Distant Supervision for Question Summarization

Tatsuya Ishigaki†, Kazuya Machida†, Hayato Kobayashi†, Hiroya Takamura‡, Manabu Okumura†
†Tokyo Institute of Technology / ‡Yahoo Japan Corporation / ‡AIST

1. Introduction

Motivation:
Questions tend to be lengthy and hard to understand. We aim to convert them easy-to-understand shorter questions.

Task: Extractive Question Summarization

Input: multi-sentence question

Output: extracted single-sentence summary

Existing Approaches (Extracrive):

Supervised:
- Classification/Regression
  - [Ishigaki+, 2017, Tamura+ 2007]
  - learning-to-rank [Higurashi+, 2018]
  - Supervised methods require costly labeled data

Unsupervised:
- Graph-based (e.g. LexRank) [Erkan+ 2004]
  - Semantic similarity [Kobayashi+, 2018]
  - Major unsupervised methods do not perform well
  [See our experiments.]

Our Approach:
This paper describes a distant supervision that creates pseudo labeled data for training a summarizer w/o labeled data.

Contributions:
1. We propose a distant supervision approach to create a pseudo labeled data for training a question summarizer.
2. Our models w/o any supervision performs competitively with respect to supervised models.
3. We release a large dataset including 2.5M sentences with pseudo labels.

2. Proposed Framework

Input: Lengthy Questions

Distant Supervision Summarizer

Output: Single-sentence Questions

1. Data Extraction

We extracted 2.5M sentences from a corpus of CQA. All sentences are labeled by our proposed heuristics:

- Positive labels: single-sentence questions
- Negative labels: individual sentences extracted from extremely long post.

Single-sentence questions have summary-like properties:
basically they are self-contained questions.
(= similar to ones that we want to include in the summary).

Individual sentences in long post are not summary-like:
basically they are not self-contained and often not a question.
(= we need information from other sentences to understand.)

1. Train Classifier

We trained a binary classifier that outputs a score that represents how likely the sentence is summary-like.

2. Sentence Selection

We score every sentence in an input. We propose several sentence selection strategies that use the scores as explained in Sec.3.

3. Experiment

Datasets:
1. Dataset with pseudo labels (2.5M sentences)
   - Labeled data created by our framework.
2. Dataset with manually annotated labels (10K sentences)
   - We used a crowdsourcing to annotate the sentences.

Compared Models:
- Our Models (trained on our data with pseudo labels)
  - DistNet: NN-based sentence tagger (LSTM + Softmax)
  - DistReg: Logistic Regression with N-gram, POS features
- Unsupervised Models
  - Lead: Simply selects the initial sentence.
  - LexRank: A graph-based algorithm for sentence selection.
  - SimEmb: Selects the sentence that has the minimum Word Movers’ Distance from the input.
  - Tfidf: Selects the sentence that has the higher Tfidf in the input.
- Supervised Models (trained on the manually annotated data)
  - SupNet: NN-based sentence tagger (LSTM + Softmax)
  - SupReg: Logistic Regression with N-gram, POS features

Sentence Selection Strategies:
- Greedy: Simply selects the highest scored sentence.
- Init: Selects the initial sentence that has higher score than a specific threshold (tuned on validation data).
- Q: Selects the highest scored question sentence.

4. Result

Accuracy = correctly selected sentences / total sentences.

<table>
<thead>
<tr>
<th></th>
<th>Greedy</th>
<th>Init</th>
<th>Q</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>DistNet</td>
<td>87.38</td>
<td>90.45</td>
<td>87.38</td>
<td>90.45</td>
</tr>
<tr>
<td>DistReg</td>
<td>86.17</td>
<td>89.05</td>
<td>86.17</td>
<td>89.05</td>
</tr>
<tr>
<td>Lead</td>
<td>81.79</td>
<td>81.79</td>
<td>88.08</td>
<td>88.08</td>
</tr>
<tr>
<td>LexRank</td>
<td>78.49</td>
<td>81.79</td>
<td>84.95</td>
<td>84.95</td>
</tr>
<tr>
<td>SimEmb</td>
<td>59.46</td>
<td>81.79</td>
<td>71.17</td>
<td>81.79</td>
</tr>
<tr>
<td>Tfidf</td>
<td>52.03</td>
<td>81.79</td>
<td>69.68</td>
<td>81.79</td>
</tr>
<tr>
<td>SupNet</td>
<td>81.67</td>
<td>86.31</td>
<td>81.67</td>
<td>86.31</td>
</tr>
<tr>
<td>SupReg</td>
<td>87.89</td>
<td>91.21</td>
<td>87.89</td>
<td>91.21</td>
</tr>
</tbody>
</table>

- Our distant supervision approach outperformed all unsupervised baselines.
- Using our pseudo data improved the performance of NN-based approach (DistNet).
- There is no statistically significant difference between the best performed model of our distant supervision approach and the best model of supervised models.

5. Conclusion

- We proposed a distant supervision for extractive summarization task.
- Our approach outperformed unsupervised baselines and performed competitively with supervised baselines.
- The data is publicly available: http://fr-www.pi.titech.ac.jp/~ishigaki/chiebukuro/