Discriminative Training of Translation Models
Outline

• Part 1: Evaluation
  • What’s a good translation?
  • Automatic translation evaluation
  • BLEU

• Part 2: Discriminative training for MT
  • Linear models
  • Minimum error rate training (MERT)
  • Training as pairwise ranking (PRO)
Translation Evaluation

- We want to compare systems
- We want to measure improvements
- We want to make scientific claims
- We want to adjust parameters $\langle \theta_1, \theta_2, \theta_3, \ldots \rangle$
What is a Good Translation?

By no means is this an easy problem to answer!

Depends on many, many factors.
BLEU

• BiLingual Evaluation Understudy

• Compare hypothesis translation against one or more reference translations

• Product of two scores

• *n*-gram precision: What proportion of \{1,2,3,4\}-grams in the hypothesis match the reference(s)

• length factor: Is translation long enough? (Why?)
hypothesis: ‘extension of isi in uttar pradesh ’

ref 1: ‘ isi ’s expansion in uttar pradesh ’
ref 2: ‘ the spread of isi in uttar pradesh ’
ref 3: ‘ isi spreading in uttar pradesh ’
ref 4: the spread of isi in uttar pradesh

prec(1) = \frac{1}{1}
hypothesis: ‘extension of isi in uttar pradesh’

ref 1: ‘isi’s expansion in uttar pradesh’
ref 2: ‘the spread of isi in uttar pradesh’
ref 3: ‘isi spreading in uttar pradesh’
ref 4: the spread of isi in uttar pradesh

\[
\text{prec}(l) = \frac{1}{1+1}
\]
hypothesis: ‘ extension of isi in uttar pradesh ’

ref 1: ‘ isi ’s expansion in uttar pradesh ’
ref 2: ‘ the spread of isi in uttar pradesh ’
ref 3: ‘ isi spreading in uttar pradesh ’
ref 4: the spread of isi in uttar pradesh

\[
\text{prec}(l) = \frac{|+|}{|+|+|+|}
\]
hypothesis: ‘extension of isi in uttar pradesh’

ref 1: ‘isi’s expansion in uttar pradesh’
ref 2: ‘the spread of isi in uttar pradesh’
ref 3: ‘isi spreading in uttar pradesh’
ref 4: the spread of isi in uttar pradesh

\[\text{prec}(1) = \frac{|+|+|}{|+|+|+|}\]
hypothesis: ‘extension of isi in uttar pradesh’
ref 1: ‘isi’s expansion in uttar pradesh’
ref 2: ‘the spread of isi in uttar pradesh’
ref 3: ‘isi spreading in uttar pradesh’
ref 4: the spread of isi in uttar pradesh

\[ \text{prec}(I) = \frac{1+1+1+1}{1+1+1+1+1} \]
hypothesis: ‘extension of isi in uttar pradesh’

ref 1: ‘isi’s expansion in uttar pradesh’

ref 2: ‘the spread of isi in uttar pradesh’

ref 3: ‘isi spreading in uttar pradesh’

ref 4: the spread of isi in uttar pradesh

\[ \text{prec}(l) = \frac{|+|+|+|+|+|}{|+|+|+|+|+|} \]
hypothesis: ‘extension of isi in uttar pradesh’

ref 1: ‘isi’s expansion in uttar pradesh’

ref 2: ‘the spread of isi in uttar pradesh’

ref 3: ‘isi spreading in uttar pradesh’

ref 4: the spread of isi in uttar pradesh

\[ \text{prec}(l) = \frac{1 + 1 + 1 + 1 + 1 + 1}{1 + 1 + 1 + 1 + 1 + 1 + 1} \]
hypothesis: ‘ extension of isi in uttar pradesh ’

ref 1: ‘ isi ’s expansion in uttar pradesh ’
ref 2: ‘ the spread of isi in uttar pradesh ’
ref 3: ‘ isi spreading in uttar pradesh ’
ref 4: the spread of isi in uttar pradesh

\[
\text{prec}(1) = \frac{1 + 1 + 1 + 1 + 1 + 1 + 1 + 1}{1 + 1 + 1 + 1 + 1 + 1 + 1 + 1} = 0.875
\]
hypothesis: ‘extension of isi in uttar pradesh’

ref 1: ‘isi’s expansion in uttar pradesh’
ref 2: ‘the spread of isi in uttar pradesh’
ref 3: ‘isi spreading in uttar pradesh’
ref 4: the spread of isi in uttar pradesh

definition of prec:

\[
\text{prec}(1) = \frac{1+1+1+1+1+1+1+1}{1+1+1+1+1+1+1+1} = 0.875
\]

\[
\text{prec}(2) = \frac{\text{some formula}}{1}
\]
hypothesis: ‘ extension of isi in uttar pradesh ’

ref 1: ‘ isi ’s expansion in uttar pradesh ’
ref 2: ‘ the spread of isi in uttar pradesh ’
ref 3: ‘ isi spreading in uttar pradesh ’
ref 4: the spread of isi in uttar pradesh

\[ \text{prec}(1) = \frac{|+|+|+|+|+|+|+|}{|+|+|+|+|+|+|+|} = 0.875 \]

\[ \text{prec}(2) = \frac{}{|+|} \]
hypothesis: ‘extension of isi in uttar pradesh’

ref 1: ‘isi’s expansion in uttar pradesh’
ref 2: ‘the spread of isi in uttar pradesh’
ref 3: ‘isi spreading in uttar pradesh’
ref 4: the spread of isi in uttar pradesh

\[ \text{prec}(1) = \frac{1+1+1+1+1+1+1+1}{1+1+1+1+1+1+1+1} = 0.875 \]

\[ \text{prec}(2) = \frac{1}{1+1+1} \]
hypothesis: ‘extension of ISI in Uttar Pradesh’

ref 1: ‘ISI’s expansion in Uttar Pradesh’

ref 2: ‘the spread of ISI in Uttar Pradesh’

ref 3: ‘ISI spreading in Uttar Pradesh’

ref 4: the spread of ISI in Uttar Pradesh

\[
\text{prec}(1) = \frac{1+1+1+1+1}{1+1+1+1+1+1+1+1} = 0.875
\]

\[
\text{prec}(2) = \frac{1+1}{1+1+1+1}
\]
hypothesis: ‘ extension of isi in uttar pradesh ’

ref 1: ‘ isi ’s expansion in uttar pradesh ’

ref 2: ‘ the spread of isi in uttar pradesh ’

ref 3: ‘ isi spreading in uttar pradesh ’

ref 4: the spread of isi in uttar pradesh

\[
\text{prec}(1) = \frac{1+1+1+1+1+1+1+1}{1+1+1+1+1+1+1+1} = 0.875
\]

\[
\text{prec}(2) = \frac{1+1+1}{1+1+1+1+1+1} = \frac{3}{6} = 0.5
\]
hypothesis: ‘extension of isi in uttar pradesh’
ref 1: ‘isi’s expansion in uttar pradesh’
ref 2: ‘the spread of isi in uttar pradesh’
ref 3: ‘isi spreading in uttar pradesh’
ref 4: the spread of isi in uttar pradesh

\[
\text{prec(1)} = \frac{1+1+1+1+1+1+1+1+1}{1+1+1+1+1+1+1+1+1+1} = 0.875
\]

\[
\text{prec(2)} = \frac{1+1+1+1+1+1}{1+1+1+1+1+1+1+1+1} = \text{ }\]
hypothesis: ‘extension of isi in uttar pradesh’

ref 1: ‘isi’s expansion in uttar pradesh’
ref 2: ‘the spread of isi in uttar pradesh’
ref 3: ‘isi spreading in uttar pradesh’
ref 4: the spread of isi in uttar pradesh

prec(1) = \frac{1+1+1+1+1+1+1+1}{1+1+1+1+1+1+1+1} = 0.875

prec(2) = \frac{1+1+1+1+1}{1+1+1+1+1+1+1+1} = 0.714
hypothesis: ‘ extension of isi in uttar pradesh ’

ref 1: ‘ isi ’s expansion in uttar pradesh ’
ref 2: ‘ the spread of isi in uttar pradesh ’
ref 3: ‘ isi spreading in uttar pradesh ’
ref 4: the spread of isi in uttar pradesh

\[
\text{prec}(1) = \frac{|+|+|+|+|+|+|+|}{|+|+|+|+|+|+|+|+|} = 0.875
\]

\[
\text{prec}(2) = \frac{|+|+|+|+|}{|+|+|+|+|+|+|+|} = 0.714
\]

prec(3) = 0.666

prec(4) = 0.6
Average Precision

\[ \text{prec}(1) = 0.875 \]
\[ \text{prec}(2) = 0.714 \]
\[ \text{prec}(3) = 0.666 \]
\[ \text{prec}(4) = 0.6 \]

\[ \overline{\text{prec}} = 0.707 \]

\[ \exp \left( \frac{1}{4} [\log 0.875 + \log 0.714 + \log 0.666 + \log 0.6] \right) \]
Pitfalls of Precision

hypothesis: isi in uttar pradesh

ref 1: ‘isi ’s expansion in uttar pradesh ’
ref 2: ‘the spread of isi in uttar pradesh ’
Pitfalls of Precision

hypothesis: isi in uttar pradesh

ref 1: ‘isi ’s expansion in uttar pradesh ’
ref 2: ‘ the spread of isi in uttar pradesh ’

prec(1) = \frac{1+1+1+1}{1+1+1+1} = 1.0
Pitfalls of Precision

hypothesis: isi in uttar pradesh

ref 1: ‘isi ’s expansion in uttar pradesh ’
ref 2: ‘the spread of isi in uttar pradesh ’

\[
\text{prec}(1) = \frac{|+|+|+|}{|+|+|+|} = 1.0
\]

\[
\text{prec}(2) = \frac{|+|+|}{|+|+|} = 1.0
\]
Pitfalls of Precision

**hypothesis:** isi in uttar pradesh

**ref 1:** ‘isi ’s expansion in uttar pradesh ’
**ref 2:** ‘ the spread of isi in uttar pradesh ’

\[
\text{prec}(1) = \frac{1+1+1+1}{1+1+1+1} = 1.0
\]

\[
\text{prec}(2) = \frac{1+1+1}{1+1+1} = 1.0
\]

\[
\text{prec}(3) = \frac{1+1}{1+1} = 1.0
\]
Pitfalls of Precision

hypothesis: isi in uttar pradesh

ref 1: ‘isi ’s expansion in uttar pradesh ’

ref 2: ‘the spread of isi in uttar pradesh ’

\[
\text{prec}(1) = \frac{1 + 1 + 1 + 1}{1 + 1 + 1 + 1} = 1.0
\]

\[
\text{prec}(2) = \frac{1 + 1 + 1}{1 + 1 + 1} = 1.0
\]

\[
\text{prec}(3) = \frac{1 + 1}{1 + 1} = 1.0
\]

\[
\text{prec}(4) = \frac{1}{1} = 1.0
\]
Pitfalls of Precision

hypothesis: isi in uttar pradesh

ref 1: ‘isi’s expansion in uttar pradesh’
ref 2: ‘the spread of isi in uttar pradesh’

\[\text{prec}(1) = \frac{1+1+1+1}{1+1+1+1} = 1.0\]
\[\text{prec}(2) = \frac{1+1+1}{1+1+1} = 1.0\]
\[\text{prec}(3) = \frac{1+1}{1+1} = 1.0\]
\[\text{prec}(4) = \frac{1}{1} = 1.0\]
Pitfalls of Precision

hypothesis: isi in uttar pradesh

ref 1: ‘isi’s expansion in uttar pradesh’
ref 2: ‘the spread of isi in uttar pradesh’

\[
\text{prec}(1) = \frac{1 + 1 + 1 + 1}{1 + 1 + 1 + 1} = 1.0
\]

\[
\text{prec}(2) = \frac{1 + 1 + 1}{1 + 1 + 1} = 1.0
\]

\[
\text{prec}(3) = \frac{1 + 1}{1 + 1} = 1.0
\]

\[
\text{prec}(4) = \frac{1}{1} = 1.0
\]

Is this a perfect translation??
Brevity Penalty

\[
BP(\hat{e}, \mathcal{E}) = \begin{cases} 
1 & \text{if } |\hat{e}| \geq |\mathcal{E}| \\
1 - e^{1 - |\mathcal{E}|/|\hat{e}|} & \text{otherwise}
\end{cases}
\]

If the translation is long enough, no penalty
...otherwise, exponential decay!
Brevity Penalty

Slide from Chris Callison-Burch
Pitfalls of Precision

hypothesis: isi in uttar pradesh \( n=4 \)

ref 1: ‘isi’s expansion in uttar pradesh’ \( n=8 \)

ref 2: ‘the spread of isi in uttar pradesh’ \( n=9 \)

\[
\text{prec} = 1.0
\]
Pitfalls of Precision

hypothesis: isi in uttar pradesh $n=4$

ref 1: ‘isi ’s expansion in uttar pradesh ’ $n=8$

ref 2: ‘the spread of isi in uttar pradesh ’ $n=9$

\[
\text{prec} = 1.0 \times \text{BP} = e^{1 - \frac{8.5}{4}} \approx 0.325
\]
Pitfalls of Precision

hypothesis: isi in uttar pradesh $n=4$

ref 1: ‘isi’s expansion in uttar pradesh’ $n=8$

ref 2: ‘the spread of isi in uttar pradesh’ $n=9$

\[
\overline{\text{prec}} = 1.0 \times \text{BP} = e^{1 - \frac{8.5}{4}} \approx 0.325
\]

\[
\text{BLEU} = 0.325
\]
Discriminative Training
Noisy Channels Again

\[ p(e) \]

source \[\rightarrow\] English
Noisy Channels Again

$p(e) \rightarrow \text{English} \rightarrow p(g \mid e) \rightarrow \text{German}$
Noisy Channels Again

\[ p(e) \xrightarrow{} \text{English} \xrightarrow{} \text{decoder} \xrightarrow{} \text{German} \]

\[ e^* = \arg \max_e p(e \mid g) \]
\[ = \arg \max_e \frac{p(g \mid e) \times p(e)}{p(g)} \]
\[ = \arg \max_e p(g \mid e) \times p(e) \]
Noisy Channels Again

\[ e^* = \arg \max_e p(e \mid g) \]

\[ = \arg \max_e \frac{p(g \mid e) \times p(e)}{p(g)} \]

\[ = \arg \max_e p(g \mid e) \times p(e) \]
Noisy Channels Again

\[ e^* = \arg \max_e p(e \mid g) \]

\[ = \arg \max_e \frac{p(g \mid e) \times p(e)}{p(g)} \]

\[ = \arg \max_e p(g \mid e) \times p(e) \]

\[ = \arg \max_e \log p(g \mid e) + \log p(e) \]
Noisy Channels Again

\[ e^* = \arg \max_e p(e | g) \]

\[ = \arg \max_e \frac{p(g | e) \times p(e)}{p(g)} \]

\[ = \arg \max_e p(g | e) \times p(e) \]

\[ = \arg \max_e \log p(g | e) + \log p(e) \]

\[ = \arg \max_e \begin{bmatrix} 1 \\ 1 \end{bmatrix}^\top \begin{bmatrix} \log p(g | e) \\ \log p(e) \end{bmatrix} \]

\[ w^\top h(g, e) \]
Noisy Channels Again

\[ e^* = \arg \max_e p(e | g) \]

\[ = \arg \max_e \frac{p(g | e) \times p(e)}{p(g)} \]

\[ = \arg \max_e p(g | e) \times p(e) \]

Does this look familiar?

\[ = \arg \max_e \left[ \begin{array}{c} 1 \\ 1 \end{array} \right]^T \left[ \begin{array}{c} \log p(g | e) \\ \log p(e) \end{array} \right] \]

\[ = \arg \max_e \left( \begin{array}{c} w^T \\ h(g, e) \end{array} \right) \]
The Noisy Channel

\[-\log p(g|e)\] vs \[-\log p(e)\]
As a Linear Model

\[\text{-log } p(g|e)\]

\[\text{-log } p(e)\]

\[\mathbf{w}\]
As a Linear Model

\[-\log p(g|e)\]

\[-\log p(e)\]

\[\mathbf{w}\]
As a Linear Model

\[-\log p(g|e)\]
As a Linear Model

\[-\log p(g|e)\] \sim w

Improvement 1:

to find better translations
As a Linear Model

\[-\log p(g|e)\]
As a Linear Model

\[-\log p(g|e) \sim w\]
As a Linear Model

\[- \log p(g \mid e)\]

\[\sim \mathbf{w}\]
As a Linear Model

Improvement 2:

Add dimensions to make points **separable**
Linear Models

\[ e^* = \arg \max_e w^\top h(g, e) \]

- Improve the modeling capacity of the noisy channel in two ways
- Reorient the weight vector
- Add new dimensions (*new features*)

Questions
- What features? \( h(g, e) \)
- How do we set the weights? \( w \)
Mann beißt Hund
Mann beißt Hund

\[ x \text{ BITES } y \]
Mann beißt Hund

man bites cat

man chase dog

dog bites man

man bites dog
<table>
<thead>
<tr>
<th>English</th>
<th>German</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>bites</td>
<td>beißt</td>
<td>BITEs</td>
</tr>
<tr>
<td>cat</td>
<td>Hund</td>
<td>dog</td>
</tr>
<tr>
<td>bite</td>
<td>man</td>
<td>bite</td>
</tr>
<tr>
<td>dog</td>
<td>man</td>
<td>dog</td>
</tr>
<tr>
<td>dog</td>
<td>man</td>
<td>man</td>
</tr>
<tr>
<td>bites</td>
<td>beißt</td>
<td>bites</td>
</tr>
<tr>
<td>man</td>
<td>Hund</td>
<td>dog</td>
</tr>
</tbody>
</table>

- **Mann** (man)
- **beißt** (bites)
- **Hund** (dog)

The diagram illustrates the relationship between the English and German words, with the action 'bites' being used to describe the action of a man biting a cat and a man chasing a dog.
Mann beißt Hund

man bites cat

Mann beißt Hund

man bite cat

Mann beißt Hund

dog bites man

Mann beißt Hund

man chase dog

Mann beißt Hund

man bite dog

Mann beißt Hund

man bites dog
Mann beißt Hund

man bites cat

man chase dog

man bite dog

dog bites man
Feature Classes

Lexical

Are lexical choices appropriate?

bank = “River bank” vs. “Financial institution”
Feature Classes

Lexical

Are lexical choices appropriate?

\[ bank = \text{“River bank” vs. “Financial institution”} \]

Configurational

Are semantic/syntactic relations preserved?

\[ \text{“Dog bites man” vs. “Man bites dog”} \]
Feature Classes

**Lexical**

Are lexical choices appropriate?

*bank* = “River bank” vs. “Financial institution”

**Configurational**

Are semantic/syntactic relations preserved?

“Dog bites man” vs. “Man bites dog”

**Grammatical**

Is the output fluent / well-formed?

“Man *bites* dog” vs. “Man *bite* dog”
What do lexical features look like?

Mann  beißt  Hund

man  bites  cat
What do lexical features look like?

Mann  beißt  Hund  
man  bites  cat
What do lexical features look like?

First attempt:

\[
\text{score}(g, e) = w^\top h(g, e)
\]

\[
h_{15,342}(g, e) = \begin{cases} 
1, & \exists i, j : g_i = \text{Hund}, e_j = \text{cat} \\
0, & \text{otherwise}
\end{cases}
\]
What do lexical features look like?

First attempt:

\[
\text{score}(g, e) = w^\top h(g, e)
\]

\[
h_{15,342}(g, e) = \begin{cases} 
1, & \exists i, j : g_i = \text{Hund}, e_j = \text{cat} \\
0, & \text{otherwise}
\end{cases}
\]

But what if a **cat** is being chased by a **Hund**?
What do lexical features look like?

Latent variables enable more precise features:

\[
\text{score}(g, e, a) = w^\top h(g, e, a)
\]

\[
h_{15,342}(g, e, a) = \sum_{(i,j) \in a} \begin{cases} 
1, & \text{if } g_i = \text{Hund}, e_j = \text{cat} \\
0, & \text{otherwise}
\end{cases}
\]
Standard Features

- **Target side features**
  - $\log p(e)$ [n-gram language model]
  - Number of words in hypothesis

- **Source + target features**
  - $\log$ relative frequency $e|f$ of each rule [log $(e,f) - \log #(f)$]
  - $\log$ relative frequency $f|e$ of each rule [log $(e,f) - \log #(e)$]
  - “lexical translation” $\log$ probability $e|f$ of each rule [\( \approx \log p_{model}(e|f) \)]
  - “lexical translation” $\log$ probability $f|e$ of each rule [\( \approx \log p_{model}(f|e) \)]

- **Other features**
  - Count of rules/phrases used
  - Reordering pattern probabilities
Parameter Learning
Hypothesis Space
Hypothesis Space
Hypothesis Space

References
Preliminaries

We assume a **decoder** that computes:

\[ \langle e^*, a^* \rangle = \arg\max_{\langle e, a \rangle} w^\top h(g, e, a) \]

And **K-best lists** of, that is:

\[ \{ \langle e^*_i, a^*_i \rangle \}_{i=1}^{K} = \arg \max_{\langle e, a \rangle} w^\top h(g, e, a) \]

**Standard, efficient algorithms exist for this.**
Learning Weights

- Try to match the reference translation **exactly**
- **Conditional random field**
  - Maximize the conditional probability of the reference translations
  - “Average” over the different latent variables
Learning Weights

- Try to match the reference translation exactly

- **Conditional random field**
  - Maximize the conditional probability of the reference translations
  - “Average” over the different latent variables

- **Max-margin**
  - Find the weight vector that separates the reference translation from others by the maximal margin
  - Maximal setting of the latent variables
Problems

- These methods give “full credit” when the model exactly produces the reference and no credit otherwise

- **What is the problem with this?**
Problems

• These methods give “full credit” when the model \textit{exactly} produces the reference and no credit otherwise

• \textbf{What is the problem with this?}
  • There are many ways to translate a sentence
  • What if we have multiple reference translations?
  • \textbf{What about partial credit?}
Cost-Sensitive Training

• Assume we have a cost function that gives a score for how good/bad a translation is

\[ \ell(\hat{e}, E) \mapsto [0, 1] \]

• Optimize the weight vector by making reference to this function

• We will talk about two ways to do this
K-Best List Example

$h_1$ $\vec{w}$ $h_2$
K-Best List Example
K-Best List Example

\[ \vec{w} \]

- \( 0.8 \leq \ell < 1.0 \)
- \( 0.6 \leq \ell < 0.8 \)
- \( 0.4 \leq \ell < 0.6 \)
- \( 0.2 \leq \ell < 0.4 \)
- \( 0.0 \leq \ell < 0.2 \)
Training as Classification

- **Pairwise Ranking Optimization**
  - Reduce training problem to **binary classification** with a **linear model**

- **Algorithm**
  - For $i=1$ to $N$
    - Pick random pair of hypotheses (A,B) from K-best list
    - Use cost function to determine if is A or B better
    - Create $i$th training instance
  - Train binary linear classifier
The diagram shows a scatter plot with points labeled from #1 to #10. The points are color-coded as follows:

- Black dots: $0.8 \leq \ell < 1.0$
- Red dots: $0.6 \leq \ell < 0.8$
- Orange dots: $0.4 \leq \ell < 0.6$
- Green dots: $0.2 \leq \ell < 0.4$
- Dark green dots: $0.0 \leq \ell < 0.2$

The axes are labeled $h_1$ and $h_2$.
Worse!
Worse!
Better!

- 0.8 ≤ ℓ < 1.0
- 0.6 ≤ ℓ < 0.8
- 0.4 ≤ ℓ < 0.6
- 0.2 ≤ ℓ < 0.4
- 0.0 ≤ ℓ < 0.2
Better!
Fit a linear model
Fit a linear model
K-Best List Example

$\mathbf{h}_1$ $\mathbf{h}_2$

- $0.8 \leq \ell < 1.0$
- $0.6 \leq \ell < 0.8$
- $0.4 \leq \ell < 0.6$
- $0.2 \leq \ell < 0.4$
- $0.0 \leq \ell < 0.2$
MERT

- **Minimum Error Rate Training**
- Directly target an automatic evaluation metric
  - BLEU is defined at the corpus level
  - MERT optimizes at the corpus level
- **Downsides**
  - Does not deal well with > ~20 features
MERT

Given weight vector \( w \), any hypothesis \( \langle e, a \rangle \) will have a (scalar) score \( m = w^\top h(g, e, a) \)

Now pick a search vector \( v \), and consider how the score of this hypothesis will change:

\[
\mathbf{w}_{\text{new}} = \mathbf{w} + \gamma \mathbf{v}
\]
Given weight vector $w$, any hypothesis $\langle e, a \rangle$ will have a (scalar) score $m = w^\top h(g, e, a)$

Now pick a search vector $v$, and consider how the score of this hypothesis will change:

$$w_{\text{new}} = w + \gamma v$$

$$m = (w + \gamma v)^\top h(g, e, a)$$
MERT

Given weight vector \( w \), any hypothesis \( \langle e, a \rangle \) will have a (scalar) score \( m = w^\top h(g, e, a) \)

Now pick a **search vector** \( v \), and consider how the score of this hypothesis will change:

\[
\begin{align*}
  w_{\text{new}} &= w + \gamma v \\
  m &= (w + \gamma v)^\top h(g, e, a) \\
  &= w^\top h(g, e, a) + \gamma v^\top h(g, e, a)
\end{align*}
\]
MERT

Given weight vector $w$, any hypothesis $\langle e, a \rangle$ will have a (scalar) score $m = w^\top h(g, e, a)$

Now pick a **search vector** $v$, and consider how the score of this hypothesis will change:

$$w_{\text{new}} = w + \gamma v$$

$$m = (w + \gamma v)^\top h(g, e, a)$$

$$= w^\top h(g, e, a) + \gamma v^\top h(g, e, a)$$

$$= a \gamma + b$$
MERT

Given weight vector \( w \), any hypothesis \( \langle e, a \rangle \) will have a (scalar) score \( m = w^\top h(g, e, a) \)

Now pick a **search vector** \( v \), and consider how the score of this hypothesis will change:

\[
\begin{align*}
  w_{\text{new}} &= w + \gamma v \\
  m &= (w + \gamma v)^\top h(g, e, a) \\
        &= w^\top h(g, e, a) + \gamma v^\top h(g, e, a) \\
        &= a \gamma + b
\end{align*}
\]
MERT

Given weight vector \( w \), any hypothesis \( \langle e, a \rangle \) will have a (scalar) score \( m = w^\top h(g, e, a) \)

Now pick a search vector \( v \), and consider how the score of this hypothesis will change:

\[
\begin{align*}
  \mathbf{w}_{\text{new}} &= \mathbf{w} + \gamma \mathbf{v} \\
  m &= (\mathbf{w} + \gamma \mathbf{v})^\top h(g, e, a) \\
  &= \mathbf{w}^\top h(g, e, a) + \gamma \mathbf{v}^\top h(g, e, a) \\
  &= a \gamma + b
\end{align*}
\]

Linear function in 2D!
Recall our k-best set \( \{\langle e_i^*, a_i^* \rangle \}_{i=1}^K \)
Recall our k-best set \( \{ \langle e_i^*, a_i^* \rangle \}_{i=1}^K \)
MERT
\[ \langle e_{162}^*, a_{162}^* \rangle \quad \langle e_{28}^*, a_{28}^* \rangle \quad \langle e_{73}^*, a_{73}^* \rangle \]
MERT

\[ \langle e_{162}^*, a_{162}^* \rangle \]

\[ \langle e_{28}^*, a_{28}^* \rangle \]

\[ \langle e_{73}^*, a_{73}^* \rangle \]
MERT

\[ \langle e_{162}^*, a_{162}^* \rangle \]

\[ \langle e_{28}^*, a_{28}^* \rangle \]

\[ \langle e_{73}^*, a_{73}^* \rangle \]

\[ \gamma \]

\[ m \]

errors
MERT

\[ \langle e^*_{162}, a^*_{162} \rangle, \langle e^*_{28}, a^*_{28} \rangle, \langle e^*_{73}, a^*_{73} \rangle \]

errors
MERT

\[ m \]

\[ \gamma \]

errors
Let $w_{\text{new}} = \gamma^* \mathbf{v} + \mathbf{w}$
MERT

• In practice “errors” are sufficient statistics for evaluation metrics (e.g., BLEU)

• Can maximize or minimize!

• Envelope can also be computed using dynamic programming

• Interesting complexity bounds

• How do you pick the search direction?
Summary

• Evaluation metrics
  • Figure out how well we’re doing
  • Figure out if a feature helps
  • But ALSO: train your system!

• What’s a great way to improve translation?
  • Improve evaluation!
Thank You!

$\langle e_{162}^*, a_{162}^* \rangle$

$\langle e_{28}^*, a_{28}^* \rangle$

$\langle e_{73}^*, a_{73}^* \rangle$