Large-scale Discriminative Training
Review

- Weighted linear model

\[ y = \arg\max_{z \in \mathcal{Y}(x)} W \cdot h_z \]

- Basic tuning strategy: MERT
Need for More Features

• We can add many more features

\[ h(e,f,a) = \begin{cases} 
1 & \text{if } f_i = "早上好", e_i = "good morning" \\
0 & \text{otherwise} 
\end{cases} \]

\[ h(e,f,a) = \begin{cases} 
1 & \text{if exists a verb in } e \\
0 & \text{otherwise} 
\end{cases} \]

Pros

• Incorporate rich human knowledge

• High-dimension $\rightarrow$ more linearly separable

Cons

• Need careful feature engineering

• Training becomes much harder

• MERT no longer suitable
MIRA

- Margin Infused Relaxed Algorithm
- Online learning algorithm
- Capable of handling millions of features
- Theoretically sound
- Easy to implement
How it works

1. Online learning setting

早上好
Morning good.
Good morning.
How are you.
Good evening.
...

非常感谢
Thank you.
How are you.
Thank you very much.
You are welcome.
...

现在几点
What time is it?
What's the date?
What day is it?
How are you doing?
...

$W_0 \xrightleftharpoons{} \text{Decoder}$
$W_1 = \text{update}(W_0)$

$W_1 \xrightleftharpoons{} \text{Decoder}$
$W_2 = \text{update}(W_1)$

$W_2 \xrightleftharpoons{} \text{Decoder}$
$W_3 = \text{update}(W_2)$
How it works

1. Online learning setting

How are you.
Thank you.
What time is it?

Morning good.
Good morning.
What's the date?

How are you.
Thank you very much.
What day is it?

Good evening.
You are welcome.
How are you doing?

非常感谢
非常感谢

W₀ \rightarrow \text{Decoder}
W₀ \rightarrow \text{Decoder}
W₀ \rightarrow \text{Decoder}

Morning good.
Good morning.
What's the date?

How are you.
Thank you very much.
What day is it?

Good evening.
You are welcome.
How are you doing?

现在几点

W₁ = update(W₀)

Compare: batch learning (eg. MERT)
## How it works

### 2. Updating strategy

The updating strategy updates the weight matrix \( W_{t+1} \) based on the previous weight matrix \( W_t \), the learning rate \( \lambda_t \), and the difference between the output hidden state \( h_t^o \) and the desired hidden state \( h_t^d \):

\[
W_{t+1} = W_t + \lambda_t (h_t^o - h_t^d)
\]

### Hypotheses

<table>
<thead>
<tr>
<th>Hypotheses</th>
<th>Phrase</th>
<th>LM</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>The sky remain clear</td>
<td>-192.3</td>
<td>-263.2</td>
<td>4</td>
</tr>
<tr>
<td>The sky remained clear</td>
<td>-176.2</td>
<td>-98.7</td>
<td>4</td>
</tr>
<tr>
<td>Sky is the clear</td>
<td>-250.5</td>
<td>-505.2</td>
<td>4</td>
</tr>
<tr>
<td>The sky is very clear</td>
<td>-187.8</td>
<td>-103.7</td>
<td>5</td>
</tr>
<tr>
<td>The sky is still clear</td>
<td>-210.6</td>
<td>-106.4</td>
<td>5</td>
</tr>
</tbody>
</table>

### Decode

The sky remain clear

天空依然十分清澈
How it works

3. Learning rate

- When loss is low, hope to decrease $\lambda_t$
- When loss is high, hope to increase $\lambda_t$
- However, $\lambda_t$ should be bounded from above. Otherwise the algorithm might diverge.

The MIRA learning rate:

$$
\lambda_t = \min \left\{ C, \frac{\rho_t - W_t \cdot \Delta h_t}{\|\Delta h_t\|^2} \right\}
$$

Where does it come from?
Theoretical Foundations

1. Concepts

- Margin

\[ M = W \cdot h^o - W \cdot h^p = W \cdot \Delta h \]

(Different in binary classification case: \( M = yW \cdot h \))

- Hinge loss

\[ \mathcal{L}(W) = \begin{cases} 
0 & \rho - M < 0 \\
\rho - M = \rho - W \cdot \Delta h & \text{otherwise}
\end{cases} \]
2. Optimization problem

At round $t$, solve:

$$
\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{w} - \mathbf{w}_t\|^2 + C\xi
$$

\[s.t.\]
\[\mathcal{L}(\mathbf{w}) \leq \xi\]
\[\xi \geq 0\]

Solution = MIRA:

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \lambda_t \Delta \mathbf{h}_t$$
$$\lambda_t = \min \left\{ C, \frac{\rho_t - \mathbf{W}_t \cdot \Delta \mathbf{h}_t}{\|\Delta \mathbf{h}_t\|^2} \right\}$$

- Passive-aggressive in nature
- Can be treated as online-SVM
Theoretical Foundations

3. Performance guarantee

**Theorem 1.** (Crammer et al., 2006) Let $\|u\|$ be an arbitrary weight vector in $\mathbb{R}^d$, $\mathcal{L}^t(w)$ be the maximum loss at round $t$ given a weight vector $w$, and assume $\forall t, \forall y \in \mathcal{Y}(x_t), \|h(x_t, y_t) - h(x_t, y)\|_2 \leq R_2$. The cumulative cost obtained by the MIRA algorithm is bounded from above by

$$
\sum_{t=1}^{T} \rho_t \leq \|u\|_2^2 R_2^2 + 2CR_2^2 \sum_{t=1}^{T} \mathcal{L}^t(u)
$$
4. A comparison of the loss function

- Direct loss is non-convex
- Hinge loss = tightest convex surrogate for direct loss
MIRA: A Summary

1. Updating strategy

\[ W_{t+1} = W_t + \lambda_t \Delta h_t \]
\[ \lambda_t = \min \left\{ C, \frac{\rho_t - W_t \cdot \Delta h_t}{\|\Delta h_t\|^2} \right\} \]

when \( \lambda_t = 1 \): Perceptron

2. Properties

- Online learning
- Large margin (hinge loss)
- Passive-aggressive
- Error upper-bound

Proper algorithm for large-scale discriminative training
MIRA in Practice

1. Choice of “oracle” and “prediction”

**“oracle”**

- Min cost:
  \[ y^* = \arg\min_{z \in \mathcal{Y}(x)} \text{cost}(y, z) \]

**“prediction”**

- Max model score:
  \[ y' = \arg\max_{z \in \mathcal{Y}(x)} \mathbf{W} \cdot \mathbf{h}_z \]

- “Hope”:
  \[ y^* = \arg\max_{z \in \mathcal{Y}(x)} (\mathbf{W} \cdot \mathbf{h}_z - \text{cost}(y, z)) \]

- “Fear”:
  \[ y' = \arg\max_{z \in \mathcal{Y}(x)} (\mathbf{W} \cdot \mathbf{h}_z + \text{cost}(y, z)) \]

- Max cost:
  \[ y' = \arg\max_{z \in \mathcal{Y}(x)} \text{cost}(y, z) \]
MIRA in Practice

2. Online learning setting

早上好
Morning good.
Good morning.
How are you.
Good evening.
...

非常感谢
Thank you.
How are you.
Thank you very much.
You are welcome.
...

现在几点
What time is it?
What's the date?
What day is it?
How are you doing?
...

Decoder

$W_t \Rightarrow$ Decoder

Morning good.
Good morning.
How are you.
Good evening.
...

$W_{t+1} = \text{update}(W_t)$

Decoder

Decoder

Decoder

$W_{t+1} \Rightarrow$ Decoder

Thank you.
How are you.
Thank you very much.
You are welcome.
...

$W_{t+2} = \text{update}(W_{t+1})$

Decoder

$W_{t+2} \Rightarrow$ Decoder

What time is it?
What's the date?
What day is it?
How are you doing?
...

$W_{t+3} = \text{update}(W_{t+2})$
2. Online learning setting

早上好
Morning good.
Good morning.
How are you.
Good evening.
…

非常感谢
Thank you.
How are you.
Thank you very much.
You are welcome.
…

现在几点
What time is it?
What's the date?
What day is it?
How are you doing?
…

$W_t \Rightarrow$ Decoder

$W_{t+1} = \text{update}(W_t)$

$W_{t+2} = \text{update}(W_{t+1})$

$W_{t+3} = \text{update}(W_{t+2})$

k-best MIRA
MIRA in Practice

3. Pseudo corpus for BLEU

• BLEU is not designed for sentence-level evaluation

• Compute BLEU in a context → create a pseudo-corpus

Morning good.
Good morning.
How are you.
Good evening.

... ... i-2

Thank you.
How are you.
Thank you very much.
You are welcome.

... ... i-1

What time is it?
What's the date?
What day is it?
How are you doing?

... ... i

\[0.9^{2}\text{BLEUStats}(i-2) + 0.9\text{BLEUStats}(i-1) + \text{BLEUStats}(i)\]

= Pseudo corpus BLEU stats
PRO

- Pairwise Ranking Optimization
- Batch learning algorithm
- Listwise ranking → pairwise ranking
- ≈ Ranking SVM
Idea: correct listwise ranking = correct pairwise ranking
How to train?

1. Sample

2. Train a binary classifier!
   (using any classifier you like)

Listwise ranking

Pairwise ranking
Performance Comparison

Urdu-English SBMT baseline feature tuning

Urdu-English SBMT extended feature tuning

15 features

2250 features
Many More Algorithms...

- Batch learning
  M3N, SSVM, MinRisk, Rampion, HOLS, ...

- Online learning
  AROW, CW, AdaGrad, ORO, ...

- Still far from reaching the oracles