의존관계 패턴에 기반한 한국어 명사구의 추출

강승식○, 이상모, 이민행
국민대학교, 한국법제연구원, 연세대학교
sskang@kookmin.ac.kr, peacekeeper@klri.re.kr, leemh@yonsei.ac.kr

The Extraction of Korean Noun Phrases based on Dependency Patterns

Seungshik Kang○, Sangmo Lee, Minhaeng Lee
Kookmin University, Korea Legislation Research Institute, Yonsei University

요 약

이 연구는 한국어 명사구를 말뭉치로부터 추출하는 방법의 하나로 의존관계 패턴에 기반한 접근방법을 제안하는 것을 목적으로 한다. 이 방법론을 활용할 경우에 명사구 추출의 정확성을 높일 수 있다. 이 논문에서는 한국어 법령 의존 말뭉치를 구축하는 단계부터 상위 명사구 목록을 생성하기 까지 거치는 5단계에 대해 상세하게 논의하는 한편, 의존구조 검색시스템을 통해 의존관계 패턴을 추출하는 절차에 대해 기술하고 이 작업을 수행하기 위한 검색식들의 특성들에 대해 검토한다.

주제어: 한국어 법조문, 기본 명사구, 명사구 패턴, 의존구조 검색, 의존관계 패턴

1. 서론

이 연구는 한국어 명사구를 추출하는 방법의 하나로 의존관계 패턴에 기반한 접근방법을 제안하는 것을 목적으로 한다. 강승식(2022)[1]에서 논의한 바와 같이 한국어 법조문에서 특정 명사구들이 반복적으로 사용되기 때문에 한국어 법령을 영어나 다른 언어로 번역하는 작업을 수행할 경우 빈번히 출현하는 명사구 목록을 확보하는 과제가 매우 중요하다. 이 문제를 해결하는 방법으로 n-gram을 이용하는 방법, 구구조 트리뱅크를 활용하는 방법 등이 있을 수 있다. 그러나 전자는 명사구 뿐만 동사구나 부사절 등 다양한 통사범주에 속하는 복합표현들이 함께 추출되는 한계를 지닌다. 한편, 한국어 구구조 트리뱅크의 경우 법령 텍스트를 대상으로 구구조 파싱을 한 결과를 검토하면 분석결과의 정확률이 높지 않기 때문에 파싱결과로부터 명사구를 추출하는 것이 효율적이지 않다. 때문에 우리는 대안으로 의존구조 파싱 결과를 활용하는 방안을 제안하고자 한다[2,3]. 한국어 법령 텍스트를 대상으로 하여 의존구조 파싱을 수행한 결과는 구구조 파싱과 비교할 때 매우 오류율이 낮다.

2. 관련 연구

한국어 법령에 나타나는 단일어 목록의 생성과 구문패턴 정보를 추출하는 연구를 수행한 구명철(2020)[4]과 달리 강승식(2022)은 ngram 추출기법을 활용하여 법령 말뭉치로부터 복합용어와 기본 명사구의 빈도 분석을 수행하였다. 그런데 ngram을 이용하여 명사구를 추출할 경우 정확성이 높은 않은 한계가 있다.

본 연구에서는 강승식(2022)과 Kang(2022)[5]의 연구결과를 보완하는 방안으로 의존 말뭉치로부터 명사구의 의존관계 패턴을 추출한 다음에 패턴들을 검색식으로 변환하여 명사구 추출에 사용하는 방법을 모색하고자 하는 것이다. 이를 통해 국문법령 번역 효율성의 증대를 기대할 수 있다. 이상모(2021)[6]에 따르면, 한국법제연구원[1]에서 법률용어를 정비하고 한글 법령의 영문번역 서비스를 제공한다.

3. 연구 내용

한국어 법령 의존 말뭉치 (Dependency Corpus)로부터 명사구 추출 과정은 다음과 같이 5단계로 구분된다.

제1단계: 한국어 법령에 포함된 63,640 문장[2]을 대상으로 미국 Stanford대에서 개발한 stanza 파이썬 코드[3]를 이용하여 의존구조 파싱을 수행한다.

제2단계: CoNLLU 포맷의 의존 말뭉치를 대상으로 4-gram(quardrigram)과 5-gram(pentigram) 어절을 추출한 다음에 그 가운데 상위빈도를 차지하는 명사구 20개를 선별한다.

제3단계: 선별된 20개 명사구들의 어절 연속체 패턴을 검토하여 출현빈도가 높은 10개 패턴을 다시 선별한다.

제4단계: 어절 연속체 패턴 10개의 의존관계 패턴을 분석한 다음 가장 출현빈도가 높은 의존관계 패턴을 분석한다.

제5단계: 출현빈도 상위의 의존관계 패턴들을 검색식으로 변환하여 Tundra[4]에서 검색하는 방법으로 명사구들을 추출한다.

---

1) https://elaw.klri.re.kr/
2) 한국어 법조문 전체는 대략 70만 조항에 달하는데, 의존구조 검색시스템 Tundra가 수용하는 파일 크기가 50기가 바이트를 넘을 수 없기 때문에 그 규모를 최대한으로 고려하여 63,640 문장을 대상으로 하여 한국어 법령 의존 말뭉치를 구축한다.
3) stanza를 이용한 의존구조 파싱에 대해서는 Qi(2020)[7] 참조.
4) 의존구조 검색시스템 Tundra를 실행할 수 있는 사이트: https://weblicht.sfs.uni-tuebingen.de/Tundra/

위의 제3단계에서 선별된 어절 연속체 패턴 10개는 다음과 같다.

다음 각 호의 __을/를
__ 및 __에 관한 법률
호의 어느 하나에 해당하는 __에는
다음 각 __의 어느 하나에
다음 각 호의 __에
어느 하나에 해당하는 __을/를
호의 어느 하나에 해당하는 __을/를
__항 각 호의 __을/를
__항에 따른 __ __
__항 각 호의 어느 하나에

이 명사구 패턴들이 각각 어떤 의존관계들로 나타나는지를 확인하기 위해 검색시스템 Tundra를 통해 검색할 수 있다. 예를 들어 여섯째 패턴 '어느 하나에 해당하는 __을/를' 에 관여하는 의존관계들을 파악하기 위해서 검색시스템에 먼저 의존 말뭉치를 탑재한 다음 아래에 제시된 복합 검색식을 실행한다.

```
#1:[token="어느"] & #2:[token="하나에"] &
#3:[token="해당하는"] & #4:[token=/.*(을|를)/] & #1 .
#2 & #2 . #3 & #3 . #4
```

이 검색식을 실행하면 932개의 명사구가 추출된다. 이 명사구들의 의존수형도를 통해 어떠한 의존관계들이 관여되어 있는 지를 확인할 수 있다. 아래에 수형도 하나가 제시되어 있다.[5]

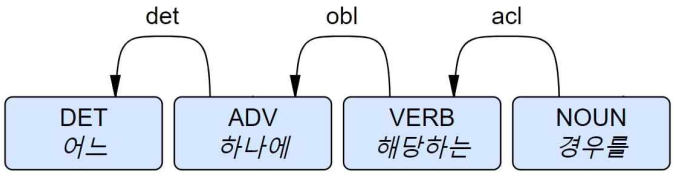

그림 1. 의존수형도-1

이 수형도를 살펴보면 첫 어절 '어느' 가 핵심어인 둘째 어절 '하나에' 에 대해 의존기능 det를 수행하고, 다시 둘째 어절 '하나에' 는 핵어인 셋째 어절 '해당하는' 에 대해 의존기능 obl을 수행하며 또 다시 셋째 어절 '해당하는' 은 핵어인 넷째 어절 '경우를' 에 대해 의존기능 acl을 수행하는 것을 확인할 수 있다.

이러한 의존관계 패턴을 보이는 명사구들을 의존 말뭉치로부터 추출하고자 할 경우 다음 검색식을 실행하면 된다.

```
#1 & #2 & #3:[pos="VERB"] & #4:[pos="NOUN"] & #2
>det #1 & #3 >obl #2 & #4 >acl #3 &  #1.#2 & #2.#3 &
#3.#4
```

이 검색식의 실행결과 2,611개의 명사구가 추출되며, 그 가운데 출현빈도가 상위에 속하는 명사구 5개를 제시하면 다음과 같다.[6]

어느	하나에 해당하는 경우에는	(555)
어느	하나에 해당하는 경우를	(221)
어느	하나에 해당하는 자는	(171)
어느	하나에 해당하는 사람은	(85)
어느	하나에 해당하는 행위를	(85)

이와 동일한 절차를 제4단계에서 확정된 나머지 6개 패턴에 적용하면 각 패턴별로 명사구들을 추출할 수 있다. 이 과정에서 필요한 작업은 각 의존관계 패턴을 Tundra에서 사용가능한 검색식으로 작성하는 것인데, 아래에 열거된 검색식들이 6가지 패턴에 대응되는 검색식들이다.

<1>	#1 & #2 & #3 & #4:[pos="NOUN"] & #4 >nmod #3 & #3 >amod #2 & #3 >compound #1 & #1.#2 & #2.#3 & #3.#4
<2>	#1 & #2 & #3 & #4:[pos="VERB"] & #5:[pos="NOUN"] & #5 >acl #4 & #4 >obl #3 & #4 >obl #1 & #3 >cc #2 & #1.#2 & #2.#3 & #3. #4 & #4.#5
<3>	#1 & #2 & #3 & #4:[pos="VERB"] & #5:[pos="NOUN"] & #5 >acl #4 & #4 >obl #3 & #3 >det #2 & #3 >nmod #1 & #1.#2 & #2.#3 & #3.#4 & #4.#5
<4>	#1 & #2 & #3 & #4 & #5:[pos="ADV"] & #5 >det #4 & #5 >nmod #3 & #3 >amod #2 & #3 >compound #1 & #1.#2 & #2.#3 & #3.#4 & #4.#5
<5>	#1 & #2 & #3 & #4:[pos="ADV"] & #4 >nmod #3 & #3 >amod #2 & #3 >compound #1 & #1.#2 & #2.#3 & #3.#4
<6>	#1 & #2 & #3:[pos="NOUN"] & #4:[pos="NOUN"] & #4 > #3 & #3 >acl #2 & #2 >obl #1 & #1.#2 & #2.#3 & #3.#4

위의 첫 검색식에 의해 추출된 명사구를 의존수형도를 통해 보이면 그림 2와 같다.

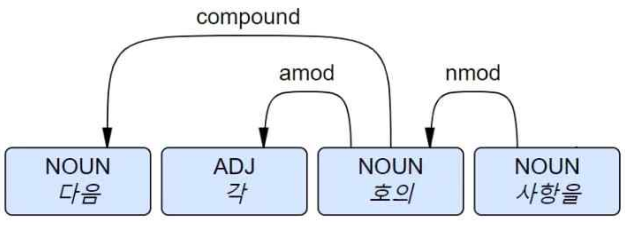

그림 2. 의존수형도-2

이 수형도에서 명사구 "각 호의 사항을" 이 담고 있는 의존관계 패턴이 위 검색식 <1>에 표현되어 있으며, 이 검색식을 이용해 추출한 명사구들 가운데 출현빈도가

5) Tundra사이트에서 제공하는 수형도는 가독성이 현저히 떨어져 다음 사이트를 이용하여 의존수형도를 생성했다: https://urd2.let.rug.nl/~kleiweg/conllu/

6) 괄호안의 숫자는 출현빈도이다.

높은 예들이 아래에 제시된다.

다음 각 호의 사항을 (829)
다음 각 호의 사항이 (547)
다음 각 호의 업무를 (184)
다음 각 호의 서류를 (162)
다음 각 호의 권한을 (112)

동일한 절차에 따라 위 검색식 <2>를 실행하여 추출한 상위 빈도의 명사구들은 다음과 같다.

계획 및 이용에 관한 법률 (21)
취득 및 보상에 관한 법률 (10)
독점규제 및 공정거래에 관한 법률 (6)
금융실명거래 및 비밀보장에 관한 법률 (4)
구매촉진 및 판로지원에 관한 법률 (4)

마찬가지 방법으로 검색식 <3>를 실행하여 추출한 상위 빈도의 명사구들은 다음에 제시된다.

호의 어느 하나에 해당하는 경우에는 (538)
호의 어느 하나에 해당하는 경우를 (213)
호의 어느 하나에 해당하는 자는 (162)
호의 어느 하나에 해당하는 사람은 (83)
호의 어느 하나에 해당하는 행위를 (74)

나머지 검색식 셋도 동일한 과정으로 Tundra에서 실행하여 해당 의존관계 패턴을 보이는 명사구들을 추출할 수 있다.

7개 검색식을 모두 실행하여 추출한 명사구 목록을 파일 하나에 통합하여 출현빈도를 기준으로 정렬한 결과 상위 순위 20위에 속하는 명사구들은 다음과 같다.

다음 각 호의 어느 하나에 (2005)
다음 각 호의 사항을 (829)
어느 하나에 해당하는 경우에는 (555)
다음 각 호의 사항이 (547)
호의 어느 하나에 해당하는 경우에는 (538)
다음 각 호의 구분에 (354)
어느 하나에 해당하는 경우를 (221)
호의 어느 하나에 해당하는 경우를 (213)
다음 각 호의 업무를 (184)
어느 하나에 해당하는 자는 (171)
호의 어느 하나에 해당하는 자는 (162)
제1항 각 호의 어느 하나에 (127)
다음 각 호의 사항에 (109)
다음 각 목의 어느 하나에 (61)
다음 각 호의 기준에 (51)
계획 및 이용에 관한 법률 (21)
취득 및 보상에 관한 법률 (10)
독점규제 및 공정거래에 관한 법률 (6)
제1항에 따른 검사 결과 (6)
하나에 해당하는 사람 중 (6)

앞서 언급한 바, 명사구의 추출에 사용한 CoNLLU 포맷의 의존 말뭉치는 데이터 구조만 변형하여 CWB 기반의 코퍼스로 구축할 수 있다. 이러한 변환 방식을 취할 경우에 매우 효율적인 검색엔진 CQP를 이용할 수 있어 유용성이 배가한다. CWB 기반 말뭉치의 구축방법과 CQP에 대해서는 이민행(2015)[8]에 상세하게 기술되어 있다.

4. 결론

본 연구에서는 한국어 법령 의존 말뭉치로부터 의존관계 패턴을 기반으로 명사구들을 추출하는 방법에 대해 논의했다. 이처럼 의존관계 패턴을 기반으로 추출한 명사구들을 모두 모아서 출현빈도를 기준으로 명사구 목록을 작성할 경우에 한국어 법령을 영어, 중국어 등 외국어로 번역하는 작업을 수행할 때 이 명사구 목록을 매우 유용하게 활용할 수 있다. 더욱 바람직한 것은 출현빈도가 높은 한국어 명사구의 대응 번역쌍을 언어별로 미리 만들어 두는 작업일 것이다.

감사의 글

이 연구는 2023년도 한국법제연구원 법령번역센터의 용역과제 지원을 받아 수행되었음.

참고문헌

[1] 구명철, 국문법령의 법률용어와 상용어구에 대한 코퍼스 분석, 한국법제연구원 연구보고서, 2020.
[2] P. Qi, T. Dozat, Y. Zhang, and C. Manning, “Universal dependency parsing from scratch”, In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, 2018.
[3] D. Zeman, J. Hajic, M. Popel, M. Potthast, M. Straka, F. Ginter, J. Nivre, and S. Petrov, “CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies”, In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, 2018.
[4] 강승식, 이민행, 국문법령의 복합용어와 기본 명사구에 대한 코퍼스 분석, 한국법제연구원 연구보고서, 2022.
[5] S. Kang, S. Lee, and M. Lee, "Word phrase analysis of high-frequency Ngrams in the large-scale Korean law corpus", In Proceedings of the 2022 IEEE International Conference on Big Data, pp.6684-6686, 2022.
[6] 이상모, "기계번역을 활용한 법령번역의 실제와 과제", T&I REVIEW, 제11권 제1호, 2021.
[7] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning, “Stanza: A Python natural language processing toolkit for many human languages”, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp.101-108, 2020.
[8] 이민행, 『빅데이터 시대의 언어연구 - 내 손 안의 검색엔진』 21세기북스, 2015.