신뢰도 점수를 이용하여 반복적으로 계산하였다. Greenwood & Stevenson [16]는 반지도식 알고리즘을 이용하여 추출한 삼진관계와의 관련성과 기존 패턴과의 유사도를 이용하여 점진적으로 복잡한 패턴을 학습하였다.

3 IS-A 관계 추출 및 순위화 방법

이 장에서는 그림 1과 같이 텍스트의 구문 구조 분석, 도메인 지식 적용 그리고 반복적으로 점수를 정제하는 방법에 기반한 패턴 매칭 기반의 IS-A 관계 추출 방법을 설명한다. (3.1) 절에서는 사용한 도메인 코퍼스에 대하여, (3.2) 절에서는 의존 구조 분석에 기반한 패턴 매칭 방법에 대하여, (3.3) 절에서는 도메인 지식의 적용 방법에 대하여, 마지막으로 (3.4) 절에서는 추출한 삼진관계를 점진적으로 순위화하는 과정을 설명한다.

표 1. 구축된 위키피디아 문서 집합에서 참조하는 고빈도 문서 이름과 빈도수

문서 이름	참조된 회수
Computer	3,731
Software	3,460
Data	3,307
Operating system	3,284
Film	3,020
Television	2,777
Linux	2,572
Radio	2,409
Set	2,409
Algorithm	2,397

3.1 도메인 코퍼스

본 연구에서 도메인 코퍼스는 다음과 같은 두 가지 작업 단계에서 이용한다. 첫 째, 패턴 매칭 방법을 적용하여 IS-A 관계를 표현하는 삼진관계를 코퍼스에서 추출하고, 둘 째, 도메인 코퍼스 기반의 통계정보를 이용하여 도메인 지식을 모델링한다.

위키피디아(http://en.wikipedia.org) 에서 추출한 100,000개의 문서 집합 (약 2천6백만 단어)을 도메인 코퍼스로 사용한다. 이 문서 집합은 '*Computer science*', '*Information technology*' 와 같은 정보 기술(IT) 분야 카테고리와 그 하위 카테고리에 포함된 문서로 구성된다. 표 1은 코퍼스에 포함된 문서들이 가장 많이 참조하는 상위 10개의 위키피디아 문서의 이름과 참조되는 회수를 나타낸다. 대부분의 고빈도 문서들이 정보통신 분야와 밀접한 관련이 있다. 위키피디아 문서를 관계 추출 대상 코퍼스로 사용할 때의 장점은 다음과 같다. 첫 째, 문서 집합의 규모가 충분히 크기 때문에 원하는 정보를 추출할 가능성이 높고, 둘 째, 문서들이 카테고리 단위로 잘 조직화되어 있기 때문에, 카테고리 단위로 문서를 처리할 수 있는 특징이 있다[1].

3.2 구문 구조 분석

Connexor 영어 의존 구문 구조 분석기[2]를 사용하여

[1] 현재 위키피디아 영어 버전은 180만 개 이상의 문서를 포함한다.

[2] http://www.connexor.com

코퍼스를 분석한다. 이 분석기는 입력문장에 대하여 단어 단위의 품사정보, 어근 정보, 단어 사이의 구문 의존 정보, 의존 관계의 종류를 생성한다. 의존 구조를 이용하면 단순한 패턴 매칭 방법에서 발견하기 어려운 동사와 간접 목적어 관계, 먼 거리 의존 관계 등을 파악할 수 있다. 그림 2는 주어진 문장을 분석하여 얻어진 의존 구조 정보를 표현한다.

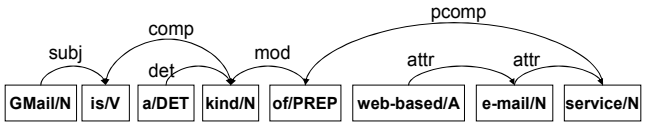

그림 2. “GMail is a kind of web-based e-mail service.”에 대한 의존 구조 분석 결과

3.3 삼진관계 추출

의존 구조가 분석된 문서에서, 의미 관계의 후보가 되는 언어적인 삼진관계 (linguistic triple)을 추출한다. 언어적인 삼진관계는 잠재적인 의미 관계를 표현하는 관계 표현 (relational expression)과 관계 표현에 의하여 연결되는 두 개의 어휘 (relation instance)로 구성된다 [2]. IS-A 관계 추출을 위하여 표 2와 같은 패턴을 적용하여 삼진관계를 추출한다. 이 패턴들은 사전의 정의문을 만들 때 일반적으로 사용되는 형식이다. NP_i는 단일 명사 또는 명사 구를 나타낸다. 삼진관계에서 첫 번째 인자는 “*is a*”, “*is a kind of*”와 같이 IS-A 관계에 대한 언어적인 표현이다. 두 번째, 세 번째 인자는 (하위어, 상위어) 또는 (인스턴스, 클래스) 관계를 표현하는 관계 인스턴스이다. 패턴 1이 패턴 2보다 제약 조건이 높기 때문에 본 연구에서는 패턴 1을 패턴 2보다 먼저 적용한다. 패턴 3은 패턴 1, 2와 관계없이 적용할 수 있다.

위의 패턴을 의존 구조로 표현하고, 패턴의 의존 구조와 실제 텍스트의 의존 구조 트리를 비교/탐색하면서 매칭 여부를 판단한다. 그림 3은 패턴 1에 대한 의존 구조 표현을 보여준다.

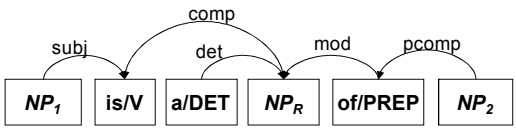

그림 3. IS-A 관계 추출을 위한 패턴의 의존 구조

패턴 매칭 과정에서 다음과 같은 언어학적인 휴리스틱이 적용된다.

- 참조 표현 (referring expression)은 문맥에 의존적인 지식을 표현하기 때문에 참조 표현을 포함하는 어휘를 가진 삼진관계는 제외한다. “*the computer*”, “*his computer*”, “*this*”와 같은 참조 표현은 문맥 내에서 다른 개체를 참조할 때 사용하기 때문에 이 표현을 포함하는 삼진관계는 문맥 의존적인 정보를 표현한다.
- 부사, 형용사 등 패턴의 구성 요소를 수식하는 경우는 매칭 과정에서 허용한다. 따라서 “*is a special type of*” 과 “*is also a kind of*”와 같은 관계 표현은 패턴 1 에 매칭한다.
- 관계 표현에서 부정의 의미를 포함하는 삼진관계는 제외한다. 예를 들어서 (*is not a kind of*, *Windows*, *computer*)은 관계 표현이 부정의 의미를 가지기 때문에 IS-A 관계를 표현하지 않는다.

표 2. IS-A 관계 추출을 위한 패턴, 패턴을 적용하여 추출되는 삼진관계 및 적용 예

	패턴	삼진관계
1	NP_1 is a NP_R of NP_2	(*be a* NP_R *of*, NP_1, NP_2)
	GMail is a *kind* of *web-based e-mail service*. → (*be a kind of*, *GMail*, *web-based e-mail service*)	
2	NP_1 is a NP_2	(*be a*, NP_1, NP_2)
	Microsoft is a software company. → (*be a*, *Microsoft*, *software company*)	
3	NP_1 refer to NP_2	(*refer to*, NP_1, NP_2)
	In an electrical system, ground loop refers to a current, generally unwanted. → (*refer to, ground loop*, *current*)	

3.4 도메인 지식 적용

기존의 패턴 기반의 관계 추출 방법은 자연 언어의 의미적인 특성 상 다음과 같은 문제점을 가진다 [7].

- **도메인에 적합하지 않은 관계**: 구문적으로 유효한 관계이지만 원하는 도메인 지식을 표현하지 못하는 관계를 추출하는 경우가 있다. 예를 들어, (*be a*, *Solaris*, *computer operating system*)는 (*be a*, *Solaris*, *novel*)보다 IT 분야에 더 적합한 관계이다. 따라서 앞의 관계 인스턴스에 더 높은 점수를 부여하여야 한다.
- **도메인에서 의미없는 관계**: 패턴 기반의 관계 추출 방법에서는 적용하는 패턴의 커버리지가 지나치게 넓은 경우, 도메인에서 의미없는 관계를 추출할 수 있다. 예를 들어 아래 문장에서 (*be a*, *source code*, *component*)를 추출하는 경우, “*component*”의 범위가 너무 넓기 때문에 “*source code*”의 상위 개념이 되기에 적절하지 않다.

예문) *Source code* is a *component* in the activity of porting software to alternative computer platforms.

이와 같은 문제점은 분포 가정 (distributional hypothesis)을 기반으로 도메인 코퍼스에서 추출한 어휘 사이의 관련성을 이용하여 줄일 수 있다. 어휘 사이의 관련성은 관계 추출 과정에서 도메인 제약의 역할을 한다. 본 연구에서는 log-likelihood ratio에 기반하여 제안한 통합 이론에 기반한 어휘 관련성 추정 방법 (syntagmatic word association measure)을 적용한다 [17]. 두 어휘가 코퍼스에서 우연히 공기하는 사건보다 더 자주 공기하는 경우, 두 어휘 사이에는 관련성이 있고 따라서 특정 어휘-구문 패턴에 의해 연결될 것이라는 가정한다. 예를 들어, IT 분야에서 전형적으로 관련이 있는 어휘 쌍은 (*problem*, *solve*), (*Unix*, *operating system*) (*image*, *jpeg*) 등이 있다.

3.5 관계 표현과 관계 인스턴스 순위화

본 연구에서는 Espresso 알고리즘 [9]을 선형적인 모델로 변형한 방법을 사용하여 관계 표현과 관계 인스턴스를 순위화한다. 이 알고리즘은 패턴과 관계 인스턴스의 신뢰도를 상호 연관성과 각자의 신뢰도를 이용하여 추정하는 방법을 제안하였다. 이 방법의 기본적인 원리는 신뢰도가 높은 패턴은 정확한 관계 인스턴스를 많이 추출할 수 있으며, 반대로 신뢰도가 높은 관계 인스턴스는 여러 개의 신뢰도가 높은 패턴과 높은 관련성을 가진다는 가정에 기반한다. 본 연구에서는 Espresso 알고리즘과 달리 관계 표현과 관계 인스턴스의 관련성을 이용하여 관계 인스턴스의 신뢰도를 계산한다. 본 연구에서는 의존 구조에 기반한 패턴 매칭 방법을 적용하기 때문에 다양한 관계 표현을 발견할 수 있다. 따라서 각각의 관계 표현에 별도의 신뢰도를 부여하면 정교한 관계 추출 방법을 모델링할 수 있다. 예를 들어 "*be a kind of*"와 "*be a informal kind of*" 모두 표 2의 패턴 1에 의해서 추출되지만 IS-A 관계를 표현하는 능력에는 차이가 있다. 관계 표현 *e*의 신뢰도 $r_e(e)$ 는 *e*와 모든 관계 인스턴스 사이의 평균 관련성으로 정의한다. 이 때 각 인스턴스의 신뢰도를 이용하여 가중치를 부여한다 (식 1). 관계 인스턴스 *i*의 신뢰도는 관계표현 신뢰도 계산방법과 유사하게 관계표현과의 관련성을 이용하여 계산한다 (식 2).

$$r_e(e) = \frac{\sum_{i \in I} \left(\frac{pmi(i,e)}{\max_{pmi}} * r_i(i) \right)}{|I|} \qquad (1)$$

$$r_i(i) = \frac{\sum_{p \in P} \left(\frac{pmi(i,e)}{\max_{pmi}} * r_e(e) \right)}{|E|} \qquad (2)$$

여기에서 $pmi(i,e)$ 는 *i*와 *e* 사이의 pointwise mutual information [18]을 나타내고, $\max_{pmi}$ 는 인스턴스와 관계표현 사이의 모든 가능한 *pmi* 중 최대값을 나타낸다. *I*와 *E*는 각각 모든 가능한 관계 인스턴스 집합과 관계 표현 집합을 나타낸다.

위의 신뢰도 계산 방법을 반복적으로 적용하여 관계표현과 인스턴스의 신뢰도를 계산하는 경우, 각 신뢰도가 급격하게 0으로 수렴하여 데이터 관찰을 어렵게 하는 문제가 있기 때문에, 선형적으로 값이 변할 수 있도록 다음과 같이 수정한 수식을 적용한다. 기본적인 가정은 Espresso 방법과 동일하고, 각 반복 단계별로 신뢰도는 다른 관계표현 또는 관계 인스턴스의 평균 값과의 차이만큼 변화한다. *k+1*번째 반복에서 관계 표현 *e*의 신뢰도 $r_e^{k+1}(e)$ 는 전단계 신뢰도에 현재 상태에서 e의 신뢰도와 전체 관계표현의 평균 신뢰도의 차이를 합한 값으로 정의한다 (식 3). 현재 상태의 신뢰도가 평균보다 높은 값인 경우 전체 신뢰도가 높아지고, 반대의 경우는 낮아진다. 초기 관계표현 신뢰도는 0.5로 지정한다.

$$r_e^{k+1}(e) = r_e^k(e) + \eta \left(r_e(e) - \frac{\sum_{e_j \in E} r(e_j)}{|E|} \right) \qquad (3)$$

$$r_e^1(e) = 0.5$$

여기에서 $r_e(e)$는 식 4를 이용하여 계산한 전단계의 인스턴스 신뢰도를 식 1에 적용하여 계산한다.

*k+1*번째 단계에서 인스턴스 *i*의 신뢰도 $r_i^{k+1}(i)$는 전단계 신뢰도에 현재 상태에서의 *i*의 신뢰도와 평균 인스턴스 신뢰도 사이의 차이를 더한 값으로 정의한다. 초기 인스턴스 신뢰도는 3.4장에서 설명한 어휘 사이의 분포가정에 의한 관련성을 더한다. 이 때, 이 관련성은 0과 0.5사이의 값으로 정규화된다.

$$r_i^{k+1}(i) = r_i^k(i) + \eta \left(r_i(i) - \frac{\sum_{i_j \in I} r(i_j)}{|I|} \right) \qquad (4)$$

$$r_i^1(i) = 0.5 + assoc(NP_1, NP_2)$$

여기에서 $r_i(i)$는 식 3을 이용하여 계산한 현재 상태의 관계표현 신뢰도를 식 2에 적용하여 계산한다.

표 3. 기본 관계 표현과 확장된 관계 표현의 수 및 각 관계 표현으로 연결된 관계 인스턴스의 수

	기본 관계 표현	확장된 관계 표현	계
관계 표현 수	812	781	1,593
관계 인스턴스 수	13,714	5,981	19,695

표 4. 첫 번째 반복과 100번째 반복 후 신뢰도가 높은 상위 패턴 10개 리스트

순위	1회 반복		100회 반복	
	관계 표현	신뢰도	관계 표현	신뢰도
1	be a	0.6663	be a	16.0608
2	refer to	0.5640	be a type of	12.6196
3	be a type of	0.5512	refer to	10.5990
4	be a form of	0.5343	be a form of	4.9978
5	be also a	0.5263	be also a	3.5924
6	be a method of	0.5218	be a kind of	3.4252
7	be a series of	0.5186	be a series of	3.3885
8	be a kind of	0.5162	as be a type of	3.1791
9	be a set of	0.5160	be a method of	3.0558
10	be a branch of	0.5141	be a branch of	2.5274

4 실험

실험에서 3.3장의 삼진관계 추출 방식을 3.1장의 코퍼스에 적용하여 19,695개의 IS-A 관계를 위한 삼진관계를 추출하였다.

먼저, 삼진관계를 추출하는 단계에서 의존 구조 기반 패턴 매칭 방법이 기존의 정규표현 기반 방법에 비하여 풍부한 관계 인스턴스를 추출할 수 있음을 보인다. IS-A 관계를 표현하는 기본 관계 표현 "*be a*", "*be a* NP_R *of*", "*refer to*"에 수식어를 부가한 형태인 "*especially be a*", "*as be a series of*", "*be also a common type of*", "*typically refer to*"와 같은 확장된 관계 표현이 포함된 삼진관계를 추출할 수 있고, 확장된 관계 표현을 이용하여 추가적인 관계 인스턴스를 추출할 수 있다. 표 3은 추출한 삼진관계 중에서 기본 관계 표현의 수와 확장된 관계 표현의 수를 보여준다. 또한 각각의 관계 표현으로 연결된 관계 인스턴스의 수를 나타낸다. 확장된 관계 표현을 적용하는 경우 추가적으로 43.6%의 후보 관계 인스턴스를 더 추출할 수 있음을 알 수 있다.

표 5. 도메인 정보를 적용한 실험 결과에서 상위 15 개 관계 인스턴스

관계 인스턴스	초기 신뢰도	최종 신뢰도	IS-A
software, computer	0.5501	0.5575	X
red, color	0.5511	0.5522	O
Megatron, Decepticon	0.5475	0.5467	O
role-playing game, game	0.5000	0.5385	O
Linux, system	0.5384	0.5376	O
backup, data	0.5344	0.5365	O
metadata, data	0.5324	0.5363	O
Canada, country	0.5369	0.5360	O
Norway, country	0.5351	0.5345	O
heat, energy	0.5327	0.5343	O
green, color	0.5344	0.5336	O
copyleft, license	0.5317	0.5334	O
data, information	0.5293	0.5294	O
ram, memory	0.5252	0.5288	O
Finland, country	0.5294	0.5286	O

표 4는 신뢰도 계산을 각각 1번과 100번 반복한 후 신뢰도가 높은 상위 10개의 관계 표현을 보여준다. 1회 반복 결과에 '*be a set of*'가 포함된 것과, 100회 반복 결과에서 '*as be a type of*'가 포함된 경우를 제외하고 두 리스트의 내용을 유사하다. '*as be a type of*' 가 '*be a set of*' 보다 IS-A 관계를 더 잘 표현하기 때문에 100회 반복 결과에서 신뢰도가 높아진 것으로 판단된다. 또한 IS-A 관계를 표현하는 전형적인 패턴인 '*be a type of*', '*be a kind of*' 의 순위가 100번 반복 후 높아졌다. 첫 번째 반복 이후의 관계 표현 사이의 신뢰도 차이가 크지 않지만, 100회 반복 후에는 패턴 사이의 신뢰도 차이가 커졌다. 이 현상은 신뢰도 있는 관계표현이 반복 프로세스를 거치면서 신뢰도 있는 인스턴스와의 높은 관련성에 의하여 더 높은 신뢰도를 가지게 됨을 보여준다.

표 5와 표 6은 각각 3.4절에서 설명한 도메인 지식을 적용한 경우와, 적용하지 않은 경우 상위 15개의 관계 인스턴스를 보여준다. 표 5에서 "*role-playing game*, *game*" 인스턴스를 제외하고 다른 모든 인스턴스는 초기 신뢰도가 0.5 이상이다. 즉 IT 분야에서 두 엔티티가 관련성이 있는 경우, 높은 초기 신뢰도를 가지고, 관계 표현 신뢰도와의 반복 계산에 의하여 지속적으로 높은 신뢰도를 가진다. 상위 인스턴스는 "*data*", "*software*", "*memory*"와 같이

표 6. 도메인 정보를 적용하지 않은 실험 결과에서 상위 15 개 관계 인스턴스

관계 인스턴스	초기 신뢰도	최종 신뢰도	IS-A
role-playing game, game	0.5000	0.5219	O
Stacheldraht, DDoS tool	0.5000	0.5155	O
email filter, user software	0.5000	0.5148	O
games, culture	0.5000	0.5144	O
black, mourning	0.5000	0.5139	X
Kuro, nobility	0.5000	0.5137	X
Winnuke, Nuke	0.5000	0.5109	O
black metal, music	0.5000	0.5103	O
Colette, melee fighter	0.5000	0.5102	O
the black parade, music	0.5000	0.5099	O
black death, pandemic	0.5000	0.5098	O
Captcha, variant	0.5000	0.5097	X
storage virtualization, abstract	0.5000	0.5077	X
Sheena, game	0.5000	0.5076	O
telephone exchanges, circuit switching	0.5000	0.5076	O

IT 분야의 기본적인 개념을 포함하거나 국가, 색깔의 관계를 포함한다. 국가와 색깔과 관련된 인스턴스가 많이 포함된 이유는 위키피디아 사이트의 특성 상 해당 어휘를 많이 표현하고 있기 때문으로 판단된다. 표 6의 관계 인스턴스는 표 5 보다 IT 분야에서 더 자세한 개념의 관계 인스턴스를 포함하거나, 음악, 영화, 애니메이션, 질병 등의 다양한 분야의 관계를 표현하고 있다. 따라서 도메인 정보를 사용하는 경우 IT 분야의 기본 개념 사이의 IS-A 관계를 추출할 수 있음을 알 수 있다.

그림 4는 각 10번째 반복결과에서 신뢰도를 기준으로 상위 100개 관계 인스턴스의 정확률을 보여준다. 두 명의 전문가가 동시에 정확하다고 판단한 경우에 정답이라고 인정하였다. 반복 회수가 증가할수록 정확률이 높아지는 것을 알 수 있다. 따라서 제안한 알고리즘이 정확한 IS-A 관계를 추출할 때 점진적으로 관계 인스턴스의 점수를 정제하여 상위 인스턴스와 하위 인스턴스 사이의 변별력을 높임을 알 수 있다. 표 4에서 관계 표현 "*be a*"의 신뢰도가 가장 높은데, 실제 상위 관계 인스턴스 중에서 "*be a*"로 연결된 인스턴스의 수가 가장 많았다.

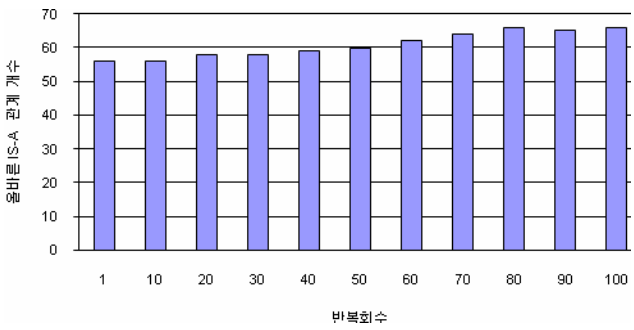

그림 4. 10회 단위의 반복 결과에서 상위 100개 인스턴스 중 정확한 관계 인스턴스의 개수

5 결론

본 연구에서는 범용의 패턴 매칭 기반의 IS-A 관계 추출 방법을 제안하였다. 이 방법은 패턴과 텍스트 사이의 의존 구조 매칭을 이용하여 IS-A 관계를 표현하는 삼진관계를 추출하고, 코퍼스 기반의 도메인 지식과, 관계 표현 및 관계 인스턴스 반복적이고 상호 작용 기반의 신뢰도 학습 방법을 적용하여 추출한 관계 인스턴스를 순위화하였다. 의존 구조 기반의 패턴 매칭 방법은 단순한 정규 표현 기반의 패턴 매칭 방법에 비하여 유연한 매칭을 가능하게 하여 추출하는 관계의 재현율을 높일 수 있었다. 분포 가정에 기반한 엔티티 사이의 관련성 정보는 해당 분야에 적합한 관계를 추출할 수 있게 하였다. 마지막으로, 반복적이고 상호 작용에 의한 관계 표현과 관계 인스턴스의 점수화 방법을 통하여 추출한 후보를 순위화하였다.

제안한 방법은 IS-A 관계에 의존적이지 않기 때문에 전체-부분 관계, 원인-결과 관계 등 다른 종류의 의미 관계 추출에도 적용할 수 있다. 한편 알고리즘의 반복 과정의 종료 조건에 대한 연구가 진행 중이다.

감사의 글

본 논문은 정보통신부 및 정보통신연구진흥원의 정보통신선도기반기술개발사업의 연구결과로 수행되었습니다.

참고문헌

[1] Alfonseca, E. et al.,: Towards Large-scale Non-taxonomic Relation Extraction: Estimating the Precision of Rote Extractors: In. Proceedings of the 2nd Workshop on Ontology Learning and Population (2006) 49-56

[2] Specia, L., & Motta, M.: A hybrid approach for extracting semantic relations from texts: In. Proceedings of the 2nd Workshop on Ontology Learning and Population (2006) 57-64

[3] Lo, K. K. & Lam, L.: Using Semantic Relations with World Knowledge for Question Answering: In. Proceedings of the 15th Text Retrieval Conference (TREC 15) (2007)

[4] Schutz, A. & Buitelaar, P.: RelExt: A Tool for Relation Extraction from Text in Ontology Extraction: In. Proceeding of International Semantic Web Conference (2005)

[5] Ciramita, M., Gangemi, A., Ratsch, E., Saric, J., Rojas, I.: Unsupervised learning of semantic relations between concepts of a molecular biology ontology: In. Proceedings of the 19th International Joint Conference on Artificial Intelligence (2005)

[6] Gamallo, P., Gonzalez, M., Agustini, A., Lopes, G., de Lima, V.S.: Mapping syntactic dependencies onto semantic relations: In. Proceedings of the ECAI Workshop on Machine Learning and Natural Language Proceeding for Ontology Engineering (2002)

[7] Gliozzo, A. M., Pennacchiotti, M. Pantel, P.: The Domain Restriction Hypothesis: Relating Term Similarity and Semantic Consistency: In. Proceedings of NAACL HLT (2007) 131-138

[8] Cimiano, P., Wenderoth, J.: Automatic Acquisition of Ranked Qualia Structures from the Web: In. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (2007) 888-895

[9] Pantel, P. Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations: In. Proceedings of the 44th Annual Meeting of the Association of Computational Linguistics (2006) 113-120

[10] Hearst, M.: Automatic Acquisition of Hyponyms from Large Text Corpora: In. Proceedings of the 14th International Conference on Computational Linguistics (1992)

[11] Morin, E.; Jacquemin, C.: Automatic Acquisition and Expansion of Hypernym Links: Journal of Computers and the Humanities, Vol. 38, Num. 4, (2004) 363-396

[12] Cimiano, P., Pivk, A., Schmidt-Thieme, L., & Staab, S.: Learning Taxonomic Relations from Heterogeneous Sources of Evidence: In Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, Amsterdam, (2005)

[13] Litkowski, K. C.: CL Research Experiments in TREC-10 Question Answering: In. Proceedings of the 10th Text Retrieval Conference (2001) 121-131

[14] Reinberger, M. L., Spyns, P.: Discovering knowledge in texts for the learning of DOGMA-inspired ontologies: In. Proceedings of the ECAI 2004 Workshop on Ontology Learning and Population (2004) 19-24

[15] Sabou, M., Learning Web service ontologies: An automatic extraction method and its evaluation: In Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, Amsterdam, (2005)

[16] Greenwood, M. A., Stevenson, M.: Improving Semi-Supervised Acquisition of Relation Extraction Patterns: In. Proceedings of the Workshop on Information Extraction Beyond The Document (2006) 29-35

[17] Rapp, R.: The Computation of Word Associations: Comparing Syntagmatic and paradigmatic Approaches: In. Proceedings of the 15th International Conference on Computational Linguistics (2002)

[18] Manning, C. D., Schütze, H.: Foundations of Statistical Natural Language Processing: MIT Press (1999) 178-183