Main Goals

1. It’s easy to implement and use!

2. Many possible loss functions
   What’s best for MT?
(Alternative) Discriminative Training

• Many new methods
  – 5 at this conference alone
    (Cherry & Foster, Gimpel & Smith, Bazrafshan et al., Watanabe, Chung & Galley)

• Margin-based learning consistently effective
  – Not yet widespread
    • Complicated
    • Complex
    • Don’t know what to implement
Let’s Bridge the Gap

• Disconnect between machine learning and MT
  – Latent variables
  – Surrogate references
    • Gimpel & Smith, McAllester et al.

• Different methods in practice
  – k-best constraints (Watanabe)
  – Cutting plane (Chiang)
    • Require QP solver
Passive-Aggressive Update

• Crammer presents simple PA update instead
  – Performing dual coordinate descent

• Similar to stochastic subgradient descent
MIRA Training

require: Training set $T = (x_i, y_i)_{i=1}^T$, w, C

1: for $j \leftarrow 1$ to $N$ do
2:     for $i \leftarrow 1$ to $T$ do
3:         $\mathcal{Y}(x_i) \leftarrow \text{Decode}(x_i, w)$
4:         $y^+ \leftarrow \text{FindOracle}(\mathcal{Y}(x_i))$
5:         $y^- \leftarrow \text{FindPrediction}(\mathcal{Y}(x_i))$
6:         margin $\leftarrow w^\top f(x_i, y^-) - w^\top f(x_i, y^+)$
7:         cost $\leftarrow$ BLEU$(y_i, y^+) -$ BLEU$(y_i, y^-)$
8:         loss $\leftarrow$ margin + cost
9:         if loss $> 0$ then
10:             $\delta \leftarrow \min \left( C, \frac{\text{loss}}{\|f(x_i, y^+) - f(x_i, y^-)\|^2} \right)$
11:             $w \leftarrow w + \delta (f(x_i, y^+) - f(x_i, y^-))$
12:         end if
13:     end for
14: end for
15: return w
MIRA Training

Require: : Training set $T = (x_i, y_i)_{i=1}^T$, w, C

1: for $j \leftarrow 1$ to N do
2: for $i \leftarrow 1$ to T do
3: \[ \mathcal{Y}(x_i) \leftarrow \text{Decode}(x_i, w) \]
4: \[ y^+ \leftarrow \text{FindOracle}(\mathcal{Y}(x_i)) \]
5: \[ y^- \leftarrow \text{FindPrediction}(\mathcal{Y}(x_i)) \]
6: margin $\leftarrow w^\top f(x_i, y^-) - w^\top f(x_i, y^+)$
7: cost $\leftarrow \text{BLEU}(y_i, y^+) - \text{BLEU}(y_i, y^-)$
8: loss $\leftarrow$ margin + cost
9: if loss $> 0$ then
10: \[ \delta \leftarrow \min \left( C, \frac{\text{loss}}{\| f(x_i, y^+) - f(x_i, y^-) \|^2} \right) \]
11: \[ w \leftarrow w + \delta (f(x_i, y^+) - f(x_i, y^-)) \]
12: end if
13: end for
14: end for
15: return w
MIRA Training

Require:  : Training set $T = (x_i, y_i)_{i=1}^T$, w, C

1: for $j \leftarrow 1$ to N do
2:   for $i \leftarrow 1$ to T do
3:     $\mathcal{V}(x_i) \leftarrow \text{Decode}(x_i, w)$
4:     $y^+ \leftarrow \text{FindOracle}(\mathcal{V}(x_i))$
5:     $y^- \leftarrow \text{FindPrediction}(\mathcal{V}(x_i))$
6:     margin $\leftarrow w^T f(x_i, y^-) - w^T f(x_i, y^+)$
7:     cost $\leftarrow \text{BLEU}(y_i, y^+) - \text{BLEU}(y_i, y^-)$
8:     loss $= \text{margin} + \text{cost}$
9:     if loss $> 0$ then
10:        $\delta \leftarrow \min \left(C, \frac{\text{loss}}{\|f(x_i, y^+) - f(x_i, y^-)\|^2}\right)$
11:        $w \leftarrow w + \delta (f(x_i, y^+) - f(x_i, y^-))$
12:     end if
13:   end for
14: end for
15: return w
MIRA Training

Require: \( \text{Training set } T = \{(x_i, y_i)\}_{i=1}^T, w, C \)

1: \textbf{for } j \leftarrow 1 \text{ to } N \textbf{ do}
2: \hspace{1em} \textbf{for } i \leftarrow 1 \text{ to } T \textbf{ do}
3: \hspace{2em} \mathcal{Y}(x_i) \leftarrow \text{Decode}(x_i, w)
4: \hspace{2em} y^+ \leftarrow \text{FindOracle}(\mathcal{Y}(x_i))
5: \hspace{2em} y^- \leftarrow \text{FindPrediction}(\mathcal{Y}(x_i))
6: \hspace{2em} \text{margin} \leftarrow w^\top f(x_i, y^-) - w^\top f(x_i, y^+)
7: \hspace{2em} \text{cost} \leftarrow \text{BLEU}(y_i, y^+) - \text{BLEU}(y_i, y^-)
8: \hspace{2em} \text{loss} = \text{margin} + \text{cost}
9: \hspace{2em} \textbf{if } \text{loss} > 0 \textbf{ then}
10: \hspace{3em} \delta \leftarrow \min \left( C, \frac{\text{loss}}{\|f(x_i, y^+) - f(x_i, y^-)\|^2} \right)
11: \hspace{3em} w \leftarrow w + \delta (f(x_i, y^+) - f(x_i, y^-))
12: \hspace{2em} \textbf{end if}
13: \hspace{1em} \textbf{end for}
14: \textbf{end for}
15: \textbf{return } w
MIRA Training

**Require:** Training set $T = (x_i, y_i)_{i=1}^{T}$, w, C

1: **for** $j \leftarrow 1$ to N **do**

2:    **for** $i \leftarrow 1$ to T **do**

3:        $Y(x_i) \leftarrow$Decode($x_i$, w)

4:        $y^+ \leftarrow$FindOracle($Y(x_i)$)

5:        $y^- \leftarrow$FindPrediction($Y(x_i)$)

6:        margin $\leftarrow w^T f(x_i, y^-) - w^T f(x_i, y^+)$

7:        cost $\leftarrow$ BLEU($y_i, y^+$) $-$ BLEU($y_i, y^-$)

8:        loss = margin + cost

9:        **if** loss $>$ 0 **then**

10:           $\delta \leftarrow \min \left( C, \frac{\text{loss}}{\|f(x_i, y^+) - f(x_i, y^-)\|^2} \right)$

11:           $w \leftarrow w + \delta (f(x_i, y^+) - f(x_i, y^-))$

12:        **end if**

13:    **end for**

14: **end for**

15: **return** w
MIRA Training

Require: Training set $T = (x_i, y_i)_{i=1}^T$, w, C

1: for $j \leftarrow 1$ to N do
2:     for $i \leftarrow 1$ to T do
3:         $\mathcal{Y}(x_i) \leftarrow$Decode($x_i, w$)
4:         $y^+ \leftarrow$ FindOracle($\mathcal{Y}(x_i)$)
5:         $y^- \leftarrow$ FindPrediction($\mathcal{Y}(x_i)$)
6:         margin $\leftarrow w^\top f(x_i, y^-) - w^\top f(x_i, y^+)$
7:         cost $\leftarrow$ BLEU($y_i, y^+$) $-$ BLEU($y_i, y^-$)
8:         loss $= \text{margin} + \text{cost}$
9:     if loss $> 0$ then
10:        $\delta \leftarrow \min \left( C, \frac{\text{loss}}{\|f(x_i, y^+) - f(x_i, y^-)\|^2} \right)$
11:        $w \leftarrow w + \delta (f(x_i, y^+) - f(x_i, y^-))$
12:     end if
13: end for
14: end for
15: return w
MIRA Training

Require: Training set \( T = (x_i, y_i)_{i=1}^{T}, w, C \)

1: for \( j \leftarrow 1 \) to \( N \) do
2:     for \( i \leftarrow 1 \) to \( T \) do
3:         \( \mathcal{Y}(x_i) \leftarrow \text{Decode}(x_i, w) \)
4:         \( y^+ \leftarrow \text{FindOracle}(\mathcal{Y}(x_i)) \)
5:         \( y^- \leftarrow \text{FindPrediction}(\mathcal{Y}(x_i)) \)
6:         margin \leftarrow w^\top f(x_i, y^-) - w^\top f(x_i, y^+) \)
7:         cost \leftarrow \text{BLEU}(y_i, y^+) - \text{BLEU}(y_i, y^-) \)
8:         loss = margin + cost
9:     if loss > 0 then
10:        \( \delta \leftarrow \min \left( C, \frac{\text{loss}}{\| f(x_i, y^+) - f(x_i, y^-) \|^2} \right) \)
11:        \( w \leftarrow w + \delta (f(x_i, y^+) - f(x_i, y^-)) \)
12:     end if
13: end for
14: end for
15: return \( w \)
Hypothesis Selection

-cost

score
Oracle Hypothesis Selection

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} (x_i) - \text{cost}(y_i, y)$$

max-BLEU

-cost

score
Oracle Hypothesis Selection

\[ y^+ \leftarrow \arg\max_{y \in \mathcal{Y}(x_i)} -\text{cost}(y_i, y) \]
Oracle Hypothesis Selection

\[ y^+ \leftarrow \arg \max_{y \in \mathcal{Y}(x_i)} -\text{cost}(y_i, y) \]

\[ y^+ \leftarrow \arg \max_{y \in \mathcal{Y}(x_i)} w^T f(x_i, y) - \text{cost}(y_i, y) \]

Cost-diminished
Candidate Hypothesis Selection

\[ y^+ \leftarrow \arg \max_{y \in Y(x_i)} -\text{cost}(y_i, y) \]

\[ y^+ \leftarrow \arg \max_{y \in Y(x_i)} w^T f(x_i, y) - \\text{cost}(y_i, y) \]

\[ y^- \leftarrow \arg \max_{y \in Y(x_i)} \text{cost}(y_i, y) \]

min-BLEU
Candidate Hypothesis Selection

\[ y^+ \leftarrow \arg\max_{y \in Y(x_i)} -\text{cost}(y_i, y) \]

\[ y^+ \leftarrow \arg\max_{y \in Y(x_i)} w^T f(x_i, y) - \text{cost}(y_i, y) \]

\[ y^- \leftarrow \arg\max_{y \in Y(x_i)} w^T f(x_i, y) + \text{cost}(y_i, y) \]
Candidate Hypothesis Selection

\[ y^+ \leftarrow \arg \max_{y \in \mathcal{Y}(x_i)} -\text{cost}(y_i, y) \]

\[ y^+ \leftarrow \arg \max_{y \in \mathcal{Y}(x_i)} w^T f(x_i, y) - \text{cost}(y_i, y) \]

\[ y^- \leftarrow \arg \max_{y \in \mathcal{Y}(x_i)} w^T f(x_i, y) \]

\[ y^- \leftarrow \arg \max_{y \in \mathcal{Y}(x_i)} w^T f(x_i, y) + \text{cost}(y_i, y) \]

Prediction-based
Ramp Losses

- Parameterized by $\gamma$ and $\beta$

\[
\ell_r = - \max_{y^+ \in \mathcal{Y}(x_i)} \left( \gamma^+ w^\top f(x_i, y^+) - \beta^+ \text{cost}(y_i, y^+) \right) \\
+ \max_{y^- \in \mathcal{Y}(x_i)} \left( \gamma^- w^\top f(x_i, y^-) + \beta^- \text{cost}(y_i, y^-) \right)
\]
Experimental Setup

• Fr-En and Cs-En Translation
  – NT08 tune, NT09 and NT10 test
• Using Constrained data
• Hierarchical PBMT in cdec
• Everything constant except choice of oracle and candidate
Cs-En Results

|       | C       | M-C     |
|-------|---------|---------|
| NT09  | 16.4    | 18.3    |

|       | C       | M-C     |
|-------|---------|---------|
| NT10  | 17      | 19.3    |
Cs-En Results

|     | C   | M-C |
|-----|-----|-----|
| NT09| 16.4| 18.3|
| NT10| 17  | 19.3|
Cs-En Results

- C: NT09
  - 16.4

- M-C: NT09
  - 18.3

- C: NT10
  - 17

- M-C: NT10
  - 19.3
Cs-En Results

C
16.4
M-C
18.3
NT09

C
17
M-C
19.3
NT10
Cs-En Results

|     | C       | M-C     |
|-----|---------|---------|
| NT09| 16.4    | 18.3    |

|     | C       | M-C     |
|-----|---------|---------|
| NT10| 17      | 19.3    |
Cs-En Results

|       | NT09 | NT10 |
|-------|------|------|
| C     | 16.4 | 17   |
| M-C   | 18.3 | 19.3 |
|      |      |      |

- C: Control
- M-C: Modified Contrast
Cs-En Results

C
M-C
NT09

16.4
18.3

C
M-C
NT10

17
19.3
Cs-En Results

|        | C    | M-C  |        | C    | M-C  |
|--------|------|------|--------|------|------|
| NT09   | 16.4 | 18.5 | NT10   | 17   | 19.1 |
|        | 18.3 | 16   |        | 19.3 | 17.5 |
Cs-En Results

|       | NT09 | NT10 |
|-------|------|------|
| C     | 16.4 | 17   |
| M-C   | 18.5 | 19.1 |
| M+C   | 18.3 | 18.4 |
|       | 18.7 | 19.3 |
|       | 16   | 17.5 |
|       | 18.7 | 19.6 |

Legend:
- M
- C
- M+C
Fr-En Results

Min/Max Cost

VS.

M±C
(Hope/Fear)
Sparse Lexical Feature Set

• Word Pair
  – \((e_i, f_j)\)

• Insertion for unaligned target word
  – \((e_i, f_j), (e_i, f_{j+1}), \ldots\)

• Target Bigram
  – \((e_i, e_{i+1})\)

• 650k features for cs-en

• 1.1M features for fr-en
Expanded Feature Results

- NT09
- NT10

Fr-En

Base C/C
C/C

Cs-En

- NT09
- NT10
Expanded Feature Results

![Bar Chart Image]

- NT09
- NT10
- Fr-En
- NT09
- Cs-En
- NT10

Legend:
- Base C/C
- Based M±C
- C/C
- M±C
Cs-En Learning Curve

BLEU

Iteration
Cs-En Learning Curve
Cs-En Learning Curve

M-C Oracle (Hope)  Min Cost Oracle
Fr-En Learning Curve
Fr-En Learning Curve
Fr-En Learning Curve
Conclusion

• Presented simple margin-based training algorithm
  – Easy to implement
  – Good performance

• Explored family of loss functions for MT
  – M±C and C/C have comparable performance
  – M±C loss is more stable

• Available in a decoder near you
Thank You!