AI 어시스턴트 플랫폼의 한국어와 중국어 음악청취 요청문 패턴구축 비교 연구

윤소은○, 이가빈 & 남지순
한국외국어대학교, DICORA 연구센터/언어인지과학과
yuns-e@hufs.ac.kr, lijiabin951103@gmail.com & jeesun.nam@gmail.com

A Comparative Study on Building Korean & Chinese Music Request Sentence Patterns for AI Assistant Platforms

Soeun Yun○, Jiabin Li & Jeesun Nam
DICORA/Dept. of LCS, Hankuk University of Foreign Studies

요 약

본 연구에서는 AI 어시스턴트의 음악청취 도메인 내 요청문을 인식 및 처리하기 위해 한국어와 중국어를 중심으로 도메인 사전 및 패턴문법 언어자원을 구축하고 그 결과를 비교분석 하였다. 이를 통해 향후 다국어 언어자원 구축의 접근 방법을 모색할 수 있으며, 궁극적으로 패턴 기반 문법으로 기술한 언어자원을 요청문 인식에 직접 활용하고 또한 주석코퍼스 생성을 통해 기계학습 성능 향상에 도움을 줄 수 있을 것으로 기대된다. 본 연구에서는 우선 패턴문법의 구체적인 양상을 살펴보기에 앞서, 해당 도메인의 요청문 유형의 카테고리를 결정하는 과정을 거쳤다. 이를 기반으로 한국어와 중국어 요청문의 실현 양상과 패턴 유형을 LGG 프레임으로 구조화한 후, 한국어와 중국어 패턴문법 간의 통사적, 형태적, 어휘적 차이점을 비교분석 하여 음악청취 도메인 요청문의 언어별 생성 구조 차이점을 관찰할 수 있었다. 구축한 패턴문법은 개체명을 변수(X)로 설정하는 경우, 한국어에서는 약 2,600,600개, 중국어에서는 약 11,195,600개의 표현을 인식할 수 있었다. 결과적으로 본 연구에서 제안한 언어자원의 언어별 차이에 대한 통찰을 통해 다국어 차원의 요청문 인식 자원과 기계학습 데이터로서의 효용을 확인하였다.

주제어: AI 어시스턴트, 요청문 유형분류, 한국어 중국어 비교, LGG 패턴문법

1. 서론

본 연구는 AI 어시스턴트(Artificial Intelligence Assistant) 플랫폼의 음악청취 도메인 내 다목적 기능들을 수행하기 위한 요청문(request sentence)을 자동으로 인식하고 분류하기 위하여, 여기 실현되는 언어 패턴유형을 다국어 언어로 구축하는 한국외국어대학교 DICORA연구센터[1] 연구프로젝트의 일환으로 진행되었다. LGG(Local-Grammar Graph)[2] 프레임을 통해 패턴문법을 구성함으로써 언어 간 호환성과 확장성을 확보하고, 이와 더불어 기계학습 데이터로서 활용될 수 있는 주석코퍼스를 생성하는 것을 가능하게 한다.

2000년대 후반부터 스마트폰 기반의 개인용 음성비서에서 챗봇(chatbot)의 기능까지 수행할 수 있는 인공지능 스피커 분야가 주목받기 시작하였다. 이로 인해 국내외 관련 시장에서는 해당 기술의 개발과 동시에 이를 활용할 수 있는 여러 분야를 확장하는 연구가 진행되었다. AI 어시스턴트의 활용 분야로는 날씨 정보 안내, 일정 관리 등과 더불어 음악 재생 관련 서비스가 대표적이라 할 수 있다. AI 어시스턴트는 이처럼 특정 과업 수행에 초점이 맞춰진 만큼, 여러 양상의 사용자 요청문들을 처리하는 것이 기술의 핵심이 된다.

AI 어시스턴트가 요청된 기능을 수행할 수 있도록 요청문의 요구사항을 인식하는 데에 두 가지 방향에서의 접근이 가능하다. 첫 번째는 대용량 코퍼스를 이용한 기계학습 방법 접근법이며, 두 번째는 요청문의 패턴(pattern)을 언어학적 분석을 통해 직접 기술하여 이를 텍스트 분석 및 생성에 적용하는 접근법이다. 본 연구에서 제안하는 패턴문법 연구는 실제로 두 가지 접근법 모두에 사용될 수 있는 정교한 언어자원을 구축하는 것이다.

특정 도메인에 나타나는 요청문 유형은 문장 구조가 일정한 틀에서 크게 벗어나지 않기 때문에 이를 패턴문법으로 구축하는 것이 가능하다. 그림 1에서 보는 바와 같이 본 연구에서는 음악청취 관련 도메인의 요청문에 해당하는 프레임을 설정하고 여기에 맞는 패턴을 기술한다. 다음으로 특정 과업 수행과 관련된 정보 유형을 XML 방식으로 마크업한 LGG 그래프문법을 구축하여 이를 활용하여 실제 요청문을 분석하는 데에 사용한다. 동시에 이를 통해 요청문 유형을 자동 생성할 수 있어, 이를 추후 기계학습을 위한 학습데이터로 활용하게 된다.

2. 관련 연구

AI 음성 어시스턴트와 관련한 연구는 실질적인 상품으로서의 필요성이 대두하기 전부터 진행되어 왔다. AI 음성 어시스턴트의 시스템을 구성하는 요소들에는 여러 가지가 있는데, 'Spoken Language Understanding(SLU)'의 경우 도메인 인식과 의도 인식, 의미 슬롯 채우기의 세 가지 요소가 중요하다. 의도 인식을 위해서는 특정 도메인으로의 한정이 중요하고, 가령 요청문 중에서도 '음악청취' 도메인과 같이 그 범위를 한정하는 것이 필요하다.

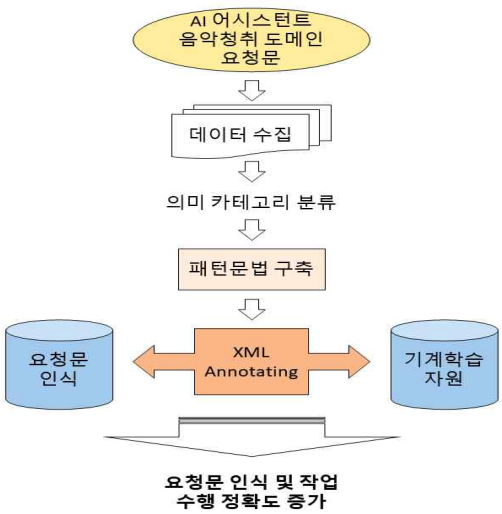

그림 1 패턴문법 구축 및 활용 과정

전통적인 의도 인식의 방법론으로는 사전과 틀에 기반한 규칙기반 의미인식 방식이 있다. 대용량의 코퍼스에서 정보를 추출하여 이들을 분류하고 사전으로 구축하는 것이다. [3]에서는 규칙과 카테고리 정보를 구축하기 위해 규칙 기반 방법론을 사용하였고 이를 이용하여 사용자 요청문을 분류할 수 있도록 하였다. [4]에서도 사용자의 의도 파악과 분류 결과의 성취 정도를 높이기 위해 의도 인식에 규칙 기반 방법을 적용하였다. [5]는 규칙 기반 방식의 정확성을 언급하면서도 무한히 생성되는 표현들을 모두 처리하기는 어렵다는 점을 설명하였다. 이로 인해 규칙을 기반으로 하는 방법은 그 자체로서 정확한 언어자원으로서 기능하면서도 기계학습과 같은 접근법의 보완이 필요하다고 판단된다.

규칙기반 외에 통계적 모델을 사용한 사례들도 제시되었다. 통계적 자질분류 알고리즘은 코퍼스 상에서 핵심 자질들을 추출한 뒤 분류기를 훈련하여 의도 분류가 가능하게 하도록 하는 것이다. [6],[7],[8] 등의 연구에서는 Naive Bayes(NB) 알고리즘과 AdaBoost, Support Vector Machine(SVM) 등의 알고리즘들을 사용하였다. [9]의 연구에서는 SVM과 NB 분류기가 의도 파악을 위해 사용되었는데, 이 방법은 자질의 정확성이 충분히 보장되지 않아 보완이 필요하다는 점이 지적되었다. [10]에서는 PSO를 사용하여 SVM 계수를 최적화하기 위해 AdaBoost 알고리즘을 사용하였다. 이 연구의 분류 수행 능력은 SVM 분류기에 비교하여 상당히 높아졌다고 평가되었으나, 이러한 연구들은 텍스트의 깊은 의미 정보를 파악하기 어려운 통계 방법론의 전형적인 문제점을 안고 있다.

요청문 파악을 위한 딥러닝 기반의 방법론의 중요성도 부상하였다. 많은 연구자가 단어 벡터(word vectors), Convolutional Neural Networks(CNN), Recurrent Neural Network(RNN) 등의 방식을 도입하기 시작했으며, 이들의 도입은 기존 머신러닝 기법에 비하여 자연어처리 성능이 향상하게 하였다. [11]에서는 CNN 방식을 사용하여 만족할 만한 연구성과를 내었고, [12]에서는 RNN과 LSTM(Long Short Term Memory)을 사용하여 의도 분류의 문제점을 해결하고자 하였다. 그 결과, 오류율이 1.48%로 나타나 RNN 기법보다 낮은 오류 비율을 보여주었다. 이외에도 분류에 초점을 맞춘 BTM-BiGRU 기법이 사용되는 등의 시도가 있어 왔다.

전통적인 규칙 기반 의미 인식 방법에서부터 딥러닝에 이르기까지 의미파악에 큰 성취를 이루어 왔지만, 텍스트의 불규칙성, 사용자 의도의 다양성 등의 문제들을 모두 해결하지는 못했다. 이러한 문제점들의 가장 큰 이유 중 하나는 주석된 데이터를 충분히 확보하지 못했다는 데에 있다고 볼 수 있다. 현재로서 주석 데이터를 확보하는 것은 크게 두 가지 측면에서 진행돼야 한다. 하나는 데이터에 직접 주석 작업을 적용하는 것과 또 하나는 주석 데이터가 준지도학습 방식으로 생성되도록 하는 것이다. 본고에서 제안하는 접근법은 패턴문법을 유한상태 트랜스듀서(Finite-State Transducer: FST)로 컴파일하여 이를 코퍼스에 주석할 수 있도록 하는 방법론으로, 위의 연구들과 그 맥을 같이 한다. [13]에서 DECO 한국어 전자사전[14]을 기반으로 LGG 패턴문법을 구축하여 유의미한 성능을 보여준 바 있는데, 이는 본 연구의 중요한 토대가 되었다. 본 연구에서는 한국어와 중국어 요청문 문장에 대한 패턴문법을 구축하는 과정을 비교하여 두 언어 간의 차이점을 분석해 봄으로써, 여기서 제안하는 방법론의 다국어 확장성을 보여주고자 한다.

3. 음악청취 도메인 요청문의 의미카테고리

3.1. 요청문 의미카테고리 분류

본 연구에서 제안하는 요청문 의미 카테고리 분류는 [13]에서 제시하였던 카테고리 분류를 토대로 보완 확장된 형태이다. 표 1에서 보이는 바와 같이 여기서는 '음악관리(Music-Administration)'와 '음악 검색(Music-Search)' 카테고리가 확장되었는데, 국내외 AI 음성 어시스턴트인 '기가지니(GiGA Genie)', '누구(NUGU)', '네이버 클로바(Naver Clova)', '헤이 카카오(Hey kakao)', '구글 어시스턴트(Google Assistant)'의 작동방법 가이드에서 공통으로 설명되거나 핵심적인 기능으로 판단되는 것들과 더불어 [13]에서 실행된 실험에서 인식되지 못한 요청문 부류를 종합분석한 것을 기준으로 분류를 확장한 것이다. 이들은 음악청취 도메인에서 상당 부분 필수로 작동되어야 하는 유형들로 구성된다.

표 1 요청문 의미 카테고리 분류

분류	내용
Music-On/Off	음악 켜기 및 끄기
Volume-Up/Down	소리 키우기 및 줄이기
Music-Control	선택 재생(Select), 반복 재생(Repeat), 랜덤 재생(Random), 일시 정지(Pause)
Music-Administration	플레이리스트에 추가하기(Add), 삭제하기(Delete), 다운로드 받기(Download), 관심 표현하기(Like)
Music-Search	검색 기능(Search), 추천 받기(Recommend)

첫 번째로 'Music-On/Off(음악 켜기/끄기)'는 음악을 켜고 끄는 작업이 수행되어야 하는 카테고리다. 음악을 켜는 작업에서는 단순히 무작위 음악 재생을 요구할 수 있고, 또는 구체적인 음악명을 대며 재생 요청을 할 수도 있다. 특정 뮤지션의 곡을 요청할 수도 있으며 음악명과 가수명을 동시에 언급하는 것도 가능하다. 특정 장르 또는 분위기를 조건으로 하는 음악을 요청하는 경우도 고려해야 할 대상이다. 이외에 개인 플레이리스트 전체 재생 기능도 가능하다.

두 번째로 'Volume-Up/Down(소리 키우기/줄이기)' 카테고리는 재생되고 있는 음악의 음량을 키우거나 줄이는 작업을 담당하는 범주다. 여기서 나타나는 개체명은 '볼륨' 부류일 수도 있고, '음악' 부류일 수도 있다. 음량의 크기를 조절하는 것이기 때문에 정도 부사와 '키우다', '줄이다'류 서술어가 결합할 수 있고, AI 어시스턴트 종류에 따라 특정 숫자를 언급함으로써 음량 조절의 정도를 정할 수도 있다.

세 번째인 'Music-Control(음악 제어)'은 하위 범주로 다시 세분류된다. 우선 '선택 재생(Select)'은 현재 음악 또는 이전, 다음 음악을 재생하게 하는 기능이다. '반복 재생(Repeat)'은 특정 곡을 반복적으로 재생하게 하는 기능 범주로 개체명은 '음악 켜기'와 '선택 재생' 범주에 나타나는 개체명과 거의 동일한 표현들이 나오게 된다. '랜덤 재생(Random)'은 '음악 켜기' 카테고리에도 포함되는 기능이지만 이 경우는 '음악을 틀어줘'의 의미를 지니는 표현들이 포함되는 반면, '랜덤 재생'에서는 '랜덤'의 의미를 내포하는 표현들이 포함된다. 다음으로 '일시 정지(Pause)'는 재생 중인 음악을 잠깐 정지시키는 카테고리로, 목적어로서 개체명이 수의적으로 나타난다.

네 번째는 'Music-Administration(음악 관리)' 범주이다. 이 카테고리도 네 가지 하위 범주로 나뉘는데 여기에는 '플레이리스트에 추가하기(Add)', '삭제하기(Delete)', '다운로드 받기(Download)', '관심 표현하기(Like)' 등이 속한다. '플레이리스트에 추가하기'는 현재 재생되고 있는 곡 또는 특정 곡을 사용자의 플레이리스트에 추가하는 기능이다. '삭제하기'도 '플레이리스트에 추가하기'와 마찬가지의 개체명이 나타난다. '다운로드 받기'도 앞선 두 범주처럼 특정 곡을 지목하여 기능을 수행해야 하기 때문에 특정 곡을 가리킬 수 있는 개체명 표현과 함께 '다운로드하다'의 의미가 들어간 표현이 서술부에 나타난다. 마지막으로 '관심 표현하기'는 어떤 곡에 '좋아요' 등의 관심을 표현하는 기능 범주다. 이는 '좋아요'를 받은 곡들만 따로 모아 플레이리스트를 구성하는 등 그 활용 범위가 더 확장될 수도 있다.

다섯 번째인 'Music-Search(음악 검색)'도 두 가지 하위 범주로 세분화된다. '검색 기능(Search)'은 특정 곡을 찾기 위해 특정 곡명과 '검색하다'류의 의미를 표현하는 서술어가 결합하여 요청문을 구성하는 것이다. '추천 받기' 범주의 요청문에는 개체명으로서 특정 곡이 나타나지 않는다. 이에 따라 단순히 노래를 추천해 달라고 요청할 수도 있고 특정 분위기, 장르 조건을 가진 노래 내에서 추천해 달라는 표현으로 실현될 수 있다.

3.2. 요청문 실현 양상

3.2.1. 한국어에서의 실현 양상

앞서 소개한 의미 카테고리는 실제 어휘적 구문으로 실현될 때에는 단순한 요청문에서부터 복잡한 문장 구조의 요청문까지 다양한 형태로 실현된다. Deco T-Crawler[16]로 AI 스피커의 호출명을 트윗 검색 쿼리로 설정하여 수집한 코퍼스에서 인용문(약 300여 개) 형태의 요청문을 일괄 추출함으로써 [13]에서 요청문의 실현 양상을 확인하고 패턴문법의 틀을 체계화하였다. 위에서 논의한 각 카테고리별 실현 문장 양상의 예를 보이면 표 2와 같다.

표 2 요청문 의미 카테고리별 한국어 문장 예시

의미 카테고리	실제 문장 예문
Music-On/Off	잔잔한 곡을 플레이해 줘. 노래 이제 꺼.
Volume-Up/Down	볼륨을 좀 올려라. 노래 크기 최대한 낮춰 줘.
Music-Control	비틀즈의 렛잇비 반복해서 재생해. 잠깐 멈춰.
Music-Administration	비틀즈의 렛잇비 플레이리스트에 넣어 봐. 지금 음악 다운 받아 줘 봐.
Music-Search	비 오는 날 들으면 좋은 노래 추천해 줘. 비틀즈 히트곡 찾아 줄래?

3.2.2. 중국어에서의 실현 양상

현재 본 연구에서 구축된 중국어 요청문 패턴문법도 앞서 서술한 5가지 의미 카테고리 분류에 기반하고 있다. 이에 대한 예를 보이면 표 3과 같다.

표 3 요청문 의미 카테고리별 중국어 문장 예시

의미 카테고리	실제 문장 예문
Music-On/Off	播放歌曲。[노래를 틀어.] 关闭一下音乐吧。[음악을 꺼.]
Volume-Up/Down	稍微调高音量。[볼륨을 약간 높여.] 帮我静音。[나를 위해 음소거 해 줘.]
Music-Control	播放之前的音乐。[이전 음악을 틀어 봐.] 随机播放歌曲。[노래를 랜덤으로 켜 줘.]
Music-Administration	下载当前正在播放的音乐。[지금 재생되고 있는 음악을 다운 받아 줘 봐.] 将甲壳虫乐队的《Let it be》放到我的播放列表中。[비틀즈의 "렛잇비"를 내 플레이리스트에 넣어.]
Music-Search	帮我播放适合在下雨天听的歌。 [비오는 날 들으면 좋은 노래 추천해 줘.] 可以搜索披头士乐队发行的歌曲吗？[비틀즈가 발표한 곡들을 검색해 줄 수 있어?]

4. LGG 패턴문법 기반 요청문 카테고리 기술

4.1. 한국어 LGG 패턴문법

AI 어시스턴트의 음악청취 도메인에서의 요청문들을 패턴문법으로 기술하기 위한 LGG 구축은 Unitex 플랫폼[15]에서 제공하는 FST-editor를 사용하였다. 이 부분문법 그래프는 방향성 비순환 그래프(DAG) 형식으로 구현되어 코퍼스 분석 시 FST 형식으로 컴파일되어 적용된다.

한국어 LGG는 앞서 설명한 카테고리 분류 범주를 따라 각각 구성되며 그 하위에 다양한 유형의 서브그래프들을 호출하는 형식으로 구조화된다.

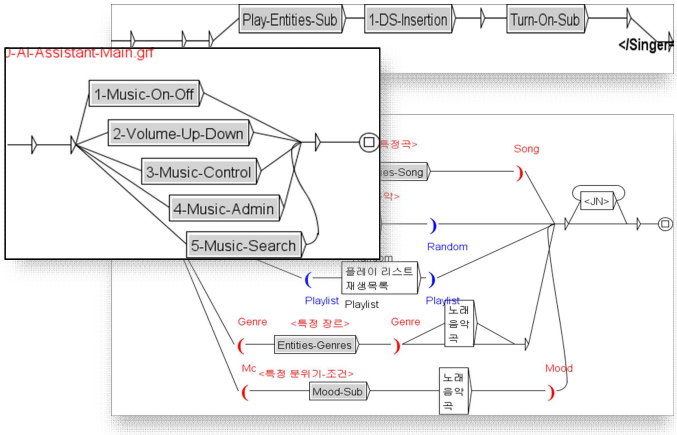

그림 2 한국어 LGG 구조 예시

그림 2의 메인그래프에서 호출하는 서브그래프는 약 40여 가지로, 각각 개체명, 부사 표현, 서술어, 보조동사가 기술되어 있다. 개체명 부분은 네 가지 대범주 간에 상당 부분 공유되는 양상을 보이지만, 서술부 부분에서는 대부분 서로 다른 서브그래프를 구성하게 된다. 이는 서술부의 목적어로서 등장해야 하는 개체명들이 곡, 플레이리스트, 장르 등으로 한정되는 데 반해, 서술부는 특정 작업을 수행하도록 요청해야 하는 부분으로서 다양한 용언 표현들이 등장할 수 있기 때문이다.

개체명은 패턴문법에 일일이 나열하는 대신, 별도의 사전자원을 활용하는 것이 더 효과적이다. 가령 고유명사로 실현되는 구체 곡명과 가수명이 여기에 해당한다. 현재 이러한 개체명들은 DECO 전자사전에 각 도메인별 별도의 개체명 목록이 표제어로 등재되어 있어 이를 활용하는 방법으로 패턴문법이 구조화된다.

4.2. 중국어 LGG 패턴문법

중국어의 LGG 패턴문법은 한국어 요청문과 같은 의미 카테고리 분류를 기반으로 구축된다. 그림 3에 제시된 중국어 요청문의 메인그래프는 50여 개의 서브그래프를 호출하는 방식으로 구성되는데, 이때 개체명으로는 특정 가수명과 곡명을 지정하는 개체명 서브그래프, 볼륨을 기술하는 서브그래프, 플레이리스트 어휘가 수록된 서브그래프 등으로 구성되었다.

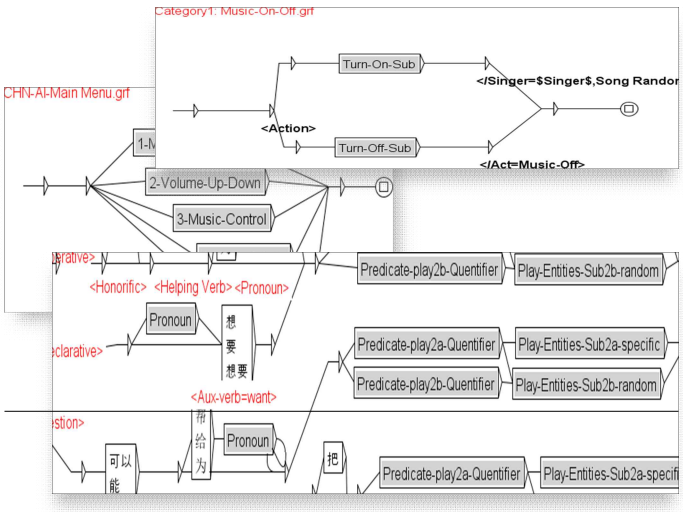

그림 3 중국어 LGG 구조 예시

서술부로는 음악을 켜고 끄는 서브그래프, 음량을 올리고 내리는 서브그래프, 음악 제어 기능 제반의 서브그래프, 그리고 재생목록 관련 관리 및 검색 기능을 수행할 수 있도록 하는 서술어들이 기술된 서브그래프 등으로 표상되었다. 이외에 시간 관련, 정도 관련 부사어를 기술하는 서브그래프들도 구축되었다. 이처럼 서브그래프 구성 방식은 한국어와 차이가 거의 없으나, 여기에 포함되는 표현들의 종류와 배치 순서에서 상이하게 된다.

중국어는 서법에 따라 포함되는 어휘의 종류가 달라지고 이들이 위치하는 곳도 다르게 적용된다. 따라서 한 그래프 안에 각 서법에 대한 패턴문법 기술이 다른 경로에 할당되는 차이가 나타난다.

5. 한국어와 중국어 패턴문법의 비교분석

5.1. 통사적 요인

한국어와 중국어의 중요 차이점 중 하나는 어순적 차이에서 설명될 수 있다. 한국어는 문장의 기본 구조가 SOV([주어]-[목적어]-[서술어]) 유형이지만, 중국어는 SVO([주어]-[서술어]-[목적어])가 기본 어순 구조가 되기 때문이다. 즉 한국어는 개체명에 서술부가 후행하게 되는 구조이나, 중국어는 서술부가 먼저 위치하고 이후에 개체명 부류가 나타난다. 이러한 어순 차이로 인해 한국어와 중국어에서만 각각 나타나는 특수한 표현들이 등장하기 때문에, 이들에 대한 별도의 기술이 두 언어의 패턴문법의 구축을 상이하게 하는 하나의 중요 요인이 된다.

어순 구조 차이와 더불어 서로 다른 성분이 등장하는 대표적인 예 중 하나는 한국어 패턴문법에서 '싶다/좋다' 보조 용언이 나타나는 경우다. 중국어에서는 '我想听音乐。(나는 음악을 듣고 싶다.)'처럼 화자를 나타내는 '주어(我)'가 실현하고 '想(생각하다)'가 본용언으로 실현한다. 따라서 직접적인 명령 요청문 형태가 아닌 서술문 형식으로 실현되므로 별도의 문형 구조로서 구조화되는 과정이 필요하다.

또한 한국어에서는 보조 용언으로서 표현될 수 있는 요소들이 중국어에서는 독립적인 어휘로 사용될 수 있는 경우들이 관찰되므로 이들을 고려하여 서로 다른 문형 구조의 LGG 유형으로 구조화하는 것이 필요하다.

또 다른 예로는 '把'가 나타나는 문장을 들 수 있다. '把'는 대상에 특정 행동을 가하여 어떠한 결과가 생겼는지에 초점을 주기 위해 사용되는 표현이다. 이 표현이 사용될 때에는 목적어와 서술어의 자리가 바뀐다. 기존 중국어 기본 어순에서는 서술어에 목적어가 후치하게 되는데, '把'가 사용될 때는 반드시 이 뒤에 서술어를 위치시킨 다음 목적어에 해당하는 개체명을 기술해주어야 한다. 예를 들면 다음과 같다.

(1ㄱ) 播放(재생하다)-古典(클래식)-音乐(음악)
[클래식 음악을 재생해.]
(1ㄴ)把-古典(클래식)-音乐(음악)-播放(재생하다)-一下(좀~하다)
[클래식 음악을 재생해.]

한국어와 중국어의 이러한 통사적 차이를 이해하는 것은 추후 다국어 간 요청문 화행을 자동 번역하는 데에 필요한 패턴문법을 LGG로 구축할 때 중요한 단서가 된다.

5.2. 형태적 요인

한국어와 중국어의 형태적 차이는 활용의 측면에서 두드러지게 나타난다. 이러한 차이는 크게 두 가지 측면에

서 패턴문법 기술에 영향을 미친다. 첫 번째는 서법 기술적 관점에서의 문제이고, 두 번째는 체언과 용언의 형태적 차이 기술 관점에서의 문제다.

한국어는 후치사 활용이 일어나지만, 중국어는 활용 현상이 나타나지 않는다. 한국어의 서법은 활용어미들로 표현될 수 있기 때문에 대부분의 용언 활용형이 앞선 표현들과 바로 이어지는 경로 처리가 가능한 반면, 중국어의 서법은 새로운 어휘가 추가되거나 일부 표현의 삭제가 수반되므로 서법에 따른 그래프 경로를 별도로 구축해야 한다.

예를 들어, 한국어에서는 명령문을 표현하는 활용어미 '어', '어라', 의문문을 표현하는 활용어미 '는가', '을래' 등이 실현된다. 이들은 패턴문법 구축에서 DECO 활용형 사전에서 사용하는 명령문 어미 태그(IMP)와 의문문 어미 태그(INT)를 활용하여 동사부 어간 뒤에 기술되는 방식으로 정규화된다. 반면 중국어 요청문 패턴문법에서는 이 모든 서법을 인식할 수 있도록 기술하기 위해서 서법마다 각각의 경로를 나누어 개별적으로 필요한 어휘들을 고유의 순서에 따라 배치해주는 작업이 수행되어야 한다. 다음은 한국어와 중국어의 활용어미 부분의 차이를 보인다.

(2ㄱ) 노래 아무거나 틀어 **줘**.
(2ㄴ) 노래 아무거나 틀어 **줄 수 있어**?

(3ㄱ) 帮我随机播放歌曲。
(3ㄴ) **可以**帮我随机播放歌曲**吗**？

한국어에서는 예시 (2ㄱ)과 (2ㄴ)에서처럼 활용어미 변화만 고려해서 패턴을 구성하면 된다. 반면 중국어에서는 (3ㄱ)과 (3ㄴ)에서처럼 앞뒤로 새로운 성분이 등장하며 경우에 따라 중간에 새로운 요소가 삽입되기도 한다. 이 성분 및 요소들은 서법에 따라서 필수적으로 실현되는 것, 수의적으로 실현될 수 있는 것이 서로 다르므로 중국어에서는 서법별로 그래프의 경로를 분리하는 것이 적절하다.

다음으로 체언과 용언의 형태 차이 관점이다. 한국어의 체언과 용언은 형태적 차이가 명시적이기 때문에 이 둘이 혼동되기 어렵지만, 중국어의 체언과 용언은 형태적 차이가 표면적으로 드러나지 않고 문장 내 위치에 따라 규정되는 것이므로 패턴문법 구축에서 이들의 구별에 유의해야 한다. 대표적이면서도 기본적인 예시로 '歌'라는 어휘를 들 수 있다. 이 표현은 '노래'라는 명사로서 사용될 수도 있지만, '노래하다'라는 동사로서 기능할 수도 있다. 해당 어휘가 음악청취 도메인에서 명사로 사용된다는 것을 판단하는 기준은 이 표현이 동사 표현들의 뒤에 위치한다는 사실에서부터 비롯되는 것이다. 물론 '노래하다'로 '唱', '唱歌' 등의 표현들이 더 자주 사용되긴 하지만, 동일한 형태가 서로 다른 품사들로 나타나는 경우를 기계 인식적 차원에서 혼동하지 않을 수 있도록 해야 한다. 이는 공기하는 성분에 따라 변수를 다르게 부여하는 방법 등으로 해결하는 것이 필요하다.

5.3. 어휘적 요인

모든 언어 사이의 어휘적 대응 관계가 그렇듯이, 한국어와 중국어에도 완벽한 일대일 대응이 되지 않는 표현들이 존재한다. 한국어의 어떤 표현은 중국어의 다단어 표현에 해당하기도 하고, 두 단어의 의미가 동일하지 않고 유사하거나, 특수한 의미의 표현이 하나의 언어에서만 사용되는 경우도 있다.

중국어 명령문에는 '请[honorific]'이 문두에 나타나는 경우가 있는데, 이 어휘는 한국어 명령문에서 대응되는 어휘가 존재하지 않는다. 또한 한국어에서 '~해 줄래?'에 해당하는 표현은 중국어에서는 '能[Auxiliary Verb=can] ~ 吗[Question Marker]'로만 기술된다. 즉 중국어에서 요청 의문문을 구성할 때는 가능성을 의미하는 조동사 부류가 문장 맨 앞에 반드시 실현되어야 한다.

두 언어의 음량 조절 카테고리에서 서술부에 해당하는 표현들에서도 차이점이 나타났다. 한국어에서는 개체명과 공기하는 서술어로 '낮추다', '높이다'와 같은 단일 서술어 표현들이 등장한다. 중국어에서는 '이동하다', '열다' 등을 함의하는 '调', '开'와 높낮이를 의미하는 '高', '低' 등이 공기하여 낮추고 높이는 의미를 표현할 수 있다. 이처럼 한국어에서는 사동의 의미를 지니는 후치사 '-추-', '-이-'가 한 동사 내에 삽입되는 형태라면 중국어에서는 해당 의미의 표현을 '调'와 '开'의 두 가지 어휘로 나타내는 차이를 보인다.

같은 방식으로, 한국어에서는 '작다'와 '크다' 표현을 사용하기 위해서 '작게 하다', '크게 하다'처럼 다단어 표현이 실현될 수 있으나, 중국어에서는 이 경우에도 '낮추다', '높이다'류 표현과 동일하게 동사 '调', '开'를 사용하는 차이를 보인다.

이와 같은 언어별 상이성이 고려되어야 추후 요청문 처리 다국어 플랫폼을 구축할 때 공통된 범주와 차별적인 범주를 정리 종합하여 음악청취 도메인 내 요청문 패턴문법 구축에 보편적 가이드라인을 제시할 수 있을 것이다.

6. 두 언어의 패턴문법의 비교 요약

6.1. 패턴문법 규모

LGG 프레임을 활용한 패턴문법 기술은 사용자 정의에 따라 특정 원소 혹은 서브그래프를 경로로 연결하여 문법을 기술하는 방식이다. 이렇게 인식된 경로를 통해 해당 시퀀스들은 각 분류태그를 할당받게 된다.

표 4는 본 연구에서 구축된 한국어 및 중국어의 패턴문법의 규모를 보인다. 여기에 수록된 패턴의 개수는 개체명을 변수(X)로 설정하여 매핑(mapping)한 것으로, 여기에 개체명이 실제로 채워지는 경우 인식되는 시퀀스의 수는 기하급수적으로 증가하게 된다. 한국어와 중국어 메인그래프에서는 각각 40여 개, 50여 개의 서브그래프들이 호출된다. 여기서 기술된 패턴문법들이 인식할 수 있는 표현의 전체 수는 한국어의 경우 약 2,600,500개, 중국어의 경우는 약 11,195,600개에 이른다. 변수(X)로 설정된 개체명을 가령 10,000개로 가정한다면, 한국어 'Music-Control' 카테고리에서 인식 가능한 패턴 수는 총 6,688,000,000여 개가 된다.

표 4 한국어 및 중국어 패턴문법 규모

언어	서브그래프	패턴 수
한국어	Music-On/Off	약 115,600
	Volume-Up/Down	약 1,159,000
	Music-Control	약 668,800
	Music-Administration	약 627,300
	Music-Search	약 29,800
중국어	Music-On/Off	약 5,748,100
	Volume-Up/Down	약 327,000
	Music-Control	약 573,500
	Music-Administration	약 493,000
	Music-Search	약 4,054,000

6.2. 패턴문법의 적용 결과

구축된 패턴문법의 입력(input) 및 출력(output) 변수에는 요청문이 요구하는 작업을 판단하는 데 필수적인 성분들이 할당되어 실현된다. 패턴문법은 FST문법 형식으로 구축되기 때문에, 여기 명시된 출력태그를 통해 요청문의 작업종류와 구체적인 작업내용을 명시하는 것이 가능하다. 그림 4는 이러한 패턴문법을 코퍼스에 적용하여 요청문의 의미 카테고리를 분류한 결과의 예를 보인다.

<Action> 조용한 노래 틀어 줘 </Singer=, Song=조용한 노래, Act=Music-On>
<Action> 播放一首安静的歌曲。</Singer=, Song=一首安静的歌曲, Act=Music-On>

그림 4 패턴문법을 적용하여 XML마크업한 예

7. 결론

본 연구에서는 AI 어시스턴트의 음악청취 도메인에서 나타나는 요청문을 인식할 수 있도록 LGG 프레임으로 패턴문법을 구조화하는 방법론을 한국어와 중국어를 대상으로 소개하였다. 그 결과 굴절어에 가까운 한국어의 활용 양상과 고립어에 가까운 중국어의 어휘 배열 양상을 각 언어적 특징에 맞추어 체계적으로 패턴문법으로 구축할 수 있었다. 두 언어자원은 패턴문법 구성 성분의 통사적 배열과 어휘·형태적 측면에서 차이를 보였으며, 이러한 차이점은 구축된 그래프의 구조와 구성성분 측면에서 서로 비교해볼 수 있었다.

본 연구에서는 특정 언어에 국한되지 않는 의미 카테고리를 설정한 후 이에 대한 개별 언어의 어휘적 실현을 기술하는 방식을 취함으로써, 다양한 언어에 호환적으로 적용될 수 있는 방법론을 제시하였다. 이는 다국어 기반의 AI 어시스턴트 플랫폼을 구축하는 데에 중요한 언어자원의 구축 방법론으로 활용될 수 있을 것으로 기대된다.

참고문헌

[1] 한국외국어대학교 DICORA. http://dicora.hufs.ac.kr/

[2] Gross, M., *The Construction of Local Grammars. Finite-State Language Processing*. Roche & Schabes(eds.). the MIT Press, 329-354. 1997.

[3] Prager, J., Radev, D. & Brown, E., The use of predictive annotation for question answering in TREC8. *Conference of the Eighth Text Retrieval*, 399-411. 1999.

[4] Ramanand, J., Bhavsa, R. K. & Pedaneka, R. N., Wishful thinking: finding suggestions and 'buy' wishes from product reviews. *Proceedings of the NAACL HLT 2010 Workshop* 54-61. 2010.

[5] Li, X. & Dan, R., Learning question classifiers : the role of semantic information. *Natural Language Engineering 12*.3, 229-249. 2015.

[6] McCallum, A. & Nigam, K. A., Comparison of event models for naive bayes text classification. *AAAI-98 of the Workshop on Learning for Text Categorization*, 41-48. 1998.

[7] Schapire, R. E. & Singer, Y., BoosTexter : a boosting-based system for text categorization. *Machine Learning 39*.2/3, 135-168. 2000.

[8] Haffner, P., Tur, G., & Wright, J.H., Optimizing SVMs for Complex Call Classification. *2003 IEEE International Conference on Acoustics Speech and Signal Processing 1*. 2003.

[9] 陈浩辰. 基于微博的消费意图挖掘. *哈尔滨工业大学*. 2014.

[10] 贾俊华.一种基于AdaBoost和SVM的短文本分类模型. *河北工业大学*. 2016.

[11] Kim, Y., Convolutional neural networks for sentence classification. *Proc of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1746-1751. 2014.

[12] Ravuri et al. A Comparative Study of Recurrent Neural Network Models for Lexical Domain Classification. *2016 IEEE International Conference on Acoustics of Speech and Signal Processing (ICASSP)*, 6075-6079. 2016.

[13] 황창회, 윤소은 & 남지순. AI 어시스턴트의 요청문 인식 및 분류 주석을 위한 음악청취 관련 문형 패턴 연구. 「언어과학」 27.1, 95-132. 2020.

[14] 남지순. *코퍼스 분석을 위한 한국어 전자사전 구축 방법론*. 출판사 역락. 2018.

[15] Paumier, S. *De la Reconnaissance de Formes Linguistiques a L'analyse Syntaxique.* Ph.D. dissertation. Univ of PEMLV in France. 2003.

[16] 황창회 & 남지순. SNS 사용자 생성문에 대한 코퍼스 수집 시스템 구축 방법론: Deco-T-Crawlers. 「DICORA-TR-2018-01」. 한국외국어대학교 DICORA 연구센터. 2018.

[17] 이경순, 황금하, 권오욱 & 김영길. 목적지향 대화 시스템을 위한 챗봇 연구. 「정보처리학회논문지 - 소프트웨어 및 데이터 공학」 6.11, 499-506. 2017.

[18] Ahmad, A. S. et al. A review on applications of ANN and SVM for building electrical energy consumption forecasting. *Renewable & Sustainable Energy Reviews 33*.2, 102-109. 2014.

[19] Gross, M. Nouvelles Applications des Graphes D'automates Finis a la Description Linguistique. *Lingvisticae Investigationes 22*.1-2, 249-262. 1999.

[20] Su, H. et al. Improving Multi-turn Dialogue Modelling with Utterance ReWriter, *ACL Proceedings of the 57th Annual Meeting of the ACL* 22-31. 2019.