Autoencoder as Assistant Supervisor: Improving Text Representation for Chinese Social Media Text Summarization

Shuming Ma (speaker), Xu Sun, Junyang Lin, Houfeng Wang
Peking University, Beijing, China
### Source:
昨晚，中联航空成都飞北京一架航班被发现有多人吸烟。后因天气原因，飞机备降太原机场。有乘客要求重新安检，机长决定继续飞行，引起机组人员与未吸烟乘客冲突。
Last night, several people were caught to smoke on a flight of China United Airlines from Chendu to Beijing. Later the flight temporarily landed on Taiyuan Airport. Some passengers asked for a security check but were denied by the captain, which led to a collision between crew and passengers.

### Summary:
航班多人吸烟机组人员与乘客冲突。
Several people smoked on a flight which led to a collision between crew and passengers.
Source:
昨晚，中联航空成都飞北京一架航班被发现有多人吸烟。后因天气原因，飞机备降太原机场。有乘客要求重新安检，机长决定继续飞行，引起机组人员与未吸烟乘客冲突。
Last night, several people were caught to smoke on a flight of China United Airlines from Chendu to Beijing. Later the flight temporarily landed on Taiyuan Airport. Some passengers asked for a security check but were denied by the captain, which led to a collision between crew and passengers.

Summary:
航班多人吸烟机组人员与乘客冲突。
Several people smoked on a flight which led to a collision between crew and passengers.
Source:
昨晚，中联航空成都飞北京一架航班被发现有多人吸烟。后因天气原因，飞机备降太原机场。有乘客要求重新安检，机长决定继续飞行，引起机组人员与未吸烟乘客冲突。

Last night, several people were caught to smoke on a flight of China United Airlines from Chendu to Beijing. Later the flight temporarily landed on Taiyuan Airport. Some passengers asked for a security check but were denied by the captain, which led to a collision between crew and passengers.

Summary:
航班多人吸烟机组人员与乘客冲突。
Several people smoked on a flight which led to a collision between crew and passengers.
Method: Autoencoder as a Assistant Supervisor

Step 1:
Build a Seq2Seq and Autoencoder

Step 2:
Supervise with reference summaries

Step 3:
Supervise Seq2Seq with Autoencoder

\[ L_s = \lambda \| z_t - z_s \|_2 \]
Method: Autoencoder as a Assistant Supervisor

Source content encoder

Summary encoder

Testing Stage:
Only using Seq2Seq
Method: Adversarial Learning

Source content encoder

Summary encoder

Discriminator

Discriminator:
Identify the representation of the autoencoder and the seq2seq
\[
L_D(\theta_D) = -\log P_{\theta_D}(y = 1|z_t) - \log P_{\theta_D}(y = 0|z_s)
\]
Experiments

Dataset

Large Scale Chinese Social Media Text Summarization Dataset (LCSTS): The dataset consists of more than 2,400,000 text-summary pairs, constructed from a famous Chinese social media website called Sina Weibo.

Evaluation Metrics

ROUGE score: The metrics compare an automatically produced summary with the reference summaries, by computing overlapping lexical units, including unigram, bigram, trigram, and longest common subsequence (LCS). We use ROUGE-1 (unigram), ROUGE-2 (bi-gram) and ROUGE-L (LCS).
## Experiments: Results

<table>
<thead>
<tr>
<th>Models</th>
<th>R-1</th>
<th>R-2</th>
<th>R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN-W (Hu et al. 2015)</td>
<td>17.1</td>
<td>8.5</td>
<td>15.8</td>
</tr>
<tr>
<td>RNN (Hu et al. 2015)</td>
<td>21.5</td>
<td>8.9</td>
<td>18.6</td>
</tr>
<tr>
<td>RNN-cont-W (Hu et al. 2015)</td>
<td>26.8</td>
<td>16.1</td>
<td>24.1</td>
</tr>
<tr>
<td>RNN-cont (Hu et al. 2015)</td>
<td>29.9</td>
<td>17.4</td>
<td>27.2</td>
</tr>
<tr>
<td>SRB (Ma et al. 2017)</td>
<td>33.3</td>
<td>20.0</td>
<td>30.1</td>
</tr>
<tr>
<td>CopyNet-W (Gu et al. 2016)</td>
<td>35.0</td>
<td>22.3</td>
<td>32.0</td>
</tr>
<tr>
<td>CopyNet (Gu et al. 2016)</td>
<td>34.4</td>
<td>21.6</td>
<td>31.3</td>
</tr>
<tr>
<td>RNN-dist (Chen et al. 2016)</td>
<td>35.2</td>
<td>22.6</td>
<td>32.5</td>
</tr>
<tr>
<td>DRGD (Li et al. 2017)</td>
<td>37.0</td>
<td>24.2</td>
<td>34.2</td>
</tr>
<tr>
<td>Seq2seq (our implementation)</td>
<td>32.1</td>
<td>19.9</td>
<td>29.2</td>
</tr>
<tr>
<td>+superAE (this paper)</td>
<td>39.2</td>
<td>26.0</td>
<td>36.2</td>
</tr>
<tr>
<td>w/o adversarial learning</td>
<td>37.7</td>
<td>25.3</td>
<td>35.2</td>
</tr>
</tbody>
</table>
Experiments: Analysis of Text Representation

<table>
<thead>
<tr>
<th>Models</th>
<th>2-class (%)</th>
<th>5-class (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq2seq</td>
<td>80.7</td>
<td>65.1</td>
</tr>
<tr>
<td>+superAE</td>
<td>88.8(+8.1)</td>
<td>71.7(+6.6)</td>
</tr>
</tbody>
</table>

Accuracy of the sentiment classification on the Amazon dataset. Our superAE model outperforms Seq2seq with a large margin of 8.1% and 6.6%.
Conclusion

- The autoencoder, as a supervisor of the sequence-to-sequence model, can learn a better internal representation for abstractive summarization.
- The adversarial learning approach is able to further improve the supervision of the autoencoder.
- Experimental results show that our model outperforms the sequence-to-sequence baseline by a large margin, and achieves the state-of-the-art performances on a Chinese social media dataset.
Thank you!

The code is available at https://github.com/lancopku/superAE