Harvesting More Answer Spans from Paragraph beyond Annotation

Qiaoben Bao, Jiangjie Chen, Linfang Liu, Jingping Liu, Jiaqing Liang, Yanghua Xiao

Fudan University

Background

- Answer span extraction (AE) focuses on identifying answer spans from paragraphs

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers […]

Answer span: Denver Broncos
Question: Which NFL team represented the AFC at Super Bowl 50?

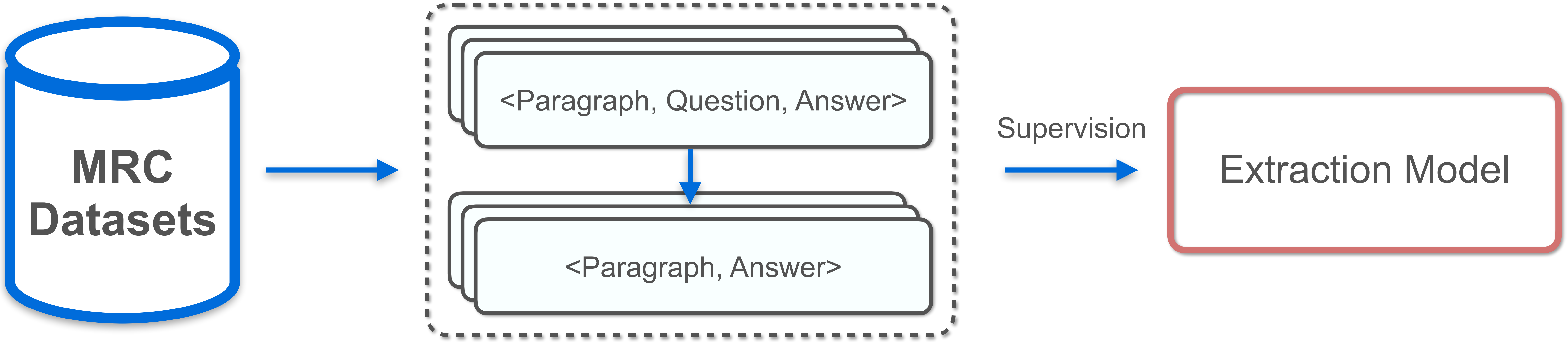
Answer span: American Football Conference
Question: What is the AFC short for?

...

  – AE has a wide range of both research and real-life applications
    - Facilitating information extraction
    - Data augmentation for MRC or QG
    - Building FAQs against documents
    - …

Background

- The current work of AE relies on the annotation from MRC datasets
  - There is currently no dataset dedicated to AE tasks
  - MRC dataset contains <Paragraph, Question, Answer> triples and can be easily converted to <Paragraph, Answer> pairs for AE



Challenge

- Is the annotation from MRC sufficient for the AE task?
  - MRC datasets are only required to extract limited answer spans (usually 5) for each paragraph
  - The unannotated candidate spans may also be valid answer spans

- We analyze two well-known MRC datasets
  - SQuAD (Rajpurkar et al.) & DROP* (Dua et al.)

* DROP contains three types of answers and we only consider the extractive examples where the answer is a span from the original paragraph.

Challenge

- Re-annotate 50 paragraphs from each dataset
  - We define the missing rate γ as

$$\gamma = \frac{|\mathcal{M}|}{\left|\mathcal{S}_p \cup \mathcal{M}\right|}$$

    - $\mathcal{S}_p$ is the positive labeled samples in paragraph
    - $\mathcal{M}$ is the unlabeled samples in paragraph

Challenge

- Re-annotate 50 paragraphs from each dataset
  - We define the missing rate γ as

$$\gamma = \frac{|\mathcal{M}|}{|\mathcal{S}_p \cup \mathcal{M}|}$$

    - $\mathcal{S}_p$ is the positive labeled samples in paragraph
    - $\mathcal{M}$ is the unlabeled samples in paragraph
- Both datasets contain a comparable number of positive answer spans not annotated among unlabeled candidate spans

Dataset	**#Sentences**	$\lvert\mathcal{S}_p\rvert$	$\lvert\mathcal{M}\rvert$	γ
SQuAD	237	219	207	48.59%
DROP	445	296	492	62.44%

Challenge

- This MRC annotation procedure ignores other detailed key information that would also be helpful for readers to understand the context.

Paragraph #1:
[…] In 1815, the British government selected Saint Helena as the place of detention of Napoleon Bonaparte. He was taken to the island […]

Golden answer spans:

Answer span: 1815
Corresponding question: What year was Napoleon Bonaparte taken to the island?

Answer span: Napoleon Bonaparte
Corresponding question: The British government detained who in Saint Helena?

Unlabeled answer spans:

Answer span: British
Corresponding question: Which government sent Napoleon Bonaparte to Saint Helena?

Answer span: Saint Helena
Corresponding question: Where was Napoleon Bonaparte imprisoned?

Paragraph #2:
[…] In 1882, Albert Zahm built an […]. Around 1899, Professor Jerome Green became the first American to […]. In 1931, Father Julius Nieuwland performed early work […]

Golden answer spans:

Answer span: 1882
Corresponding question: In what year did Albert Zahm begin comparing aeronatical models at Notre Dame?

Answer span: Professor Jerome Green
Corresponding question: Which professor sent the first wireless message in the USA?

…

Unlabeled answer spans:

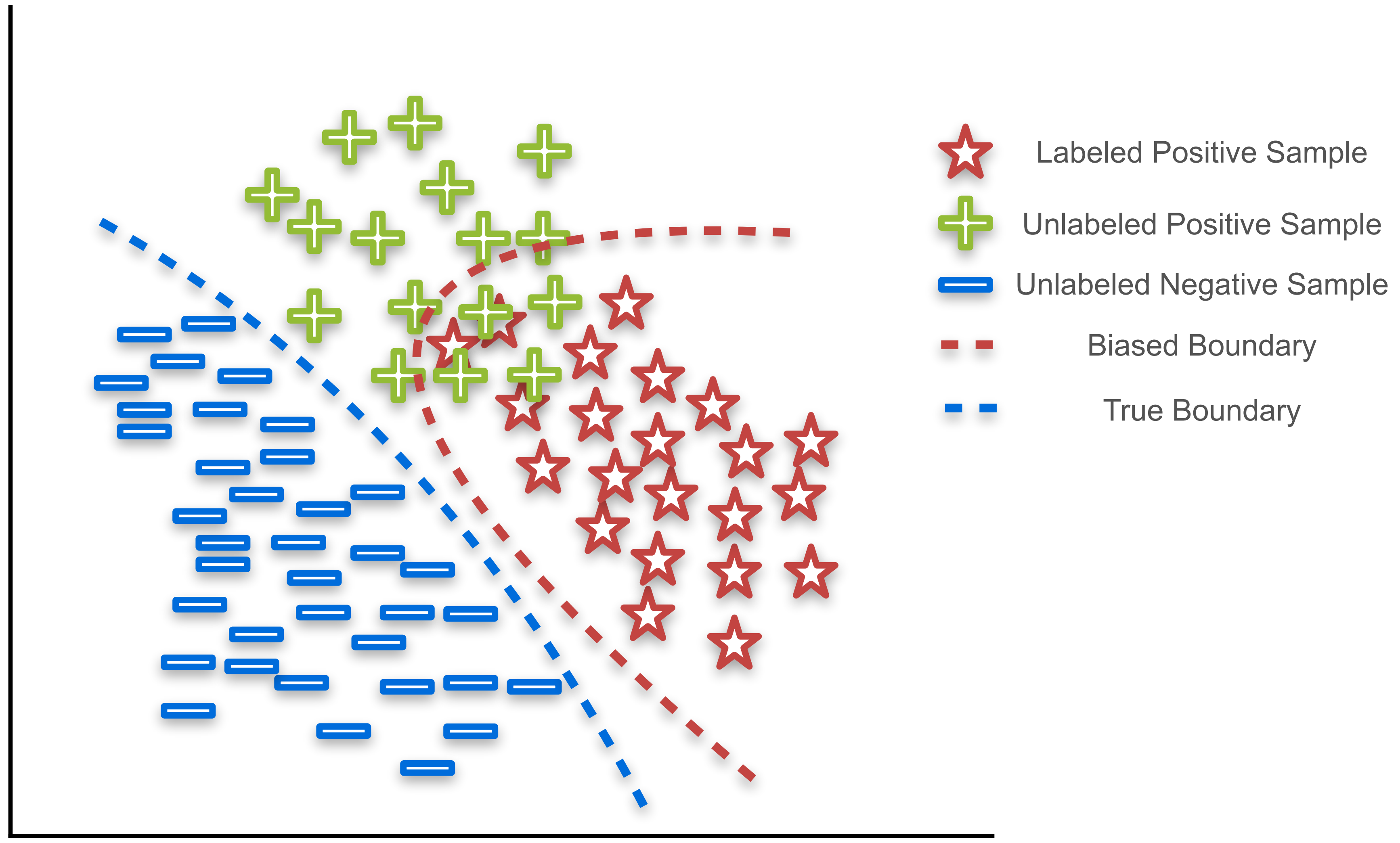
Answer span: 1931

Answer span: Albert Zahm

…

Challenge

- This MRC annotation procedure ignores other detailed key information that would also be helpful for readers to understand the context.
  - Previous work
    - Directly treating unlabeled data as negative one may lead to the wrong decision boundary



Challenge

- This MRC annotation procedure ignores other detailed key information that would also be helpful for readers to understand the context.
  - Previous work
    - Directly treating unlabeled data as negative one may lead to the wrong decision boundary
  - Our goal
    - Narrows the discrepancy between AE and MRC to extract more answer spans from paragraph
    - For answer span that is not labeled, we can automatically evaluate its quality

Method

- Basic idea
  - Formulate AE task as a positive-unlabeled learning problem
    - Split the risk estimator into the positive part and the negative part

$$R_\ell = \pi \mathbb{E}_{\mathbf{x}, y=1} \ell(f(\mathbf{x}), 1) + (1-\pi) \mathbb{E}_{\mathbf{x}, y=0} \ell(f(\mathbf{x}), 0)$$

  - f is the classifier
  - ℓ is loss function
  - π is the prior distribution of positive samples

    - We do not have labeled negative samples for calculating $(1-\pi)\mathbb{E}_{\mathbf{x}, y=0} \ell(f(\mathbf{x}), 0)$

Method

- Basic idea
  - Formulate AE task as a positive-unlabeled learning problem
    - Split the risk estimator into the positive part and the negative part

$$R_\ell = \pi \mathbb{E}_{\mathbf{x}, y=1} \ell(f(\mathbf{x}), 1) + (1 - \pi) \mathbb{E}_{\mathbf{x}, y=0} \ell(f(\mathbf{x}), 0)$$

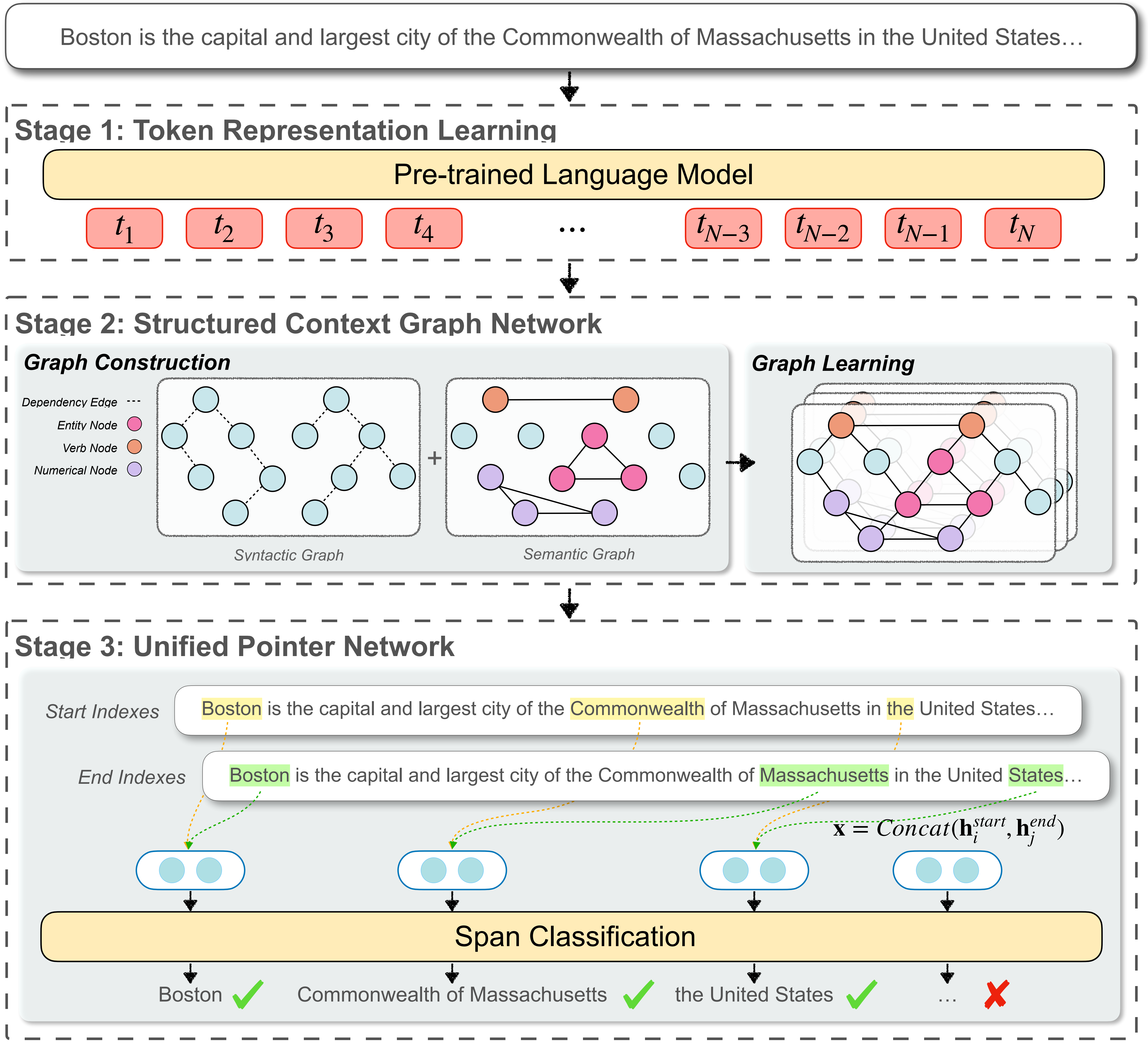
    - Re-estimate the negative part with positive samples and unlabeled samples

$$(1 - \pi) \mathbb{E}_{\mathbf{x}, y=0} \ell(f(\mathbf{x}), 0) = \mathbb{E}_{x} \ell(f(\mathbf{x}), 0) - \pi \mathbb{E}_{x, y=1} \ell(f(\mathbf{x}, 0))$$

    - Finally, we can calculate R_ℓ by estimating the prior distribution π

Method

- Framework: SCOPE
  - PLM-based token representation
    - Encode sub-token with PLM
    - Token representation is averaged by sub-tokens
  - GNN-based information prorogation
    - Syntactic edges
    - Semantic edges
  - PU classifier
    - Unified pointer network
    - PU loss



Evaluation

- Question-worthy score: Questionability + Worthiness
  - Questionability

Questionability
If a span is askable, there exists at least one question that can be answered by this span with a high probability.

  - Evaluated by a QG-QA model
    - The question generation model first generates questions based on the given paragraph and extracted answer span
    - The question answering model then scores the answer spans against the generated question

  - Worthiness

Evaluation

- Question-worthy score: Questionability + Worthiness
  - Questionability
  - Worthiness

Worthiness
If a span is worthy of asking, it contains more information for people to ask a question.

- Evaluated by an extractive summarization model
  - Extractive summarization model score each sentence an informative score
  - For each candidate span, we define its worthiness as the informative score of the sentence it locates

Experiments

- Conventional metrics
  - Our proposed framework extracts more high-quality answer spans on both datasets

Model	SQuAD				DROP			
	Precision	Recall	F1	Avg. spans	Precision	Recall	F1	Avg. spans
ENT	*13.63*	*40.41*	*20.39*	*12.82*	*6.31*	*52.58*	*11.27*	*43.90*
ENT Classifier (BERT$_{base}$)	**48.55**	20.37	28.70	1.81	36.90	19.10	25.17	2.73
⌞ (SpanBERT$_{base}$)	47.90	21.09	29.29	1.90	**38.62**	20.52	26.80	2.80
⌞ (RoBERTa$_{base}$)	47.57	20.61	28.76	1.87	35.54	20.90	26.32	3.10
Sequence Tagger (BERT$_{base}$)	44.39	25.96	32.76	2.53	30.07	20.60	24.45	3.61
⌞ (SpanBERT$_{base}$)	46.96	25.98	33.45	2.39	35.24	20.52	25.94	3.07
⌞ (RoBERTa$_{base}$)	48.43	25.19	33.14	2.25	34.91	19.33	24.88	2.92
Boundary-aware NER (Zheng et al., 2019)	32.80	18.67	23.79	2.46	34.40	7.23	11.95	1.11
BiFlaG (Luo and Zhao, 2020)	36.50	25.97	30.35	3.08	38.13	20.03	26.27	2.77
MRC NER (BERT$_{base}$) (Li et al., 2020)	37.71	25.84	30.67	2.96	29.53	21.20	24.68	3.78
⌞ (SpanBERT$_{base}$)	37.04	27.94	31.85	3.26	31.79	20.52	24.94	3.40
⌞ (RoBERTa$_{base}$)	39.59	27.39	32.38	2.99	32.38	22.43	26.50	3.65
SCOPE (BERT$_{base}$)	36.96	39.99	38.41	4.68	30.74	32.32	31.51	5.54
⌞ (SpanBERT$_{base}$)	41.15	39.70	**40.41**	4.17	33.95	35.84	34.87	5.56
⌞ (RoBERTa$_{base}$)	36.10	**45.19**	40.14	**5.41**	33.51	**37.08**	**35.21**	**5.83**

Experiments

- Is the extracted span high-quality?
  - Automatic metrics
    - When SCOPE extracts more new spans, it also sightly outperforms baselines on question-worthy score

Model	Avg. $score_q$	Avg. $score_w$	Avg. ϵ
Golden	*80.07 (0.21)*	*38.09 (0.13)*	*59.08 (0.13)*
ENT Classifier	**78.52** (0.21)	32.93 (0.13)	55.72 (0.12)
Sequence Tagger	74.34 (0.21)	35.66 (0.12)	55.00 (0.12)
Boundary-aware NER	70.02 (0.25)	35.30 (0.13)	52.66 (0.14)
BiFlaG	75.95 (0.22)	34.83 (0.13)	55.39 (0.13)
MRC NER	75.07 (0.21)	**36.09** (0.12)	55.58 (0.12)
SCOPE	76.94 (0.21)	35.60 (0.12)	**56.27** (0.12)

  - Performance boost on down-stream QA tasks

Backbone Model	Exact Match	F1
BERT_{large} (Devlin et al., 2019)	78.7	81.9
BERT_{large} (Our implementation)	77.9	81.3
⌞ ENT Classifier	77.8 (−0.1)	81.1 (−0.2)
⌞ Sequence Tagger	79.0 (+1.1)	82.2 (+0.9)
⌞ Boundary-aware NER	77.3 (−0.6)	80.8 (−0.5)
⌞ BiFlaG	79.1 (+1.2)	82.1 (+0.8)
⌞ MRC NER	79.3 (+1.4)	82.6 (+1.3)
⌞ SCOPE	**79.9 (+2.0)**	**83.2 (+1.9)**

Analysis

- Is the model sensitive to prior distribution?
  - SCOPE has a consistent performance gain with different backbone PLMs and prior distribution π

	BERT_{base}				SpanBERT_{base}				RoBERTa_{base}			
	Precision	Recall	F1	Avg. spans	Precision	Recall	F1	Avg. spans	Precision	Recall	F1	Avg. spans
$\pi' \times 1.50$	**39.93**	35.02	37.31	3.79	**44.67**	35.33	39.46	3.42	**42.18**	36.83	39.32	3.78
$\pi' \times 1.75$	39.21	36.25	37.67	4.00	41.77	38.32	39.97	3.97	40.28	38.62	39.43	4.15
$\pi' \times 2.00\dagger$	36.96	39.99	**38.41**	4.68	41.15	39.70	**40.41**	4.17	36.10	45.19	**40.14**	5.41
$\pi' \times 2.25$	29.32	47.65	36.30	7.03	37.99	43.04	40.36	4.90	33.31	**47.63**	39.20	6.18
$\pi' \times 2.50$	29.41	**49.17**	36.81	7.23	34.39	**46.64**	39.59	5.86	32.18	47.06	38.22	6.32

Summarize

- Re-formulate current AE task as a PU learning problem
- Propose SCOPE, a Structured Context graph network with Positive-unlabeled learning, to extract more answer spans from paragraphs
- Propose question-worthy score for automatically evaluate the quality of answer spans

Thanks for listening