Cross-View Training

This repository contains code for *Semi-Supervised Sequence Modeling with Cross-View Training*. Currently sequence tagging and dependency parsing tasks are supported.

Requirements

- Tensorflow
- Numpy

This code has been run with TensorFlow 1.10.1 and Numpy 1.14.5; other versions may work, but have not been tested.

Fetching and Preprocessing Data

Run `fetch_data.sh` to download and extract pretrained GloVe vectors, the 1 Billion Word Language Model Benchmark corpus of unlabeled data, and the CoNLL-2000 text chunking dataset. Unfortunately the other datasets from our paper are not freely available and so can't be included in this repository.

To apply CVT to other datasets, the data should be placed in `data/raw_data/<task_name>/(train|dev|test)`. For sequence tagging data, each line should contain a word followed by a space followed by that word's tag. Sentences should be separated by empty lines. For dependency parsing, each tag should be of the form `<index_of_head>-<relation>` (e.g., `0-root`).

After all of the data has been downloaded, run `preprocessing.py`.

Training a Model

Run `python cvt.py --mode=train --model_name=chunking_model`. By default this trains a model on the chunking data downloaded with `fetch_data.sh`. To change which task(s) are trained on or model hyperparameters, modify base/configure.py. Models are automatically checkpointed every 1000 steps; training will continue from the latest checkpoint if training is interrupted and restarted. Model checkpoints and other data such as dev set accuracy over time are stored in `data/models/<model_name>`.

Evaluating a Model

Run `python cvt.py --mode=eval --model_name=chunking_model`. A CVT model trained on the chunking data for 200k steps should get at least 97.1 F1 on the dev set and 96.6 F1 on the test set.

Citation

If you use this code for your publication, please cite the original paper:

```
@inproceedings{clark2018semi,
  title = {Semi-Supervised Sequence Modeling with Cross-View Training}},
  author = {Kevin Clark and Minh-Thang Luong and Christopher D. Manning and Quoc V. Le},
  booktitle = {EMNLP},
  year = {2018}
}
```

Contact

- Kevin Clark (@clarkkev).
- Thang Luong (@lmthang).