Distant Supervision for Extractive Question Summarization

Tatsuya Ishigaki [1], Kazuya Machida [1], Hayato Kobayashi [2]
Hiroya Takamura [1,3] and Manabu Okumura [1]

[1] Tokyo Institute of Technology,
[2] Yahoo Japan Corporation/RIKEN AIP,
[3] AIST

Extractive Question Summarization

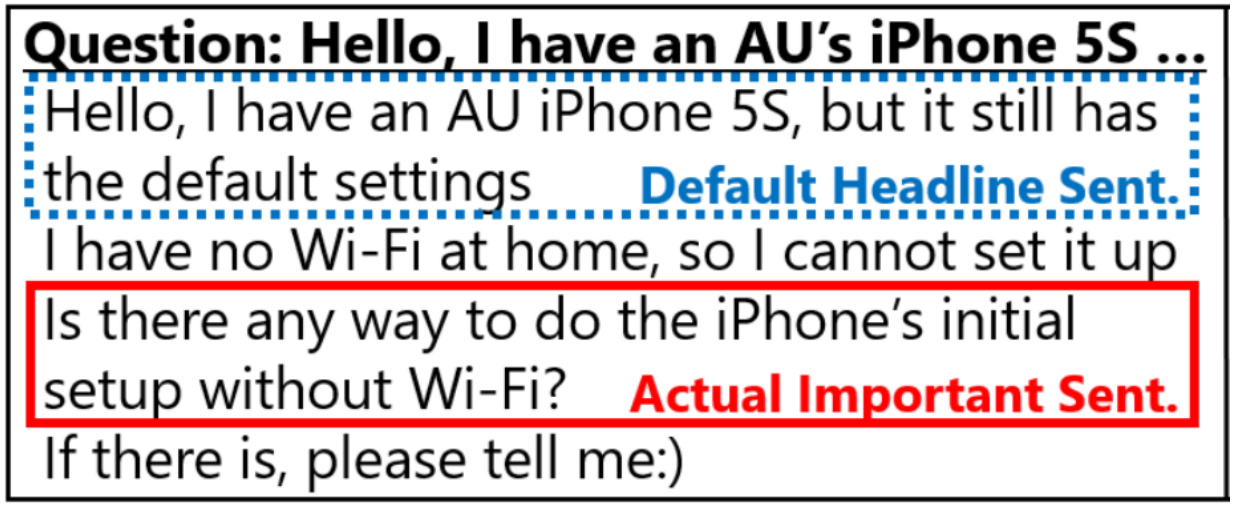
Tokyo Tech

Motivation:

Community question answering (CQA) sites tend to display the first sentence as the headline.

😣 They do not necessarily represent the most important.

→ Users face difficulties to efficiently search questions.



Task: Input : question post

Output: single-sentence summary

Approaches for Question Summarization

- Previous approaches are mostly supervised.
  - Feature-based classifiers [Tamura+2005];
  - Learning-to-rank approach [Higurashi+2018]
    😣 costly annotated data

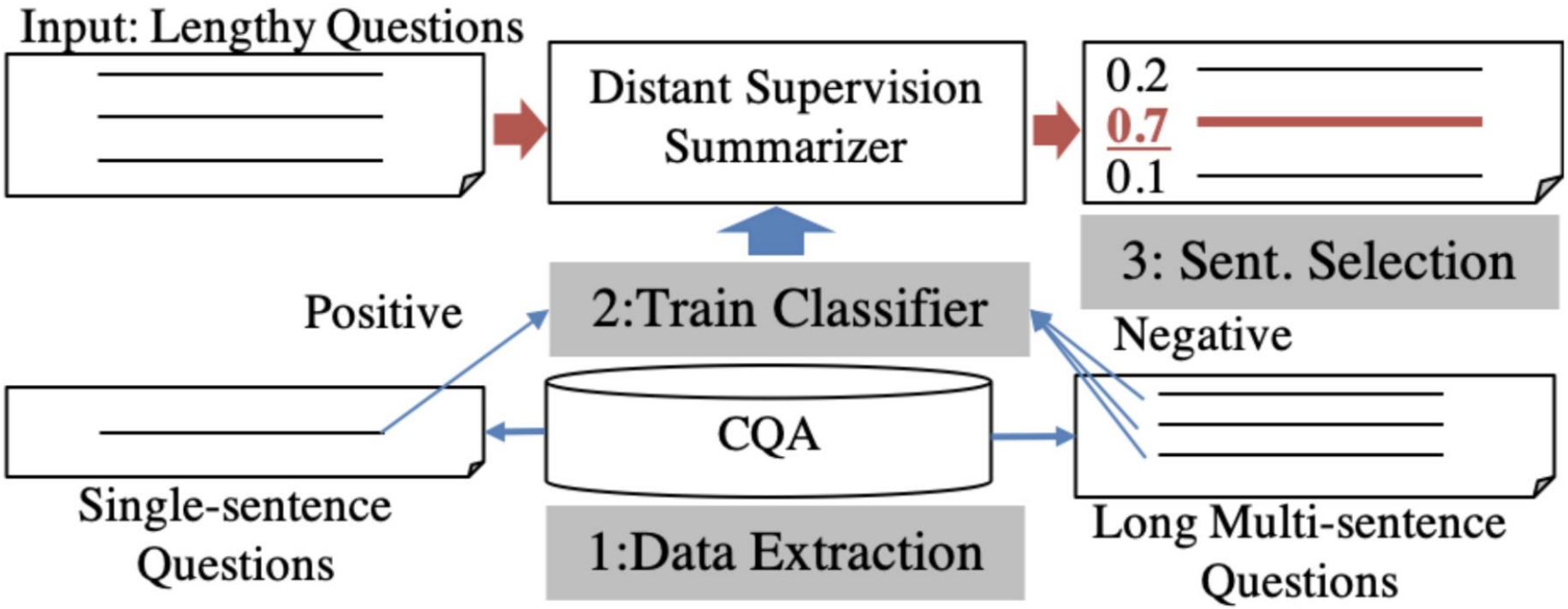
- Unsupervised models also can be applied to this task.
  - Graph-based model, e.g. LexRank [Erkan+2004]
  - Embedding similarity-based model [Kobayashi+2015]

We propose a distant supervision-based approach for this task.

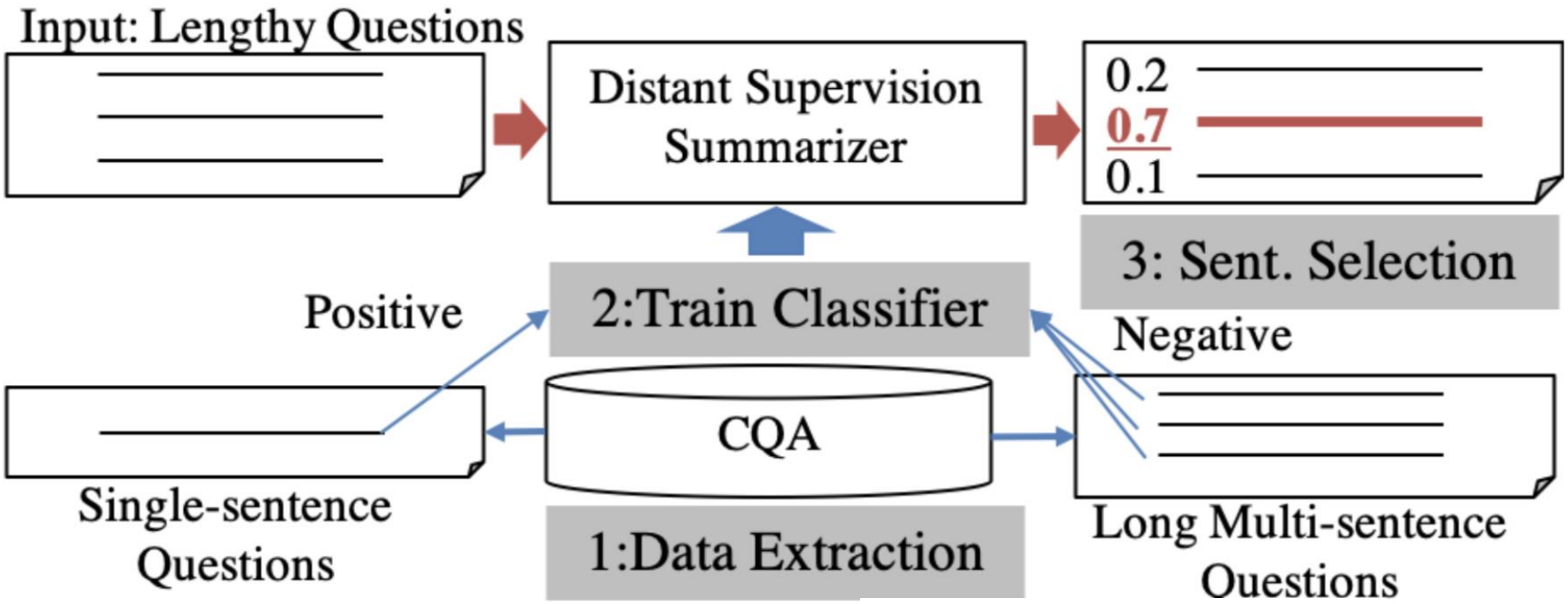
Automatically labels training instances by using rules/heuristics.
[Mints+2009]

We compare our distant supervision-based models with various supervised and unsupervised models.

Our Model


Input: Lengthy Questions
Distant Supervision Summarizer
0.2
0.7
0.1
3: Sent. Selection
Positive
2:Train Classifier
Negative
CQA
Single-sentence Questions
1:Data Extraction
Long Multi-sentence Questions

Our Model



Single-sent Q

Where can I fix my broken iPhone? (POSITIVE)

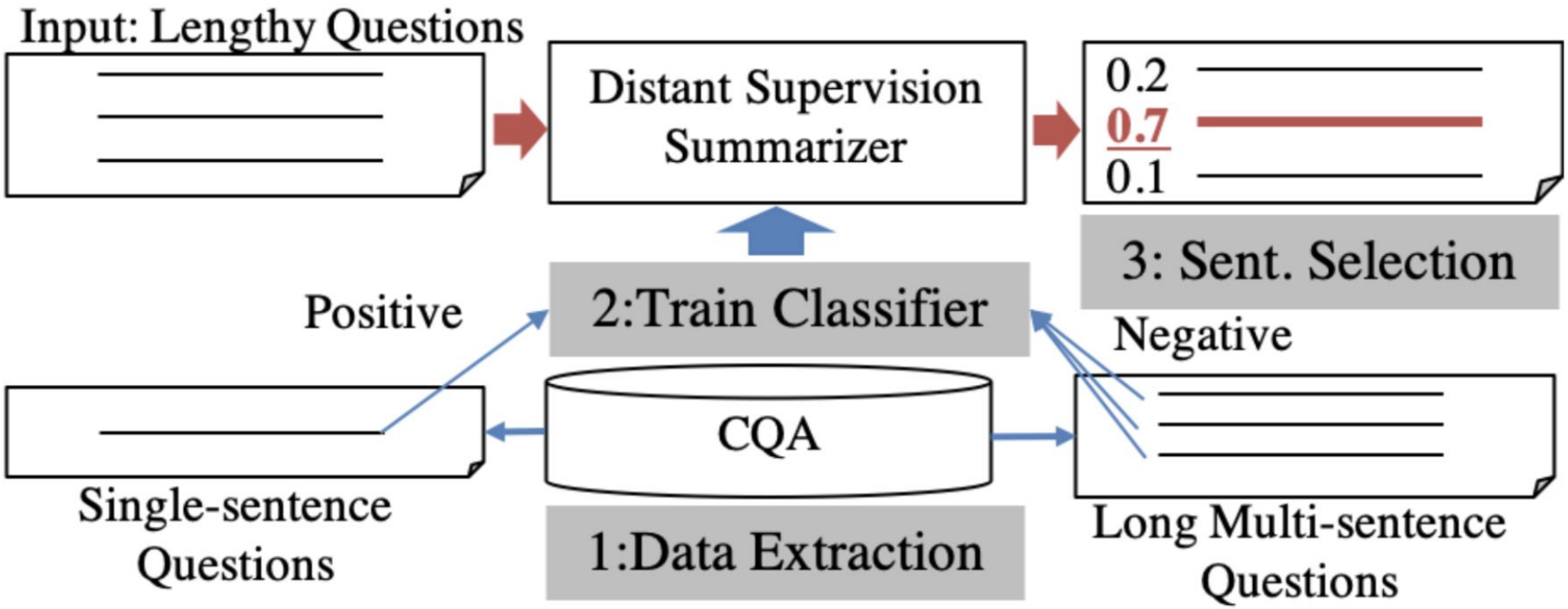
→ Summary-like property!

Long multi-sents Q

Hi! (NEGATIVE)
My iPhone has been broken.(NEGATIVE)
I live in Tokyo. (NEGATIVE)
….
Does anyone know good repair shop? (NEGATIVE)

→ Non summary-like property!

Our Model



Step2 and Step3

We train a classifier that determines how likely a sentence is a single-sentence question.

The confidence scores from the classifier are used as importance scores for sentences.

Experiments

- Datasets:
  - Pseudo: 800K positive and 1.7M negative sentences.
    (automatically annotated by our framework.)
  - Label: Manually annotated 12K sentences.

- Evaluation:
  - Label is used for the evaluation data.
  - Accuracy Measure
    (# of correctly predicted questions/ # of total questions)
  - Supervised models are evaluated by 5-fold cross validation
- Results:
  1. Our approaches outperformed unsupervised baselines.
  2. Our approaches performed competitively with supervised baselines.

	Greedy
DistNet	**87.38**
DistReg	86.17
Lead	81.79
LexRank	78.49
SimEmb	59.46
TfIdf	52.03
SupNet	81.67
SupReg	87.89

Conclusion

- We proposed a distant supervision–based approach for extractive question summarization.
- Our models performed competitively with supervised models even without manually annotated data.
- Our data will be publicly available.
- Please come to the slack channel if you have any questions! Please also check our another accepted paper!
  - **Semi-Supervised Extractive Question Summarization Using Question-Answer Pairs.** Kazuya Machida, Tatsuya Ishigaki, Hayato Kobayashi, Hiroya Takamura and Manabu Okumura