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

Chinese Social Media Text Summarization

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.

Chinese Social Media Text Summarization

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.

Chinese Social Media Text Summarization

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。

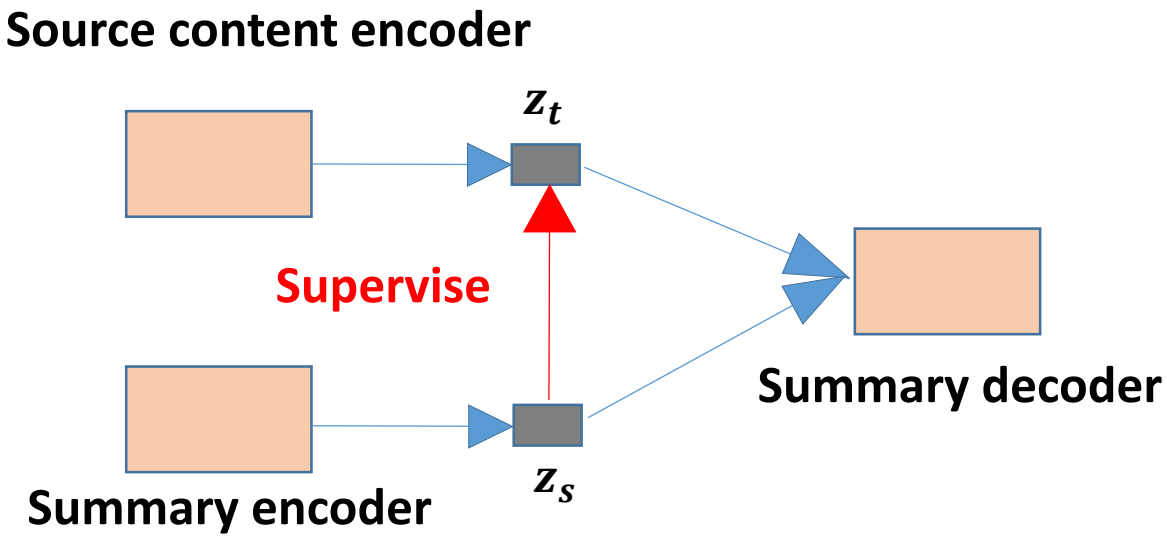
Last night, several people were caught to smoke on a flight of China United Airlines from Chendu to Beijing. Later the flight temporarily landed on Taiyuan Airport. Some passengers asked for a security check but were denied by the captain, which led to a collision between crew and passengers.

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:

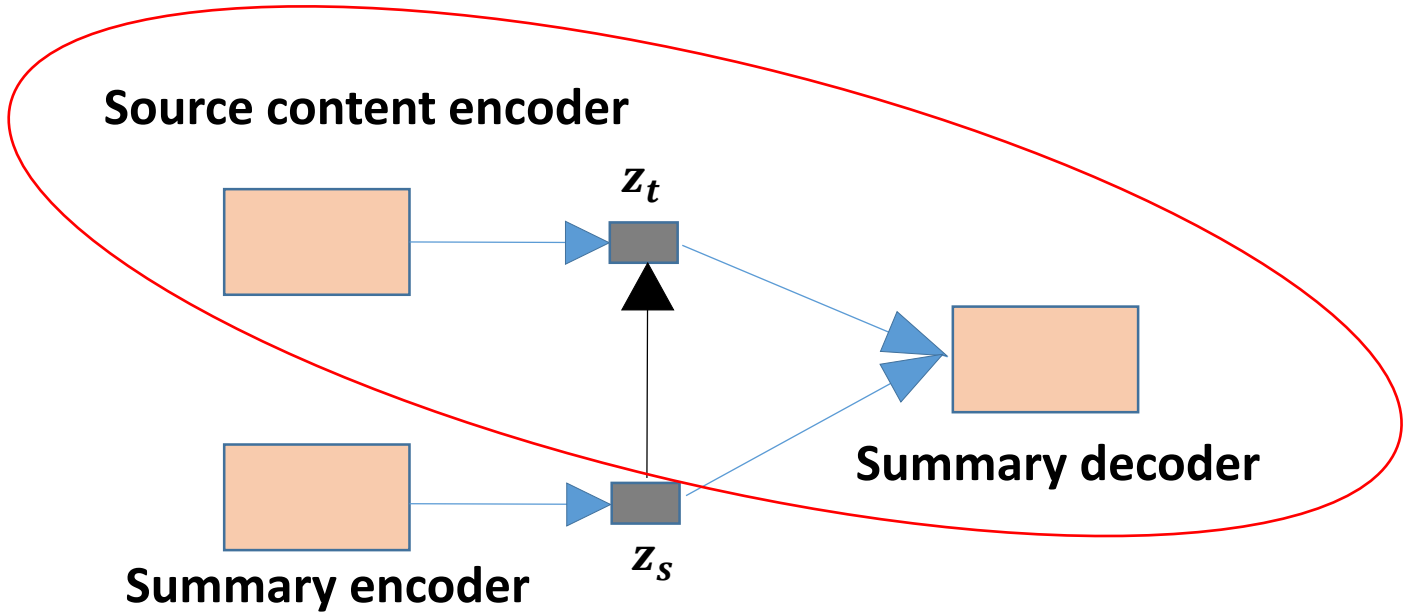
Supervise with reference summaries

Step 3:

Supervise Seq2Seq with Autoencoder

$$L_s = \lambda \|z_t - z_s\|_2$$

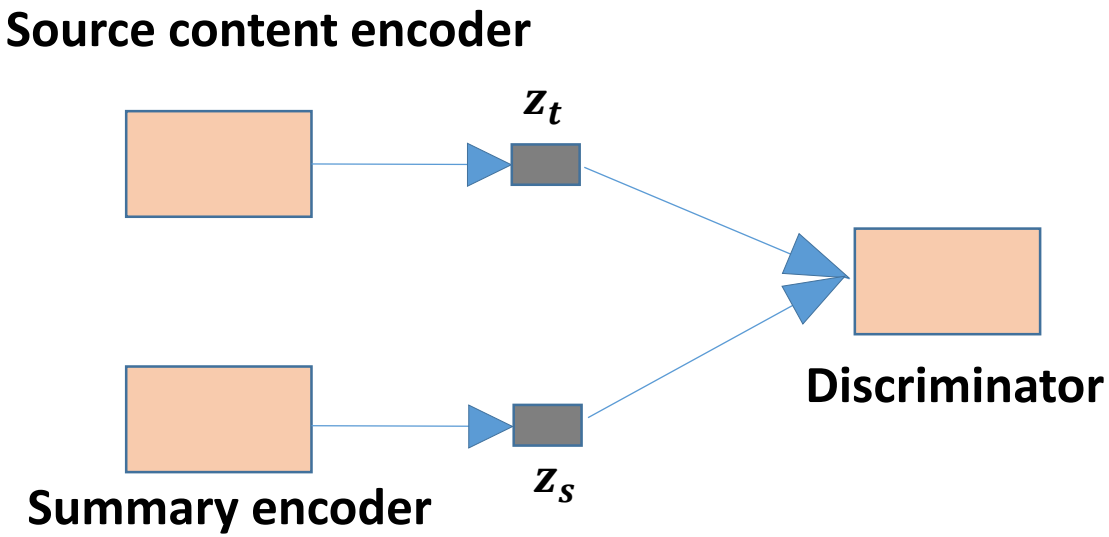
Method: Autoencoder as a Assistant Supervisor



Testing Stage:

Only using Seq2Seq

Method: Adversarial Learning



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.

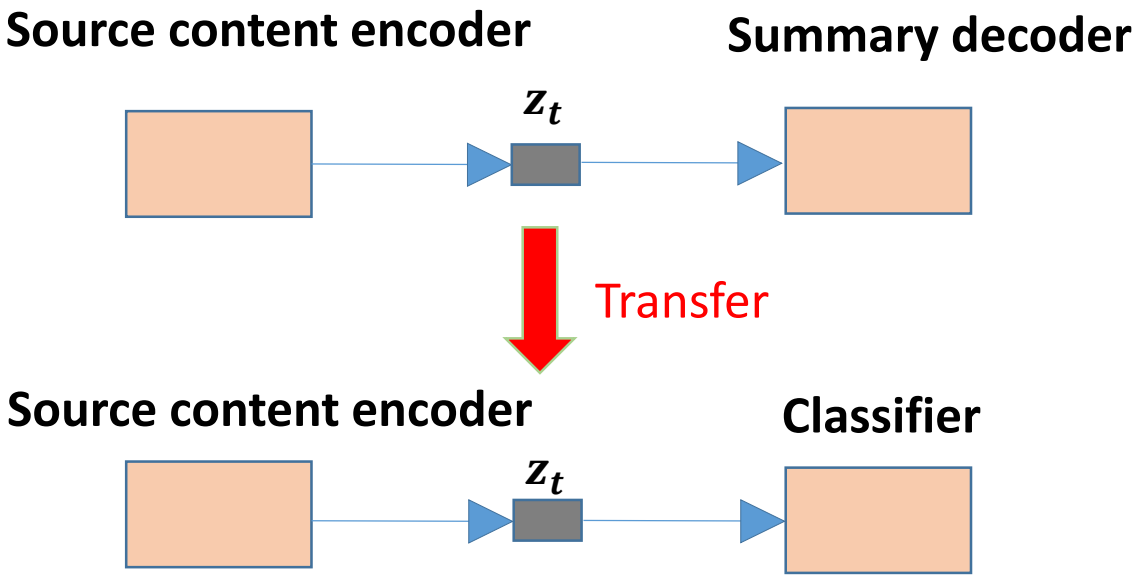
Evaluation Metrics

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

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

Experiments: Analysis of Text Representation



Models	2-class (%)	5-class (%)
Seq2seq	80.7	65.1
+superAE	88.8(+8.1)	71.7(+6.6)

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

北京大学 PEKING UNIVERSITY