오픈도메인 질의문 자동 분류를 위한 주석 말뭉치 구축 연구

안애림○, 이서진, 최동현, 김응균, 남지순†

(주)카카오, 한국외국어대학교†

{eileen.an, kaylee.sj, heuristic.085, Jason.eg}@kakaocorp.com, namjs@hufs.ac.kr†

A study on the Construction of Annotated corpora for the Automatic Classification of Open Domain Queries

AeLim Ahn○, SeoJin Lee, DongHyun Choi, EungGyun Kim, JeeSun Nam

Kakao corporation, Hanguk Univesity of Foreign Studies

요 약

본 연구는 오픈도메인 자연어 질의문 유형을 '질문 초점(Question Focus)'에 따라 분류하고, 기계학습 기반 질의문 유형 분류기의 성능 향상을 위한 주석 말뭉치 구축을 목표로 한다. 오픈도메인 질의문 분석을 통해 의문사 등의 키워드 기반 질의문 유형 분류의 한계를 설명하고, 질의문 내의 비명시적인 의미자질을 고려한 질문 초점 기반 질의문 유형 분류 기준을 정의하였다. 이 기준에 따라 구축된 112,856 문장의 주석 말뭉치를 기계학습(CNN) 기반 문장 분류 시스템의 학습 데이터로 사용하여 실험한 결과 F1-Score 97.72%성능을 보였다. 또한 이를 카카오 오픈도메인 질의응답시스템에 적용하여 질의문 확장을 위한 의미 자질로 사용하였고 그 결과 전체 시스템 성능을 1.6%p 향상시켰다.

주제어: 질의문 분류, 질의문 분석, 질문 초점, 오픈 도메인 질의 응답 시스템, 주석 말뭉치

1. 서론

질의응답시스템은 자연어 질의를 분석하여, 문서 내에 정답과 가장 가까운 답을 자동으로 추출하여 제공한다. 정답 추출 시에 웹검색을 통해 정답 후보 문서를 선택하는데, 이때 자연어 질의를 그대로 사용하여 검색할 경우 원하는 문서를 얻는데 한계가 있다. 질의문에 포함된 어휘들이 문서 내에 존재하지 않는 형태이거나, 의미적으로 불필요한 어휘일 수 있기 때문이다. 이때 정보의 손실 없이 질의자의 의도를 잘 드러낼 수 있는 질의 확장이 요구되는데, 이를 위해 질의문을 의미 유형에 따라 구분할 필요가 있다. 질의문 유형 분류는 기계학습 기반의 질의문 유형 판별기에 의해서 자동으로 이루어지는데, 이 시스템의 성능 향상에 있어 주석 말뭉치 구축은 매우 중요한 과제이다.

질의문 자동 분류를 위한 주석 말뭉치 구축을 위해 먼저 한국어 자연어 질의문의 언어학적 분석을 선행하였다. 모든 질의문에는 '질의자가 요구하는 정보'인 '질문 초점(Question Focus)[1][2][3][1]'이 존재한다. 질문 초점은 정답 유형(Answer Type)과 달리 정답에 대한 사전 정보 없이 질의문에 나타난 어휘의 의미적 속성에 따라서 구분 지을 수 있다. 질의문 내에서 질문 초점을 잘 보여주는 어휘 범주는 의문사로, 보통 이들은 질의문 유형을 분류하는 핵심 키워드로 사용된다. 그러나 자연어 질의문의 모든 경우에 있어 의문사가 질의문 유형을 명시적으로 결정짓는 것은 아니다.

(1) a. 대한민국 대통령이 누구야?
b. 누가 대한민국 대통령이야?
c. 대한민국 대통령 이름이 뭐야?
d. 대한민국 대통령 알려줘.
e. 대한민국 대통령 이름은?

(1)은 모두 '사람'을 묻는 질의문으로 육하원칙 분류 유형에 따라 '누구'에 해당한다. 이들은 동일한 질문 초점을 가지지만 문장의 형태는 다르다. (1a)는 '사람'을 묻는 명시적인 의문사 '누구'가 사용되었고, (1b)는 '누구'가 문장의 주어 역할을 하며 '누가'로 실현되었다. (1c)는 의문사 '무엇'이 사용되었는데 선행어 '대통령 이름'에 의해 질문 초점이 명확해진다. (1d)와 (1e)는 의문사를 사용하지 않았지만 질의문 내 어휘들에 인해 질문 초점을 충분히 파악할 수 있다. 따라서 한국어 자연어 질의문 분류는 의문사를 활용한 단순 키워드 기반의 분류 방식으로

[1] 성시형(1987)과 박홍원(1997)에서 '의문의 초점'의 개념으로 '발생한 의문에 대한 발화된 화자의 불확실성'이라 소개된 바 있고, Dan Moldovan(2003)에서는 질의문의 의문사 그 자체로 정답 유형(Answer Type)을 예측해내기 어려움을 설명하며 질문 초점(Question Type)이라는 개념을 소개하였다.

접근해서는 안 되고, 의문사 외의 비명시적인 의미 자질을 충분히 고려해야 할 것이다.

본 연구를 통하여 한국어 자연어 질의문 분석을 바탕으로 기계 학습 기반의 질의문 분류기의 성능 향상에 기여할 수 있는 질의문 유형 분류 기준을 제안한다. 또한 이 기준으로 구축된 112,856 문장의 자연어 질의문 주석 말뭉치를 학습한 문장 분류 시스템을 카카오 한국어 오픈도메인 질의응답시스템[2]에 적용하고자 한다.

2. 관련 연구

질의응답시스템의 자연어 질의를 처리하기 위한 언어학적 기초를 둔 연구는 많이 수행되지 않았다. 전통적인 한국어 의문문 및 의문사에 대한 연구들이 대부분인데[4][5], 이들은 한국어 의문문의 문장 형식이나 의문사의 의미적 특성에 집중하고 있어, 질의응답시스템의 질의 분류 및 확장의 관점에서 적용하기에는 어려움이 있다.

질의응답시스템을 위한 질의문 분석에 있어 [6]은 의문사로 질문 유형(Question Type)을 결정한 후 질문 유형과 개체명 정보를 활용하여 정답 유형(Answer Type)을 예측하였다. 이때 의문사가 질문의 정확한 정답유형을 결정짓기 어려우나, 제약 정보로 중요하게 활용될 수 있음을 강조하였다. [3]에서도 의문사 유형에 따라 질문유형과 정답 유형을 결정하고, 질문 유형으로 정답 유형을 유추하기 어려운 경우를 보완하기 위해 질문 초점(Question Focus)를 제안하였다.

앞선 연구에서 질의문 분석은 의문사에 따라 질문 유형을 결정짓는 것으로부터 시작하는데, 의문사가 생략된 형태나 청유형 문장에 대한 구체적인 처리 방안은 이야기되고 있지 않다. 또한 정답 유형을 인식하는 문제는 (2)와 같이 동일한 개체명, 통사 구조, 어휘 등이 사용된 문장이라도 의미적 속성에 따라 그 유형이 달라질 수 있는데, 이는 기계학습 기반의 시스템에서 일관성있는 학습을 어렵게 한다.

(2) a. 영화 라이언킹 주인공이 누구야? (정답유형: 동물)
b. 영화 기생충 주인공이 누구야? (정답유형: 사람)

본 연구에서는 실제 자연어 질의문의 다양한 패턴을 제시하고, 기계학습을 위한 말뭉치 구축을 위해 정답 유형이 아닌 질문 초점의 관점에서 문장 분류 기준에 대해 자세히 설명할 것이다.

3. 오픈 도메인 질의문의 3 가지 형태

질의문의 의미 속성을 잘 보여주는 대표적인 어휘 범주인 의문사의 사전적 정의는 '의문의 초점이 되는 사물이나 사태를 지시하는 말'로 풀이된다. 이들은 품사 분류의 기준이 되는 의미, 기능, 형태적 속성은 같지 않기 때문에 하나의 품사로 분류하지 않지만, 설명의문문(Explicative Question)에서 실현된다. 한국어 의문사는 표 1과 같다[7].

표 1 한국어 의문사 유형

품사	의문사
의문대명사	누구, 무엇, 뭐, 언제, 어디, 얼마, 몇
의문형용사	어떻다, 어떠하다
의문부사	어떻게, 얼마나, 왜, 어째서, 언제
의문동사	어찌하다
의문관형사	어떠한, 어떤, 어느, 무슨, 몇, 웬

국어 문법에서 설명하는 의문사와 의문문의 특성은 대부분의 자연어 질의문에서도 나타나지만 실제 쓰임에 있어 매우 복잡한 의미 특성을 보인다. 3 장에서는 의문사를 이용한 키워드 기반 자동 분류의 어려움 설명하기 위해 자연어 질의문의 특징을 세 가지 형태로 나누어 이야기하고자 한다.

3.1 명시적인 의문사를 포함한 질의문

질의문 내에 의미가 명시적인 의문사 '누구(누가), 언제, 어디, 왜, 어떻게'를 포함하는 경우이다. 의문사가 명시적이기 때문에 이를 기반으로 질의문 분류가 가능할 것이라 생각할 수 있으나, 하나의 의문사가 두 개 이상의 질문 초점을 표현할 수 있으므로 질의 확장의 차원에서 상세하게 나눠야한다.

(3) 어디
a. 사드가 배치되는 곳이 어디야? (장소: WHERE)
b. 아이유 소속사가 어디야? (기관/단체: WHAT)
(4) 누구
a. 빌게이츠가 누구야? (사람: WHO)
b. 국내 건조기 시장 점유율 1위는 누구야? (기관/단체: WHAT)
(5) 언제
a. 전주 영화제 개막식이 언제야? (날짜/시간: WHEN)
b. 교정이 필요할 때는 언제지? (조건: WHAT)
(6) 어떻게/어떠하다
a. 조선시대에는 왕세자비를 어떻게 골랐을까? (방법: HOW)
b. 김영란법 시행에 대해 산업계에서는 어떻게 생각해? (생각/의견: WHAT)
c. 바닷물을 마시면 어떻게 돼? (결과: WHAT)
(7) 왜
a. 커피를 마시면 왜 졸리지 않을까요? (이유: WHY)

(3a)과 (3b)의 의문사는 모두 '어디'이지만 질문의 초점이 되는 대상은 다르다. (3a)는 지리적인 장소를 묻는 반면, (3b)는 장소가 아닌 소속사의 이름을 묻는 질의로 볼 수 있다. 질의 확장의 관점에서 보았을 때, (3a)는 '사드 배치 위치', (3b)는 '아이유 소속사 (이름)'

[2] https://nlp-api.kakao.com/

가 될 것이다. (4)~(6)의 경우도 마찬가지이고, (7)의 의문사 '왜'만이 질문 초점이 되는 대상이 유일하다.

3.2 비명시적인 의문사를 포함한 질의문

질의문 내에 의문사가 존재하나, 의문사 자체로 의미가 명시적이지 않은 '무엇, 얼마, 얼마나'와 의문관형사를 포함한 경우이다. 이들은 공기하는 어휘들에 의해 질문 초점이 드러난다.

(8) 무엇
a. 페미니즘이 뭐야? (정의/속성: WHAT)
b. 발렌타인데이 날짜가 뭐지? (날짜/시간: WHEN)
c. 동지에 팥죽을 먹는 이유가 뭘까? (이유: WHY)
d. 계좌통합관리서비스 이용 방법이 무엇이야? (방법: HOW)

(9) 얼마/얼마나
a. 경의선 책거리에서 하는 야외조각전은 얼마나 했어? (시간의길이: WHAT)
b. 스타벅스 라떼 얼마야? (가격: WHAT)
c. 백두산 높이가 얼마야? (정의/속성: WHAT)
e. 아이유 나이가 얼마나 돼? (나이: WHAT)

(10) 의문관형사: 어떠한, 어떤, 어느, 무슨, 몇
a. 김유신은 어떤 사람이야? (인물속성: WHO)
b. 지금 몇 시야? (날짜/시간: WHEN)
c. 무슨 이유로 방역 당국에 비상이 걸렸어? (이유: WHY)
e. 대구와 부산 중 어느 곳에 인구가 더 많아? (장소: WHERE)

(8)~(10)을 통해 질의문 내 어휘들과 비명시적 의문사의 쓰임이 어떻게 나타나는지 파악할 수 있다. 비명시적 의문사의 경우 각각의 질문 초점에서 어떤 어휘들과 공기하여 나타나는지를 상세하게 구분해야할 것이다.

3.3 의문사가 없는 질의문

자연어 질의문에는 '-을/를 알려줘' 등의 청유형 문장이거나, 의문사가 생략되어 보조사 '-은/는'으로 종결되는 문장을 쉽게 볼 수 있다. 이들은 의문사가 없는 대신 질문의 초점을 드러내는 어휘들이 문장 내에 나타난다.

(11) a. 라면 업계 1위 알려줘. (기관/단체: WHAT)
b. 할리우드 여배우 수입 1위는? (사람: WHO)

(11)의 문장은 의문사가 없는 질의문 형태이다. 두 질의문 모두 '1위'를 묻지만 질문 초점은 각각 '라면 업계 회사(회사)'와 '할리우드 여배우(인물)'로 그 유형이 다르다. 이처럼 의문사가 없다면 문장 내 어휘들의 의미가 질문 초점을 결정하는데 영향을 미친다.

4. 질문 초점에 따른 6가지 질의문 유형 분류

4 장에서는 질문 초점에 따라 질의문의 유형을 여섯 가지로 나누고 상세 의미와 문장 패턴을 기술하고자 한다.

4.1 사람(WHO)

'사람(WHO)' 범주는 다음의 상세 의미를 포함한다.

표 2 사람(WHO)의 상세 유형

번호	의미	문장 패턴
1-1	인명	X-은/는/이/가 {누구} X-의 이름은 {무엇}
1-2	인물속성	X-은/는/이/가 {어떤}-Y이다

(12) a. 훈민정음을 쓴 왕이 누구지?, 미국 대통령 이름이 뭐야?, 민주당 원내대표는?
b. 베토벤은 어떤 사람이야?

<인명>은 (12a)의 예시처럼 '속성 X'(ex. 훈민정음을 쓴 왕, 미국 대통령 등)를 가지는 사람의 이름을 질문하는 것으로 'X-인 사람은 이름은 무엇?'로 치환될 수 있다. <인물속성>은 (12b)와 같이 '특정 인물 X'(ex. 베토벤 등)의 속성을 질문하는 것으로 Y 의 자리에는 인물 명사(ex. 사람, 선수, 학생, 대표자 등) 등이 보여진다.

그러나 이름을 묻는 경우라도 (13)의 예처럼 알고자 하는 대상이 사람이 아님을 나타내는 명사(ex. 펭귄, 반려견 등)와 공기한다면 '사물/개념(WHAT)'으로 분류한다.

(13) a. 유해진 반려견 이름이 뭐야?
b. 뽀로로에 나온 펭귄 캐릭터 이름은?

4.2 장소(WHERE)

'장소(WHERE)' 범주는 다음의 상세 의미을 포함한다.

표 3 장소(WHERE)의 상세 유형

번호	의미	문장 패턴
2-1	위치	X-은/는/이/가 {어디} X-은/는/이/가 {어디}-에 [있다/위치하다 등] X-의 주소는 {무엇}
2-2	사건	X-은/는/이/가 {어디}-에서 [일어나다/열리다 등] X-은/는/이/가 [일어나다/열리다 등]-ㄴ Y-은/는 {어디} X-의 Y-은/는 {어디}
2-3	이동	{어디}-에/에서/로/를 [가다/오다/떠나다 등] [가다/오다/떠나다 등]-ㄴ Y-은/는/이/가 {어디}
2-4	기타행위	{어디}-에서 팔다/사다/놀다/만나다 등

(14) a.올림픽공원은 어디에 있어?, 터키 수도 어디야?, 바르셀로나 홈구장은 어디야?

b. 올해 아시안 게임 어디서 열렸어?, 축제 개막 공연이 열린 곳은?, 2002 년 월드컵의 개최지는?
c. 어디로 휴가를 떠났는가?, 세월호 최종 목적지는?
d. 호가든 레몬을 파는 나라는 어디야?

<위치>는 (14a)와 같이 특정 장소의 위치를 묻는 질의로 'X 의 위치는 어디?'로 치환될 수 있으며, X 의 자리에 장소개체명(ex. 판교역, 올림픽공원 등)이나 행정구역 단위명사(ex. 나라, 시, 수도 등), 장소성 일반명사(ex. 집, 홈구장 등) 등이 나타날 수 있다. <사건>은 (14b)처럼 특정 사건의 발생 장소를 질문하는 것으로 '사건 X 가 일어난 장소는 어디?'로 치환된다. Y 에는 '장소'의 유사 의미 단어(ex. 곳, 지역, 개최지 등)가 올 수 있다. <이동>은 (14c)의 예처럼 '이동의 목적지 또는 출발지' 등을 이동 동사를 사용하여 질의한 형태로 '이동한 장소는 어디?'의 의미로 치환될 수 있는 특징을 가진다. <기타행위>는 장소성 논항과 공기하는 서술어를 포함한 질의로 '어떤 행위가 발생한 장소는 어디?'로 치환될 수 있다.

장소를 표현하는 대표적인 의문사 '어디'는 빈번하게 '무엇/무슨' 대신하여 사용된다. 다음을 보자.

(15) a. 세계에서 가장 깊은 블루홀은 어디야?
b. 태조 왕건이 건국한 나라는 어디야?
c. 도르트문트가 어디 나라 팀이야?
(16) a. 류덕환은 군제대 후 어디에 출연하나?
b. 김미화가 사과문을 올린 곳은 어디지?

(15)에 나타난 '어디'는 'A-IS-B'의 지정사 구문의 보어 위치에 실현된 것으로, '무엇/무슨'과 같은 의문사로 치환될 수 있는 사물/개념(WHAT) 유형의 질의문을 유도한다. (15)의 서술어 '출연하다, 올리다 등'과 공기하는 '어디'는 지리적인 위치를 가지는 장소를 뜻하는 것으로 보기 어렵다. 이들 또한 '무엇'으로 치환되므로 사물/개념(WHAT)으로 분류된다.

4.3 날짜/시간(WHEN)

'날짜/시간(WHEN)' 범주의 상세 유형은 아래와 같다.

표 4 장소(WHEN)의 상세 유형

번호	의미	문장 패턴
3-1	날짜/시간	X-은/는/이/가 {언제} X-은/는/이/가 {몇/무슨/어떤/어느} Y-이다 X-의 Y-은/는 {언제}
3-2	일정	X-은/는/이/가 {언제}-까지 X-의 Y-은/는 {언제/무엇}
3-3	시간적 배경	X-의 [시대/시간]-적 배경 {언제/무엇} X-은/는/이/가 {어느/어떤/무슨} [시대/시간]이다

(17) a. 광복절이 언제야?, 신미양요는 몇 년에 일어났어?, 다음 대통령 선거일은?
b. 정부 예산안 국회 제출이 언제까지야?, '위대한 유혹자' 편성 시간표는?
c. '걸 프롬 예스터데이'의 시대 배경은?, 맹자가 어느 시대 사람이야?

이 범주는 '구체적인 시점'의 유무를 의미적 분류 기준으로 삼았다. (17)의 예와 같이 특정 사건(ex. 광복절, 대통령 선거, 예산안 국회 제출, 드라마 편성 등)의 발생 시점을 묻거나, 사건이나 인물의 시대적 배경을 묻는 질문이다. 이들은 '사건 X 의 시간/날짜는 언제?' 또는 '인물 X 의 시대적 배경은 언제?' 등으로 치환 가능하다. Y 에는 시간단위명사(ex. 년, 시, 요일 등) 또는 시간성 명사(ex. 때, 시점, 일정, 기간, 선거일, 탄생일 등) 등이 보여질 수 있다.

(18) a. '북극의 눈물' 촬영기간은 며칠 걸렸어?
b. 탄저병과 역병은 언제 많이 나타나?

그러나 표면적으로 시간 또는 날짜를 묻는 질문처럼 보이기도 하지만 (18a)와 같이 '시간의 길의', (18b)의 '사건의 발생 조건'을 묻는 경우라면 사물/개념(WHAT) 으로 분류한다.

4.4 방법(HOW)

'방법(HOW)' 범주는 표 5 와 같다.

표 5 방법(HOW)의 상세 유형

번호	의미	문장 패턴
4-1	방법	X-을/를 {어떻게} 하다 X-의 Y-은/는 {무엇} X-을/를 {어떤/어느/무슨} Y-로/으로 하다

(19) a. 조선시대에는 왕세자비를 어떻게 골랐을까?, 치과 공포증을 극복하는 방법은 뭐가 있어?, 강수진은 국립발레단을 어떤 식으로 이끌었어?

이 범주는 특정 사건의 발생 방식/방법을 묻는 질문으로, '사건 X 의 방법은 무엇?'으로 치환될 수 있다. Y 에는 '방법'의 유사 의미 단어(ex. 방식, 방안 등)가 보여진다.

(20) a. 바닷물을 마시면 어떻게 돼?
b. 제시 가족 농구 결과 어떻게 나왔어?

방법을 묻는 대표적인 의문사 '어떻게'가 (20a)와 같이 용언 '되다'와 함께 쓰이거나, (20b)에서 처럼 '결과'와 공기하여 사용되는 경우는 '사건의 결과'를 질문 초점으로 가지므로 '사물/개념(WHAT)'으로 분류한다.

4.5 이유(WHY)

'이유(WHY)' 범주는 표 6 과 같다.

표 6 이유(HOW)의 상세 유형

번호	의미	문장 패턴
5-1	이유	X-은/는/이/가 {왜} [일어나다/발생하다 등] X-을/를 {왜} [하다/먹다/주다 등] X-의 Y-은/는 {무엇/뭐} X-은/는 {어떻게 하다/어쩌다} [죽다, 놓치다 등]

(21) a. 지진은 왜 일어나요?, 그는 왜 범죄를 저지를 수밖에 없었지?, 김정남이 사망한 이유가 뭐야?, 엘비스는 어쩌다 죽었어?

이 범주는 특정 사건의 원인을 묻는 것으로 '사건 X 의 이유는 무엇?'의 문장으로 치환할 수 있다. Y 는 '이유'의 유사 의미 단어(ex. 근거, 원인, 까닭, 요인 등)이 보여진다. 또한 '어떻게 하다'와 그의 준말인 '어쩌다'가 특정 문맥에서 사건 발생의 이유를 묻는 질문 의도를 나타내기도 하는데 이 경우는 이유(WHY)로 분류한다.

4.6 사물/개념(WHAT)

'사물/개념(WHAT)' 범주는 앞서 설명한 다섯 개의 질문 초점에 해당하지 않는 대부분이 포함된다. 상세 유형은 표 7 과 같다.

표 7 사물/개념(WHAT)의 상세 유형

번호	의미	문장 패턴
6-1	사물/개념의 이름	X-은/는/이/가 {무엇}
6-2	사물 속성	X-은/는/이/가 {무엇} X-의 Y-이/가 {무엇} X-은/는/이/가 {무슨/어떤/어느} Y-이다
6-3	개념적 장소명	X-은/는/이/가 {어디} X-은/는/이/가 {어느/어떤} Y-이다
6-4	그룹/기업/단체	X-은/는/이/가 {누구/어디}
6-5	인물 프로필	X-의 [나이/고향/별명/소속/직업/출신학교 등]
6-6	극중 역할	X-의 [이름/역할] {무엇}이다 X-는 {누구}-을/를/의 [연기하다, 역할하다 등]
6-7	시간의 길이	X-은/는/이/가 {얼마 동안/얼마나} 무엇하다 X-의 기간이 {얼마/얼마나}이다 X-을/를/은/는 {몇} [년/월/일/시간/박 등]이다
6-8	사건의 결과	X-은/는/이/가/면 {어떻게} 되다 X-의 결과 {어떻게} 되다
6-9	조건	X-의 [적용/후원/관람 등] 대상 {무엇}
6-10	생각/의견	X-은/는/이/가 {어떠하다} X-은/는/이/가 {어떻게} [생각하다, 말하다 등]
6-11	기타	신체부위: X-은/는/이/가 {어디}-에/을/를 [다치다/착용하다/뽀뽀하다 등] 형식: X-은/는/이/가 {어떤/어떠한} [형식/형태 등]이다 계절: X-은/는/한 계절 {무엇}, {어느} 계절이 X-하다

(22) a. 이란의 공용어가 뭐야?, 황정민 데뷔작은?
b. 페미니즘이 뭐야?, 독립출판 뜻이 뭐야?, CEO 가 무슨 뜻이야?

<사물/개념의 이름>은 질문하는 대상의 속성(ex. 이란의 공용어, 황정민 데뷔작 등)를 통해 알고자하는 대상의 이름을 묻는 것이고, <사물 속성>은 이미 알고있는 대상 이름(ex. 페미니즘, 독립출판, CEO 등)으로 그 속성을 물어보는 질문이다. Y 에는 '뜻'의 유사 의미 단어(ex. 정의 등)가 보여진다. 그밖의 세부 의미에 대한 예시는 (23)를 참고하자.

(23) a. <개념적 장소명>: 세계에서 가장 높은 산은 어디야?, 갈라파고스가 어느 나라 거야?
b. <그룹/기업/단체>: 피땀눈물 부른 그룹이 누구야?, 작년 한국시리즈 우승팀이 어디야?, 국권회복을 위한 비밀결사단체가 어디지?
c. <인물 프로필>: 에드시런은 어느 나라 사람이야?, 박옥선 할머니 연세가 어떻게 되지?, 방탄소년단 지민이 몇 년생이야?
d. <극중 역할>: '화랑'에서 이광수 이름 뭐야?, '미씽나인'에서 백진희는 누구를 연기해?
e. <시간의 길이>: 서울시 청년수당은 최대 얼마 동안 받을 수 있어?, 소멸시효완성 채권의 최대 채무 기간은 몇 년이지?
f. <사건의 결과>: 조현병에 걸리면 사람이 어떻게 돼?, U23 챔피언십 예선 2 차전 결과는 어떻게 됐어?
g. <조건>: 대출 계약 철회권 적용 대상은?, 희망가게 후원 대상 알려줘.
h. < 생각/의견>: 마크롱 대통령 정치 성향이 어떻게 돼, 박근혜 대통령 탄핵에 대해 천주교는 어떻게 생각해?, 아이폰 품질이 어때?
I. <기타>: 지난 경기에서 네이마르가 어디를 다쳤어?, 지드래곤의 새 앨범은 어떤 형식으로 발매돼?, 꽃이 피는 계절은?

5. 질의문 주석 말뭉치 구축 및 성능 평가

4 장에서 제시한 질문 초점 기반의 자연어 질의문 분류 기준에 따라 실제 쓰임을 잘 보여주는 112,856 개의 질문 문장을 포함하는 주석 말뭉치를 구축하였다. 표 8 은 각 유형별로 구축된 질의문의 개수를 보여준다.

표 8 유형별 질의문 개수

WHAT	WHO	WHEN	WHERE	HOW	WHY
55,852	25,402	10,328	7,992	2,533	10,749

이 말뭉치가 실제 시스템 성능 향상에 영향을 주는지 확인하기 위하여 [8]에서 제시된 Convolutional Neural Network(CNN)을 이용하는 기계학습 기반의 문장 분류기 사용하여 주어진 질의문을 육하 원칙 중 하나로 분류하는 모델을 구축하였다. 전체 말뭉치의 각 유형별 문장들은 8:1:1 의 비율로 훈련 데이터셋, 검증 데이터셋, 테스트 데이터셋으로 나뉘어졌다. 또한 이와 비교하기

위해 의문사의 유무를 보고 육하원칙을 파악하는 규칙 기반 육하원칙 분류기가 Baseline 시스템으로 설정되었다. 표 9 는 Baseline 시스템에서 각 육하원칙 카테고리를 분류하기 위해 사용된 키워드를 나타낸다.

표 9 Baseline의 유형 분류 키워드

WHAT	WHO	WHEN	WHERE	HOW	WHY
뭐/무엇	누구	언제	어디	어떻게	왜

표 10 은 두 시스템의 성능 평가 결과이며 질문 초점 기반 질의문 분류 말뭉치를 적용한 시스템의 성능은 97.72%의 상당히 높은 결과를 보였다. 반면 키워드 기반의 Baseline 시스템의 결과는 77.77%를 보였다. 이 실험 결과를 토대로, 본 논문에서 제시한 분류 방법론으로 구축된 주석 말뭉치가 기계 학습 기반의 시스템 성능 향상에 도움을 주고 있음을 확인할 수 있었다.

표 10 문장 분류기 성능 비교

문장 분류기	성능(F1-Socre)
규칙 기반 육하원칙 분류기	77.77
질문 초점 기반 육하원칙 분류기	97.72

6. 결론 및 활용 방법

본 연구는 질문 초점을 기반으로 자연어 질의문 유형 분류 기준을 정의하였다. 이는 다양한 형태와 의미의 자연어 질의문을 고려한 심도있는 언어학적 고찰이라는 점에서 의의가 크다. 또한 본 연구에서 마련된 분류 기준으로 구축된 주석 말뭉치가 기계 학습 기반의 문장분류기 성능 향상에 영향을 주고 있음을 확인할 수 있었다.

본 연구의 효용성을 좀 더 입증하기 위해 관리자의 허락을 얻어 '카카오 오픈 도메인 한국어 질의응답시스템'에 이를 적용하여 보았다. 이 시스템은 사용자 질의가 주어졌을 때 해당 질의로 웹 검색 후, 검색된 문서를 기계 독해를 통해 분석하고, 이를 통합하여 사용자에게 정답으로 제시한다. 이 과정에서 질의문의 육하원칙 분류 결과는 웹 검색을 위한 쿼리 확장으로 사용된다. 앞서 실험에서 설정했던 키워드 기반의 Baseline 시스템과 본 연구에서 제안한 질문 초점 기반 분류 방식의 시스템을 질의응답시스템에 적용하여 성능을 비교한 결과 1.6%p 더 나은 성능을 확인했다.

표 11 '카카오 오픈 도메인 질의 응답 시스템' 성능 비교

질의응답시스템	성능(F1-Socre)
규칙 기반 육하원칙 분류기 적용	68.13
질문초점 기반 육하원칙 분류기 적용	69.76

자연어 질의문 유형을 정하고 수작업으로 주석 말뭉치를 구축하는 일련의 작업들은 각 유형별 패턴을 구축하는 기초 작업으로 볼 수 있다. 본 연구에서 정의된 유형과 패턴들은 향후 형식화를 통해서 반자동으로 주석 말뭉치를 확장하는데 활용될 수 있을 것이다.

참고문헌

[1] 박홍원, "의문의 초점을 고려한 자연어 기반의 정보검색 시스템", 한국정보과학회 논문집, pp. 37-43, 1997

[2] 성시형, "의문의 초점과 의문문의 유형에 관한 연구", 한양대학교 석사학위 논문, 1987

[3] Dan Moldovan, Marius Pasca, Sanda Harabagiu, and Mihai Surdeanu. Performance Issues and Error Analysis in an Open-Domain Question Answering System. ACM Transactions on Information Systems, vol.21, no.2, pp.133-154, 2003.

[4] 류현미, "국어 의문문의 연구", 충남대학교 박사학위 논문, 1999

[5] 김충효, "국어의 의문사와 부정사 연구", 박이정, 1999

[6] 허정, 류범모, 장명길, 김현길, "오픈 도메인 질의응답을 위한 검색문서 제약 및 정답유형 분류기술", 정보과학회논문지, 제 39 권 제 2 호, pp. 118-132. 2012

[7] 남지순, "질의응답시스템을 위한 속성질의문의 의문사 및 속성명사에 대한 연구", 2010

[8] 최동현, 박일남, 임재수, 백슬예, 이미옥, 신명철, 김응균, 신동렬, "한국어 대화 엔진에서의 문장 분류", 제 30 회 한글 및 한국어 정보처리 학술발표 논문집, pp. 210-214, 2018.