Motivation

- Open-source tool for large-scale structured learning
  1) Decoder agnostic online large-margin learning
     - Plug in any inference procedure
     - Simple communication between learner and decoder
     - Perform cost augmented training
     - Passive-Aggressive optimization
     - Use task-specific cost function
  2) Large-scale training with Hadoop MapReduce
     - Efficient parallelization architecture with Hadoop streaming

Evaluation

- Machine Translation training
  - 3k-50k tuning set size
    - Dense (16 features)
    - Sparse (up to 200k features)
  - Increase mappers with input size

Learner-Decoder Communication

1. Receive tab delimited input
2. Spawn decoder sub-process and connect input/output pipes
3. Send sgml encoded input to decoder
4. Receive k-best output
5. Compute update
6. Output final weights after feature selection and average

MapReduce

1. Receive input
2. Process all sentences, outputting model best for each
3. Output final mapper weights and total processed
4. Sort by key: all with -1 are sent to reducer
5. 1-best output
6. Output final weights after feature selection and average

#examples #mappers #features time/iter (sec) #features time/iter (sec)
3k 120 16 119 85k 232
5k 120 16 289 159k 610
10k 120 16 432 200k 1136
25k 300 16 942 200k 2395
50k 600 16 1802 200k 4495

https://github.com/kho/mr-mira