RICT at the NTCIR-14 QALab-PoliInfo Task

Jiawei Yong, Shintaro Kawamura, Katsumi Kanasaki,
Shoichi Naitoh, and Kiyohiko Shinomiya

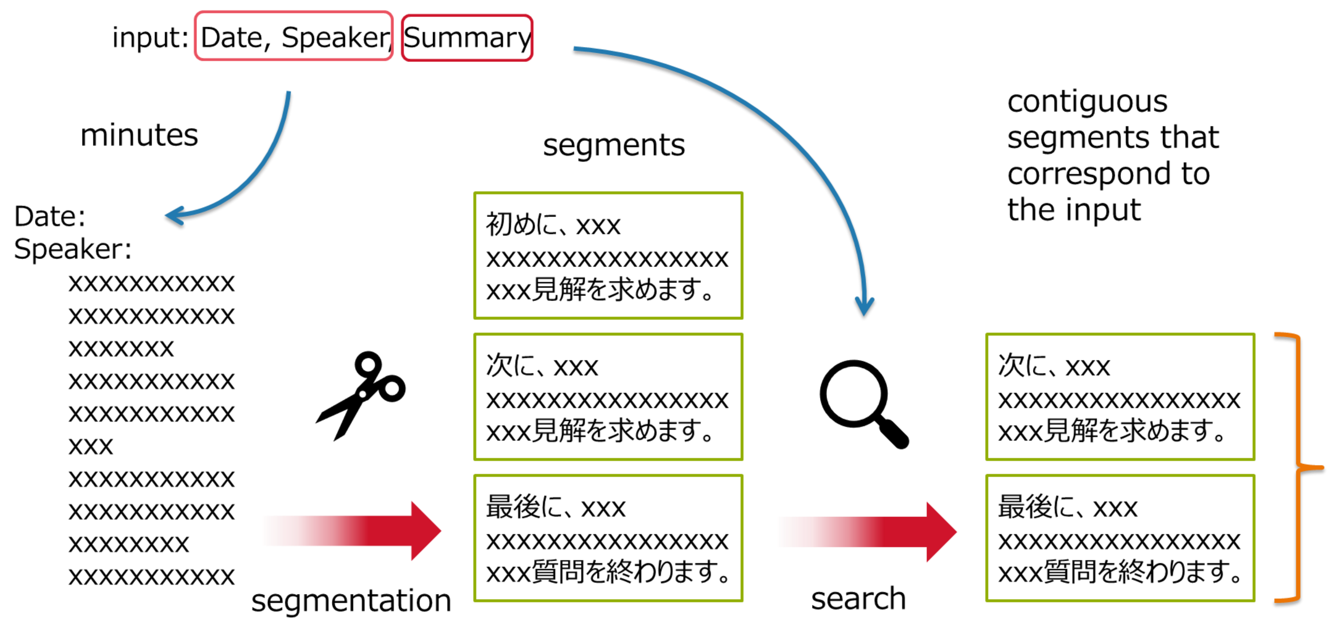
Ricoh Company, Ltd.

RICOH imagine. change.

Segmentation

Segmentation subtask in 2 steps

1. The segmentation step splits the assembly utterances into segments.
2. The search step finds a sequence of segments that corresponds to a given summary.

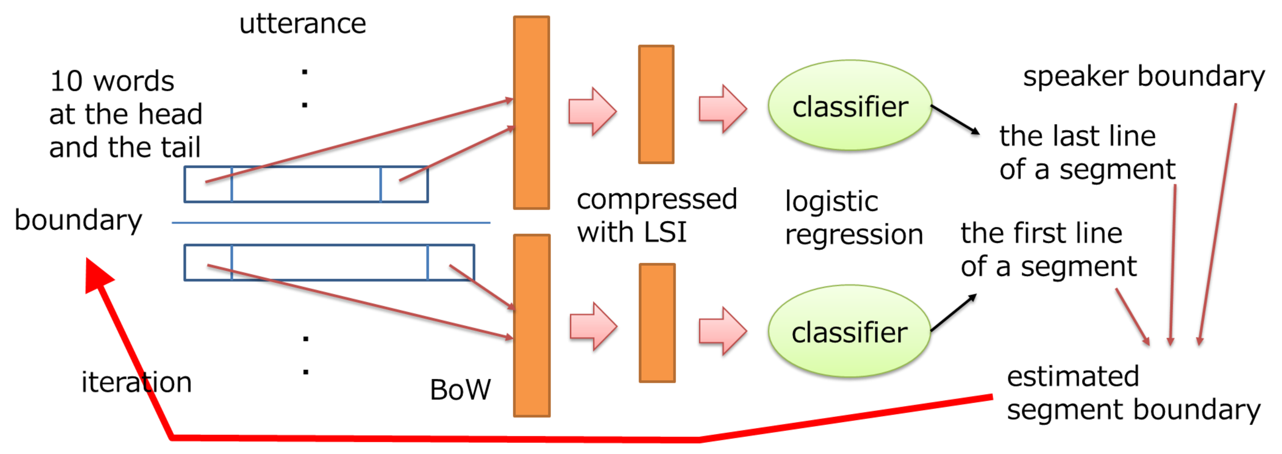


Data set preparation

- The training data set provided by the task organizer was used as the development data for the search step.
- Additional training/development data sets were prepared for the segmentation step.

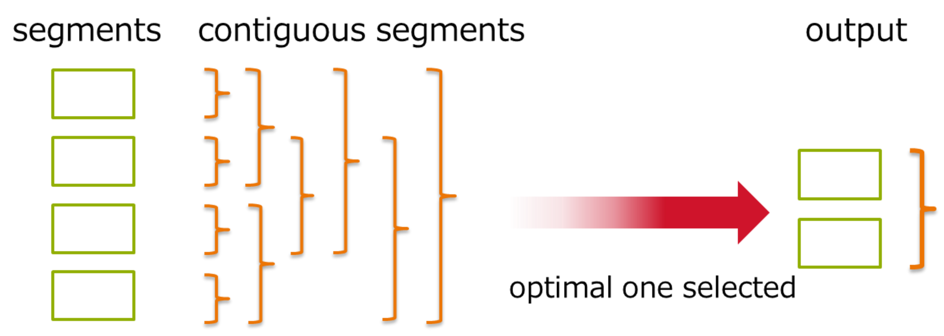
Cue-phrase-based segmentation step

- Tried methods include a novel semi-supervised method, which iteratively learns features of cue phrases at segment boundaries through bootstrapping.



Probability-model-based search step

- The sequence of segments that maximizes $\sum_{i=1}^{k} idf(t_i) - \lambda k \log(n)$ is selected, where the sequence contains words $t_i (i = 1, \ldots, k)$ in the summary, and the sequence consists of n utterances.



$\lambda = 0.4$ for questions and 0.7 for answers (determined by the development data)

Evaluation results

- The rule-based segmentation was the best during the formal run (rank 1 in F1). The method using a hierarchical attention network, which was tried after the formal run, also shows good performance.

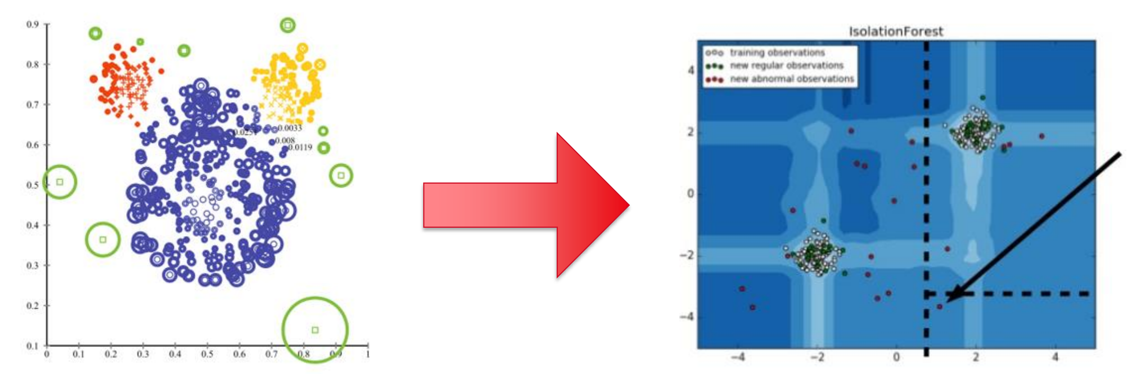
Segmentation method	Question			Answer		
	Recall	Precision	F1	Recall	Precision	F1
rule-based (pattern matching)	0.851	0.913	0.881	0.949	0.903	0.925
BoW + SVM	0.819	0.851	0.834	0.913	0.939	0.925
pre-trained word2vec + LSTM	0.916	0.690	0.780	0.909	0.925	0.914
word embeddings + HAN	0.871	0.874	0.873	0.949	0.921	0.934
semi-supervised	0.836	0.760	0.796	0.907	0.814	0.858
no segmentation	0.828	0.715	0.767	0.680	0.839	0.751

Classification

Challenge1: Imbalance

Relevance(“1”):Irrelevance(“0”) = 9390:901 = 10:1

1. We regard Majority class as normal data, minority class as outlier.
2. We have constructed a tree by random feature selection, and then divided it into several sub trees. The average value of “shallowness" is regarded as the final abnormal score (threshold).

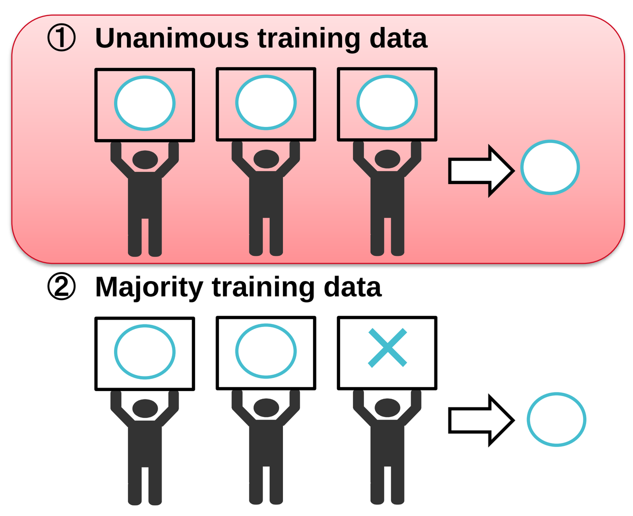


outlier detection Isolation Forest

Challenge2: Low Kappa Statistic

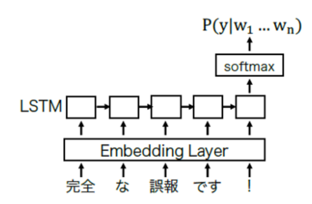
The low kappa statistic shows that there are different understandings about labeling thus leading to a doubt of label correctness.

1. **Unanimous training data** prefers to quality rather than quantity. Here we only pick up the data with same labels from all workers.
2. **Majority training data** places a higher value on quantity than quality.



Learning Model

News Detection Support for Fact Check(NLP2018)



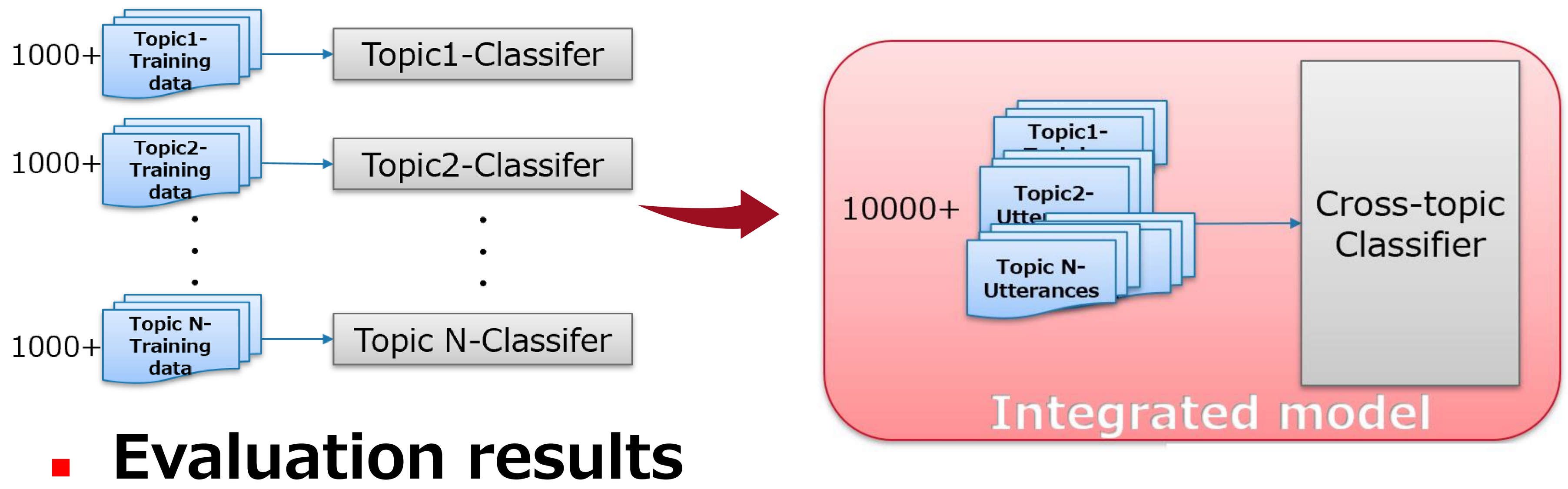
LSTM Classifier

Suspicious News Detection Using Micro Blog Text (2018)

SVM Classifier

Challenge3: Underfitting

Learning for each individual topic would easily cause underfitting problem. Here we apply integrated model.



Evaluation results

The high F1 scores indicate that the above problems have been alleviated effectively, but still need some more tuning.

<table>
<tr><th rowspan="2">Classification Subtasks</th><th colspan="7">Top Values of RICT Runs for each criteria</th></tr>
<tr><th>Accuracy</th><th>1-Recall</th><th>1-Precision</th><th>1-F1</th><th>0-Recall</th><th>0-Precision</th><th>0-F1</th></tr>
<tr><td>1. Relevance</td><td>0.857 (rank 7)</td><td>0.99</td><td>0.865</td><td>0.923 (rank 7)</td><td>0.524</td><td>0.332</td><td>0.406 (rank 2)</td></tr>
<tr><td>2. Fact-checkability</td><td>0.729 (rank 3)</td><td>0.693</td><td>0.476</td><td>0.564 (rank 3)</td><td>0.899</td><td>0.738</td><td>0.811 (rank 3)</td></tr>
<tr><td rowspan="3">3. Stance</td><td rowspan="3">0.808 (rank 1)</td><td>0.295</td><td>0.63</td><td>0.40 (rank 3)</td><td rowspan="3">0.962</td><td rowspan="3">0.827</td><td rowspan="3">0.889 (rank 2)</td></tr>
<tr><th>2-Recall</th><th>2-Precision</th><th>2-F1</th></tr>
<tr><td>0.194</td><td>0.579</td><td>0.290 (rank 4)</td></tr>
</table>