KorSciQA: 한국어 논문의 기계독해 데이터셋

함영균[1O], 정용빈[1], 정희석[2], 황혜경[2], 최기선[1]

한국과학기술원[1], 한국과학기술정보연구원[2]

hahmyg@kaist.ac.kr, kuonom@kaist.ac.kr, hsjeong@kisti.re.kr, hkhwang@kisti.re.kr, kschoi@kaist.ac.kr

KorSciQA: A Dataset for Machine Comprehension of Korean Scientific Paper

Younggyun Hahm[1O], Youngbin Jeong[1], Heeseok Jeong[2], Hyekyong Hwang[2], Key-Sun Choi[1]

KAIST[1], KISTI[2]

요 약

본 논문에서는 한국어로 쓰여진 과학기술 논문에 대한 기계독해 과제(일명 KorSciQA)를 제안하고자 하며, 그와 수반하는 데이터 구축 및 평가를 보고한다. 다양한 제약조건이 부가된 크라우드소싱 디자인을 통하여, 498개의 논문 초록에 대해 일관성 있는 품질의 2,490개의 질의응답으로 구성된 기계독해 데이터셋을 구축하였다. 이 데이터셋은 어느 논문에서나 나타나는 논박 요소들인 논의하는 문제, 푸는 방법, 관련 데이터, 모델 등과 밀접한 질문으로 구성되고, 각 논박 요소의 의미, 목적, 이유 파악 및 다양한 추론을 하여 답을 할 수 있는 것이다. 구축된 KorSciQA 데이터셋은 실험을 통하여 기존의 기계독해 모델의 독해력으로는 풀기 어려운 도전과제로 평가되었다.

주제어: 자연언어처리, 기계독해, 질의응답

1. 서론

기계독해(Machine reading comprehension)란 기계가 텍스트를 읽고 이해하는 능력을 평가하는 자연언어처리의 주요 과제(task)이다. 최근의 연구들은 기계독해 과제를 텍스트에 대한 질문들에 대해 답변하는 질의응답의 형식으로 다루고 있다[1][2]. 대표적인 기계독해 과제인 SQuAD는 위키피디아 텍스트에 대한 질의응답 데이터로서 공개 과제(shared task)화 되었다. 최근에는 인공지능에 의한 기계독해 능력이 사람의 독해 능력보다도 높은 결과를 보이고 있다[3]. 한국어 텍스트에 대한 기계독해를 평가하기위한 데이터 또한 구축되어 공개 및 평가된 바 있다[4][5].

한편 학술 논문들의 양이 급속도로 증가함에 따라, 학술 논문에 대한 분석과 이해 역시도 자연언어처리 분야에서 주목받고 있다. 이러한 관심에 따라 최근에는 학술 논문 분석을 위한 다양한 공개 과제들이 활성화 되고 있는 추세에 있다[6][7][8][9][10]. 그러나 학술 논문의 경우, 위키피디아와 같은 일반적인 도메인의 웹 코퍼스와 달리 해당 학술 도메인에 대한 지식이 없는 비전공자의 경우 쉽게 이해하기 어려운 내용을 담고 있다. 이러한 이유로 학술 논문에 관련된 연구는 기존의 자연언어처리 분야에 비해 도전적인 과제로 여겨지며, 특히 학술 논문에 대한 기계독해 연구는 국내외 연구에서 거의 이루어지지 않고 있다. 특히 한국어 논문에 대한 자연언어처리 연구는 영어권 연구에 비해 데이터의 부재 등으로 인하여 매우 미흡한 것이 사실이다.

본 논문은 한국어 논문의 기계독해 과제(KorSciQA)를 제안하고, 이를 위한 데이터셋 구축 방법과 결과를 논의한다. KorSciQA는 주어진 한국어 논문의 초록에 대해, 기계가 주요 내용을 이해하였는가를 질의응답의 형식으로 평가하는 과제이다. KorSciQA의 예시는 그림 1과 같다.

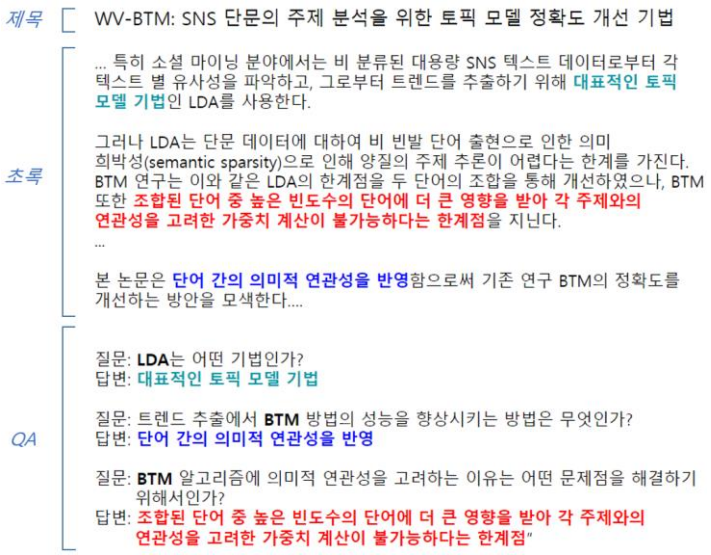

그림 1. KorSciQA의 예시. 논문의 제목과 초록, 그리고 논문과 관련된 질의응답쌍들로 구성되어 있다. 논문에서 다루는 논박 요소들(예: LDA, BTM)에 관련된 질문들로 구성되고, 각 답변은 논문 초록의 특정 범위 텍스트이다.

KorSciQA는 주어진 한국어 논문의 초록에 대해, 기계가 주요 내용을 이해하였는가를 질의응답의 형식으로 평가하는 과제이다. KorSciQA는 주어진 질문에 대해 논문 초록 텍스트의 범위(span)을 내어주는 과제라는 점에서 기존의 기계독해 과제들의 형식과 유사하다. 그러나 논문에서 다루는 논박 요소들(논의하는 문제, 푸는 방법, 관련 데이터, 모델, 결과 등)에 관련된 질문들로 구성되어 있다는 점, 그리고 문장에 대한 이해 능력만을 평가하는 것이 아니라 논문 초록 텍스트의 전반적인 구성과 내용을 이해하였는가를 평가한다는 점에서 기존의 기계독해 과제들보다 고차원의 과제이다. 본 논문에서 언급하고 있는 과제와 함께 구축된 데이터는 과제의 타당성 조사를 위한 원형으로서의 역할을 한다.

2장에서는 학술 논문에 관련된 자연언어처리 연구들을 소개하고, 본 논문이 제안하는 연구의 위치를 기술한다. 3장에서는 본 논문이 제안하는 한국어 논문의 기계독해 과제를 정의하고 발생할 수 있는 이슈를 소개한다. 4장에서는 이를 위한 데이터셋 구축 방법을 논의하고, 그 결과를 5장에서 여러 유형으로 나누어 논의한다. 결론은 6장에서 기술한다.

2. 관련 연구

학술 논문은 국내의 NDSL[1], DBpia[2], 해외의 ArXiv, PubMed, Science.gov와 같은 호스팅 서비스를 통해 웹에 공개되어 있다. 이러한 호스팅 서비스들은 논문 컨텐츠를 제공하고 논문의 키워드 기반 검색을 지원한다. 반면 Google Scholar, Semantic Scholar, Microsoft Academic 과 같은 인공지능 기반의 학술 논문 검색 시스템은 논문 컨텐츠를 호스팅하고 있지는 않으나, 논문에 대한 추천이나 요약 등의 서비스를 제공하며 학술 컨텐츠에 대한 활용 가치에 주목하고 있다.

학술 컨텐츠의 가치는 다양한 연구 분야에서 주목받고 있으며, 자연언어처리 연구가 활성화 되도록 유도하는 공개 과제들이 최근 많아지고 있다. BioASQ[6]의 경우, 훈련된 전문 인력에 의한 PubMed 문서의 수작업 분류의 비용을 줄이기 위해 의미적 인덱싱(semantic indexing)에 의한 자동 분류, 그리고 검색 시스템 기반의 질의응답 과제를 2013년부터 매년 진행하고 있다. CL-SciSumm[7] 공개 과제의 경우 논문의 자동 요약의 연구가 진행되었으며, 특히 논문의 인용 네트워크를 활용한 검색 기반의 자동 요약 과제를 진행한 바 있다. 자연언어처리 분야의 대표적 공개 과제인 SemEval[8][9]에서는 논문에서 등장하는 주요 어휘들을 인식하고, 이 주요 어휘들 사이의 관계를 추출하는 공개 과제를 진행하였다. SCiDTB[10]의 경우에는 학술 논문의 담화 구조를 바탕으로 논의의 구성요소 및 논의 구조를 분석하고 이를 자동으로 추출하기 위한 연구를 수행하였다.

한편, 기계독해를 평가하기 위한 과제는 SQuAD[1]가 대표적이다. SQuAD는 일반적인 도메인인 위키피디아 텍스트를 사용하여 질의응답 데이터를 구축하였으나, 대부분의 질문이 하나의 문장에서 답변할 수 있는 등 난이도가 쉬워 기계독해를 충분히 평가하지 못한다고 지적되기도 한다[11]. 이러한 단점을 보완하기 위해, MS-MARCO[2]의 경우는 인위적 질문이 아닌 실제 존재하는 질문들을 사용하여 상대적으로 높은 난이도의 과제가 제안되었다. 이와 유사하게 AI2 Reasoning[12]의 경우에는 배경지식을 사용한 추론을 바탕으로 한 객관식 질문으로 구성된 질의응답 과제를 제안하기도 하였다.

상기의 학술 논문을 위한 공개 과제 및 연구들은 모두 영어 논문 컨텐츠를 대상으로 연구가 수행되었으며, 한국어 논문을 대상으로 한 공개 과제나 연구는 데이터의 부재 등으로 상대적으로 미흡하다. 특히 기계독해의 관점에서, 일반적인 도메인인 위키피디아나 초등학교 수준의 과학적 상식 등에 대한 연구는 활발히 진행되고 있지만 학술 논문을 대상으로 한 연구는 아직 활성화된 바 없다고 보여진다.

본 논문이 제안한 KorSciQA의 연구 위치는 표 1과 같다.

표 1. 기존 연구와 비교한 KorSciQA의 연구 위치

공개 과제	도메인	과제 유형	언어
BioASQ	학술 논문	검색형 QA	영어
CL-SciSumm	학술 논문	요약	영어
SemEval-SciIE	학술 논문	정보 추출	영어
SciDTB	학술 논문	논의 구조	영어
SQuAD	위키피디아	기계독해	영어
MS-Marco	웹 문서	기계독해	영어
AI2 Reasoning	과학 문서	추론형 QA	영어
KorSciQA	**학술 논문**	**기계독해**	**한국어**

본 논문이 제안한 KorSciQA는 한국어 학술 논문을 대상으로 하는 연구라는 점 이외에도, 학술 논문에 대한 기계독해 라는 도전적 과제를 다룬다는 점에서 상기의 연구들과 차별적이며 또한 보다 고차원의 과제라고 볼 수 있다.

3. 문제정의

본 논문을 통해 제안된 KorSciQA는 한국어 논문의 기계독해를 위한 질의응답 과제이다. 주어진 한국어 논문의 초록 텍스트에 대해, 관련된 질문에 대해 답변을 내어주는 과제이다. 해당 질의응답 데이터셋은 SQuAD의 기계독해 패러다임을 따라 질문에 대하여 주어진 논문 초록 텍스트에서 답변에 해당하는 범위를 선택하는 형식으로 정의한다.

SQuAD의 경우는 일반적 도메인의 위키피디아 텍스트를

[1] https://www.ndsl.kr/

[2] https://www.dbpia.co.kr/

다루었기 때문에 일반 작업자에 의한 크라우드소싱을 통하여 14만개 이상의 질의응답으로 구성된 대규모의 데이터셋을 구축할 수 있었다. 그러나 논문 초록의 경우에는 크라우드소싱으로 데이터를 구축할 경우 다음에 기술된 어려움이 있다.

첫 째로, 크라우드소싱에 의해 구축된 질문이 논문의 내용과 관련 있는지를 보증할 수 없다는 문제가 있다. 예를 들어, SQuAD의 질문들은 대부분 문장의 일부 단어나 구를 찾는 문제로 구성되어 있어 전체 문서를 보지 않고도 하나의 문장 내에서 답변이 가능한 질문들이 많다. 그런데 학술 논문의 텍스트는 논문에서 다루고자 하는 논박 요소들을 중심으로 기술되어 있다는 특징이 있다. 이러한 논박 요소들과 무관한 질문들을 구축할 경우, 기존의 기계독해 과제들과 유사하게 문장에 대해 이해하였는가를 평가하는 데이터가 될 우려가 있다. 예를 들어, "*… LDA는 단문 데이터에 대하여 비 빈발 단어의 출현으로 인한 의미 희박성으로 인해 양질의 주제 추론이 어렵다 …*" 와 같은 논문의 일부 문장을 생각해 보자. 이 때, LDA와 같은 논박 요소에 관련된 질문이 아닌 "주제 추론의 난이도는 어떠한가?" 와 같은 단순한 질문이 크라우드소싱에 의해 구축되는 것을 방지할 필요가 있다.

두 번째로는, 크라우드소싱에 의해 구축된 질문들의 난이도가 균형 있게 구축되는 것을 보증할 수 어렵다는 문제가 있다. 크라우드소싱에 의한 질의응답 데이터 구축은 짧은 시간에 인위적으로 작성된 쉬운 문제들로 구축되는 경향이 있는데 논문 초록의 경우 크라우드소싱 작업자들이 이해하기 어렵다는 이유로 더욱 단순하고 쉬운 질문들만으로 구성될 가능성이 높아 이를 방지할 필요가 있다.

본 논문에서는 KorSciQA 데이터셋 구축에 있어서 상기의 두 가지 이슈를 해결하기 위한 데이터 구축 방법을 제안한다. NDSL의 컴퓨터공학 관련 학회 및 학술지의 2018년 논문들 중 한국어 초록이 존재하는 임의의 데이터(논문 498편) 각각에 대해 5개의 질의응답 데이터를 크라우드소싱으로 구축하였다. 그리고 구축된 데이터에 대한 정성적 분석과 함께 기계학습 성능을 평가하였다.

4. KorSciQA 데이터셋 구축 방법

KorSciQA 데이터셋은 크라우드소싱에 의해 구축하였다. 주어진 한국어 논문 초록에 대해 크라우드 작업자는 질문을 직접 타이핑하여 작성하고, 해당 질문에 대한 답을 논문 초록 텍스트에서 드래그하여 범위를 선택하는 방식으로 구축하였다. 해당 크라우드소싱 환경의 UI는 아래 그림 2와 같다.

KorSciQA 데이터셋을 크라우드소싱으로 구축하면서, 3장에서 논의한 두 가지 이슈인 1) 논문과 관련있는 질의응답 데이터셋 구축과 2) 일관성있는 난이도의 데이터셋 구축의 문제를 다루기 위해 본 논문에서는 다음의 방법을 사용하였다.

그림 2. KorSciQA 데이터셋 구축을 위한 크라우드소싱 UI의 예. 화면 좌측에는 논문 제목과 초록을 제공하고, 크라우드 작업자는 화면 우측에서 질문을 직접 작성하고, 이에 대한 답변을 논문 초록에서 드래그 하여 질의응답 쌍을 구축한다.

먼저, 질문 작성은 다음의 두 단계를 거치도록 설계하였다.

- **1단계**: 주어진 논문 초록에서 논박 요소를 지칭하는 주요 어휘 선택 (논의하는 문제, 푸는 방법, 관련 데이터, 모델, 결과 등)
- **2단계**: 주요 어휘가 포함된 질문 작성

각각의 작업자는 모든 초록에서 5개의 질문을 작성하도록 하였는데, 질문들은 다음의 제약조건을 통해 세 가지의 난이도로 구성되도록 하였다.

- **Easy**: 선택된 주요 어휘에 대한 부가 설명 질문
- **Normal**: 선택된 주요 어휘가 포함된 문장 내에서 답변이 가능한 질문
- **Challenge**: 선택된 주요 어휘가 포함된 문장 밖에서 답변이 가능한 질문

위와 같은 두 단계와 제약조건을 통해 기대한 결과는 다음과 같다.

먼저, 논문에서 다루는 논박 요소들(논문이 다루는 문제와 데이터, 이를 해결하기 위해 사용되는 기존의 모델, 혹은 논문에서 제안된 모델 등)에 대해서 질문이 작성되는 것을 기대하였다. 이를 위하여 작업자들은 먼저 논문에서 논박 요소를 지칭하는 주요 어휘(keyphrase)를 드래그로 선택하고, 이 어휘가 포함된 질문을 작성하도록 안내하였다. 이때 주요 어휘들은 논문에서 다루는 문제, 방법, 데이터, 모델, 결과 등을 지칭하는 어휘이다. 이를 통해 논문에서 논의되는 주제에 대해 관련 있는 질문들이 구축되는 것을 기대하였다.

또한 각각의 질문들에 대해 제약조건을 줌으로서, 한 문장만으로도 답변이 가능한 쉽게 질문할 수 있는 내용(easy, normal) 이외에도, 논문 초록 전체를 읽어야만 답변이 가능한 질문(challenge)들을 작성하도록 유도하여 인위적인 쉬운 질문들이 아닌, 논문 초록의 전반적인 내

용을 기계가 이해하였는가를 평가하는 기계독해 질의응답 데이터셋을 구축하도록 하였다.

각각의 난이도에 대하여, easy와 normal은 2개씩, 그리고 challenge에 대해서는 1개씩을 작성하도록 하였다. 이렇게 구성한 이유는, 크라우드소싱 파일럿테스트를 진행해 본 결과 challenge에 해당하는 질문을 2개 이상 작성하기 어려운 논문 초록들이 상당수 존재했기 때문이다.

각각의 논문에 대해 한 명의 크라우드소싱 작업자가 5개의 질문을 작성하고, 다른 한 명의 크라우드소싱 작업자가 각각의 질문들이 위의 가이드라인을 충실히 지켰는지 여부를 검수하는 방식으로 진행하였다. 질문을 작성하는 작업자는 하나의 논문에 대해 1,000원이 지급되었고, 검수하는 작업자에게는 1,200원이 지급되었다.

5. KorSciQA 데이터셋 평가

5.1. 질문 유형 분석

4장에서 기술된 방식에 의해, 본 논문에서는 KorSciQA의 초기 데이터로서 498개의 논문에 대해 2,490개의 질의응답쌍을 구축할 수 있었다. 본 장에서는 구축된 KorSciQA 데이터에서 나타나는 대표적인 현상을 예시와 함께 보이고, 해당 데이터 전반에 대하여 개괄한다.

전체 질문 2,490개의 질문 중, 1,262개의 질문들은 '~은 무엇인가?' 와 같은 형태로 작성되었고, 나머지 질문들은 의문사를 제외한 형태(예: "~에 영향을 주는 호흡기 질환은?", "~방법의 목적은?", "~가 개발된 이유는?")로 작성되거나, 다양한 형식의 질문들(예: "~할 수 있는 예시를 나열하시오", "~시스템의 한계를 기술하시오")로 작성되었다. 각 질문들은 논문의 주요 어휘인 문제나 모델에 대한 목적, 구현, 역할, 적용, 실험의 결과 등의 키워드를 포함하여 구축되었다.

유형 1. 주요 어휘의 구체적 의미를 묻는 질문.

학술 논문들은 주요 어휘 및 전문용어에 대한 정의를 기술하거나, 각 논문에서 사용되는 구체적인 의미를 상술하고 있다. 유형 1의 경우는 특히 easy 난이도의 질문에서 구축되기를 기대한 질문들로, 주요 어휘들이 논문에서 사용된 맥락의 의미를 이해하였는가를 평가하는 질문이라고 볼 수 있다. 이에 대한 예시는 그림 3과 같다

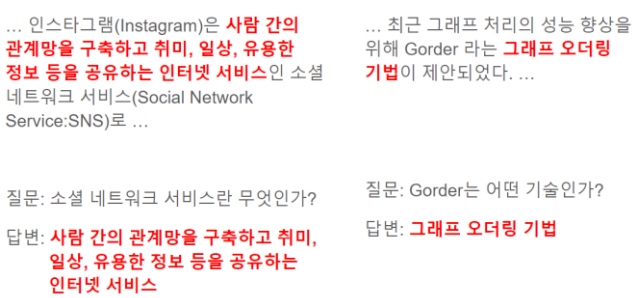

그림 3. 주요 어휘에 대한 설명을 요구하는 질문의 예

유형 2. 문제 및 방법의 목적이나 이유를 묻는 질문.

학술 논문에는 각 논문이 다루는 문제나 논문에 의해 제안된 혹은 기존의 방법들이 기술되어 있다. 그리고 그 문제나 방법들이 사용된 목적이나 이유 등이 기술되어 있다. KorSciQA의 많은 질문들이 다양한 방식으로 기술되어 있었지만, 목적에 대한 질문이 상당수를 차지하고 있었다. 이에 대한 예시는 그림 4와 같다.

유형 2에 해당하는 질문들은 다양한 형식으로 작성되어 있다. 예를 들어, "~의 목적은 무엇인가?" 와 같은 직접적인 질문들 이외에도, "~은 어떤 용도로 활용될 수 있는가?", "~가 필요한 이유는 무엇인가?" 등의 다양한 표현으로 작성되었다. 이러한 질문들은 난이도에 상관없이 고르게 분포하고 있는 것으로 보였다.

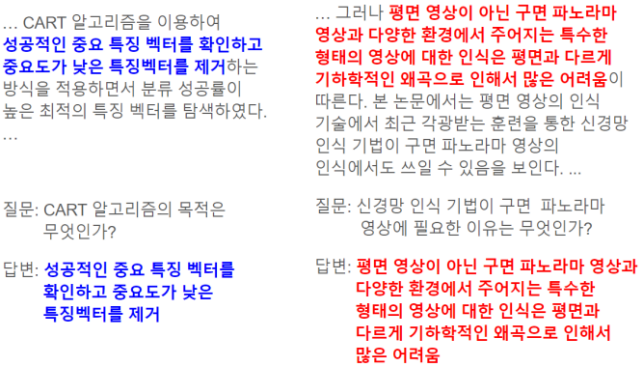

그림 4. 문제 및 알고리즘의 목적에 대한 질문의 예

유형 3. 주요 어휘에 대한 세부 사항을 묻는 질문.

유형 3의 질문들은 주요 어휘에 대한 세부 사항에 대한 질문들이다. 특히 각 논문에서 사용된 주요 어휘인 모델이나 방법론에 대한 구체적 내용을 질문하는 경우이다. 이러한 구체적 내용은 해당 주요 어휘가 포함된 문장 내부에서 기술되는 경우도 있으나, 그림 5와 같이 논문 초록의 전체를 이해하였을 때 답변이 가능한 경우도 상당수 존재하였다. 이는 단순히 문장에 대한 의미를 이해하는 방법이 아닌 논문의 담화구조분석이나 상호 참조분석 등의 접근법이 필요할 것으로 보인다.

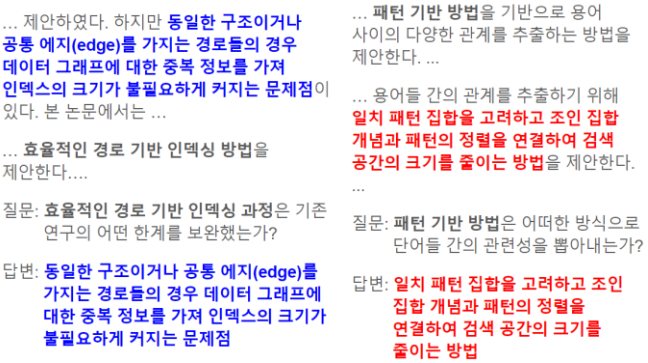

그림 5. 주요 어휘에 대한 세부 사항 질문의 예.

유형 4. 추론이 필요한 질문.

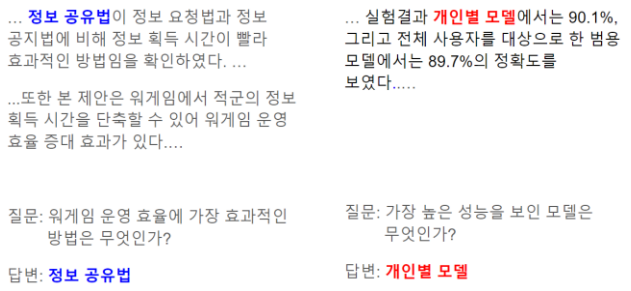

그림 6. 추론이 필요한 질문의 예

대부분의 질문들은 상기의 유형 중에 속하지만, 몇몇 질문들은 상당한 수준의 추론 기법이 요구되는 것으로 보이는 경우가 있었다.

예를 들어, 그림 6의 좌측 질문의 경우, 질문의 내용이 명시적으로 기술되지는 않았지만 맥락에 의한 추론이 필요하다. 해당 질문에 답변하기 위해서는, "*정보 획득 시간*" 이 빠르다는 것이 "*워게임*"에서 효과적이라는 점이 파악되어야 한다. 또한 맥락에 의한 추론 이외에도, 산술 추론이 필요한 질문도 존재하였다. 그림 6의 우측 질문의 경우, 개인별 모델과 범용 모델의 성능 간의 수치를 비교하여, "*가장 높은 성능*"을 보인 모델을 선택해야 하는 질문이다.

유형 5. 기타 유형들.

상기의 유형들 이외에도 다양한 형태의 질문들이 작성되었다. 예를 들어, 그림 7의 좌측 예시는 특정 조건에 해당하는 어휘들을 나열하는 질문이다. 해당 질문에 답하기 위해서는 "*오픈소스*"에 해당하는 키워드들이 무엇인지를 파악하여야 한다. 그림 7의 우측의 예시는 여러 성능 결과 중 특정 조건에 해당하는 값을 선택해야 하는 질문이다. 해당 질문에 답하기 위해서는 문장의 구조에서 특정 조건에 해당하는 어휘(혹은 숫자)가 무엇인지를 파악해야 한다.

그림 7. 특정 조건에 해당하는 키워드를 묻는 질문의 예

5.2. 기계독해 성능 평가

본 장에서는 다음의 두 가지 실험을 수행하였다. 먼저, 기존의 일반적 도메인의 기계독해 데이터셋인 KorQuAD를 학습하고, KorSciQA에 대해 독해 능력을 평가해 보았다. 이는 새롭게 구축된 KorSciQA가 보다 난이도가 높고 어려운 과제임을 보이기 위함이다. 두 번째로, KorSciQA를 임의의 9:1로 학습과 평가 데이터로 나눈 뒤 실험해 보았다. 사용한 모델은 BERT에서 공개한 multilingual이고, 학습은 BERT 논문의 fine-tuning 방식을 사용하였다. 학습 횟수는 3번, 학습률은 5e-5, 배치사이즈는 6이다.

첫 째로, 표 2는 KorQuAD를 학습데이터로 사용하고, 이로부터 KorSciQA에 대한 기계독해 능력을 평가한 결과이다. 표 2에서 볼 수 있듯, 해당 모델은 KorQuAD 평가 데이터에 대해서는 89.16%의 높은 F1성능을 보였으나, 학술 논문에 대한 기계독해인 KorSciQA의 경우 45.15%의 낮은 성능을 보였다. 특히 KorSciQA는 4장에서 논의된 바와 같이 세 개의 난이도로 구성되어 있는데, 쉬운 문제인 경우 56.34%의 성능을 보였으나, 어려운 문제의 경우는 24.88%로 낮은 성능을 보였다. 3장에서 논의한 바와 같이, 본 논문은 KorSciQA의 데이터셋이 쉬운 문제들로만 구성되지 않도록 균형있는 데이터셋을 구축하고자 하였고, 실험 결과를 통해 질문의 난이도가 의도된 바와 같이 균형 있고 또한 어려운 문제들로 구축되었음을 확인할 수 있었다. 특히 Challenge의 경우, 4.1장에서 기술된 바와 같이, 하나의 문장이 아닌 문서 전체적인 의미를 독해하여야만 답변할 수 있는 문제들로 구성되어 있다는 점에서 다른 문제들보다 난이도가 높다.

표 2. KorQuAD를 학습데이터로 사용한 기계독해 성능

평가 데이터		**F1**
KorQuAD (test)		89.16
KorSciQA	All	45.15
	Easy	56.34
	Normal	44.09
	Challenge	24.88

표 3. KorSciQA 기계독해 성능

학습데이터	**F1**
KorQuAD (train)	45.62
KorSciQA	**66.52**

두 번째로, 표 3은 KorSciQA의 90%를 학습데이터로 사용하여 나머지 10%를 평가데이터로 사용한 기계독해 성능 결과이다. 비교를 위하여 KorQuAD를 학습데이터로 사용한 모델로 KorSciQA의 평가데이터에 대한 성능을 비교하였다. KorSciQA를 학습데이터로 사용하였을 경우가 KorQuAD를 학습데이터로 사용하였을 경우보다 약 21.37% 정도의 F1 성능이 향상되었음을 보았다.

그러나, KorSciQA가 컴퓨터 공학 학술 도메인에 한정되어 있고, KorQuAD가 위키피디아라는 일반적 도메인을 다룬다는 점에서 도메인의 차이에서 기인한 결과일 수

있다는 점도 고려되어야 한다. 이는 향후 연구의 영역으로 남긴다. 그럼에도 불구하고, 학술 논문에 대한 기계독해는 기존의 기계독해 모델의 독해 능력으로는 한계가 명확한 것으로 보인다는 점에서 의의를 찾을 수 있다.

5.3. 논의

4장에서 논의한 바와 같이, 기존의 기계독해 데이터셋과 달리 KorSciQA의 데이터셋은 다음의 두 가지 측면을 고려하여 구축하였다: 1) 논문에 대한 관련성 높은 질의응답 데이터셋 구축, 그리고 2) 일관성 있는 난이도의 데이터셋 구축.

5.1장에서 기술된 바와 같이, 모든 질문들은 논문에서 중점적으로 다루는 논박 요소들(문제, 방법, 데이터, 모델, 결과 등)에 대한 다양한 유형의 질문들(논박 요소들에 대한 의미, 목적, 이유, 세부사항 등)로 구축되었다. 또한 난이도 디자인을 통해 쉬운 문제들은 물론, 논문 초록 전체를 읽어야만 답변 할 수 있는 challenge 난이도의 문제를 포함하여 의미적 추론, 산술적 추론, 조건 추론 등의 다양한 난이도의 데이터셋이 구축되었다.

이렇게 구축된 KorSciQA는 5.2장의 실험을 통하여 기존의 일반적인 기계독해 모델의 독해 능력으로는 풀지 못하는 과제라고 평가되었다. 특히 문서 전반에 대한 이해를 필요로 하는 어려운 질문(challenge 난이도)에 대해서는 새로운 접근 방법이 필요한 것으로 보인다. 양질의 KorSciQA 데이터셋 구축은 학술 논문을 위한 기계독해의 성능을 향상시킬 수 있는 방안으로 모색된다. 본 논문에서 사용된 KorSciQA는 기존의 기계독해 데이터셋 [1][5]과 비교하여 상대적으로 적어 충분한 성능 평가가 이루어지지 못하였다는 점에 본 연구의 한계가 있다.

6. 결론

본 논문은 한국어 논문의 기계독해 과제인 KorSciQA를 새롭게 제안하고, 이를 위한 데이터셋을 구축하는 방법론을 제안하고 분석을 수행하였다. KorSciQA는 기존의 기계독해 과제들과 달리, 난이도가 높은 학술 논문에 대한 기계독해를 목적으로 하며, 또한 기존의 학술 논문 관련 연구들에서 다루지 않았던 새로운 과제이다. 본 논문에서 제안된 방법에 의해 구축한 질의응답셋은 논문에서 다루는 중요한 논박 요소들에 관련된 질문들로 충실히 작성되었고, 또한 다양한 유형들과 난이도로 구성되었다. 이를 통해 498개 논문 초록에 대해 총 2,490개의 질의응답셋을 구축할 수 있었다. 본 데이터셋은 기존의 기계독해 모델의 독해력으로는 풀지 못하는 어려운 과제이다. 향후 본 논문에 의해 제안된 데이터 구축 방법론을 적용하여 기계학습에 사용될 수 있는 충분한 양의 데이터를 구축하고, 이를 공개 과제화 할 계획이다.

사사

본 연구는 2019년도 한국과학기술정보연구원(KISTI) 주요사업 과제로 수행한 것입니다.

참고문헌

[1] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, “SQuAD: 100,000+ Questions for Machine Comprehension of Text”, Proceedings of EMNLP, 2016.

[2] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, “A human generated machine reading comprehension dataset”, In Workshop on Cognitive Computing at NIPS, 2016.

[3] J. Devlin, M. Chang, K. Lee, K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, ArXiv preprint arXiv:1810.04805, 2018.

[4] C. Park, C. Lee, L. Hong, Y. Hwang, T. Yoo, J. Jang, Y. Hong, K. Bae, H. Kim, “S2-Net: Machine reading comprehension with SRU-based self-matching networks”, ETRI Journal, Volume 41, Issue 3, 2019.

[5] 임승영, 김명지, 이주열, “KorQuAD: 기계독해를 위한 한국어 질의응답 데이터셋”, 한국정보과학회 학술발표논문집, 539-541, 2019.

[6] A. Nentidis, A. Krithara, K. Bougiatiotis, G. Paliouras, I. Kakadiaris, “Results of the sixth edition of the BioASQ Challenge”, Proceedings of the 2018 EMNLP Workshop BioASQ: Large-scale Biomedical Semantic Indexing and Question Answering, 2018.

[7] M. K. Chandrasekaran, M. Yasunaga, D. Radev, D. Freitag, M. Kan, “Overview and Results: CL-SciSumm Shared Task 2019”, Proceedings of BIRNDL 2019 at SIGIR 2019, 2019.

[8] I. Augenstein, M. Das, S. Riedel, L. Vikraman, A. McCallum, “SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications”, ArXiv preprint arXiv:1704.02853, 2017.

[9] K. Gábor, D. Buscaldi, A. Schumann, B. QasemiZadeh, H. Zargayouna, T. Charnois, “Semeval-2018 Task 7: Semantic relation extraction and classification in scientific papers”, Proceedings of The 12th International Workshop on Semantic Evaluation, 2018.

[10] P. Accuosto, H. Saggion, “Discourse-Driven Argument Mining in Scientific Abstracts”, International Conference on Applications of Natural Language to Information Systems, 2019.

[11] P. Rajpurkar, R. Jia, P. Liang, “Know What You Don't Know: Unanswerable Questions for SQuAD”, Proceedings of ACL, 2018.

[12] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, O. Tafjord, “Think you have solved question answering? try arc, the ai2 reasoning challenge”, ArXiv preprint arXiv:1803.05457, 2018.