Harvesting More Answer Spans from Paragraph beyond Annotation

Qiaoben Bao, Jiangjie Chen, Linfang Liu, Jingping Liu, Jiaqing Liang, Yanghua Xiao
Fudan University
Answer span extraction (AE) focuses on identifying answer spans from paragraphs.

- 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
  - ...

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?

...
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
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

- Reannotate 50 paragraphs from each dataset
  - We define the missing rate $\gamma$ as
    \[ \gamma = \frac{|M|}{|S_p \cup M|} \]
    - $S_p$ is the positive labeled samples in paragraph
    - $M$ is the unlabeled samples in paragraph
• Re-annotate 50 paragraphs from each dataset
  – We define the missing rate $\gamma$ as

  $\gamma = \frac{|M|}{|S_p \cup M|}$

  $\cdot S_p$ is the positive labeled samples in paragraph
  $\cdot M$ is the unlabeled samples in paragraph

• Both datasets contain a comparable number of positive answer spans not annotated among unlabeled candidate spans

| Dataset | #Sentences | $|S_p|$ | $|M|$ | $\gamma$   |
|---------|------------|-------|-------|------------|
| SQuAD   | 237        | 219   | 207   | 48.59%     |
| DROP    | 445        | 296   | 492   | 62.44%     |
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 [...]
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
      \[ R_\ell = \pi \mathbb{E}_{x, y=1} \ell(f(x), 1) + (1 - \pi) \mathbb{E}_{x, y=0} \ell(f(x), 0) \]
      – \( f \) is the classifier
      – \( \ell \) is loss function
      – \( \pi \) is the prior distribution of positive samples
    › We do not have labeled negative samples for calculating \( (1 - \pi) \mathbb{E}_{x, y=0} \ell(f(x), 0) \)
Method

• Basic idea
  – Formulate AE task as a positive-unlabeled learning problem
    ▶ Split the risk estimator into the positive part and the negative part
      \[
      R_\ell = \pi \mathbb{E}_{x,y=1} \ell(f(x), 1) + (1 - \pi) \mathbb{E}_{x,y=0} \ell(f(x), 0)
      \]
    ▶ Re-estimate the negative part with positive samples and unlabeled samples
      \[
      (1 - \pi) \mathbb{E}_{x,y=0} \ell(f(x), 0) = \mathbb{E}_{x} \ell(f(x), 0) - \pi \mathbb{E}_{x,y=1} \ell(f(x), 0)
      \]
    ▶ Finally, we can calculate $R_\ell$ by estimating the prior distribution $\pi$
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
    • 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

• 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

<table>
<thead>
<tr>
<th>Model</th>
<th>SQuAD</th>
<th>DROP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>ENT</td>
<td>13.63</td>
<td>40.41</td>
</tr>
<tr>
<td>ENT Classifier (BERT&lt;sub&gt;base&lt;/sub&gt;)</td>
<td>48.55</td>
<td>20.37</td>
</tr>
<tr>
<td>(\mathbf{\ominus}) (SpanBERT&lt;sub&gt;base&lt;/sub&gt;)</td>
<td>47.90</td>
<td>21.09</td>
</tr>
<tr>
<td>(\mathbf{\ominus}) (RoBERTa&lt;sub&gt;base&lt;/sub&gt;)</td>
<td>47.57</td>
<td>20.61</td>
</tr>
<tr>
<td>Sequence Tagger (BERT&lt;sub&gt;base&lt;/sub&gt;)</td>
<td>44.39</td>
<td>25.96</td>
</tr>
<tr>
<td>(\mathbf{\ominus}) (SpanBERT&lt;sub&gt;base&lt;/sub&gt;)</td>
<td>46.96</td>
<td>25.98</td>
</tr>
<tr>
<td>(\mathbf{\ominus}) (RoBERTa&lt;sub&gt;base&lt;/sub&gt;)</td>
<td>48.43</td>
<td>25.19</td>
</tr>
<tr>
<td>Boundary-aware NER (Zheng et al., 2019)</td>
<td>32.80</td>
<td>18.67</td>
</tr>
<tr>
<td>BiFlaG (Luo and Zhao, 2020)</td>
<td>36.50</td>
<td>25.97</td>
</tr>
<tr>
<td>MRC NER (BERT&lt;sub&gt;base&lt;/sub&gt;) (Li et al., 2020)</td>
<td>37.71</td>
<td>25.84</td>
</tr>
<tr>
<td>(\mathbf{\ominus}) (SpanBERT&lt;sub&gt;base&lt;/sub&gt;)</td>
<td>37.04</td>
<td>27.94</td>
</tr>
<tr>
<td>(\mathbf{\ominus}) (RoBERTa&lt;sub&gt;base&lt;/sub&gt;)</td>
<td>39.59</td>
<td>27.39</td>
</tr>
<tr>
<td>SCOPE (BERT&lt;sub&gt;base&lt;/sub&gt;)</td>
<td>36.96</td>
<td>39.99</td>
</tr>
<tr>
<td>(\mathbf{\ominus}) (SpanBERT&lt;sub&gt;base&lt;/sub&gt;)</td>
<td>41.15</td>
<td>39.70</td>
</tr>
<tr>
<td>(\mathbf{\ominus}) (RoBERTa&lt;sub&gt;base&lt;/sub&gt;)</td>
<td>36.10</td>
<td>45.19</td>
</tr>
</tbody>
</table>
Experiments

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Avg. score_q</th>
<th>Avg. score_w</th>
<th>Avg. e</th>
</tr>
</thead>
<tbody>
<tr>
<td>Golden</td>
<td>80.07 (0.21)</td>
<td>38.09 (0.13)</td>
<td>59.08 (0.13)</td>
</tr>
<tr>
<td>ENT Classifier</td>
<td>78.52 (0.21)</td>
<td>32.93 (0.13)</td>
<td>55.72 (0.12)</td>
</tr>
<tr>
<td>Sequence Tagger</td>
<td>74.34 (0.21)</td>
<td>35.66 (0.12)</td>
<td>55.00 (0.12)</td>
</tr>
<tr>
<td>Boundary-aware NER</td>
<td>70.02 (0.25)</td>
<td>35.30 (0.13)</td>
<td>52.66 (0.14)</td>
</tr>
<tr>
<td>BiFlaG</td>
<td>75.95 (0.22)</td>
<td>34.83 (0.13)</td>
<td>55.39 (0.13)</td>
</tr>
<tr>
<td>MRC NER</td>
<td>75.07 (0.21)</td>
<td>36.09 (0.12)</td>
<td>55.58 (0.12)</td>
</tr>
<tr>
<td>SCOPE</td>
<td>76.94 (0.21)</td>
<td>35.60 (0.12)</td>
<td>56.27 (0.12)</td>
</tr>
</tbody>
</table>

– Performance boost on down-stream QA tasks

<table>
<thead>
<tr>
<th>Backbone Model</th>
<th>Exact Match</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT&lt;sub&gt;large&lt;/sub&gt; (Devlin et al., 2019)</td>
<td>78.7</td>
<td>81.9</td>
</tr>
<tr>
<td>BERT&lt;sub&gt;large&lt;/sub&gt; (Our implementation)</td>
<td>77.9</td>
<td>81.3</td>
</tr>
<tr>
<td>v. ENT Classifier</td>
<td>77.8 (−0.1)</td>
<td>81.1 (−0.2)</td>
</tr>
<tr>
<td>v. Sequence Tagger</td>
<td>79.0 (+1.1)</td>
<td>82.2 (+0.9)</td>
</tr>
<tr>
<td>v. Boundary-aware NER</td>
<td>77.3 (−0.6)</td>
<td>80.8 (−0.5)</td>
</tr>
<tr>
<td>v. BiFlaG</td>
<td>79.1 (+1.2)</td>
<td>82.1 (+0.8)</td>
</tr>
<tr>
<td>v. MRC NER</td>
<td>79.3 (+1.4)</td>
<td>82.6 (+1.3)</td>
</tr>
<tr>
<td>v. SCOPE</td>
<td>79.9 (+2.0)</td>
<td>83.2 (+1.9)</td>
</tr>
</tbody>
</table>
• Is the model sensitive to prior distribution?
  – SCOPE has a consistent performance gain with different backbone PLMs and prior distribution $\pi$

| $\pi' \times 1.50$ | 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