담화 성분을 활용한 지시 발화의 키프레이즈 추출: 한국어 병렬 코퍼스 구축 및 데이터 증강 방법론

조원익○, 문영기†, 김종인‡, 김남수○

서울대학교 전기정보공학부 및 뉴미디어통신공동연구소○

인하대학교 컴퓨터공학과†

서울대학교 인지과학협동과정‡

wicho@hi.snu.ac.kr, ykmoon0814@gmail.com, prows12@gmail.com, nkim@snu.ac.kr

Keyphrase Extraction of Directive Utterances via Discourse Component: Construction and Data Augmentation of Korean Parallel Corpus

Won Ik Cho○, Young Ki Moon†, Jong In Kim‡, Nam Soo Kim○

Seoul National University, Department of Electrical and Computer Engineering and INMC○

Department of Computer Engineering, Inha University†

Interdisciplinary Program in Cognitive Science, Seoul National University‡

요 약

문서 요약, 키프레이즈 추출과 패러프레이징은 인간이, 혹은 기계가 문서를 보다 원활히 이해하는 데에 도움을 주는 방법론들이다. 우리는 본 연구에서 질문/요구 등의 지시성 발화를 대상으로, 핵심 내용을 추출하는 간단한 방법론을 통해 한국어 병렬 코퍼스를 구축한다. 또한, 우리는 인적 자원을 활용한 효율적인 데이터 증강 전략을 통해 부족하거나 필수적인 유형의 발화의 양을 보강하고, 약 5만 쌍 크기의 코퍼스를 제작하여 이를 공개한다.

주제어: 담화 성분, 키프레이즈, 패러프레이징, 한국어 병렬 코퍼스

1. 서론

문장 혹은 도큐먼트의 핵심적인 내용을 추출하는 것은 해당 텍스트의 이해를 위해 필수적인 과정 중 하나이다. 이는 문서 단위에서는 종종 추출적인(extractive), 혹은 추상적인(abstractive) 요약(summarization)과 같은 방식으로 진행되며, 문장 단위에서는 키워드(keyword) 혹은 키프레이즈(keyphrase) 추출이라는 방식으로 수행된다. 핵심내용 추출은 주로 문서 단위에서 고려되는 개념이지만, 문장 단위의 핵심 내용 추출을 생각해 본다면 그 예시는 다음과 같이 볼 수 있을 것이다.

(1) 많이들 궁금해하셨던 내용을 알려드리면, 올해에는 시월 십이일부터 십삼일까지 카이스트에서 한글 및 한국어 정보처리 학술대회가 개최됩니다.
→ 올해 시월 십이일부터 십삼일까지 카이스트에서 한글 및 한국어 정보처리 학술대회 개최

(2) 오늘 저녁 여덟 시에 서울대입구 풍경소리에서 동아리 뒷풀이가 있을 예정입니다.
→ 오늘 이십 시 서울대입구 풍경소리에서 동아리 뒷풀이 예정

위의 예시를 통해 보면, 문장 단위에서의 핵심 내용 추출은 일종의 패러프레이징, 즉, 같은 내용을 좀 더 간결한 형태로 나타내는 것으로 해석할 수 있다. 하지만, 어떤 내용이 문장이 전달하고자 하는 핵심 내용인지에 대해 이견이 있을 수 있고, 그것을 기술하는 방법에 대한 합의 역시 필요하다. 이런 점에서, 자연어를 자연어로 요약하는 것보다는 WikiSQL[1]과 같은 시맨틱 파싱이 더 효과적인 측면이 있다. 예컨대 (1)과 같은 문장을,

{기간: 올해 시월 십이일부터 십삼일, 장소: 카이스트, 이벤트: 한글 및 한국어 정보처리 학술대회}

이런 식으로 표현하는 것이다. 해당 방법론은 요즈음의 기술 체계 하에서는 일반적으로 여러 술어, 논항 및 도메인을 고려하여 해당하는 내용들을 결정하는 다중 분류의 형식으로 널리 사용되고 있다. 또한 그러한 분석 방법은 (1, 2)와 같은 서술뿐 아니라 질문, 요구 등의 지시 발화를 대상으로 하는 경우도 종종 있다.

그럼에도 불구하고, 우리는 자연어를 통한 직접적인 핵심 내용 추출의 필요성을 주장하려 한다. 정보나 행위를 요구하는 문장들을 시맨틱 웹 검색 등에서뿐 아니라 일상 생활의(real-life) 발화에서도 드물지 않게 볼 수

있다. 물론 이러한 경향은 인간과 인간의 대화, 혹은 인간과 챗봇의 대화보다는, 인공지능 스피커나 스마트 비서와 같은 음성 기반 자연어 이해 서비스에서 좀 더 많이 나타난다. 이러한 이유로, 요즈음에는 많은 사람들이 인공지능 대상의 발화, 즉 확실한 목적을 갖고 표현을 하는 형태의 발화에 익숙해져 있다. 하지만, 모든 사용자가 그러한 표현에 익숙한 것은 아닐뿐더러, 일상 대화에서의 표현을 활용하더라도 자연스럽게 대화를 이어나가고 지시 사항을 수행하는 시스템이 궁극적으로 우리가 목표로 해야 할 시스템이 될 것이다. 또한, 이는 해석의 가능성이 한정되어 있는 분류 기반의 시맨틱 파싱보다 더 다양한 분석의 가능성을 열어줄 것이다.

본 연구에서 지시 발화는 여러 가지 형태로 표현된, 질문과 요구의 목적을 가진 문장들을 지칭한다. 이는 단순히 평서문, 의문문, 그리고 명령문의 기준으로 구별 가능한 특질은 아니다. 예컨대 (3, 4)와 같은 발화의 경우 평서문이나 명령문의 형태를 가지지만 모두 청자에게 어떤 정보를 요청하는 특징을 가지고 있다.

(3) 이번 회의에서 미처 다루지 못했던 내용들을 자네가 찾아서 보고했으면 좋겠네.
→ 질문: **이번 회의에서 다루지 못한 내용**

(4) 내일 서울에 비 얼마나 올지 좀 검색해봐.
→ 질문: **내일 서울 강수량**

실제 대화들에서는 이보다 훨씬 다양한 스타일과 주제의 지시 발화가 사용되며, 입력 문장에 대해 화행(speech act)을 결정하고 핵심 내용을 추출하는 것이 질문이나 요구의 목적을 가진 발화들을 올바르게 이해할 수 있는 시작점이 될 것이다. 우리는 한국어 발화에서 이러한 점을 고려한 화행 분류 체계를 제시한 논문[2] 및 코퍼스[3]를 참고하여 연구를 진행하였다. 해당 연구에서 제시되는 화행의 구별은 '담화 성분'이라는 통사/의미론적 특질[4]을 화행 층위로 확장한 것을 기반으로 한다. 또한, 질문/요구 각각에 대한 보다 상세한 유형화를 통해 기존의 문형 기반 화행 어노테이션보다 더 포괄적인 언어 이해를 목표로 하였다. 여기서 담화 성분은 질문에 대해서는 질문 모음(question set), 요구에 대해서는 요구 사항(to-do-list)을 의미한다.

본 연구에서는 지시 발화의 키프레이즈라 할 수 있는 담화 성분의 개념을 통해 한국어 병렬 코퍼스를 만드는 방법과 데이터를 증강하는 직관적이고 간편한 방법에 대해 설명한다. 우리는 해당 방법론을 통해 약 3만 문장 가량의 지시 발화의 키프레이즈를 구조화된 방식으로 추출하였다. 또 기존의 코퍼스를 어노테이션하는 과정에서 얻어지는 질문/요구 문장 유형들 간의 불균형을 해결하기 위해 비교적 손쉬운 데이터 증강 방식을 제안하고, 실제로 그를 수행하여 총 5만 문장 가량의 한국어 병렬 코퍼스가 되도록 하였다. 다음 장에서는 관련 연구에 대하여 알아보고, 이후에 코퍼스 구축 및 데이터 증강에 대해 순차적으로 소개하고자 한다.

2. 관련 연구

본 연구와 가장 관련 깊은 분야는 정보 검색 중에서도 문서 요약, 키프레이즈 추출, 패러프레이징 등이다. 그 중에서도 한국어 문서 요약, 키프레이즈 추출 및 패러프레이징 데이터셋은 공개된 자료가 많지 않다. 한국어 문서 요약 데이터셋으로는 비상업적 활용을 전제로 공개된 네이버 뉴스 요약 데이터가 있지만[5], 문장 단위의 요약 데이터셋은 더욱 적다. 연구 측면에서 보면, 키프레이즈 추출의 경우는 특허와 과학 기술 분야에서 진행된 구문 분석 기법 연구[6,7]가 있으며, 패러프레이징의 경우는 어휘 빈도 분석을 이용하여 진행된 연구[8]가 있다. 그러나, 키프레이즈 추출의 경우 별도의 데이터셋이 존재하지 않으며, 패러프레이징의 경우 데이터셋이 존재하나[9] 주로 서술문에 대한 것이고 단어/어구 수준의 교체가 일어난 발화들을 대상으로 한다. 이에 우리는, 문장 단위의 지시 발화들에 대해 핵심 내용을 추출할 수 있는 트레이닝 코퍼스가 필요하다는 사실에 주목하였다.

3. 코퍼스 구축

우리가 하고자 하는 것은 지시 발화, 즉 질문과 요구의 목적을 가진 문장들로부터 핵심 내용을 추출하는 것이다. 이 때 지시 발화는 앞서 말했듯 (언어학적으로 의미를 가지는) 문장, 문장의 일부, 혹은 토픽이 동일한 문장과 문장의 연결체(sentence-like expressions)을 포괄적으로 지칭하는 발화들 중 질문과 요구의 목적을 가진 것들을 의미하며, 문장 형식에 구애받지 않는다. [2]에서 해당 발화들은 질문/요구 각각 17,689개, 12,968개이며, 우리는 이를 판정의문문(yes/no question), 선택의문문(alternative question), 설명의문문(wh-question)의 세 가지 질문 유형과 금지(prohibition), 요구(requirement), 강한 요구(strong-voiced requirement)의 세 가지 요구 유형으로 분류하고 그로부터 핵심 내용을 추출하였다.

	지시발화 유형		대응하는 어구
질문	판정의문문		-(인)지, - 여부
	선택의문문		-랑 -중 -한/할 (것)
	설명의문문	누구	사람, 정체
		무엇	의미
		어디	위치, 장소
		언제	시간, 기간, 시각
		왜	이유
		어떻게	방법, 대책
요구	금지		-지 않기
	일반 요구		-(하)기
	강한 요구		-지 않고 -(하)기

표 1. 핵심 내용을 프레이즈로 만드는 간단한 원리

핵심 내용을 프레이즈로 만드는 과정은 표 1과 같은 원리를 기반으로 하였다. 어노테이션은 2인의 한국인 화자

들의 태깅 및 중복 검증을 통해 이루어졌으며, 전반적으로 판정의문문, 설명의문문, 그리고 일반 요구의 비율이 높게 나타났다. 각 문장 유형에 대한 간단한 예시를 아래에 제시하였다. 코퍼스의 자세한 구성은 다음 절에서 제공된다.

(5) **[판정의문문]** 너 의료봉사 신청했어?
→ 질문: 의료봉사 신청 여부

(6) **[선택의문문]** 버스로 올거야 택시로 올거야?
→ 질문: 버스와 택시 중 타고 올 것

(7) **[설명의문문, 누구]** 오늘은 어떤 친구들이 왔니?
→ 질문: 오늘 온 친구들의 정체

(8) **[설명의문문, 무엇]** 스톡옵션이 뭔 줄 아니?
→ 질문: 스톡옵션의 의미

(9) **[설명의문문, 어디]** 너 지금 어디 있니 로비야?
→ 질문: 로비의 위치

(10) **[설명의문문, 언제]** 대구 몇시에 도착해?
→ 질문: 대구 도착하는 시각

(11) **[설명의문문, 왜]** 요즘 왜 이렇게 춥지?
→ 질문: 요즘 추운 이유

(12) **[설명의문문, 어떻게]** 챌린지 신청 어떻게 해?
→ 질문: 챌린지 신청하는 방법

(13) **[금지]** 태풍 오니까 밖에 나가지 마
→ 요구: 밖에 나가지 않기

(14) **[일반 요구]** 인적사항 확인 부탁드립니다.
→ 요구: 인적사항 확인하기

(15) **[강한 요구]** 욕심부리지 말고 코인 지금 팔아
→ 요구: 코인 지금 팔기

4. 데이터 증강

우리는 앞서 기존 데이터를 활용하여 질문과 요구 문장들에 대해 키프레이즈를 추출하였다. 하지만, 일차적으로 각 문장 유형 별로 데이터가 충분하지 않으며, 또 가장 유용하게 현실에 활용할 수 있을 만한 설명의문문에 대해 충분히 병렬 코퍼스의 양이 많지 않다고 판단하였다. 따라서, 우리는 키프레이즈로부터 다양한 유형의 문장들을 얻기 위하여 인적 자원을 활용한 문장 생성을 진행하였다.

일단, 각 문장 유형 별로 데이터를 균형 있게 확보하기 위해서는 선택의문문과 금지, 그리고 강한 요구의 발화가 필요하다. 이를 위하여 우리는 세 개의 각 유형 별로 400개의 키프레이즈를 제작하였다. 키프레이즈를 제작하는 과정에서 생성될 문장의 주제(topic) 역시 고려 대상이 되었는데, 구체적으로는 메일, 스케쥴, 하우스컨트롤, 날씨, 그리고 자유 주제에 대하여 각각 1:1:1:1:4의 비율로 문장을 생성하였다. 이는 3장에서 사용된 데이터셋에서의 주제 특성을 어느 정도 반영한 것이며, 스마트 비서 등에서의 활용에도 효과적인 코퍼스를 구축하는 데에 그 목적이 있다.

두번째 목적인 설명의문문의 추가적인 확보를 위하여, 800개의 키프레이즈가 제작되었다. 이 과정에서 고려된 문장의 주제는 위 문단과 동일하며, 설명의문문에서 일어나는 의문사-대응구 간의 원활한 변환을 위해 의문사의 사용을 전적으로 배제하고 대응구만을 사용하여 키프레이즈를 제작하였다. 이러한 경향은 위에서 선택의문문에 관한 키프레이즈를 제작할 때도 동일하게 나타났다.

이렇게 구축된 2,000개의 키프레이즈를 이용하여, 우리는 키프레이즈 하나당 열 개의 발화를 자유롭게 생성하도록 주석자들에게 주지하였다. 이 경우 다음과 같은 사항들이 권장되었다.

- 열 개의 문장은 최대한 서로 다른 스타일로 작성할 것. 이 때, 스타일은 존대 여부, 어조 등을 모두 포함.
- 꼭 키프레이즈에 있는 말을 반복할 필요 없고, 상황에 맞는 다른 단어/어구/술어를 넣어도 됨. 구어로 발화하기 적합한 표현일 것.
- 도치를 통해 문장 형태의 다양성을 추구하는 것 역시 권장됨.
- 설명의문문의 경우 의문사가 필수적으로 들어가야 하며 선택의문문도 경우에 따라 삽입될 수 있음. 두 문장 유형 모두 의문문으로 작성될 필요 없음.
- 금지 문장의 경우 청자가 할 수 있는 어떤 행위를 하지 않도록 하는 문장이어야 하며, 안 해도 괜찮다는 의미보다는 더 강제성을 지녀야 함. 그 행동을 금지하는 것이 다른 행동을 요구하는 것과 실질적으로 동치일 경우, 해당 표현으로 대체해도 크게 문제되지 않음.
- 금지와 강한 요구 문장 모두 명령문일 필요 없지만, 청자의 행동을 막거나 강제하는 목적을 지녀야 함. 강한 권유도 가능함.
- 화자/청자가 포함된 키프레이즈의 경우 각각 그에 상응하는 대명사 표현을 활용할 것. 이를 통해 화자/청자의 표현이 포함된 코퍼스와 포함되지 않은 코퍼스를 모두 구축.

이로써 우리는 2,000개의 키프레이즈들로부터 총 20,000개의 키프레이즈 - 질문/요구 문장 쌍을 얻었다. 이 과정에서 얻은, 한 프레이즈에 대한 다양한 질문 및 강한 요구 표현의 예시들을 아래에 공개하며, 이와 같은

방식으로 만들어진 전체 데이터셋의 구성 및 기존 데이터와의 합산 결과는 표 2와 같다.

(16) 키프레이즈: 대수학에서 가장 중요한 개념
(토픽: 자유주제, 질문 유형: 설명의문문문)
(16-1) 대수학에서 제일 중요한 개념 하나만 뽑아보면?
(16-2) 대수학에서 가장 중요한 개념이 뭐라고 생각해
(16-3) 대수학 에서 제일 중요한 개념이 뭐지
(16-4) 대수학 시험 전에 개념 하나만 보라면 뭐 보겠냐
(16-5) 개념 정리 중인데 대수학은 뭐부터 시작하는 게 좋을까
(16-6) 개념 복습 중인데 대수학은 뭐를 까먹지 않는게 제일 중요하지?
(16-7) 대수학 보려는데 제일 핵심 개념이 뭐예요
(16-8) 대수학에 무슨 개념이 제일 중요한 거 같냐
(16-9) 대수학 개념 너무 많은데 그 중 꼭 기억해야 될 게 뭐야
(16-10) 자네는 대수학에서 어떤 개념이 제일 중요하다고 생각하나

(17) 키프레이즈: 국내방송 모니터링하기
(토픽: 하우스컨트롤, 요구 유형: 강한 요구)
(17-1) 해외 말고 국내 방송부터 모니터링해
(17-2) 해외는 됐고 국내 방송부터 모니터링하는게 어때?
(17-3) 국내 방송 모니터링해 해외에 시간 쓰지 말고
(17-4) 국내 방송 모니터링이 아무래도 해외 모니터링보다는 중요하지
(17-5) 국내 방송 모니터링이 먼저야 해외 말고
(17-6) 해외는 이제 됐으니까 국내 방송 모니터링해
(17-7) 국내 방송 보지만 말고 모니터링도 좀 해
(17-8) 자꾸 보고만 있네 국내 방송 모니터링도 좀 하라니까
(17-9) 해외 방송 운운하지 말고 국내 방송부터 제대로 모니터링 해봐
(17-10) 이제 해외 그만하고 국내 방송 좀 모니터링해

우리는 문장 유형 간의 비율을 어느 정도 의미있게 맞춤으로써, 쉽게 일상 생활에서 사용되지만 코퍼스에서는 많이 발견되지 않았던 발화들의 특징이 학습과정에 잘 반영될 수 있도록 하였다. 또한, 우리는 가장 해당 방법론이 활발하게 활용될 수 있을 설명의문문에 대해 발화의 수를 증대함으로써, 설명의문문만 따로 떼어 훈련할 때에도 양적인 부족함이 해소되도록 하였다(12K → 20K). 추가적으로, 문장들의 주제가 속하는 도메인(토픽)을 고려하면, 새로 증강된 데이터는 토픽의 분류를 학습하는 데에도 사용할 수 있다. 메일, 스케쥴, 하우스컨트롤, 날씨, 그리고 자유 주제 각각 2,500, 2,500, 2,500, 2,500, 10,000개의 발화가 있으며, 해당 토픽 내에는 선택의문문, 설명의문문, 금지, 강한 요구가 같은 비율로 포함되어 있다.

의도	발화 유형	기존	증강	총계
질문	판정의문문	5,715	-	**5,715**
	선택의문문	229	4,000	**4,229**
	설명의문문	11,988	8,000	**19,988**
요구	금지	478	4,000	**4,478**
	일반 요구	12,302	-	**12,302**
	강한 요구	125	4,000	**4,125**
	총계	30,837	20,000	**50,837**

표 2. 기존 코퍼스의 어노테이션, 증강된 데이터셋, 그리고 이를 반영한 전체 데이터의 구성

마지막으로, 같은 담화 성분으로부터 생성된 문장들은 모두 같은 핵심 내용을 갖는다는 점을 고려하면 서로가 서로에게 패러프레이즈가 되기 때문에, 이를 활용하여 패러프레이즈 데이터셋을 구축할 수 있다. 중복을 방지하여 프레이즈 하나당 열 개의 순서쌍만 취한다고 해도 총 20K 사이즈의 데이터셋을 얻을 수 있으며, 이를 위한 추가적인 작업을 진행 중이다. 현재까지 구축한 초기 데이터셋(3장) 및 증강된 데이터셋(4장) 은 모두 온라인에서 확인할 수 있다. github.com/warnikchow/sae4k

5. 결론

본 연구의 의의는 기존에 구축이나 공개가 잘 되지 않은 문장 단위의 요약 및 패러프레이징에 대해 제작 및 증강 방법론을 확립하고 실제로 이를 시행하여 공개한다는 데에 있다. 본 논문에서는 초기 데이터셋의 구축, 증강 및 패러프레이즈 데이터셋 확보에 대한 어플리케이션만 제시하였으나, 해당 개념을 이용한 자동 질문/요구 생성, 문장 유사도 판별 등의 구현 역시 가능할 것으로 생각된다. 또한, 우리는 추후의 연구에서 데이터셋을 활용한 비정형 질문/요구 키프레이즈 자동 추출에 대해 다룰 예정이며, 이를 다른 언어로도 구현 가능함을 보임으로써 보다 일반적인 코퍼스 생성 및 시스템 구축 방법론으로 확장할 계획을 가지고 있다. 공개한 데이터셋을 통해 한국어를 대상으로 자동 요약, 키프레이즈 추출 및 패러프레이징에 대한 연구가 활성화되기를 희망한다.

감사의 말

본 연구는 2019년도 산업통상자원부 및 산업기술평가관리원(KEIT) 연구비 지원(10076583)이 있었기에 가능했습니다. 또한, 연구의 개념적인 유용성을 확인해 주신 이규환, 정지오, Reinald Kim Amplayo님과 데이터 증강에 도움을 주신 고은아, 기경서, 김상현, 류기민, 이동호, 이윤경, 정민화, 그리고 정예슬 님에게 이 자리를 빌어 감사의 말씀을 전합니다.

참고문헌

[1] Victor Zhong, Caiming Xiong, and Richard Socher, “Seq2SQL: Generating Structured Queries from Natural Language Using Reinforcement Learning”, arXiv preprint arXiv:1709.00103., 2017.

[2] 조원익, 김남수, “담화성분 기반의 한국어 화행 분류를 통한 텍스트 의도 파악의 모호성 해소: 전산언어학적 접근”, 담화와 인지, 제26권, 제3호, pp. 227-247, 2019.

[3] Won Ik Cho, Hyeon Seung Lee, Ji Won Yoon, Seok Min Kim, and Nam Soo Kim, “Speech Intention Understanding in a Head-final Language: A Disambiguation Utilizing Intonation-dependency,” arXiv preprint arXiv:1811.04231, 2018.

[4] Paul Portner, “The semantics of imperatives within a theory of clause types”, In: Semantics and Linguistic Theory, Vol. 14, pp. 235-252, 2004.

[5] 설진석, sci-news-sum-kr-50, Github Repository https://github.com/theeluwin/sci-news-sum-kr-50 (최종확인 2019.09.09)

[6] 조태민, “LDA 모델을 이용한 잠재 키워드 추출”, 한국지능시스템학회 논문지, 제25권, 제2호, pp.180-185, 2015.

[7] 전영실, “특허 문서 텍스트로부터의 기술 트렌드 탐지를 위한 언어 모델 및 단서 기반 기계학습 방법”, 정보과학회논문지: 소프트웨어 및 응용, 제36권, 제5호, pp. 420-429, 2009.

[8] 김종명, “어휘 빈도를 이용한 패러프레이즈 문장의 대중성 측정”, 한국정보과학회 학술발표논문집, pp. 981-983, 2015.

[9] Hancheol Park, Kyo-Joong Oh, Ho-Jin Choi, and Gahgene Gweon, “Constructing a paraphrase database for agglutinative languages”, Data & Knowledge Engineering, 2017.