Pretraining Sentiment Classifiers with Unlabeled Dialog Data

Jul. 18, 2018

Toru Shimizu*, Hayato Kobayashi*, Nobuyuki Shimizu*

*1 Yahoo Japan Corporation, *2 RIKEN 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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>training</th>
<th>validation</th>
<th>test</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford Sentiment Tree Bank</td>
<td>8,544</td>
<td>1,101</td>
<td>2,210</td>
<td>11,855</td>
</tr>
<tr>
<td>Large Movie Review Dataset</td>
<td>25,000</td>
<td>-</td>
<td>25,000</td>
<td>50,000</td>
</tr>
<tr>
<td>SemEval 2014 Task 9 Subtask B</td>
<td>9,684</td>
<td>1,654</td>
<td>5,666</td>
<td>17,004</td>
</tr>
</tbody>
</table>
Previous Work

• Semi-supervised approaches
  – Language model

![Diagram showing semi-supervised approaches with LSTM-RNN models and pretraining and fine-tuning stages.]
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**
  
  - Good luck
  
  - I won't forgive you, never 😑
  
  - I got home really tired
  
  **Response**
  
  - Thank you!
  
  - •°・(ノД`)・°・° (crying emoticon)
  
  - 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

<table>
<thead>
<tr>
<th></th>
<th>training</th>
<th>validation</th>
<th>test</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialog data</td>
<td>22,300,000</td>
<td>10,000</td>
<td>50,000</td>
<td>22,360,000</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>training</th>
<th>validation</th>
<th>test</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentiment data</td>
<td>80,591</td>
<td>4,000</td>
<td>15,000</td>
<td>99,591</td>
</tr>
</tbody>
</table>
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**

- $\phi^{enc}$
- Recurrent layer $h^{enc}_t$
- Embedding layer
- Token ID $u_t$

**Decoder**

- $\psi^{dec}$
- Readout layer
- Recurrent layer $h^{dec}_t$
- Embedding layer
- Token ID $x_t$

- Output layer $o_t$
Model: Classification Model

• 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
The proposed method: Dial

- Dialog data
- Sentiment data
Baselines with LSTM-RNNs

- **Default**
  - No pretraining
  - Directly trained by the sentiment data
Baselines with LSTM-RNNs

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

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

```
LSTM-RNN
```

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

transfer

```
LSTM-RNN
```

```
congratulations!
```

pretraining

```
LSTM-RNN
```

```
positive
```

fine-tuning

unpaired tweet data

sentiment data
Baselines with LSTM-RNNs

• **SeqAE**
  - Pretrain an LSTM-RNNs as a sequence autoencoder (Dai and Le 2015)
Baselines with LSTM-RNNs

- Emoji and emoticon-based distant supervision
  - Prepare large-scale datasets utilizing emoticons or emoji as pseudo labels (Go+ 2009)
  - Positive emoticon examples
    - 😊 😏 😍 😍 😞 😏 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍 😍
• **Emo2M** and **Emo6M**
  – Pretrain models as classifier models using pseudo-labeled data

![Diagram showing pretraining and fine-tuning with LSTM-RNNs](image)
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
<table>
<thead>
<tr>
<th>Original text</th>
<th>Generated reply</th>
<th>English translations</th>
</tr>
</thead>
<tbody>
<tr>
<td>明日は待ちに待ったコンサートだよ</td>
<td>いいね！</td>
<td>Tomorrow I have a concert I’ve been really looking forward to</td>
</tr>
<tr>
<td>私もっ🎵（’∀’）人（’∀’）🎵</td>
<td>(<em>^_^</em>)</td>
<td>Me too 🎵 (high five emoticon) 🎵</td>
</tr>
<tr>
<td>残念だったね</td>
<td>(´・ω・`)</td>
<td>I’m sorry to hear that</td>
</tr>
<tr>
<td>後でそっちに行くよ</td>
<td>おっけー！</td>
<td>I’m coming later</td>
</tr>
<tr>
<td>頭痛いよ</td>
<td>うそっ、お大事に…</td>
<td>I have a headache</td>
</tr>
<tr>
<td>アメトーク見たかった～</td>
<td>おもう〜いよね〜</td>
<td>I missed Ame Talk (a TV program)</td>
</tr>
<tr>
<td>もう、ごめんじゃ済まされないだろ、呆れる</td>
<td>それはそれで困る。</td>
<td>Sorry doesn’t cut it anymore. I gave up on you.</td>
</tr>
<tr>
<td>大学合格したよ！</td>
<td>おめでとう！！</td>
<td>I was admitted by the university!</td>
</tr>
<tr>
<td>もうだめだ</td>
<td>そんなことないよ(´・ω･`)</td>
<td>It’s all over for me</td>
</tr>
<tr>
<td>嘘つきめ。</td>
<td>ひどい</td>
<td>You liar.</td>
</tr>
<tr>
<td>ちょっといいね</td>
<td>まじかーありがとうございます！！！</td>
<td>That’s just right</td>
</tr>
<tr>
<td>それ、すごい好き</td>
<td>うん、かっこいいよね</td>
<td>I really like it</td>
</tr>
</tbody>
</table>

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