질의응답시스템을 위한 시맨틱 프레임 정제

서지우○, 함영균, 최기선
한국과학기술원, Semantic Web Research Center
{jiwoo35, hahmyg, kschoi}@kaist.ac.kr

Semantic-Frame Granularity Refinement for QA system

Jiwoo Seo○, YoungGyun Hahm, Key-sun Choi
KAIST, Semantic Web Research Center

요 약

본 연구는 한국어 프레임 파서를 개발하기 위한 기초 단계로서, 질의응답시스템 구축을 위하여 한국어 텍스트를 시맨틱 프레임으로 표현하고자 한다. 본 연구에서는 표본 질문셋을 기반으로 필요한 프레임을 선별하고, 프레임넷 온톨로지를 이용하여 선별된 프레임을 통합함으로써 프레임 인덱스를 정제하였다. 그 결과 질의응답에 대응하는 최적화된 시맨틱 프레임이 구해진다. 본 연구에서 발생한 이슈 및 향후 과제는 다음과 같다—프레임뿐만 아니라 프레임 요소 고려, 시맨틱 프레임 정제 과정을 표본 질문셋 이외의 질의텍스트에 적용하기 위한 일반화된 알고리즘 구현, 정제된 시맨틱 프레임 결과에 대한 평가

1. 서론 및 문제 정의

질의응답 시스템을 구현하기 위해서 최근 DBpedia, Freebase와 같은 LOD (Linked Open Data) 형태의 지식베이스를 활용하는 연구개발이 늘어가고 있다[1]. 이미 서비스가 되고 있는 LOD로 부족한 경우에는 우선적으로 자연언어 텍스트 또는 웹데이터로부터 질의에 대응할 수 있는 의미 있는 정보를 추출하여 지식 데이터베이스를 생성해야 한다.

본 연구에서 활용한 프레임넷 (FrameNet)[1]은 텍스트 원문에 대하여 의미 정보 및 구문 특성들을 추출한 데이터베이스를 말한다[2]. 선행 연구에서 영어 프레임넷 코퍼스를 번역하여 한국어 프레임넷 코퍼스[2]를 자체적으로 구축하였다[3].

프레임넷과 달리 디비피디아 (DBpedia)는 위키피디아의 Infobox와 같은 구조화된 정보들을 활용한 지식 데이터베이스이기 때문에 텍스트의 비구조화된 정보를 표현하기에는 한계가 있다. 따라서 프레임넷을 활용하여 디비피디아로 표현되지 않는 지식들을 추가적으로 표현하고자 하는 것이 연구의 장기적인 목표이다[3].

질의응답시스템을 구축하기 위해서 한국어 텍스트 질문에 대하여 프레임을 자동으로 부여할 수 있는 파서를 개발할 필요가 있다. 이를 위해서 기존에 번역된 한국어 프레임넷 코퍼스의 4025개의 문장에 대하여 결합 패턴 (valence pattern)과 구문 정보 등을 파악하여 추가적인 프레임 인덱스 (Frame Index, 이하 프레임)와 프레임 요소 (Frame Element)를 주석화하는 과정이 필요하다.

질의응답시스템의 성능을 향상시키기 위해서 기존 1179개의 프레임을 모두 사용하는 것은 한계가 있다고 판단된다. 2장의 관련 연구에서 알 수 있듯이 다른 연구에서도 모든 프레임을 사용하지 않고 목적에 맞게 프레임을 선별하여 사용하고 있다[5]. 따라서 불필요한 프레임들을 제거하고 비슷한 의미를 나타내는 프레임들을 통합하여 질의응답시스템에 필요한 프레임을 최적화하는 것이 본 연구의 목적이다.

본 연구에서는 한국어 자연어 질의 텍스트에 대하여 필요한 프레임을 선별하고 프레임넷 온톨로지를 적용하여 비슷한 프레임들을 통합함으로써 프레임을 정제하였다. 2장에서는 관련 연구들을 소개하고, 3장에서는 프레임 정제 과정을 설명한다. 4장에서는 시맨틱 프레임 정제 결과를 보여주고, 5장에서는 결론 및 향후 연구 계획에 대하여 논의한다.

2. 관련 연구

본 연구와 관련하여 프레임넷을 활용하여 텍스트로부터 의미 정보를 추출한 라트비아 프레임넷의 FrameNet CNL이 있다[5]. FrameNet CNL은 뉴스 기사를 대상으로 26개의 프레임을 선별하여 별도의 프레임 온톨로지를 구현한 뒤 프레임 시맨틱 파서를 개발하였다. 이와 마찬가지로 본 연구에서도 한국어 질의응답 텍스트로부터 필요한 프레임을 선별하고 프레임넷 온톨로지를 적용하여 질문 텍스트를 프레임으로 표현하고자 한다.

또한, 한국어 프레임넷 코퍼스의 문장들에 대한 결합 패턴을 생성하는 과제와 관련하여 스웨덴 프레임넷과 영어 프레임넷의 문장 패턴과 결합 패턴을 비교한 연구가 있다[6].

[1] https://framenet.icsi.berkeley.edu/fndrupal/
[2] http://framenet.kaist.ac.kr/framenet/

3. 시맨틱 프레임 정제과정

프레임넷에 기반하여 텍스트에 대하여 용언(Predicate)에 해당하는 프레임과 프레임과의 관계를 표현하기 위해 의미 정보를 부여한 프레임 요소를 주석화한 사례가 있다[7]. 본 연구에서는 일차적으로 한국어 텍스트에 대하여 프레임을 부여하였고, 향후 프레임만으로 표현되지 않은 어휘요소(Lexical Unit)[3]들은 프레임 요소를 적용하여 추가적으로 주석화할 계획이다.

한국어 질의응답시스템을 구축하기 위해서 프레임을 선별하기 위한 평가셋이 필요하였다. 이번 연구에서는 50개의 표본 질문셋(이하 NLQ50)을 기반으로 하여 프레임을 선별하였다. NLQ50은 장학퀴즈, 퀴즈대한민국, 도전골든벨 등에서 출제된 질의응답텍스트를 질문, 정답, 문제유형, 도메인으로 정리한 평가셋이다. 문제유형은 기본형, 괄호/대명사 채우기, 조합/연상형, 다지선다형, 순서제시형으로 되어 있으며, 도메인은 사회문화, 과학, 시사상식, 예술 등의 다양한 분야를 담고 있다.

3.1 시맨틱 프레임 선별 과정

한국어 프레임넷 코퍼스의 어휘요소들에 부착된 프레임을 참고하여 NLQ50의 한국어 질의 텍스트에 대하여 해당되는 프레임들을 선별하였다. 추출된 프레임들 중에서 중복된 프레임을 제거하고, 각각의 프레임들이 NLQ50에서 사용된 횟수를 파악하였다. 그 결과 NLQ50에 대해 중복되지 않고 유일한 프레임 개수는 총 156개로 선정되었다(부록 1 참조).

문제유형: **기본형** / 도메인: 과학(물리) / 정답: 뢴트겐
이 **사람이**People **발견한**Becoming_aware X선은 각종 **질병을**Medical_Condition **발견하는데**
Becoming_aware 획기적인 **기여를**Giving 했고, 영상 **의학**Medical_specialties **시대를**Calendric_unit 연
계기가 됐다. **최초의**Ordinal_numbers 노벨 물리학**상을**Finish_competition **받은**Receiving/Getting 독
일의 이 **과학자는**People_by_vocation

그림 1 NLQ50에 대한 프레임 부여 예시

그림 1은 NLQ50의 기본형 질문에 대하여 프레임을 부여한 예시를 보여준다. 각각의 어휘요소에 대하여 문자열 매칭을 통해 프레임을 부여하였다. 그림 1에서 붉은 글씨는 프레임이 존재하는 어휘요소들을 의미하고, 파란 글씨는 각 어휘요소에 해당하는 프레임을 나타낸다. 예를 들어, '사람이'라는 어휘요소에 대하여 "People"이라는 프레임이 부여되었다.

3.2 프레임넷 온톨로지를 활용한 프레임 통합 과정

질의와 응답을 표현하는 시맨틱 프레임 집합을 구하는 과정을 본 연구에서는 "프레임 통합 과정"이라고 한다. 계층구조가 표현된 프레임넷 OWL 온톨로지[4]를 사용하여 비슷한 프레임들을 상위 프레임으로 통합하는 정제과정을 설명하고자 한다.

프레임넷 온톨로지에는 상속 (Inheritance) 관계 계층구조와 서브프레임 (Subframe) 계층구조가 표현되어 있다. 그림 2는 상속 관계 계층구조의 일부를, 그림 3은 서브프레임 계층구조의 일부를 보여준다.

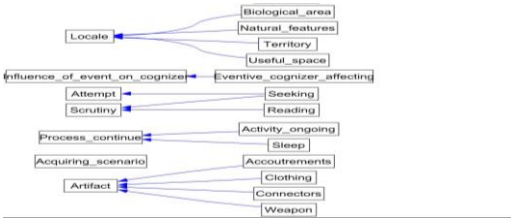

그림 2 프레임넷 상속 관계 계층구조

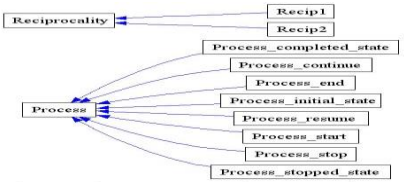

그림 3 프레임넷 서브프레임 계층구조

프레임넷 온톨로지를 활용하여 추출된 프레임들을 상위 프레임으로 통합한 과정은 크게 다음과 같다.

첫째, 같은 어휘요소에 대하여 두 개 이상의 프레임이 부여된 경우, 해당 프레임들 간의 계층구조를 고려하여 상위 프레임으로 연결하였다. 예를 들어, 그림 1에서 '받은'이라는 어휘요소에 대하여 "Receiving"과 "Getting" 프레임이 부여되었는데, 프레임넷 상속 관계 계층구조를 적용하여 "Receiving"은 "Getting"으로부터 상속되었기 때문에 상위 프레임인 "Getting"으로 정제하였다. 상위 프레임 "Getting"을 사용함으로써 한국어 질의 텍스트에서 나타나는 "받은", "얻은" 등과 같은 비슷한 의미의 어휘요소들을 포괄적으로 표현할 수 있다는 장점이 있다.

둘째, 프레임 개수를 축소하기 위하여 최대한 하위 프레임들을 통합하여 상위 프레임으로 연결하고자 하였다. 예를 들어, "Placing", "Political_locales", "Locale_by_use" 등과 같은 지역, 위치와 관련된 프레임들은 상위 프레임인 "Locale"로 통합하였다. 상위 프레임 "Locale"을 사용하여 "위치한", "도시", "항구" 등과 같은 장소를 나타내는 어휘요소들을 통합된 프레임으로 표현할 수 있다. 또한, 프레임넷 온톨로지에 비슷한 의미의 하위 프레임들을 통합할 수 있는 계층정보가 표현되어 있지 않은 경우, 이들을 통합할 수 있는 새로운 상위 프레임을 정의하였다(표 1). 예를 들어, 의학 분야와 관련된 프레임을 하나의 "Medical_frame"으로 정의함으로써 질의 텍스트를

[3] 텍스트 원문에서 표현되는 실제 문자열을 의미한다.

[4] http://rhizomik.net/ontologies/2005/07/FrameNet_1.1.owl

통합된 시맨틱 프레임으로 표현할 수 있다.

표 1 새로 정의한 상위 프레임

비슷한 의미의 하위 프레임	새로운 상위 프레임
Medical_specialites, Medical_professionals, Medical_conditions, Medical_instruments	Medical_frame
Relative_time, Time_vector, Timespan	Time_frame

4. 시맨틱 프레임 정제 결과

3절에서 설명한 시맨틱 프레임 정제과정을 통하여 3.1절에서 예시된 NLQ50 질의 텍스트에서 선별된 156개의 프레임들 중에서 두 번 이상 사용된 프레임들에 대하여 그 결과 총 53개의 프레임이 선별되었다. 최종적으로 선별된 53개의 프레임들은 표 2와 같다. 볼드체는 NLQ50에서 사용된 횟수가 높은 상위 25개의 프레임을 의미한다.

최종적으로 정제된 프레임을 2절의 관련 연구에서 언급한 FrameNet CNL에서 사용된 26개의 프레임과 비교하였다. 표 2에서 별표로 표시된 프레임들은 정제된 프레임 결과 중에서 FrameNet CNL에서도 사용된 프레임을 의미한다. 괄호 안의 프레임은 해당 프레임에 대하여 FrameNet CNL에서 같은 의미로 사용된 프레임을 의미한다.

표 2 정제된 시맨틱 프레임 리스트

Accompaniment	***Earnings_and_losses**	**Locale**
Age	Evidence	Manufacturing
Aggregate	Existence	**Medical_frame**
Becoming	Experiencer_focus	**Natural_features**
***Being_born**	Fields	**Ordinal_numbers**
Being_named	***Finish_competition**	**Organization**
Building	(Win_prize)	**Origin**
***Businesses**	**Frequency**	Part_whole
(Being_employed,	**Getting**	***People**
Hiring,	**Hostile_encounter**	(People_by_origin,
Employment_end,	Impact	People_by_vocation)
Product_line)	Increment	**Posession**
Calendric_unit	Instance	Process
Causation	**Intentionally_act**	**Purpose**
Cause_change	***Intentionally_create**	**Quantity**
Choosing	*Kinship	Research
Coming_up_with	(Personal_relationship)	***Text**
Containing	Law	Time_frame
Craft	***Leadership**	Type
Desirability	(Change_of_leadership)	
Dimension	Likelihood	

정제된 시맨틱 프레임들을 분석하는 과정에서 예상되는 이슈 및 향후 과제는 다음과 같다. 첫째, 향후 프레임뿐만 아니라 프레임 요소를 고려하여 한국어 질문 텍스트에 대하여 추가적인 의미 정보를 부여해야 한다. 둘째, NLQ50에 대하여 프레임을 선별하였기 때문에 다른 한국어 질의 텍스트에 대해서도 통합된 프레임을 적용할 수 있는 일반화된 알고리즘이 필요하다. 셋째, 현재 정제된 53개의 프레임 개수가 적정한가에 대한 평가과제가 남아 있다.

5. 결론 및 향후 연구

본 연구에서는 질의응답시스템 구축을 위하여 한국어 질문 텍스트를 시맨틱 프레임으로 표현하고자 하였다. 그 과정에서 필요한 프레임을 선별하고 프레임넷 온톨로지를 사용하여 프레임 개수를 축소하고 최종적으로 사용할 프레임을 정제하였다.

정제된 프레임을 바탕으로 향후 질의응답시스템 구축을 위하여 결합 패턴을 파악하여 한국어 텍스트에 대하여 프레임을 사용하여 의미 정보를 부여할 수 있는 파서를 개발할 계획이다.

사사

본 연구는 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음 [10044494, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발]

참고문헌

[1] Mohamed Yahya, Klaus Berberich, Shady Elbassuoni, Maya Ramanath, Volker Tresp, Gerhard Weikum, “Natural Language Questions for the Web of Data”, EMNLP-CoNLL, 2012

[2] Charles J. Fillmore, Christopher R. Johnson, Miriam R.L. Petruck, “Background to FrameNet”, Int J Lexicography, Vol. 16, Issue 3, 235-250, 2003

[3] 함영균, 서지우, 황도삼, 최기선, “프레임넷을 통한 디비피디아 온톨로지 인스턴스 생성의 커버리지 개선”, 한글 및 한국어 정보처리 학술대회, 2014

[4] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann. "DBpedia-A crystallization point for the Web of Data" Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 7, Issue 3, 154-165, 2009

[5] Guntis Barzdins, “FrameNetCNL: a Knowledge Representation and Information Extraction Language”, Controlled Natural Language, Vol. 8625, 90-101, 2014

[6] Dana Dannélls, Normunds Grūzītis, “Extracting a bilingual semantic grammar from FrameNet-annotated corpora”, LREC, 2466-2473, 2014

[7] Marco Fossati, Sara Tonelli, Claudio Giuliano, “Frame Semantics Annotation Made Easy with DBpedia”, ISWC, 2013