텍스트 기반 지식요소 추출을 위한 온톨로지 활용 방안에 관한 연구

강유리
한화시스템

yuri.k54@hanwha.com

A Study on the utilization of ontology for extracting knowledge elements from text data

Yu Ri Kang
Hanwha Systems

요 약

다양한 종류의 정보가 실시간으로 발생하는 현대 전장환경에서는 적군의 거짓 정보, 아군의 자원 오류 요인 등으로 인해 혼재된 전장환경 정보 발생 가능성이 높다. 이러한 리스크가 있는 상황에서 지휘관이 올바른 상황판단을 하도록 지휘결심을 지원하는 인공지능 기술의 필요성이 높아지고 있다. 본 논문은 DARPA 의 AIDA 프로그램 중 GAIA 의 연구 분석을 통해 현 국방 환경에서 효과적으로 다종의 데이터로부터 추출한 지식요소 통합 및 정확한 상황 분석을 위한 지식요소 추출 단계들을 분석한다. 이 분석을 바탕으로 온톨로지를 활용한 텍스트 기반 지식요소 추출 방안의 국방 분야의 적용을 제안한다.

Ⅰ. 서 론

현대 전장환경에서는 적군의 거짓 정보와 아군 정찰 자원의 오류 등으로 인해 혼재된 전장환경 정보가 지휘관에게 전달된다. 또한 다종의 빅데이터가 초단위로 발생하여 축적되는 상황에서 이를 최대한 활용해 지휘관의 올바른 지휘결심을 지원할 인공지능 기술이 필요하다.

GAIA(GAIA: A Fine-grained Multimedia Knowledge Extraction System)는 DARPA AIDA 프로그램 중 멀티미디어 및 다국어 환경에서 지식요소를 추출하는 시스템이다. 추출한 지식요소들은 텍스트와 이미지에 표현된 상황을 분석하여 상황 분석 결과를 제시하는데 사용된다. GAIA 는 데이터 종류로는 텍스트, 이미지와 언어로는 영어, 러시아어, 우크라이나어를 다룬다. [1]

본 논문은 GAIA 의 AIDA 온톨로지와 GAIA 가 다루는 조건 중 영어 텍스트 데이터에서의 지식요소 추출 단계 분석을 통해, 온톨로지를 활용한 텍스트 지식요소 추출이 다종의 데이터 통합과 정확한 상황 분석에 효과적인 요소임을 제안한다.

Ⅱ. 본 론

본론 Ⅱ.1.에서는 2018 년부터 2020 년까지의 GAIA 온톨로지 구성을 보인다. Ⅱ.2.에서는 온톨로지를 세분화하여 다양화했을 때 지식요소 추출을 통해 상황 분석에서 얻을 수 있는 이점을 소개하며, 온톨로지를 활용한 영어 텍스트에서의 지식요소 추출 과정을 분석한다.

Ⅱ.1. AIDA 온톨로지

온톨로지는 지식요소 추출을 위해 추출할 정보를 정의하기 때문에 효과적으로 정보를 표현할 수 있는 구조가 필요하다. GAIA 는 Entity, Relation, Event 를 상황 표현을 위한 온톨로지 기본 구성요소로 두며, 특히 서로 다른 데이터 종류에 대해서도 상황을 표현할 수 있는 멀티모달 온톨로지를 구현 목표로 설정했다.

그러나 GAIA 는 기존의 온톨로지는 너무 대분화(coarse-grained)되어 있거나 너무 세분화(fine-grained)되어 있어 그대로 차용하지 않고, 여러 단계의 기준을 두고 YAGO 온톨로지를 정제하여 Entity 온톨로지를 구현하거나, 다수 온톨로지들에 대해 여러 단계의 기준을 두고 추가하여 Event 온톨로지를 구현함으로써 AIDA 온톨로지를 제작했다. [3] AIDA 온톨로지 예시는 Figure1 과 같다. [4]

```
/** Entity classes **/
IdcOnt:Person  a        owl:Class ;
      rdfs:subClassOf  aidaDomainCommon:CanHaveName ,
aidaDomainCommon:EntityType .

/** Event types **/
IdcOnt:Conflict.Attack
      a              owl:Class ;
      rdfs:subClassOf  aidaDomainCommon:EventType ;
      rdfs:subClassOf  [ a              owl:Restriction ;
                  owl:allValuesFrom  [ a          owl:Class ;
                              rdfs:label   "Attacker" ;
                             owl:unionOf  ( IdcOnt:GeopoliticalEntity
IdcOnt:Organization IdcOnt:Person )
                              ] ;
                  owl:onProperty     IdcOnt:Conflict.Attack_Attacker
                     ...
                ] .

/** Relation/Event arguments **/
IdcOnt:Conflict.Attack_Attacker
      a              owl:ObjectProperty ;
      rdfs:label        "Attacker" ;
      rdfs:subPropertyOf  owl:topObjectProperty .
```

Figure 1: AIDA 온톨로지 예시

GAIA 가 온톨로지를 세분화할 때는 각 Type 에 하위 레벨을 만드는 type.subtype.subsubtype 형태로, Type 을 세분화하는 방법을 이용했다. 그 형태는 Figure2 와 같다. [5]

1. top level type for the most coarse-grained level (e.g., PER)
2. type.subtype for the next level (e.g., PER.Politician)
3. type.subtype.subsubtype for the finest-grained level (e.g., PER.Politician.Governor)

Figure 2: AIDA 온톨로지 세분화 형태

서로 다른 종류의 데이터가 하나의 상황에 대해 서로 상충되는 정보를 포함할 때, GAIA 는 AIDA 온톨로지를 활용해 추출한 각 지식요소를 최신 개발한 Visual Grounding System 을 통해 단일의 일관된 정보를 표현하도록 멀티미디어 결과들을 통합한다. 이로써, 데이터 종류를 넘어서 Entity 상호참조해결을 통한 다종의 데이터로부터의 지식요소 추출 결과 통합이 가능하다.

GAIA 는 Entity 세분화 type 수를 2018 년 163 개에서 2019 년 187 개로, Event 세분화 type 수를 2018 년 114 개에서 2019 년 139 개, 2020 년 144 개까지 확장하며 AIDA 온톨로지를 정교하게 세분화했다. 가장 최신 발표된 2020 년 GAIA 의 AIDA 온톨로지 구성은 Figure3 과 같다. [2]

	Coarse-grained Types	Fine-grained Types
Entity	7	187
Relation	23	61
Event	47	144

Figure 3: 2020 년 AIDA 온톨로지 구성

Ⅱ.2. 온톨로지를 활용한 텍스트 기반 지식요소 추출

GAIA 는 Ⅱ.1.과 같이 온톨로지를 구성하여 지식요소 추출 시스템에 적용했다. GAIA 가 2018 년 연구부터 2020 년까지 온톨로지 세분화 변화를 기반한 지식요소 추출을 통해 상황 분석에서 얻을 수 있던 이점은 크게 2 가지이다.

1. 상황 이해도 향상
2. 이벤트 예측도 향상

Type 이 세분화 되면, 발생한 Event 에 어떠한 대상이 이어서 오는지에 따라 정확한 상황 이해를 지원하며, 또한 뒤에 어떠한 Event 가 등장할 가능성이 높을지 역으로 예측 확률도 높아진다. [2]

Figure4 는 GAIA 멀티미디어 지식요소 추출 전체 아키텍처이며, 영어 텍스트에서 지식요소 추출 단계와 각 방법은 아래와 같다. [1]

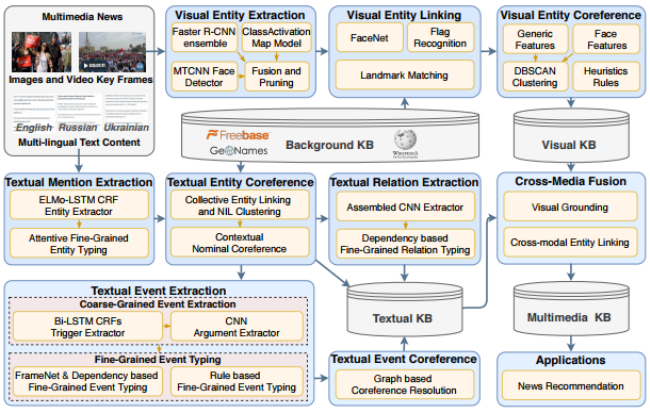

Figure 4: GAIA 멀티미디어 지식 추출 아키텍처

1. 텍스트 Entity 추출과 상호참조해결
   1.1. 대분화된 Mention 추출
      - ELMo LSTM-CRF 모델을 이용하여 Entity Mention 추출 및 "대분화 Type" 할당
   1.2. Entity Linking 과 상호참조해결
      - 대용량의 외부 데이터베이스를 포함한 지식베이스에 동일할 Entity 가 존재하는지 확인 및 상호참조해결
      - 존재하지 않을 시, 룰 이용해 NIL 클러스터링
   1.3. Entity Typing 세분화
      - 상호참조해결 정보, 지식베이스, 분류기를 이용하여 대분화 Type 을 바탕으로 "세분화 Type" 재할당
   1.4. Entity 중요도 랭킹
      - 가중치에 따라 Entity Mention 중요도 랭킹
2. 텍스트 Relation 추출
   - CNN 이용해 "대분화 Relation Type 할당" 후, Entity Type 제약 조건과 언어 의존 패턴 기반 혹은 룰 기반으로 "세분화 Relation Type" 재할당
3. 텍스트 Event 추출과 상호참조해결
   - Bi-LSTM CRF 와 CNN 사용해 "대분화 Event Type" 할당 후, 룰 비교를 통해 "세분화 Event Type" 재할당
   - 그래프 알고리즘을 통해 Event 상호참조해결

Ⅲ. 결 론

기존 인공지능 모델은 데이터 종류에 의존적이기 때문에 지식요소를 추출하여도 통합적으로 활용하여 분석하는데 어려움이 있었다. GAIA 는 AIDA 온톨로지를 구현해 지식요소 추출 및 데이터 통합에 활용함으로써 다종의 데이터로부터 추출한 지식요소를 통합하여 상황 분석을 위한 Event 정보를 보다 정확하고 풍부하게 표현하는 시스템을 만들었다.

현대 전장환경은 대량의 다종 데이터가 발생하며, Event 를 중심으로 정확한 상황 분석이 중요하다. 이 점은 혼재된 대량의 전장환경 정보를 통합하고 분석하여 지휘관의 지휘결심을 지원하는 지능형 전장인식 기술 개발의 필요성으로 이어진다.

본 논문은 GAIA 연구 분석을 통해, 현 국방 분야의 지능형 전장인식 기술 개발에 GAIA 의 온톨로지를 활용한 지식요소 추출 방법 적용이 적합한 것을 확인하였다. 향후 본 논문의 연구 결과를 기반으로 온톨로지를 활용한 텍스트 기반 지식요소 추출 시스템을 개발할 예정이다.

참 고 문 헌

[1] Li, Manling, et al. "Gaia at sm-kbp 2019-a multi-media multi-lingual knowledge extraction and hypothesis generation system." Proceedings of TAC KBP (2019).

[2] Li, Manling, et al. "Gaia: A fine-grained multimedia knowledge extraction system." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020.

[3] Zhang, Tongtao, et al. "GAIA-A Multi-media Multi-lingual Knowledge Extraction and Hypothesis Generation System." TAC. 2018.

[4] NIST "Text Analysis Conference" NIST TAC, 30 August 2018, https://tac.nist.gov/tracks/SM-KBP/2018/ontologies/SeedlingOwlOntology, accessed 27 November 2020

[5] NIST "Text Analysis Conference" NIST TAC, 28 June 2019, https://tac.nist.gov/2019/SM-KBP/guidelines/SM-KBP_2019_Evaluation_Plan_V1.5.pdf, accessed 27 November 2020