

AMBIGQA: Answering Ambiguous Open-domain Questions

Sewon Min, Julian Michael, Hannaneh Hajishirzi, Luke Zettlemoyer

EMNLP 2020

인공지능학과 2021246064 김수형

저자 소개



First Author

Sewon Min (Ph.D. student in Washington Univ.)

Interests - NLU, QA

Publications

- Joint Passage Ranking for Diverse Multi-Answer Retrieval, EMNLP(2021)
- NeurIPS 2020 EfficientQA Competition: Systems, Analyses and Lessons Learned, PMLR(2021)



Co-Author

Hannaneh Hajishirzi (Assistant Professor Washington Univ.)

Interests - QA, reading comprehension, representation learning

Publications

- One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval, NeurIPS(2021)
- DIALKI: Knowledge Identification in Conversational Systems through Dialogue-Document Contextualization, EMNLP(2021)



목차

- Motivation
- Introduction
 - AMBIGQA task
 - Evaluation metrics
 - AMBIGNQ dataset
- Model Overview
- Experiments
- Ablation Study
- Summary

Motivation

기존의 QA method : one question have single well-defined answer

BUT



AMBIGQA Input

When did harry potter and the sorcerer's stone movie come out?

Harry Potter and the Philosopher's Stone (film)

From Wikipedia, the free encyclopedia

The film had its world premiere at the Odeon Leicester Square in London on 4 November 2001, with the cinema arranged to resemble Hogwarts School. (...) The film was released to cinemas in the United Kingdom and United States on 16 November 2001.

Ambiguous(모호한) questions
(prompt question)

Multiple answers

Motivation

Split	# data	# QAs %			
		1	2	3	4+
Train	10,036	53	24	14	10
Dev	2,002	49	23	14	13
Test	2,004	44	24	16	16

Table 2: Data statistics. For the number of QA pairs (# QAs), the minimum is taken when there are more than 1 accepted annotations.

Open-domain version of Natural Questions (NQ)

- 50% 이상이 모호한 문장
- **ambiguity**를 제거하여 질문-답변 **pair**를 만들었을때 3개 이상의 **pair**가 만들어 질 정도로 **dataset**이 모호함

Introduction - AMBIGQA task



AMBIGQA Input

When did harry potter and the sorcerer's stone movie come out?

Harry Potter and the Philosopher's Stone (film)

From Wikipedia, the free encyclopedia

The film had its world premiere at the Odeon Leicester Square in London on **4 November 2001**, with the cinema arranged to resemble Hogwarts School. (...) The film was released to cinemas in the United Kingdom and United States on **16 November 2001**.

AMBIGQA Output

Q: When did harry potter and the sorcerer's stone movie come out at the Odeon Leicester Square?

A: 4 November 2001

Q: When did harry potter and the sorcerer's stone movie come out in cinemas?

A: 16 November 2001



Answering Ambiguous Open-domain Questions (AMBIGQA)

- 모호한 질문에 대해 그럴듯한 답변들을 찾고,
- 각 답변을 설명할 수 있는 질문을 **ambiguous**한 질문으로부터 **rewrite** 한다.

Figure 1: An AMBIGQA example where the prompt question (top) appears to have a single clear answer, but is actually ambiguous upon reading Wikipedia. AMBIGQA requires producing the full set of acceptable answers while differentiating them from each other using disambiguated rewrites of the question.

Introduction - AMBIGQA task

- AMBIGQA subtasks

1. Multiple Answer Prediction

- 주어진 질문 q 에 대해 의미적으로 구별되면,
그럴듯한 답변들 y_1, \dots, y_n 을 생성

2. Question Disambiguation

- 주어진 질문 q 와 답변들 $y_1 \sim y_n$ 에 대해 명확한
질문 x_1, \dots, x_n 을 생성
- 각각의 x_i 는 q 를 최소한으로 수정하여 y_i 가
답변이 되는 명확한 질문으로 생성한다.
- 같은 i 가 아닌 나머지 답변들 y_j 에 대해서는
전부 틀린 질문이 되도록한다.

Introduction - Evaluation metrics

prediction pairs = $(x_1, y_1), \dots, (x_m, y_m)$

gold pairs = $(\bar{x}_1, \bar{\mathcal{Y}}_1), \dots, (\bar{x}_n, \bar{\mathcal{Y}}_n)$

$\bar{\mathcal{Y}}_i$ = 여러가지 답변이 가능한 경우가 있으므로,
가능한 답변의 **string set**으로 구성

$$c_i = \max_{1 \leq j \leq n} \mathbb{I}[y_i \in \bar{\mathcal{Y}}_j] f(x_i, \bar{x}_j)$$

y_i 가 $\bar{\mathcal{Y}}_j$ 에 속할 때 x_i 와 \bar{x}_j 사이의 similarity function(f)의 점수중 max 값 = c_i

Similarity function $f = [0, 1]$

Introduction - Evaluation

$$c_i = \max_{1 \leq j \leq n} \mathbb{I}[y_i \in \bar{\mathcal{Y}}_j] f(x_i, \bar{x}_j)$$

$$\text{prec}_f = \frac{\sum_i c_i}{m}, \quad \text{rec}_f = \frac{\sum_i c_i}{n},$$

$$\text{F1}_f = \frac{2 \times \text{prec}_f \times \text{rec}_f}{\text{prec}_f + \text{rec}_f}.$$

1. F1_{ans} : 항상 gold와 같을때 (f-score가 항상 1)
2. F1_{BLEU} : similarity function으로 BLEU metric을 사용
3. $\text{F1}_{\text{EDIT-F1}}$: prompt question과 비교하였을때, 추가 또는 제거된 unigrams을 비교

Introduction - Evaluation metrics

$F1_{\text{EDIT-F1}}$: prompt question과 비교하였을때, 추가 또는 제거된 unigrams을 비교

prompt question : "Who made the play the crucible?"

gold question : "Who wrote the play the crucible?"

predicted question: "Who made the play the crucible in 2012?"



gold question edit : { -made, +wrote }

predicted question edit { +in, +2012 }

수정한 결과가 서로 다르기 때문에 $F1_{\text{EDIT-F1}} = 0$

Introduction - AMBIGNQ dataset

1. generation
2. validation
3. Quality control
4. Inter-annotator agreement

init setting

- prompt question -> NQ-open dataset
- evidence corpus -> English Wikipedia
- crowd sourcing -> Amazon Mechanical Turk(AMT)

Introduction - AMBIGNQ dataset

1. generation
2. validation
3. Quality control
4. Inter-annotator agreement

prompt question , Wikipedia



workers(generators)



(edited question, answer) pairs

조건

1. 같은내용의 답변이라면 여러 방식으로 표현
(e.g. Michael Jordan and Michael Jeffrey Jordan)
2. 기간에 의존(time dependency)되는 질문은 2018.1.1 이전에 가장
최근 event 3개만 가능
(e.g. drama season15, 14, 13)

Introduction - AMBIGNQ dataset

1. generation
2. validation
3. Quality control
4. Inter-annotator agreement

(edited question, answer) pairs



validator



correct or incorrect
or add new set(q,a) pair

조건

1. generator와 마찬가지로 wikipedia에서 찾고, 추가로 generator가 봤던 page도 제공한다.
2. 모든 generator의 annotation이 일치하면 skip

Introduction - AMBIGNQ dataset

1. generation
2. validation
3. Quality control
4. Inter-annotator agreement

- 높은 질의 **data**를 제공한 **workers(generator, validator)**를 재고용 하여 진행
- **dev**와 **test dataset**에 대해서 2명의 **generator**, 1명의 **validator**로 진행하고, **train dataset**에 대해서 질문당 1명의 **generator**로 진행한다.
- **train dtaset**은 **validation**은 skip

> 질문의 76%의 **annotation**은 **validation**을 통과했고 나머지 24%는 수정 또는 제외 되었다.

> 공동저자와 **workers**의 **validation sample**의 평균 **F1ans**는 89%로 대부분의 경우 인간이 유효한 답변과 잘못된 답변의 경계에 합의가 된다는것을 알 수 있다.

Introduction - AMBIGNQ dataset

Split	# data	# QAs %			
		1	2	3	4+
Train	10,036	53	24	14	10
Dev	2,002	49	23	14	13
Test	2,004	44	24	16	16

Table 2: Data statistics. For the number of QA pairs (# QAs), the minimum is taken when there are more than 1 accepted annotations.

dev와 test dataset 에서 50% 이상이 multiple qa pair를 가지고 있다.
(train dataset 에서는 47%)

=> 이전 연구들은 질문 하나에 단일 답변이 있다는 가정으로
연구되어왔음에도 불구하고 NQ-open dataset에 높은 비율의 모호성이
있다는 것을 나타낸다.

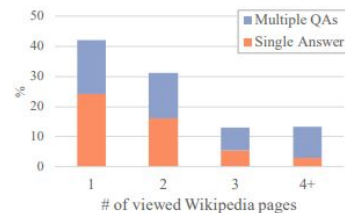
Type	Example
Event references (39%)	What season does meredith and derek get married in grey's anatomy? Q: In what season do Meredith and Derek get informally married in Grey's Anatomy? / A: Season 5 Q: In what season do Meredith and Derek get legally married in Grey's Anatomy? / A: Season 7
Properties (27%)	How many episode in seven deadly sins season 2? Q: How many episodes were there in seven deadly sins season 2, not including the OVA episode? / A: 25 Q: How many episodes were there in seven deadly sins season 2, including the OVA episode? / A: 26
Entity references (23%)	How many sacks does clay matthews have in his career? Q: How many sacks does Clay Matthews Jr. have in his career? / A: 69.5 Q: How many sacks does Clay Matthews III have in his career? / A: 91.5
Answer types (16%)	Who sings the song what a beautiful name it is? Q: Which group sings the song what a beautiful name it is? / A: Hillsong Live Q: Who is the lead singer of the song what a beautiful name it is? / A: Brooke Ligertwood
Time-dependency (13%)	When does the new family guy season come out? Q: When does family guy season 16 come out? / A: October 1, 2017 Q: When does family guy season 15 come out? / A: September 25, 2016 Q: When does family guy season 14 come out? / A: September 27, 2015
Multiple sub-questions (3%)	Who was british pm and viceroi during quit india movement? Q: Who was british viceroy during quit India movement? / A: Victor Hope Q: Who was british pm during quit India movement? / A: Winston Churchill

Table 1: Breakdown of the types of ambiguity in 100 randomly sampled items from the AMBIGNQ development data. Each example may fall into multiple categories.

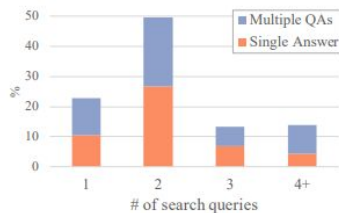
ambiguity type
ambiguity in entity references, event references,
properties, answer types

=> 이전 연구에서 ambiguity in entity references를 의도적으로
유도한 것과 달리 다양한 source로부터 의도하지 않은 모호성이
나타난다.

Introduction - AMBIGNQ dataset



(a) Number of unique Wikipedia pages visited by crowdworkers.[†]



(b) Number of search queries written by crowdworkers.



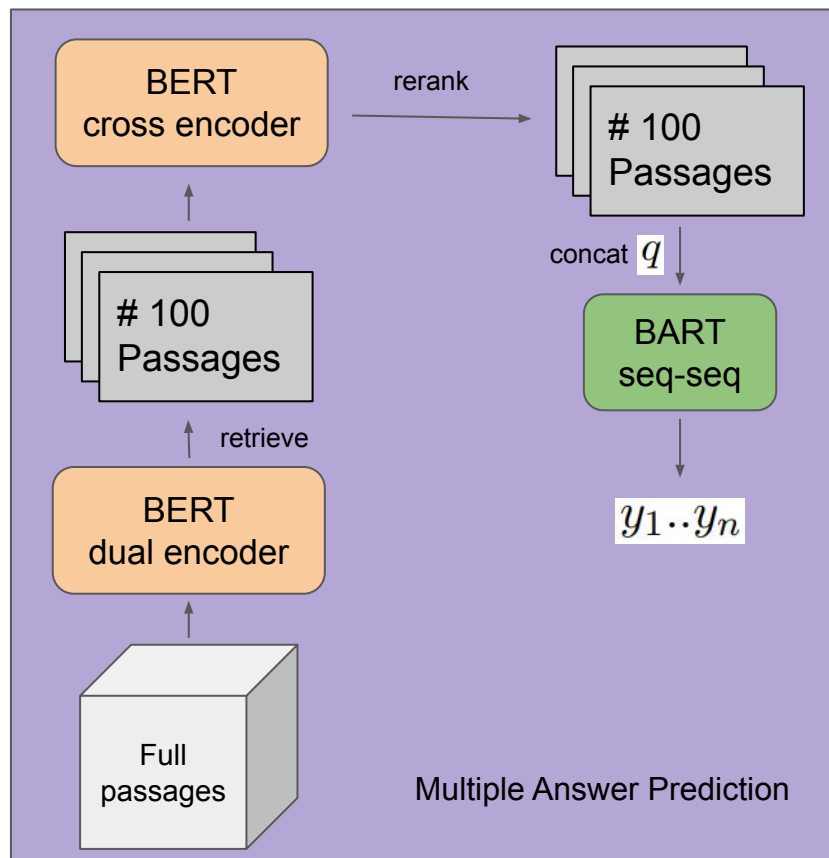
(c) Word cloud of the edits made in questions; ■ and ■ indicate added and deleted unigrams, respectively.

wh-word를 제외한
불용어는 제거하고,
숫자는 자릿수로
치환했다

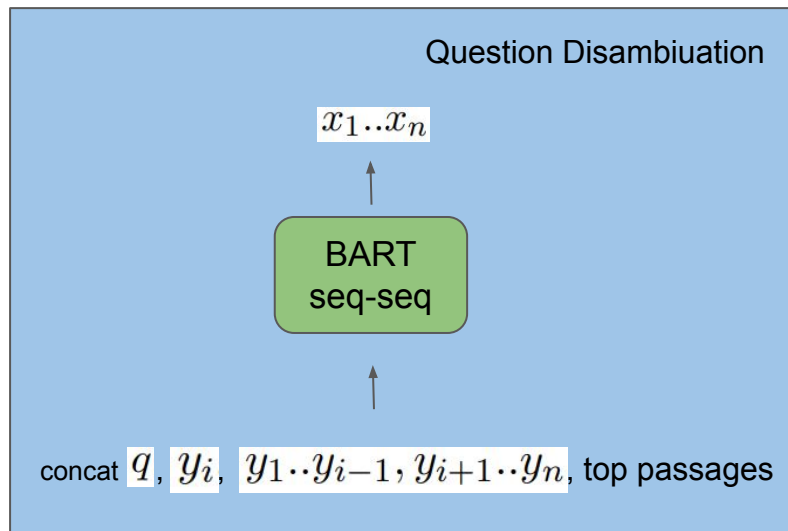
Figure 2: Data Analysis on the development data. [†]This is actually an underestimate; we could not track when annotators viewed pages by following hyperlinks for technical reasons.

1. worker들은 여러 검색어와 여러 **page(document)**를 탐색하여 **open-domain qa**의 **retrieval** 단계에서 **ambiguity**를 찾는 방법으로, 이전 연구에서 미리 지정된 **evidence document**가 있다는 가정의 접근법에서 놓친 부분이다.
2. **dev dataset**에서 **edit-question**을 만들때 수정한 **unigram**을 분석
=> 숫자를 추가하는것은 쉽게 모호성을 제거하고, 시간 의존성을 제거하는 방법이고, **wh-word**를 수정하는것으로도 답을 특정지을 수 있다. (수정된 상위 100개가 36%, 상위 1000개가 69%를 차지함)

Model Overview



prompt question = q
predict answer = $y_1 \dots y_n$
edited-question = $x_1 \dots x_n$



Model Overview - Democratic co-training

Algorithm 1 Democratic co-training with weak supervision (Section 5).

```
1: // Each question in  $D_{full}$  has an answer list annotated
2: // Each question in  $D_{partial}$  has one answer annotated
3:  $\hat{D}_{full} \leftarrow D_{full}$ 
4: for  $iter \in \{1..N\}$  do
5:   // Train  $C$  sequence-to-sequence QA models
6:   for  $i \in \{1..C\}$  do
7:      $\phi_i \leftarrow train(\hat{D}_{full})$ 
8:    $\hat{D}_L \leftarrow D_{full}$ 
9:   for  $(q^j, y^j) \in D_{partial}$  do
10:    // Get predictions by using  $y_j$  as prefix
11:     $\hat{Y}^j \leftarrow \{\hat{y} \mid \hat{y} \neq y^j, \text{ and}$ 
12:       $|\{i \mid \hat{y} \in \phi_i(q^j|y^j), 1 \leq i \leq C\}| > \frac{C}{2}$ 
13:     $\}$ 
14:    if  $|\hat{Y}^j| > 0$  then
15:      // Add it as a multiple answer case
16:       $\hat{D}_{full} \leftarrow \hat{D}_L \cup \{(q^j, \{\hat{Y}^j\})\}$ 
17:    else if  $\forall i = 1..C, |\phi_i(q^j) - \{y^j\}| = 0$  then
18:      // Add it as a single answer case
19:       $\hat{D}_{full} \leftarrow \hat{D}_L \cup \{(q^j, \{y^j\})\}$ 
```

D_{full} : AMBIGNQ

$D_{partial}$: NQ-OPEN

C : Multi Answer prediction model #

→ 추가 답변을 생성한 모델이 과반수라면 D_{full} 에 추가한다.


Experiments - Baselines

1. DISAMBIG-FIRST

- prompt question을 가능한 답변 또는 reference passages 없이 명확하게 만든다.
 - i. BERT binary classifier를 통해 prompt question (q)이 모호한지 판별
 - ii. 모호하다면, BART 모델의 input으로 넣어서 edit-question (x_i)을 생성
(각 edit-question은 [SEP]으로 분리)
 - iii. edit-questions를 state-of-the-art 모델(DPR)에 입력으로 넣어서 answer (y_i) 생성

2. Thresholding + QD

- Multiple answer prediction 과 question disambiguation 을 위해 DPR 모델에 threshold를 추가한 baseline
- hyperparameter γ 보다 높은 유효한 spans을 $y_1..y_n$ 으로 얻는다
(gold answer set $\bar{y}_1..\bar{y}_n$ 에 가까워지도록 학습한다)
- NQ-OPEN dataset 에 대해서 pretrain하고, AMBIGNQ 에 대해서 train한다.



Experiments - Baselines

Experiments - Results

Model	F1 _{ans} (<i>all</i>)		F1 _{ans} (<i>multi</i>)		F1 _{BLEU}		F1 _{EDIT-F1}	
	dev	test	dev	test	dev	test	dev	test
DISAMBIG-FIRST	28.1	24.8	21.9	18.8	4.2	4.0	2.7	2.2
Thresholding + QD	37.1	32.3	28.4	24.8	13.4	11.3	6.6	5.5
SPANSEQGEN + QD	39.7	33.5	29.3	24.5	13.4	11.4	7.2	5.8
SPANSEQGEN [†] + QD	41.2	35.2	29.8	24.5	13.6	10.6	7.4	5.7
SPANSEQGEN [†] (Co-training) + QD	42.3	35.9	31.7	26.0	14.3	11.5	8.0	6.3

Table 3: Results on AMBIGNQ. The *multi* measure only considers examples with multiple question-answer pairs. [†] indicates ensemble. See Appendix B for details on the discrepancy between development and test.

- DISAMBIG-FIRST의 classifier 성능이 67% 특히 낮았고, rewrite을 하더라도 겉보기에는 그럴듯해도, 사실에 맞지 않았다.
- SPANSEQGEN이 Thresholding을 크게 능가하진 못했다.
 - a. 답변들이 확률에 대해 경쟁할지라도 multi-answer를 출력하기에 thresholding이 매우 효과적이었다.
 - b. SPANSEQGEN(BART) 같은 seq-seq 모델에서는 잘 보정된 결과가 생성되지 않을 수 있다. (output 시퀀스 길이가 평균 3.0 token 인데 gold는 6.7 token으로 multi answer를 생성할때 낮은 recall이 문제가 될 것이다. (best - pre(49.6) / rec(25.3) / F1_{ans} (31.7))

Experiments - Results

Prompt question #1: Where was snow white and the huntsman filmed?

Reference:

Q: Where were beach scenes for snow white and huntsman predominantly filmed? / A: Marloes Sands Beach

Q: Where was principal photography for snow white and huntsman filmed? / A: United Kingdom

Q: Where was castle in snow white and huntsman filmed? / A: Gateholm island

Prediction of DISAMBIG-FIRST: ($F1_{ans}=0.40$, $F1_{EDIT-F1}=0.00$)

Q: Where was snow white and the huntsman filmed in 2017? / A: Marloes Sands Beach

Q: Where was snow white and the huntsman filmed during the filming of Season 1 of the TV series? / A: Marloes Sands Beach

Prediction of SPANSEQGEN: ($F1_{ans}=0.80$, $F1_{EDIT-F1}=0.69$)

-> Snow White and the Huntsman

Q: Where was snow white and huntsman principal photography filmed / A: United Kingdom

Q: Where were beach scenes for snow white and huntsman mostly filmed / A: Marloes Sands Beach

Prompt question #2: When was the city of new york founded?

Reference:

Q: When was city of new york founded by dutch and initially called new amsterdam? / A: 1624

Q: When was city of new york under english control and renamed to new york? / A: 1664

Prediction of SPANSEQGEN: ($F1_{ans}=1.00$, $F1_{EDIT-F1}=0.67$)

Q: When was city of new york city founded with dutch protection? / A: 1624

Q: When was city of new york city founded and renamed with english name? / A: 1664

Table 5: Model predictions on samples from the development data. (#1) DISAMBIG-FIRST generates questions that look reasonable on the surface but don't match the facts. SPANSEQGEN produces the reasonable answers and questions, although not perfect. (#2) SPANSEQGEN produces correct answers and questions.

Ablation study - disambiguation

Model	q	y_i	$y_1..y_{i-1},$ $y_{i+1}..y_n$	<i>Full task</i>		<i>Gold answers given</i>	
				F1 _{BLEU}	F1 _{EDIT-F1}	F1 _{BLEU}	F1 _{EDIT-F1}
QD model	✓	✓	✓	14.3	8.0	40.1	19.2
- prompt question	-	✓	✓	6.7	7.7	15.1	19.2
- untargeted answers	✓	✓	-	14.2	7.3	41.2	17.2
Always prompt question	✓	-	-	15.9	0.0	47.4	0.0

Table 4: Ablations on question disambiguation (development data, multiple answers only). QD model refers to the question disambiguation model described in Section 5. For multiple answer prediction, we use SPANSEQGEN[†] with co-training (*Full task*) or the gold answers (*Gold answers given*).

- 최소한의 수정으로 **edit-question**을 만들기 때문에 **prompt question**을 그대로 출력하는것이 F1BLEU 점수가 가장 높은건 당연하다(F1_{EDIT-F1} 을 정당화 해준다.)
- 모든 **context**를 받는 QD 모델이 가장 성능이 좋다.

Ablation study - disambiguation

- gold answer를 주어도 성능이 낮은 이유
 - a. output sequence의 likelihood를 높이면 QD 모델이 답변끼리의 차이의 정보를 놓칠 수 있다.
(gold output의 확률에 치중하여 높은 확률의 question을 생성하고, 답변끼리의 차이를 학습하지 않게 된다?)
 - b. NQ-OPEN으로 weakly-supervised를 하는 이점이 없는 question disambiguation annotation data가 매우 부족하다.



Q&A



Appendix