ExoBrain을 위한 한국어 의미역 가이드라인 및 말뭉치 구축

임수종[O], 권민정[1], 김준수[1], 김현기
한국전자통신연구원 자동통역인공지능연구센터, ㈜솔샘넷[1]
{isj, hkk}@etri.re.kr, {huristic1030, jskim}@solsam.net

Korean Proposition Bank Guidelines for ExoBrain

Soojong Lim[O], Minjung Kwon[1], Junsu Kim[1], Hyunki Kim
Automatic Speech Translation and AI research Center ETRI, SolsamNet[1]

요 약

본 논문은 한국어 의미역을 정의하고, 기계학습에 기반하여 한국어 의미역 인식 기술을 개발할 때 필요한 학습 말뭉치를 구축할 때 지켜야할 가이드라인을 제시하고자 한다. 한국어 의미역 정의는 전세계적으로 널리 쓰이고 있는 Proposition Bank를 따르면서, 한국어의 특성을 반영하였다. 또한 정의된 의미역 및 태깅 가이드라인에 따라 반자동 태깅 툴을 이용하여 말뭉치를 구축하였다.

주제어: 한국어 의미역 정의, PropBank, 의미역 태깅 가이드라인

1. 서론

동사나 형용사는 한 문장을 완성하기 위해 필수적으로 요구하는 성분들이 있다. 이를 필수 보어라 한다. 필수 보어는 서술어와 통사적 관계를 맺을 뿐만 아니라 특정한 의미적 관계를 맺게 되는데 이것을 보어의 의미역이라 부른다[1].

통사적 관계의 경우 언어적인 특성을 제외하면 공통 분모가 존재할 수도 있지만, 의미역의 경우 한국어의 경우 세종전자사전에서 제시한 15개 의미역을 포함해서 많은 연구에서 서로 다른 의미역을 정의한다[2, 3, 4]. 영어권 또한 예외는 아니지만, 기계학습이 가능한 수준의 학습데이터를 구축하여 이를 기반으로 CoNLL shared task를 4번이나 개최하는데 기여한 Proposition Bank (PropBank) [5]에서 정의한 방식이 널리 쓰이고 있다.

본 논문에서는 언어학적인 관점이 아닌 기계학습을 전제로 한 학습 데이터를 구축한다는 관점에서 의미역을 결정하고 이를 기반으로 하여 반자동 태깅 툴을 구축하여 질의응답 시스템에서 정답을 찾는데 사용하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 한국어 의미역 가이드라인을 제시하고, 3장에서는 의미역별로 자세한 태깅 가이드라인을 설명하며, 4장에서는 정의한 의미역을 이용하여 의미역 태깅 말뭉치를 구축하는 과정에 대해서 설명한다. 마지막으로 5장에서 결론을 기술한다.

2. 한국어 의미역 가이드라인

본 논문에서는 의미역 정의는 PropBank[5]의 정의를 따르지만 이와 유사하게 한국어를 대상으로 구축된 Korean Proposition Bank(KPB)[6]의 경우 한국어의 실정을 제대로 반영하지 못하는 측면이 있어서 ExoBrain 과제를 위한 학습 및 평가셋을 구축하는 과정에서 수립한 원칙 및 특징을 소개한다.

2.1 기본 원칙

KPB for ExoBrain 가이드라인의 기본 원칙은 다음과 같다.

- 자연언어처리를 위한 일관성 유지와 효율성 제고에 초점을 두되, 일반 언어학적 관점에서 크게 벗어나지 않도록 한다.
- 영어권에서 전산학적 언어처리를 위해 수립한 PropBank의 의미역 표지와 원칙을 바탕으로 분석하며 한국어 특성을 반영한다.
- 의존 의미역 분석의 단위로 어절을 사용한다.
- 서술어에 연결된 의미역에 대해서만 태깅하며, 세종구문 태그 중 기능 태그에 해당하는 주어(SBJ), 목적어(OBJ), 부사어(AJT)에 대해서는 필수적으로 태깅한다.
- 한 문장 내에 대상의 의미역이 중복될 때, 구문 분석 결과가 직접 연관된 것에 태깅한다.
- 한국어 특징 중 하나인 빈번한 생략 현상에 의해 주어, 목적어, 부사어 성분이 생략된 경우에는 구문적으로 관계가 없더라도 작업자가 판단하여 필수적으로 태깅한다.
- 동사와 동사의 관계에 대하여 태깅할 때 필수격인 ARG-N은 태깅하지 않고 부가격인 ARG-M만을 태깅한다.

- 구문 태깅 결과, 세종 구문 태그 VP_MOD로 설정된 동사의 경우 다른 태깅 요소가 없어도 동사를 삭제하지 않는다.
- KPB의 Frame Set에서 사동/피동에 따라 태깅에 혼동이 있을 경우 동사를 사동으로 변환하여 의미적으로 판단한다.

2.2 의미역 구분

본 논문에서는 PropBank 의미역을 사용하는데 PropBank의 의미역 태그셋은 필수격인 ARG-N, 부가격인 ARG-M의 두 가지로 분류한다. ARG-N 형태의 표지는 필수격으로 불리우며, 해당 서술어에 대해 필수적이며, Frame Set에 기술 대상이 된다. 한국어의 경우에는 생략이 빈번하게 일어나기 때문에, 필수격이라고 하더라도 실제 문장에서는 기술되어 있지 않는 경우도 많다. 필수격은 논항 뒤에 숫자가 붙어 있으며, 구문 태그는 주로 주격, 목적격에 해당하지만, 부사어도 용언에 따라 필수적인 경우 해당할 수 있다. 필수격에 해당하는 의미역은 표1과 같다.

표 1 KPB for ExoBrain 필수격

의미역	정의
ARG0	서술어의 동작주, 행위자
ARG1	서술어의 피동작주, 대상
ARG2	시작점, 수익자 등
ARG3	착점

위 정의는 절대적인 기준이 아니면, KPB Frame Set의 기준에 따라 태깅하며, Frame Set이 없는 경우 다음에 설명할 ARG-N 태그셋 별 태깅 가이드라인의 내용을 참조하여 위 정의를 따른다.

표 2 KPB for ExoBrain 부가격

의미역	정의
ARGM-LOC	장소 (locatives)
ARGM-DIR	방향 (directional)
ARGM-CND	조건 (condition)
ARGM-MNR	방법 (manner)
ARGM-TMP	시간 (temporal)
ARGM-EXT	범위 (extent)
ARGM-PRD	보조 서술 (secondary predication)
ARGM-PRP	목적 (purpose clauses)
ARGM-CAU	발생 이유 (cause clauses)
ARGM-DIS	담화 연결 (discourse)
ARGM-NEG	부정 (negation)
ARGM-INS	도구 (instrument)

부가격에 해당하는 ARG-M은 모든 서술어에 필수적이지는 않지만 부가적인 의미 관계에 대한 것이다. 주로 수식어, 한정어, 서술어 별 필수적이지는 않은 시간, 장소, 조건, 방법 등에 해당한다. 부가격에 해당하는 의미역은 표2와 같다.

3. 의미역별 태깅 가이드라인

3.1 필수격:ARG-N

PropBank의 필수격은 ARGA와 ARG 0~5, 그리고 R, C와 조합 형태로 다양하지만, 한국어 ARG-N 태그셋은 ARG0부터 ARG3까지 4가지의 의미역 표지만 해당한다. 각각의 의미역은 서술어 별로 다른 정의를 갖고 있고, 특정 번호를 규정하는 일반적인 원칙은 표1과 같지만, 절대적인 대원칙은 존재하지 않고 Frame Set에서 정의하는 기준을 바탕으로 태깅하고, 존재하지 않을 경우 아래의 원칙을 기준으로 태깅한다.

- ARG0(동작주, 행위자)

(1) 문장에서 사건의 동작주, 행위자에 해당하는 논항을 ARG0으로 분석함

예) **쑨원은(ARG0)** 삼민주의를 내세웠다.
진흥왕은(ARG0) 화랑도를 개편했다.

- ARG1(피동작주, 대상)

(1) 문장에서 사건의 피동작주, 대상에 해당하는 논항을 ARG1으로 분석함.
(2) 구문 태그는 주로 목적격에 해당하나 절대적이지는 않음
(3) 이동 사건에 의해 처소 변화를 겪거나 산출 사건이나 소멸 사건의 결과로 생기는 논항 혹은 소멸되는 논항 역시 ARG1으로 분석함

예) 범인은(ARG1) 사거리에서 발견되었다.
밤거리에는 **인적이(ARG1)** 드물다.

- ARG2(시작점, 수혜자)

(1) 행위의 시발점을 가리키는 기점 논항을 ARG2으로 분석함.
(2) 행위에 의해 수혜를 받는 대상을 ARG2로 분석함
(3) Frame Set에서 장소를 필수격인 ARG2로 정의한 경우, 명확한 장소라고 하더라도 필수격인 ARG2로 분석함

예) 비행기가 인천공항에서(ARG2) 출발했다.
영희가 **철수에게서(ARG2)** 그 선물을 받았다.

- ARG3(착점)

(1) 문장에서 행위의 도착점을 가리키는 착점 논항을 ARG3로 분석함

예) 근이가 학교에(ARG3) 갔다.
연이가 **동창회에(ARG3)** 참석했다.

3.2 부가격: ARG-M

부가격은 서술어에 따라 달라지는 필수격 숫자와는 다르게

절대적으로 정의할 수는 있지만, 방법과 도구처럼 명확하게 기준을 정의하지 않으면 작업자에 따라 태깅 결과가 달라지는 의미역이 있기 때문에 그 기준으로 아래와 같이 수립하였다.

- ARGM-LOC(장소)

  (1) 사건이 발생하는 상황적 공간을 가리키는 처소 논항.
  (2) 동사의 의미에 이동성이 없고, '~에서/~에' 조사와 함께 쓰이는 경우 분석함.
  (3) Frame Set에서 ARG-N으로 정의되어 있지 않고, 명확한 지명이나 장소를 뜻하는 경우

  예) 친구들이 **서울에(ARGM-LOC)** 많이 산다.

- ARGM-DIR(방향)

  (1) 동사의 의미가 이동성을 가질 때, 방향격 조사 '~로, ~으로'와 함께 나타나는 논항.
  (2) '오른쪽', '왼쪽', '위쪽', '아래쪽','앞으로','뒤로','동서남북'에 해당되는 논항.

  예) 달이 **서쪽으로(ARGM-DIR)** 기울었다.

- ARGM-CND(조건)

  (1) 인물이나 사물의 자격이나 서술어 발생 조건을 가리키는 논항.
  (2) ~중에, ~가운데에(범위), ~보다(비교조건), ~에 대해서는(명확한 수치가 나타나지 않는 범위로 한정될 경우)

  예) 과세 대상 금액이 많을수록(**ARGM-CND)** 높은 세율을 적용한다.

- ARGM-MNR(방법)

  (1) 서술어를 수행하는 방법에 대한 논항.
  (2) 서술어가 '언어에 의해' (한자로, 영어로, 티베르어로 등)인 경우

  예) 그는 큰 **소리로(ARGM-MNR)** 떠들었다.

- ARGM-TMP(시간)

  (1) 서술어의 발생 시간과 같이 서술어와 관계 된 시간에 대한 논항.
  (2) 명확한 날짜, 시기, 시대를 나타내는 경우.
  (3) 단, "~부터 ...까지"와 같이 기간을 나타내는 경우, Frame set과 상관없이, 시점과 착점으로 구분, 각각 ARG2(시점), ARG3(착점)으로 분석함.

  예) 진달래는 이른 **봄에(ARGM-TMP)** 핀다.

- ARGM-EXT(범위)

  (1) 크기 또는 높이 등의 수치와 정도를 의미하는 논항.
  (2) '가장', '최고', '더욱', '매우' 등의 정도를 나타내는 논항

  예) 그 악기는 **4개의(ARGM-EXT)** 현을 가진다.

- ARGM-PRD(보조서술)

  (1) 대상과 같은 의미이거나 대상의 상태를 나타내면서 서술어를 수식하는 논항
  (2) 주로 '~로서'의 조사를 가지는 논항
  (3) '말자로','최초로'등 대상이 서술어에 대해 행해진 순서를 나타내는 논항

  예) 석회암 지대에서 깔때기 **모양으로(ARGM-PRD)** 파인 웅덩이가 생겼다.

- ARGM-PRP(목적)

  (1) 서술어의 주체가 목표를 가리키는 논항.
  (2) 행위의 의도가 분명히 드러나는 논항.
  (3) '~를 위해'의 논항.

  예) 주나라의 '백이'와 '숙제'는 절개를 **지키고자(ARGM-PRP)** 수양산에 거처했다.

- ARGM-CAU(발생 이유)

  (1) 서술어가 발생한 이유로 원인 논항이 방향격 표지와 함께 나타남.
  (2) "~때문에" 넣었을 때, 문장의 의미가 통하는 경우.
  (3) '~하여' 술어를 다른 서술어에 연결하여 태깅할 때 PRP(목적)와 불분명하다면, 무조건 CAU로 태깅함.

  예) 지난 밤 **강풍으로(ARGM-CAU)** 가로수가 넘어졌다.

- ARGM-DIS(담화 연결)

  (1) '그러나', '그리고', '즉' 등의 문장 접속 부사.
  (2) PropBank 지침에 따르면, '담화 연결'은 앞의 문맥과 뒤의 문맥을 연결할 경우에 해당하지만, 여기서는 명확한 문장 접속 부사만을 대상으로 함

  예) **하지만(ARGM-DIS)** 여기서 동, 서는 중국과 유럽을 뜻한다.

- ARGM-ADV(부사적 어구)

  (1) '마치', '물론', '역시', 와 같이 부사적 어구에 해당하는 어휘를 선정하여, M-ADV로 분석

  예) 산의 능선이 **마치(ARGM-ADV)** 닭벼슬을 쓴 용의 형상을 닮았다.

- ARGM-NEG(부정)

  (1) 서술어에 대해 부정의 의미를 가지는 논항.

  예) 산은 불에 타지 **않았다(ARGM-NEG).**

- ARGM-INS(도구)

  (1) 서술어를 행할 때 사용하는 도구에 대한 논항
  (2) 서술어를 수행하는 방법인 ARGM-MNR보다 구체적인 '사물'이 있는 논항, '물리적 도구'를 나타내는

논항.

(3) '물리적 도구'가 아니더라도 '이용하다'를 대입하여, 문장이 어색하지 않을 경우.

예) 하얀 **천으로(ARGM-CAU)** 상자를 덮었다.

3.3 한국어 특화 기준

한국어에서 서술적인 성격을 갖거나 공동격 등 기준이 영어와 다른 경우에 대해서 다음과 같이 기준을 정하였다.

- 서술격 조사 '-이다'
  : 일괄적으로 서술어로 인정하지 않을뿐더러, 논항으로도 인정하지 않음
- 공동격 조사 '와/과', '~나', '또는', '이나' 등으로 연결된 어절
  : 동반격에 해당하는 조사로 연결된 어절은 실제 구문 분석 결과가 연결(NP_CNJ 등)된 것만 태깅함
  : ',' 등으로 연결된 경우는 의미역을 설정하지 않음
- 따옴표나 괄호 안에서 문장을 이룰 때
  : 따옴표 안의 술어에 대해 따옴표를 벗어나지 않는 범위에서 태깅한다
  : 따옴표 안의 문장이 바깥문장에 의미역으로 잡힐 때 따옴표 안 문장의 마지막 어절에 의미역을 태깅
- '하다' 동사의 구분
  : '이름 지어 부르다' 라는 의미의 '하다'는 하.02 (say)로 태깅
  예) 저 **꽃은(ARG1) 금강초롱이라고(ARG2)** 한다.
  : '특정한 대상을 어떤 특성이나 자격을 가지는 것으로 만들거나 삼다'는 하.3(regard)로 태깅
  예) 먼 친척 **아이를(ARG1) 양자로(ARG2)** 하다.

4. 의미역 말뭉치 구축

의미역을 태깅한 말뭉치는 질의응답(Question Answering) 시스템 개발을 위하여 구축된 말뭉치로, 퀴즈 질문 문서 및 그 정답 단락 문서의 쌍으로 구성되어 있다. (각 질문 및 정답 단락은 복수의 문장으로 구성될 수 있으며, 각 질문 및 정답 단락 별로 별도의 문서로 구성된다.) ExoBrain 언어분석 말뭉치의 세부적인 통계정보는 표3과 같다.

표 3. ExoBrain 언어분석 말뭉치 통계

	질문	정답단락	총계
문서 수	117	322	439
문장 수	182	542	724
어절 수	2,004	6,527	8,531

ExoBrain 언어분석 말뭉치는 형태소분석, 개체명인식, 구문분석, 의미역인식에 대한 언어분석 정답을 제공한다[8]. 포맷은 JSON 포맷으로 제공되고, 이 중에서 의미역에 해당하는 부분은 표4와 같다.

표 4. ExoBrain 언어분석 결과 중 의미역 예

```
...
"SRL" : [{
  "verb": "태어나", "sense": 1, "word_id": 3, ...
  "argument": [
    {"type": "ARG2","word_id": 2,"text": "사이에서",...},
    {"type": "ARG1","word_id": 4, "text": "새끼", ...}
  ]
}],
...
```

ExoBrain 언어분석 말뭉치는 언어분석 기술 개발을 위한 학습용으로는 그 양이 많지는 않으나, 동일 문장에 대해서 형태소분석부터 개체명인식, 구문분석, 의미역인식까지의 언어분석 정답을 포함하고 있기 때문에, 세부 언어분석 기술뿐 아니라 전체 언어분석 파이프라인을 평가하기 위한 용도로 활용이 가능할 것이다.

이러한 말뭉치를 사용하여 의미역을 태깅하기 위하여 반자동 태깅 툴을 개발하였다. ETRI에서 개발한 의미역 인식기[9]를 이용하여 태깅 대상 문장(문서)에 자동으로 의미역을 인식한 결과를 그림 1과 같은 편집 화면에서 수작업으로 편집을 하며, 각각의 기능은 아래와 같다.

① FrameSet 조회/선택/추가
②, ③ Link를 추가/삭제 기능
④ 구문 분석 편집 기능
⑤ 서술어의 편집 기능.FrameSet 과 연동
⑥ 술어와의 Role을 추가/편집 기능
⑦, ⑧ 기존/신규 저장 기능

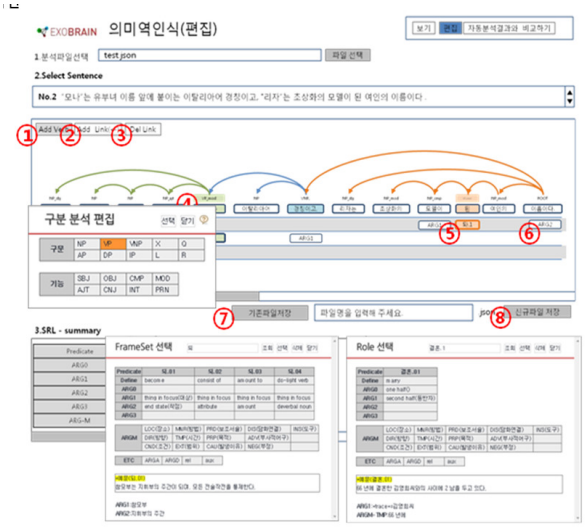

그림 1 의미역 반자동 툴:편집 화면

이 외에도 자동으로 분석한 결과와 수작업 태깅된 결과를 그림2와 같이 한 화면에 배치하여, 오류 분석 및 태깅 말뭉치에 대한 검증을 보다 효율적으로 할 수 있는 기능을 갖췄다.

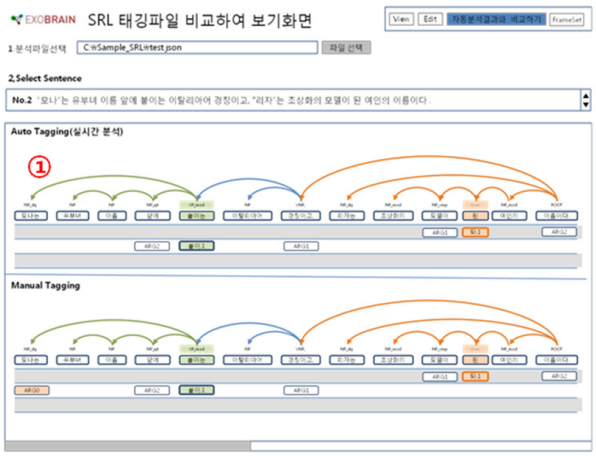

그림 2 의미역 반자동 툴:자동/수동 결과 비교 화면

5. 결론

본 논문에서는 한국어 의미역에 대한 서로 다른 정의가 존재하였으나, 세계적으로 널리 쓰이고 있는 PropBank 의미역을 기반으로 하여 한국어에 특성을 고려하여 의미역을 수정하고, 태깅 가이드라인을 제시하였다. 이를 준수하여 ExoBrain 과제에서 질의응답 시스템인 WiseQA를 개발 및 평가를 하기 위해 구축한 질문 및 정답 후보 문장으로 구성된 평가셋에 반자동 태깅 툴을 이용하여 한국어 의미역이 태깅된 학습 데이터를 구축하였다. 구축하는 과정에서 태깅 가이드라인 불분명하거나 추가적으로 필요한 의미역에 대해서는 반영이 필요하다.

감사의 글

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.R0101-15-0062, 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발)

참고문헌

[1] 이성범, "언어와 의미", 태학사, 1999년
[2] Myung-Chul Shin, "Integration of Case-Frame Dictionary into Machine Learning Techniques for Semantic Role Assignment of Korean Adverbial Cases,", MS Thesis, Pohang University of Science and Technology, 2006
[3] S.B. Park, "Decision Tree Based Disambiguiation of Semantic Roles for Korean Adverbial Postposition," IEICE Transaction Information and System, Vol.E86-D,No.8, 2003
[4] 김윤정, 옥철영, "한국어 서술어와 논항들 사이의 의미역", 제26회 한글 및 한국어 정보처리 학술대회 논문집, pp.143-148, 2014.
[5] M. Palmer, D.Gildea and P. Kingsbury, "The proposition bank: An annotated corpus of semantic roles," Computational Linguistics 31(1), 2005.
[6] M.Palmer et al., "Korean Propbank," Linguistic Data Consortium, Philadelphia, 2006.
[7] http://exobrain.re.kr/onedintro
[8] 최미란, "형태소 태깅 말뭉치 작성용 품사 태그 세트," TTAK.KO-11.0010/R, 2015
[9] 임수종, 김현기, "의미 정보를 이용한 한국어 의미역 인식 연구,", 제27회 한글 및 한국어 정보처리 학술대회 논문집, 2015.