Pretraining Sentiment Classifiers with Unlabeled Dialog Data

Jul. 18, 2018

Toru Shimizu*¹, Hayato Kobayashi*¹,*², Nobuyuki Shimizu*¹
*¹Yahoo Japan Corporation, *²RIKEN AIP
Problem

• The amount of labeled training data
  – You will need at least 100k training records to surpass classical approaches (Hu+ 2014, Wu+ 2014)
  – Large-scale labeled datasets of document classification

| Dataset                              | training | validation | test  | total  |
|--------------------------------------|----------|------------|-------|--------|
| Stanford Sentiment Tree Bank         | 8,544    | 1,101      | 2,210 | 11,855 |
| Large Movie Review Dataset           | 25,000   | -          | 25,000| 50,000 |
| SemEval 2014 Task 9 Subtask B        | 9,684    | 1,654      | 5,666 | 17,004 |
Previous Work

- Semi-supervised approaches
  - Language model

```
how are you </s>
LSTM-RNN
<s> how are you
pretraining

LSTM-RNN

congratulations!
positive
fine-tuning
```
Previous Work

• Semi-supervised approaches
  – Sequence autoencoder (Dai and Le 2015)
Our Contributions

• Pretraining strategy with *unlabeled dialog data*
  – Pretrain an encoder-decoder model for sentiment classifiers

• Outperform other semi-supervised methods
  – Language model
  – Sequence autoencoder
  – Distant supervision with emoji and emoticons

• Case study based on...
  – Costly labeled sentiment dataset of 99.5K items
  – Large-scale unlabeled dialog dataset of 22.3M utterance-response pairs
Key Idea

- Emotional conversations in a dialog dataset

| Utterance                  | Response                      |
|---------------------------|-------------------------------|
| Good luck                 | Thank you!                    |
| I won't forgive you, never | °・(ノД`)・°・° (crying emoticon) |
| I got home really tired   | Good job today!               |

- Implicitly learn sentiment-handling capabilities through learning a dialog model
Overview of the Proposed Method

• Datasets
  – Large-scale dialog corpus: a set of a large number of unlabeled utterance-response tweet pairs
  – Labeled dataset: a set of a moderate number of tweets with a sentiment label

• Pretraining

• Fine-tuning
Data Preparation

• Dialog data
  – Extract 22.3M pairs of an utterance tweet and its response tweet from Twitter Firehose data

|                      | training  | validation | test  | total     |
|----------------------|-----------|------------|-------|-----------|
| Dialog data          | 22,300,000| 10,000     | 50,000| 22,360,000|

• Sentiment data
  – Positive: 15.0%, Negative: 18.6%, Neutral 66.4%

|                      | training  | validation | test  | total     |
|----------------------|-----------|------------|-------|-----------|
| Sentiment data       | 80,591    | 4,000      | 15,000| 99,591    |
Model: Dialog Model

- Dialog model
  - One-layer LSTM-RNN encoder-decoder
  - Embedding layer: 4000 tokens, 256 elements
  - LSTM: 1024 elements
  - Representation which encoder gives: 1024 elements
  - Decoder's readout layer: 256 elements
  - Decoder's output layer: 4000 tokens
  - LSTMs of the encoder and decoder share the parameter
Model: Dialog Model

**Encoder**

- $\phi^{\text{enc}}$
- Recurrent layer $h_t^{\text{enc}}$
- Embedding layer
- Token ID $u_t$

**Decoder**

- $\psi^{\text{dec}}$
- Output layer $o_t$
- Readout layer
- Recurrent layer $h_t^{\text{dec}}$
- Embedding layer
- Token ID $x_t$

- $\alpha^{\text{dec}}$

**Network Structure**

- Encoder RNN
- Decoder RNN
Classification model

- The architecture of the encoder RNN part is identical to that of the dialog model
- Produce a probability distribution over sentiment classes by a fully-connected layer and softmax function
Training: Dialog Model

- Model pretraining with the dialog data
  - MLE training objective
  - 1 GPU (7 TFLOPS)
  - 5 epochs = 15.9 days
  - Batch size: 64
  - Optimizer: ADADELTA
  - Apply gradient clipping
  - Evaluate validation costs 10 times per epoch and pick up the best model
  - Theano-based implementation
Training: Classification Model

• Classifier model training with the sentiment data
  – Apply 5 different data sizes for each method
    • 5k, 10k, 20k, 40k, 80k (all)
  – 5 runs for each method/data size with varying random seeds
  – Evaluate the results by the average of f-measure scores
  – Adjust the duration so that the cost surely converges
    • Pretrained models converge very quickly but those trained from scratch converge slowly
  – The other aspects are the same with pretraining
Proposed Method

• The proposed method: Dial

[Diagram showing a sequence of steps involving LSTM-RNNs:
- Transfer learning from dialog data (how are you)
- Pretraining
- Fine-tuning with sentiment data (congratulations!)]
Baselines with LSTM-RNNs

- **Default**
  - No pretraining
  - Directly trained by the sentiment data

![Diagram](image.png)
Baselines with LSTM-RNNs

- **Lang**
  - Pretrain an LSTM-RNNs as a language model

```
I'm great! </s>

LSTM-RNN

<s> I'm great

transfer

pretraining

LSTM-RNN

congratulations!

positive

fine-tuning

unpaired tweet data

sentiment data
```
• **SeqAE**
  - Pretrain an LSTM-RNNs as a sequence autoencoder (Dai and Le 2015)
• Emoji and emoticon-based distant supervision
  – Prepare large-scale datasets utilizing emoticons or emoji as pseudo labels (Go+ 2009)
  – Positive emoticon examples
    • 😊😊😊😊😊❤️👍❤️❤️❤️✨ (^_^) (∩_∩) (.DrawLine) o(^-^)o
  – Negative emoticon examples
    • 😞😡😢😭😢😢😢❤️❤️❤️(ТДТ) (‘^ ’ *) (／-- ) (．шло) ( ‘ △ ’) orz
Baselines with LSTM-RNNs

- **Emo2M** and **Emo6M**
  - Pretrain models as classifier models using pseudo-labeled data

![Diagram showing pretraining and fine-tuning stages for LSTM-RNNs]

- **I'm tired 😞**
  - Transfer learning
  - Pretraining

- **congratulations!**
  - Fine-tuning

- **negative**
- **positive**
- **pseudo-labeled data**
- **sentiment data**
Baselines with Linear Models

- **Data**
  - Use only the sentiment data

- **Preprocessing**
  - Segment text with a defact-standard morphological analyzer, MeCab
  - 50,000 unigrams and 50,000 bigrams
  - +233 emoji and emoticons

- **LogReg**
  - Logistic regression (LIBLINEAR)

- **LinSVM**
  - Linear SVM (LIBLINEAR)
Results: F-measure

![Graph showing F-measure vs. number of training records for different models. The graph includes lines for Default, Dial, Lang, SeqAE, Emo2M, Emo6M, LogReg, and LinSVM. The x-axis represents the number of training records ranging from 5000 to 80000, and the y-axis represents the F-measure ranging from 0.45 to 0.75. Each model has a distinct line color and marker.](image-url)
| Source tweet                                      | Generated reply | Source tweet                                      | Generated reply |
|--------------------------------------------------|-----------------|--------------------------------------------------|-----------------|
| 明日は待ちに待ったコンサートだよ                   | いいね！         | Tomorrow I have a concert I’ve been really looking forward to | That’s nice!    |
| 私もっと（'∀`)人（'∀`) (%)♪                          | (*^__^*)        | Me too ♪ (high five emoticon) ♪                  | :)              |
| 残念だったね                                         | (’・ω・`)       | I’m sorry to hear that                            | :(              |
| 後でそっちに行くよ                                   | おっけー！       | I’m coming later                                  | OK!             |
| 頭痛いよ                                            | うそ、お大事に… | I have a headache                                 | Really? Take care of yourself... |
| アメトーク見たかった～                                | おもろいよね～   | I missed Ame Talk (a TV program)                  | Watching it is fun |
| もう、ごめんじゃ済まされないだろう、呆れる              | それはそれで困る。。 | Sorry doesn’t cut it anymore. I gave up on you. | That’s too bad... |
| 大学合格したよ！                                     | おめでとう！！ | I was admitted by the university!                | Congratulations!! |
| もうだめだ                                          | そんなことないよ（’・ω・’） | It’s all over for me                            | I don’t think so :( |
| 嘘つきめ。                                          | ひどい           | You liar.                                        | You nasty       |
| ちょうどいいね                                        | まじかーありがとう！！！ | That’s just right                                | Really? Thanks!!! |
| それ、すごい好き                                     | うん、きっかけよいね | I really like it                                 | Yeah, it’s so cool |

Replies generated by the pretrained encoder-decoder model
Conclusion

• Effectiveness of the pretraining strategy using paired dialog data for sentiment analysis
  – Even more effective in extremely low-resource situations
  – Character-based processing

• Future work
  – Explore combinations of a large-scale unlabeled dataset and a supervised task
  – Exploit other kinds of structures