

‘회의록 Question Answering’ 데이터셋 구축 가이드라인

본 가이드라인은 회의록 Question Answering(이하 QA) 데이터셋 구축 가이드라인에 대해 다루고 있습니다. Annotation에 참여하기 전에 반드시 숙지해주시기 바랍니다.

1. 데이터 구축 가이드

1.1. 데이터 구축 개요

Question Answering 태스크 자연어 처리 분야에서 인간이 제기하는 질문에 자동으로 대답하는 시스템을 구축하는 것과 관련이 있습니다. 이번 데이터셋 구축 작업은 회의록 데이터셋을 활용한 기계독해 모델 생성을 위한 지문(Passage) - 질문(Question) - 답변(Answer) 데이터셋을 구축한다.

1.2. 문제 정의

1.2.1. 임무정의

정제 작업을 거친 국회 회의록을 대상으로 작업자들이 **지문(Passage) - 질문(Question) - 답변(Answer)**으로 구성된 1500 개의 기계 독해용 질의응답 데이터셋을 구축한다. 이 중 기계독해 모델의 성능 향상을 위해 질문 중 답이 없는 경우를 포함한다.

공정명	건수(비율)	데이터셋 구성
정답이 있는 질문	1350 건(90%)	지문-질문-정답 데이터셋으로 작성
정답이 없는 질문	75~150 건(5~10%)	지문-질문-빈칸 데이터셋으로 작성

1-3. 획득-정제 절차

원시 데이터는 문서 요약에 위해 가공된 의회 데이터를 사용하기로 했다. ‘완주’, ‘음성’, ‘청주’ 의회 데이터 중 구축 작업 소요시간과 데이터 주제 간 통일성을 위해 크기를 고려해 ‘완주’의 회의 데이터 165 건을 안전별로 나눠 861 건의 지문을 작업 데이터로 정했다.

기계 독해 데이터셋은 일정 길이의 텍스트로 이루어진 지문, 지문에 내포된 정보를 묻는 질문, 그에 대한 답변으로 구성된 데이터셋으로, 수집된 원시데이터는 다음과 같은 정제 절차를 거쳐 회의에 대한 정보가 담긴 지문으로 변환되어야 한다.

공정명	세부내용
지문제작	회의록에서 발화자의 정보를 제외한 텍스트로 변환한다. 텍스트는 UTF-8 방식으로 인코딩한다.
지문 정제	변환된 텍스트에서 질의응답 세트 제작에 적합하지 않은 부분을 삭제하고, 회의를 분절해 일정 길이의 회의 지문으로 제작한다. 태깅의 편의를 위해 지문 서두에 회의 제목을 같이 작성한다.
지문 정비	지문의 메타데이터의 이름을 회의 아이디와 지문로 구분하여 원천 데이터를 정비한다.

1.4. 어노테이션/라벨링 절차

공정명	세부내용	산출물
지침서 작성	유형별로 정리된 질문-답변 데이터셋을 기준으로 작업 지침서 작성	QA 데이터 제작 지침서
질문-답변 작성	- 준비된 정제데이터를 작업자별로 분배 - 관리도구를 이용하여 질문 및 답변 작성	워크시트, 입력 데이터 파일
질문-답변 유형화	- 지문의 유형, 질문의 난이도, 답의 유형, 답이 없는 질문 유형 등을 종합적으로 고려하여 입력된 질문과 답변의 수량과 품질이 균등하게 구축	워크시트, 입력 데이터 파일

1.4.1.제작 유형 기획

어노테이션의 질의 데이터는 보통 질문을 육하원칙 의문사에 기반하여 구분하지만, 회의록 특성상 나올 수 있는 질문에 제한이 있을 것이라고 가정해 장소(Where)관련 질문은 제외했다. 그래서 언제(When), 누가(Who), 왜(Why) 등을 위주로 질문을 생성하고자 한다. 또한 회의록 ODQA 라는 Task 의 특성 상 특정 사건을 다루고 있는 회의 자체를 찾고자 할 수 있기 때문에, 회의를 찾는 질문 패턴을 추가했다. 정답 유형을 결정하는 단계에서는 질문한 의문사 패턴에서 단서를 찾아 그에 따른 정답 유형을 분류한다. 데이터셋이 특정 질문 유형에 편중 되지 않도록 다양한 질문 유형으로 구성한다.

질문유형	질문 패턴	답변 유형
When(기간)	<사건>의 기간은?	기간(날짜, 시간)
When(일시)	<사건>는 언제 진행되는가?	일시(날짜, 시간)

Who(능동주, 행동)	〈행동〉을 수행한 사람은 누구인가?	사람(들)
Who(피동주, 역할)	〈역할〉을 맡게된 사람은 누구인가?	사람
Why(행동, 의견)	〈행동〉을 한 이유는 무엇인가?	명사구/기타
What(지표)	〈사건〉의 나온 〈지표〉는 얼마인가?	숫자(금액, 단위)
What(회의)	〈사건〉를 다루고 있는 회의는?	명사구

질문 데이터는 지문의 내용을 활용해 작성하고 **지문의 내용과 동일한 뜻을 가진 비슷한 형태의 문장을 패러프레이즈**해 다양한 질의에 대응하도록 데이터를 구성한다. 수집된 회의록 데이터는 해당 분야의 전문적인 용어가 사용되었기 때문에 전문 용어를 직접적으로 활용한 질문을 생성하는 동시에 전문 용어와 의미가 유사하지만 일상적으로 자주 사용되는 어휘나 어절로 대체하여 질문을 제작한다. 단, 전문 용어를 대체하기 위해 용어의 정의를 긴 어절로 서술하는 것은 지양한다.

지문예시

의사일정 제 3 항 2020 년도 제 2 회 추가경정예산안 제안설명의 건을 상정합니다. 먼저 박성일 군수님은 나오셔서 일괄 제안설명 해주시기 바랍니다."안녕하십니까? 완주군수 박성일입니다. 제안설명에 앞서 의정활동에 노고가 많으신 최등원 의장님을 비롯한 여러 의원님들께 깊은 감사의 말씀을 드립니다. 2020 년도 제 2 회 추경예산안은 코로나 사태의 장기화로 인하여 극도로 위축된 소비시장 회복과 지역경제의 활성화를 위해 모든 군민에게 10 만원씩 추가 지급하는 완주군 2 차 긴급재난지원금을 편성하는 예산입니다. 2020 년 제 2 회 추경예산안 규모는 총 8,089 억원입니다. 일반회계는 7,649 억원, 특별회계는 440 억원으로 지난 1 회 추경과 변동이 없습니다. 금번 추경의 재원은 기존 사업의 구조조정을 통해 마련된 것으로 일반회계 세입의 변동은 없으며, 일반회계 세출 규모는 기존 1 회 추경과 동일하나, 긴급재난지원금 92 억원의 재원 마련과 12 억원의 예비비를 추가 확보하기 위하여 재난 예비비와 내부유보금 67 억원 및 자체 절감액 37 억원을 활용하여 총 예산 범위 내에서 재편성한 내용입니다. 공기업특별회계는 308 억원, 기타 특별회계는 131 억원으로 변동이 없습니다. 이상으로 2020 년 제 2 회 추가경정예산안에 대한 제안설명을 모두 마치겠습니다. 감사합니다. 박성일 군수님 수고하셨습니다. 해당 상임위원회에서는 완주군의회 회의규칙 제 69 조의 규정에 따라 예산안 예비심사 결과를 6 월 3 일까지 의장에게 보고하여 주시기 바랍니다.

질문 유형 1	제작 예시 1
지문 내 어휘 사용	예산안 예비심사 결과는 언제까지 보고 해야 해?
동의어/유사어 사용	예산안 예비심사 결과 에 대해 언제까지 알려줘 야 하지?
상식적으로 통용되는 어휘 사용	예산안 예비심사 결과 는 언제 ?

질문 유형 2	제작 예시 2
지문 내 어휘 사용	완주군 2 차 긴급재난지원금은 얼마씩 지급해?
동의어/유사어 사용	완주군은 2 차 긴급재난지원금 얼마씩 줘?
상식적으로 통용되는 어휘 사용	완주군 2 차 긴급재난지원금은 얼마야?

1.4.2. 어노테이션/라벨링 기준

질문 제작 기준

질문 제작 방법

① 주어진 지문을 보고 지문 당 가능한 1 개 이상 4 개 이하의 최대한 다른 유형의 질문을, 질문유형을 참고하여 생성.

② 질문의 유형을 구분하기 위해, 질문 앞 단에 질문유형 번호와 언더바를 함께 넣어 질문을 생성.

(예시)

▼ AGENDA_2:

0: "의사팀장 수고하셨습니다. 먼저 의사일정 제1항 제207회 완주군의회 임시회 회기 결정의 건을 상정합니다. 제207회 완주군의회 임시회 회기 결정의 건에 대하여는 서남용 의원님 외 4인의 의원님이 발의한대로 8월 26일부터 9월 4일까지 10일간의 회기를 결정하고자 합니다. 의원 여러분 이의 있으십니까?"

1: " ("없습니다,하는 의원 있음) "

2: "이의가 없으므로 가결되었음을 선포합니다."

위와 같은 지문에서도 다음과 같은 QA 데이터를 생성할 수 있음.

- 질문 1: 3_제 207 회 완주군의회 임시회 회기 결정의 건은 누가 제안했나?

- 답안 1: 서남용 의원님 외 4 인의 의원님

- 질문 2: 2_제 207 회 완주군의회 임시회 회기는 언제인가?

- 답안 2: 8 월 26 일부터 9 월 4 일까지 10 일간

※ 질문 제작 시 유의사항

① 사용자가 물어볼 만한 내용, 궁금해할 만한 내용으로 질문을 제작

② 경어를 사용하지 않고, 반말체의 간결한 대화체 문장으로 작성하며 질문의 길이가 공백 포함 80 자를 넘는 질문은 지양

③ 본문에서 사용한 표현을 활용하되, 어순과 어휘를 있는 그대로 베끼는 질문은 지양

④ 문장 전체가 답이 될 수 있는 보편적인 질문이나 여러 문장이 답이 될 수 있는 질문은 지양. 특히 '어떻게'와 '왜' 유형의 질문은 문장 안에서 가능한 짧은 어절로 답변할 수 있도록 구체적인 질문으로 유의해서 작성

⑤ 질문에 구두점 외의 특수문자, 오탈자가 없도록 작성

No	질문 유형	상세 내용
1	<사건>	252 회 의회 회기 / 제 199 회 완주군의회 휴회 / 예산결산특별위원회의 활동
2	<행동>	이인숙 의원님 외 네 분의 의원님으로부터 정례회 집회 요구/ 5 분 자유발언 신청/ 제 210 회 완주군의회 임시회 회기 발의
3	<역할>	특별예산결정위원장/ 회의록 서명의원/ 예산결산위원회 간사
4	<지표>	몇 명 / 예산 / 며칠

답변 제작 기준

답변 제작 방법

- ① 질문 제작 유형과 답변의 유형을 참고해 다양한 질문-답변 세트를 제작한다.

※ 답변 제작 시 유의사항

- ① 주어진 지문 안의 문구만 사용하며, 임의 편집 불가
- ② 질문에 포함된 단어를 답변에 포함하는 것을 지양
- ③ 답변의 좌우에 인용 마크나 괄호 등 특수문자가 있는 경우 특수문자를 제외하고 공백이 포함되지 않도록 주의 (단, 답변 우측의 단위부호(% + - ₩ \$ ¥)는 허용)
- ④ 불필요한 수식어를 제외한 핵심어를 중심으로 정답을 선정

No	답변 유형	상세 내용
1	기간	5 월 19 일부터 21 일까지 3 일 동안
2	일시	2022 년 12 월 20 일/ 3 월 5 일 오후 2 시
3	사람(들)	박종원 의원 (직위 포함 지향) / 권항화 의원, 서제일 의원 / 완주군수
4	숫자(단위)	1 억/ 4 인/ 5 분/
5	명사구	제 268 회 음성군의회 제 1 차 정례회 제 6 차 본회의 /

2. Annotation 가이드라인

질문 유형 별 예시

유형 1. <사건>의 기간은?

질문	대답
1_특별위원회의 현지확인언은 언제까지?	5 월 19 일부터 21 일까지 3 일 동안

유형 2. <사건>는 언제 진행되는가?

질문	대답
2_음성군 행정기구설치조례 일부개정조례안은 언제 제출되었어?	2022 년 12 월 20 일

유형 3. <행동>을 수행한 사람은 누구인가?

질문	대답
3_5 분자유발언을 한 사람은 누구야?	박종원 의원

고려사항

- 행위를 수행하는 주체에 대한 질문
- 가능하면 직위까지 태깅

유형 4. <역할>을 맡게된 사람은 누구인가?

질문	대답
4_특별예산결정위원장을 맡게된 사람은 누구야?	특별예산결정위원장을 맡게된 사람은 누구인가?

고려사항

- 회의 내, 원래 있던 직책이 아닌, 새로 부여된 직책, 직위에 대해서만 QA 데이터셋 생성
- 연속된 복수의 대답도 가능
- 직위, 직책은 제외하고 사람의 이름만 정답으로

유형 5. <행동>을 한 이유는 무엇인가?

질문	대답
5_김은숙 의원이 5 분자유발언을 한 이유는 무엇이야?	직지코리아 국제페스티벌이 고유의 정체성을 가지며 직지의 소중한 가치를 담을 수 있는 국제적인 행사가 되기 위해서는 근본적인 변화가 필요하다는 말씀을 드리고자

고려사항:

- 길이는 하나의 문장으로 제한하고, 문맥을 변형시키지 않아야 한다.
- 대답이 긴 경우에는 핵심적인 부분만, 명사구로 끝나는 부분을 정답으로 선택
(위 예시의 빨간색 부분)

유형 6. <사건>의 나온 <지표>는 얼마인가?

질문	대답
6_2014 년도 예비비 지출 승인의 건의 표결 결과 몇 명이 찬성했어?	5 명

유형 7. <사건>를 다루고 있는 회의는?

질문	대답
7_"2014 년도 회계 세입 · 세출 결산 승인의 건과 2014 년도 예비비 지출 승인의 건"을 다룬 회의는 무슨 회의야?	제 268 회 음성군의회 제 1 차 정례회 제 6 차 본회의

고려사항:

- 해당 질문은 사건을 통해 회의를 검색하는 질문이기에, Passage 앞에 제시된 Title 을 대답으로 선택한다.

유형 8. No Answer

모델의 성능을 위해 생성하는 No Answer 질문은 위의 7 가지 유형의 질문 형태를 띄지만 답을 찾을 수 없는 질문으로 생성한다.

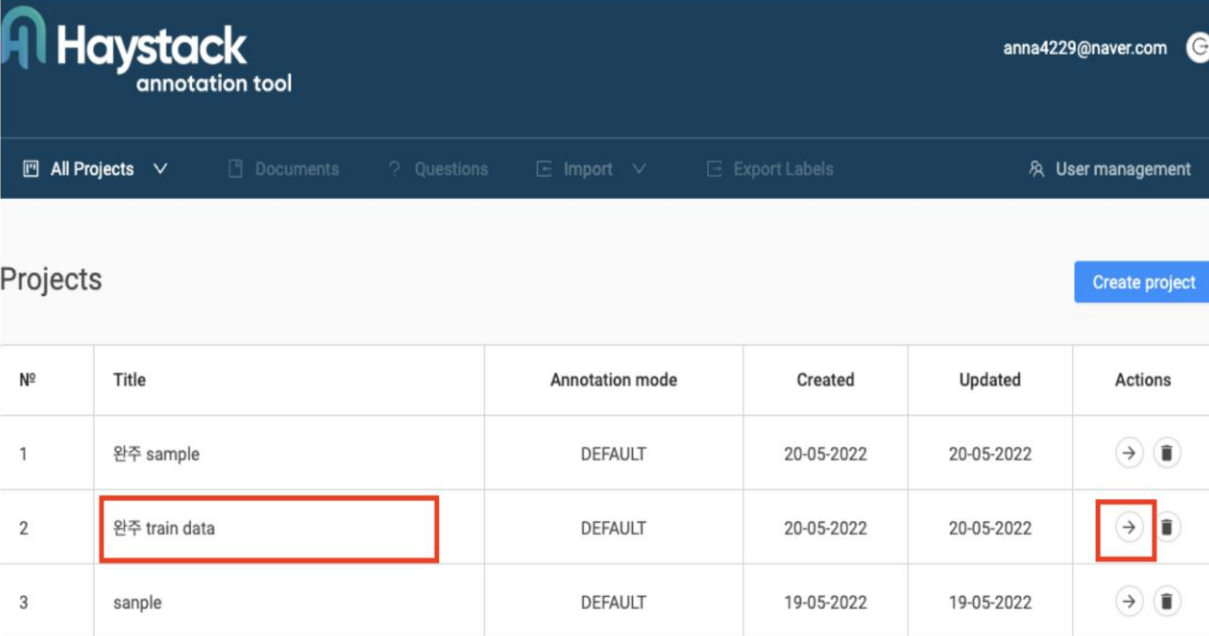
예를 들어 지문 내 OO 안건의 예산이 없을 때, “2014 년도 OO 안건의 예산은 얼마야?”라는 질문이 들어왔다고 가정해보자. 해당 질문은 유형 6.의 질문 유형과 비슷하지만, 지문에 답이 없는 질문이 된다. 위와 같이 질문의 유형을 지키지만, 답이 없는 질문을 No Answer 질문으로 생성한다.

3. Annotation 도구 사용법

3-1) 작업자 환경 준비

- [haystack 웹사이트](#)에 로그인해 문서가 업로드되어있는 프로젝트에 들어갑니다. (문서가 없을 시 업로드가 필요합니다)

1. 문서가 들어 있는 프로젝트로 들어갑니다.









Haystack annotation tool

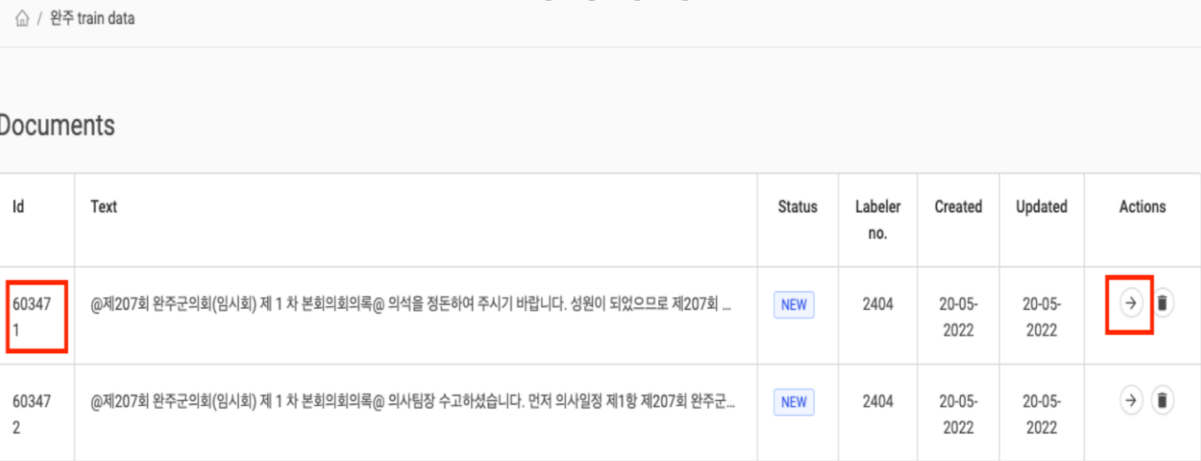
anna4229@naver.com

All Projects Documents Questions Import Export Labels User management

Projects [Create project](#)





Nº	Title	Annotation mode	Created	Updated	Actions
1	완주 sample	DEFAULT	20-05-2022	20-05-2022	 
2	완주 train data	DEFAULT	20-05-2022	20-05-2022	 
3	sample	DEFAULT	19-05-2022	19-05-2022	 

2. 라벨링할 문서를 선택합니다.



🏠 / 완주 train data

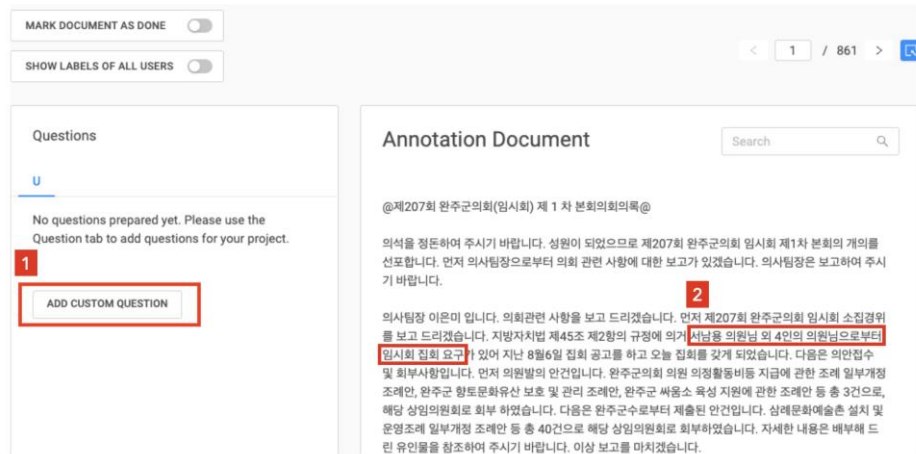
Documents

Id	Text	Status	Labeler no.	Created	Updated	Actions
60347 1	@제207회 완주군의회(임시회) 제 1 차 본회의회의록@ 의석을 정돈하여 주시기 바랍니다. 성원이 되었으므로 제207회 ...	NEW	2404	20-05-2022	20-05-2022	 
60347 2	@제207회 완주군의회(임시회) 제 1 차 본회의회의록@ 의사팀장 수고하셨습니다. 먼저 의사일정 제1항 제207회 완주군...	NEW	2404	20-05-2022	20-05-2022	 

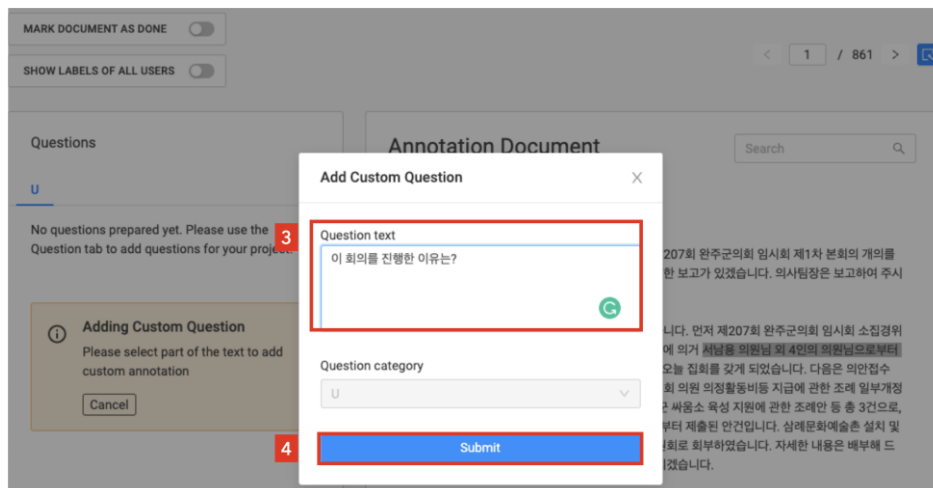
3-2) 문서에 대답이 있는 경우 라벨링 방법

- 가이드라인의 질문 유형과 예시를 참고해 질문을 생성합니다.

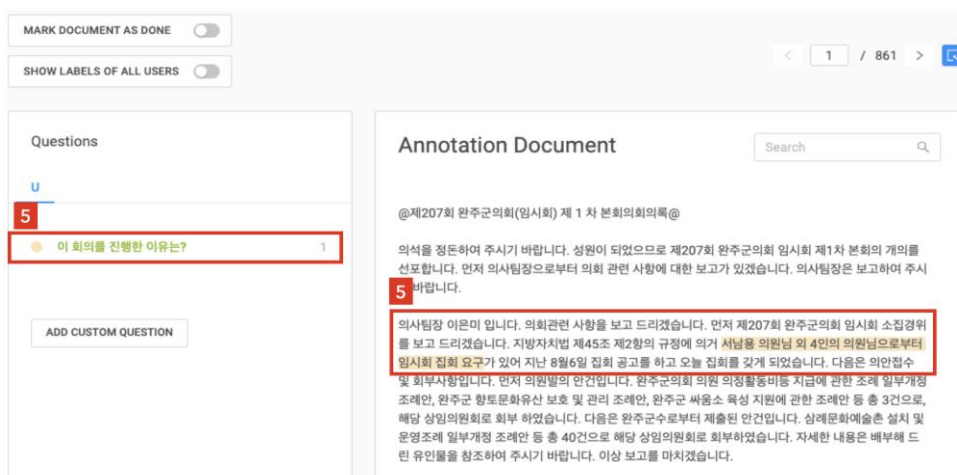
1. ADD CUSTOM QUESTION 버튼을 누릅니다.
2. 마우스로 드래그해 정답을 먼저 선택합니다.



3. 질문을 작성합니다.
4. Submit 버튼을 눌러 완료합니다.

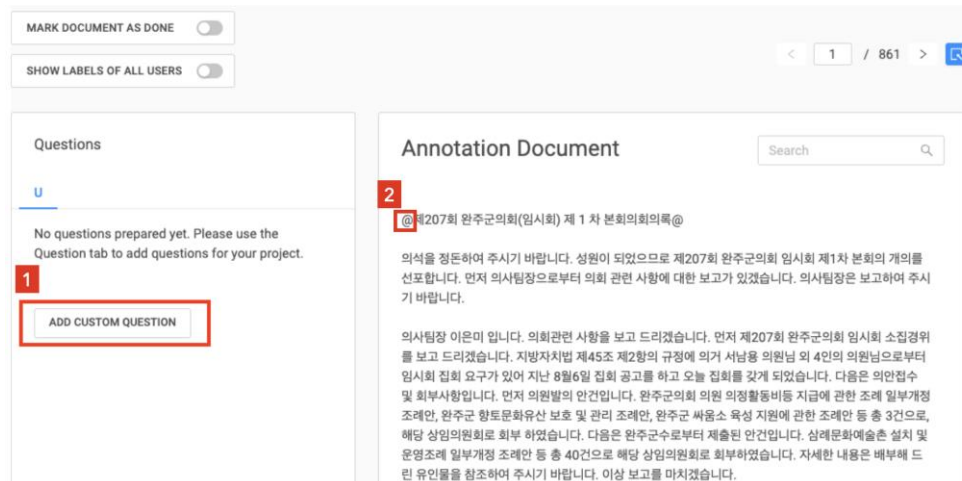


5. 생성한 질문과 답변을 확인합니다.
6. 1-5.의 과정을 반복합니다.

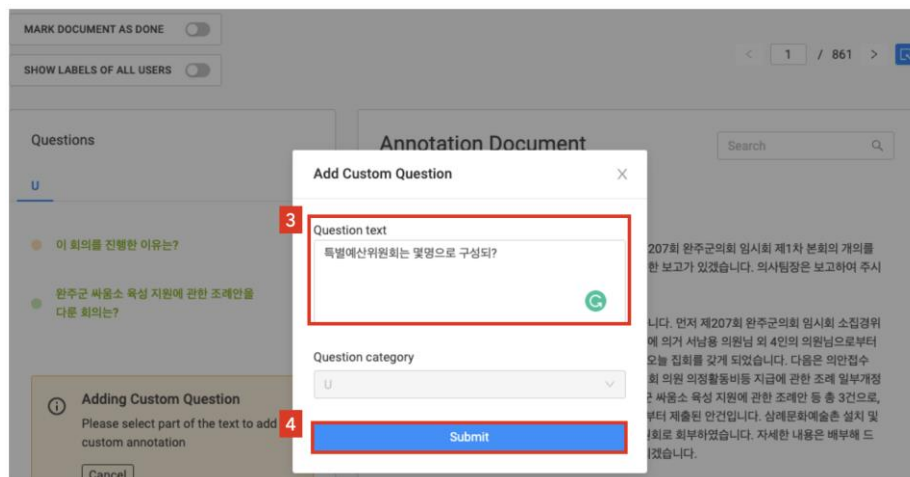


3-3) 문서에 대답이 없는 경우 라벨링 방법

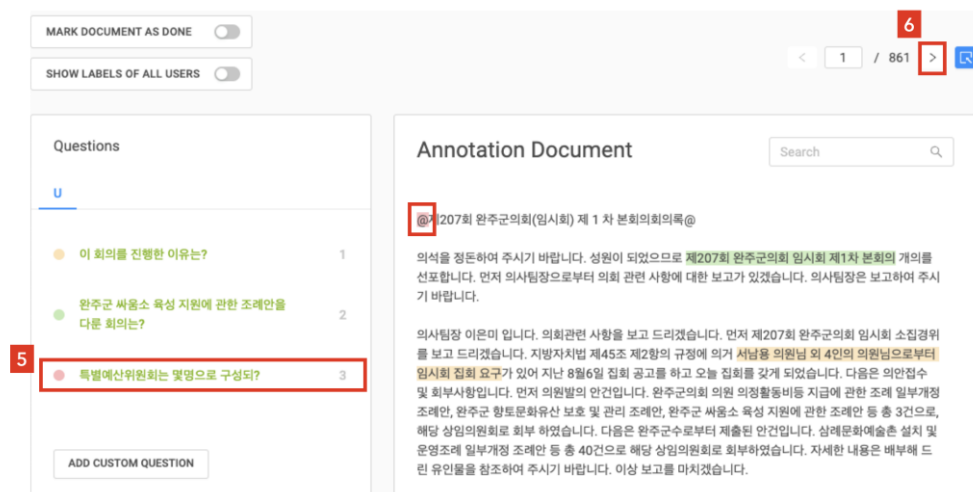
1. ADD CUSTOM QUESTION 버튼을 누릅니다.
2. no_answer 데이터를 라벨링하고 싶은 경우에는 대답으로 title 의 첫 번째 문자인 @를 선택합니다.



3. 동일하게 질문을 작성합니다.
4. Submit 버튼을 눌러 완료합니다.

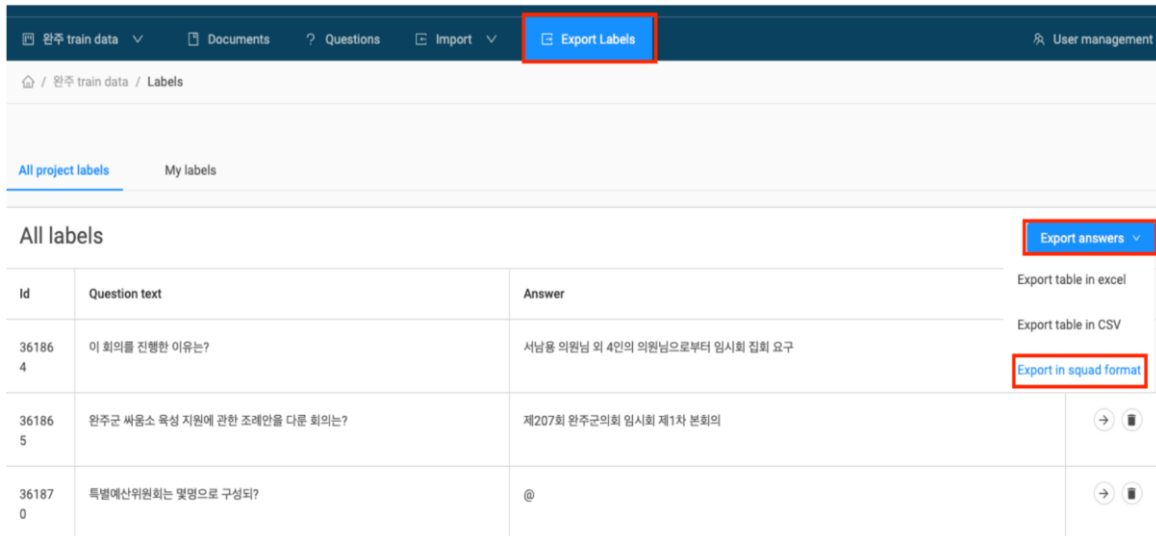


5. no_answer 유형의 질문-대답도 동일하게 생성된 것을 확인할 수 있습니다.
6. 현재 문서에 대한 라벨링을 완료했다면, > 버튼을 눌러 다음 문서로 이동합니다.



3-4) 라벨링 완료 이후

- Export Labels >> Export answers >> Export in squad format 버튼을 순서대로 눌러 .json 파일을 다운받습니다.
- 다른 확장자도 잘 다운되지만, 어차피 .json 으로 읽을꺼니까 .json 으로 다운받습니다.



3-5) 참고 자료

사용법 관련 유튜브 영상

- [Easiest way to Annotate Data for Question Answering Task \(NLP\) | Data Science | Deep Learning](#)
- [haystack github](#)

4. Annotation FAQ

Q1. 간결체의 질문은 어떤 것을 의미하는가?

A1. 실제 대화할 때 사용하는 어투를 생각하면 된다. 시리나 빅스비와 같은 가상비서에 질문하는 느낌으로 생각하면된다.

(잘못된 예시) 완주군 싸움소 육성 지원에 관한 조례안을 다룬 회의는 무엇인가?

(올바른 예시) 완주군 싸움소 육성 지원에 관한 조례안을 다룬 회의는 뭐야? / 뭘까?

Q2. no_answer 질문은 얼마나 생성해야 하는가?

A2. no_answer 질문은 지문 8 개마다 1 개씩 만든다는 느낌으로 생성을 해 대략 5-10%의 개수를 맞추고자 한다.

Q3. title 로 제시된 회의 이름을 태깅할 때는 '회의록'도 포함해야 하는가?

A3. 회의 이름을 title(제목)에서 태깅하는 경우에는 회의 이름만 태깅한다.

(잘못된 예시) @제 207 회 완주군의회(임시회) 제 1 차 본회의회의록@

(올바른 예시) @제 207 회 완주군의회(임시회) 제 1 차 본회의회의록@

Q4. 회의 이름의 경우 title 말고, 본문에 등장하는 경우도 있는데 이때는 어느 위치에 태깅을 하면 되는가?

A4. 회의 이름을 태깅하는 경우 title 혹은 본문 둘 다 상관없다. 하지만 본문에 회의 이름이 등장하는 경우를 권장한다.

Q5. 사건에 해당하는 ‘△△△조례안과 □□□제안에 대한 심사’처럼 여러가지를 다루는 사건의 경우 어떻게 태깅해야 하는가?

A5. 여러가지를 다루는 사건이더라도 그 길이가 너무 길지 않고(공백 포함 80 자 이상), 의미상 이어진다면 하나로 태깅해도 된다.

Q6. ‘5 분 사전 발언, 감표위원’과 같은 단어는 회의 특성 상 많이 나오는 것 같은데 어떻게 질문을 생성해야 하는가?

A6. 다양한 회의에서 나올 수 있을만한 용어에 대해서는 회의를 특정지을 수 있을만한 정보를 앞에 표시해줘야 한다.

(예시 1) △△△내용으로 5 분 사전발언을 수행한 사람은 누구야?

(예시 2) ◇◇◇의원이 ♡♡♡건으로 제안설명을 한 회의는 무슨 회의야?

(예시 3) ☆☆☆ 회의에서 5 분 사전발언을 한 사람은 누구야?

Q7. 사람에 대한 답변을 해야 할 때, 소속, 직위, 이름 모두 나온면 다 태깅하면 되는가?

A7. 소속은 제외하고 직위+이름만 태깅한다. 소속, 이름, 직위가 나오더라도 이름+직위만 태깅한다.

Q8. 완주군수와 같은 직위를 사람으로 태깅할 수 있을까?

A8. 문맥상 직위가 특정 사람임을 지칭한다고 해석될 수 있으면, 직위를 <사람>으로 태깅할 수 있다. 완주군수의 경우 회의를 참여한 모두가 알고 있는 존재이며, 한 회의에서 한 사람만 존재한다고 가정할 수 있기에 태깅 가능하다.

Q9. When[일정, 기간]질문의 경우 현재 시점에 대해서만 질문을 생성해야 하는가?

A9. When[일정, 기간]에 대한 질문은 과거/현재/미래 시점에 상관없이 답변으로 설정이 가능하다. 따라서 “안건은 언제 진행되었어?(과거)”, “안건을 다루고 있는 날짜는?”(현재), “안건은 언제까지 진행될 예정이야? (미래)와 같은 질문이 모두 가능하다.

5. 데이터 구축 결과

‘완주군’의 회의 데이터 165 건을 주제별로 나눠 861 개의 지문으로 데이터 구축 작업을 수행한 결과 지문(Passage) - 질문(Question) - 답변(Answer)으로 구성된 총 1539 개의 기계 독해용 질의응답 데이터셋을 구축했다. 이 중 기계독해 모델의 성능 향상을 위해 질문 중 답이 없는 경우가 7%를 차지한다.

공정명	건수(비율)	데이터셋 구성
정답이 있는 질문	1429 건(93%)	지문-질문-정답 데이터셋으로 작성
정답이 없는 질문	110 건(7%)	지문-질문-빈칸 데이터셋으로 작성

질문의 유형별 개수를 다음 표에서 확인할 수 있다. 이전 질문 유형별로 총 130~261 까지의 개수를 가진다. 가장 높은 개수의 질문은 ‘What(회의)’로 회의 자체를 찾는 질문으로, 가장 질문을 생성하기 쉬웠기 때문인 것으로 보인다. 다음으로 가장 낮은 개수의 질문은 ‘What(지표)’로 수치를 다루고 있는 회의는 많았지만, 질문을 생성하기 쉽지 않았기 때문에 수가 적은 것으로 보인다.

색인	질문유형	질문 패턴	답변 유형	개수
1	When(기간)	<사건>의 기간은?	기간(날짜, 시간)	234
2	When(일시)	<사건>는 언제 진행되는가?	일시(날짜, 시간)	217
3	Who(능동주, 행동)	<행동>을 수행한 사람은 누구인가?	사람(들)	289
4	Who(피동주, 역할)	<역할>을 맡게된 사람은 누구인가?	사람	216
5	Why(행동, 의견)	<행동>을 한 이유는 무엇인가?	명사구/기타	192
6	What(지표)	<사건>의 나온 <지표>는 얼마인가?	숫자(금액, 단위)	130
7	What(회의)	<사건>를 다루고 있는 회의는?	명사구	261

해당 지표들을 개수 순으로 정렬하면 다음과 같다. 분포를 보면 각 유형별로 큰 차이가 없으며, 해당 데이터셋의 분포는 KorQuAD 데이터셋의 분포와 비교했을 때도 유사한 분포를 보였다. 따라서 분포면에서 실험을 진행하기에 균등한 질문 생성이 되었다고 할 수 있다.

