Phrase Based Translation

Yejin Choi

Slides from Philipp Koehn, Dan Klein, Luke Zettlemoyer
Phrase-Based Systems

Sentence-aligned corpus

Word alignments

Phrase table (translation model)
Phrase Translation Tables

- Defines the space of possible translations
  - each entry has an associated “probability”
- One learned example, for “den Vorschlag” from Europarl data

- This table is noisy, has errors, and the entries do not necessarily match our linguistic intuitions about consistency...
Probability Model

- Bayes rule

\[ e_{\text{best}} = \arg\max_e p(e|f) = \arg\max_e p(f|e) p_{LM}(e) \]

- translation model \( p(e|f) \)
- language model \( p_{LM}(e) \)

- Decomposition of the translation model

\[ p(f_1^I|e_1^I) = \prod_{i=1}^{I} \phi(f_i|e_i) \ d(start_i - end_{i-1} - 1) \]

- phrase translation probability \( \phi \)
- reordering probability \( d \)
Distance-Based Reordering

Distortion Model

\[ d(x) = \alpha^{|x|} \] — exponential with distance
Extracting Phrases

- We will use word alignments to find phrases

- Question: what is the best set of phrases?
Extracting Phrases

- Phrase alignment must
  - Contain at least one alignment edge
  - Contain all alignments for phrase pair

- Extract all such phrase pairs!
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green)
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch)
(Maria no daba una bofetada, Mary did not slap), (no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch)
(Maria no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde, slap the green witch)
(Maria no daba una bofetada a la bruja verde, Mary did not slap the green witch)
Linguistic Phrases?

- Model is not limited to linguistic phrases (noun phrases, verb phrases, prepositional phrases, ...)

- Example non-linguistic phrase pair
  
  spass am → fun with the

- Prior noun often helps with translation of preposition

- Experiments show that limitation to linguistic phrases hurts quality
Phrases do help
- But they don’t need to be long
- Why should this be?
Bidirectional Alignment

english to spanish

spanish to english

intersection
Size of Phrase Translation Table

- Phrase translation table typically bigger than corpus
  ... even with limits on phrase lengths (e.g., max 7 words)

→ Too big to store in memory?

- Solution for training
  - extract to disk, sort, construct for one source phrase at a time

- Solutions for decoding
  - on-disk data structures with index for quick look-ups
  - suffix arrays to create phrase pairs on demand
Why not Learn Phrases w/ EM?

EM Training of the Phrase Model

• We presented a heuristic set-up to build phrase translation table (word alignment, phrase extraction, phrase scoring)

• Alternative: align phrase pairs directly with EM algorithm
  – initialization: uniform model, all $\phi(\bar{e}, \bar{f})$ are the same
  – expectation step:
    * estimate likelihood of all possible phrase alignments for all sentence pairs
  – maximization step:
    * collect counts for phrase pairs $(\bar{e}, \bar{f})$, weighted by alignment probability
    * update phrase translation probabilities $p(\bar{e}, \bar{f})$

• However: method easily overfits
  (learns very large phrase pairs, spanning entire sentences)
Phrase Scoring

\[ g(f, e) = \log \frac{c(e, f)}{c(e)} \]

\[ g(\text{les chats, cats}) = \log \frac{c(\text{cats, les chats})}{c(\text{cats})} \]

- Learning weights has been tried, several times:
  - [Marcu and Wong, 02]
  - [DeNero et al, 06]
  - … and others

- Seems not to work well, for a variety of partially understood reasons

- Main issue: big chunks get all the weight, obvious priors don’t help
  - Though, [DeNero et al 08]
“Also knowing nothing official about, but having guessed and inferred considerable about, the powerful new mechanized methods in cryptography—methods which I believe succeed even when one does not know what language has been coded—one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography.

When I look at an article in Russian, I say: ‘This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.’”

- Warren Weaver (1955:18, quoting a letter he wrote in 1947)
Translation is hard!

zi    zhu     zhong   duan
自助终端

self help terminal device
help oneself terminating machine

(ATM, “self-service terminal”)

Examples from Liang Huang
Translation is hard!

Examples from Liang Huang
Translation is hard!

Examples from Liang Huang
Translation is hard!

Examples from Liang Huang
Translation is hard!

Examples from Liang Huang
or even...
Scoring:

- Basic approach, sum up phrase translation scores and a language model
  - Define \( y = p_1 p_2 \ldots p_L \) to be a translation with phrase pairs \( p_i \)
  - Define \( e(y) \) be the output English sentence in \( y \)
  - Let \( h() \) be the log probability under a tri-gram language model
  - Let \( g() \) be a phrase pair score (from last slide)
  - Then, the full translation score is:
    \[
    f(y) = h(e(y)) + \sum_{k=1}^{L} g(p_k)
    \]
- Goal, compute the best translation
  \[
  y^*(x) = \arg \max_{y \in \mathcal{Y}(x)} f(y)
  \]
Phrase Scoring

\[ g(f, e) = \log \frac{c(e, f)}{c(e)} \]

\[ g(\text{les chats, cats}) = \log \frac{c(\text{cats, les chats})}{c(\text{cats})} \]

- Learning weights has been tried, several times:
  - [Marcu and Wong, 02]
  - [DeNero et al, 06]
  - … and others

- Seems not to work well, for a variety of partially understood reasons

- Main issue: big chunks get all the weight, obvious priors don’t help
  - Though, [DeNero et al 08]
### Table 1: 

<table>
<thead>
<tr>
<th>Phrase-Based Translation</th>
<th>Scoring: Try to use phrase pairs that have been frequently observed. Try to output a sentence with frequent English word sequences.</th>
</tr>
</thead>
<tbody>
<tr>
<td>the 7 people including by some and the Russian the astronauts ,</td>
<td></td>
</tr>
<tr>
<td>it 7 people included by France and the Russian international astronomical of rapporteur .</td>
<td></td>
</tr>
<tr>
<td>this 7 out including the from the French and the Russian the fifth .</td>
<td></td>
</tr>
<tr>
<td>these 7 among including from the French and of the Russian of space members .</td>
<td></td>
</tr>
<tr>
<td>that 7 persons including from the of France and to Russian of the aerospace members .</td>
<td></td>
</tr>
<tr>
<td>7 include from the of France and Russian astronauts . the</td>
<td></td>
</tr>
<tr>
<td>7 numbers include from France and Russian of astronauts who .</td>
<td></td>
</tr>
<tr>
<td>7 populations include those from France and Russian astronauts .</td>
<td></td>
</tr>
<tr>
<td>7 deportees included come from France and Russia in astronomical personnel ;</td>
<td></td>
</tr>
<tr>
<td>7 philtrum including those from France and Russia a space member</td>
<td></td>
</tr>
<tr>
<td>including representatives from France and the Russian astronaut</td>
<td></td>
</tr>
<tr>
<td>include came from France and Russia by cosmonauts</td>
<td></td>
</tr>
<tr>
<td>include representatives from French and Russia cosmonauts</td>
<td></td>
</tr>
<tr>
<td>include came from France and Russia’s cosmonauts</td>
<td></td>
</tr>
<tr>
<td>includes coming from French and Russian’s cosmonaut</td>
<td></td>
</tr>
<tr>
<td>french and Russian’s astronaut’s astronaut</td>
<td></td>
</tr>
<tr>
<td>french and Russia’s special rapporteur</td>
<td></td>
</tr>
<tr>
<td>, and Russian special rapporteur</td>
<td></td>
</tr>
<tr>
<td>, and Russia rapporteur</td>
<td></td>
</tr>
<tr>
<td>, and Russia rapporteur .</td>
<td></td>
</tr>
<tr>
<td>or Russia’s</td>
<td></td>
</tr>
</tbody>
</table>

The seven-member crew includes astronauts from France and Russia.
Phrase-Based Translation

Scoring: Try to use phrase pairs that have been frequently observed. Try to output a sentence with frequent English word sequences.
Scoring: Try to use phrase pairs that have been frequently observed.
Try to output a sentence with frequent English word sequences.
### Phrase-Based Translation

<table>
<thead>
<tr>
<th>7人</th>
<th>中包括</th>
<th>来自</th>
<th>法国</th>
<th>和</th>
<th>俄罗斯</th>
<th>的</th>
<th>宇航</th>
<th>员</th>
</tr>
</thead>
<tbody>
<tr>
<td>the 7 people including by some and the russian the astronauts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>it 7 people included by france and the russian international astronautical of rapporteur.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>this 7 that including the from the french and the russian the fift</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>these 7 among including from the french and of the russian of space members</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>that 7 persons including from the of france and to the russian of the aerospace member</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 include from the of france and astronaut</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 numbers include from france and russian of astronauts who</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 populations include those from france and russian astronauts.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 deportees included come from france and in the Astronauts Aerospace personnel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 philtrum including those from france and the russia astronaut</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>include came from france and russia by cosmonauts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>include representatives from french and russia by cosmonauts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>includes coming from french and russia's cosmonauts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>includes coming from french and russia's's astronavigation member.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7</th>
<th>人</th>
<th>中包括</th>
<th>来自</th>
<th>法国</th>
<th>和</th>
<th>俄罗斯</th>
<th>的</th>
<th>宇航</th>
<th>员</th>
</tr>
</thead>
<tbody>
<tr>
<td>the 7 people including by some and the russian the astronauts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>it 7 people included by france and the russian international astronautical of rapporteur.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>this 7 that including the from the french and the russian the fift</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>these 7 among including from the french and of the russian of space members</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>that 7 persons including from the of france and to the russian of the aerospace member</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 include from the of france and astronaut</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 numbers include from france and russian of astronauts who</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 populations include those from france and russian astronauts.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 deportees included come from france and in the Astronauts Aerospace personnel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 philtrum including those from france and the russia astronaut</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>include came from france and russia by cosmonauts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>include representatives from french and russia by cosmonauts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>includes coming from french and russia's cosmonauts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>includes coming from french and russia's's astronavigation member.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| **Table 1:** #11# the seven - member crew includes astronauts from france and russia |

**Scoring:** Try to use phrase pairs that have been frequently observed. Try to output a sentence with frequent English word sequences.
The Pharaoh Decoder

Scores at each step include LM and TM
The Pharaoh Decoder

- Space of possible translations
  - Phrase table constrains possible translations
  - Output sentence is built left to right
    - but source phrases can match any part of sentence
  - Each source word can only be translated once
  - Each source word must be translated

Morgen → Tomorrow
fliege → I will fly
ich → will fly
nach Kanada → to the conference
zur Konferenz → in Canada
Scoring:

- In practice, much like for alignment models, also include a distortion penalty
  - Define \( y = p_1 p_2 \ldots p_L \) to be a translation with phrase pairs \( p_i \)
  - Let \( s(p_i) \) be the start position of the foreign phrase
  - Let \( t(p_i) \) be the end position of the foreign phrase
  - Define \( \eta \) to be the distortion score (usually negative!)
  - Then, we can define a score with distortion penalty:

\[
f(y) = h(e(y)) + \sum_{k=1}^{L} g(p_k) + \sum_{k=1}^{L-1} \eta \times |t(p_k) + 1 - s(p_{k+1})|
\]

- Goal, compute the best translation

\[
y^*(x) = \arg \max_{y \in \mathcal{Y}(x)} f(y)
\]
Paths are the sums of all features and weights on each phrase pair.

Further, all features have corresponding weights.

Language Model(s) – look up a trigram: reordering costs, phrase/word count costs, phrase translation probability.

Hypothesis Expansion

Maria did not give a slap to the witch green by a la verde.

Translation Options

<table>
<thead>
<tr>
<th>p: 1</th>
<th>f: ---------</th>
<th>e: Mary</th>
</tr>
</thead>
<tbody>
<tr>
<td>p: 1</td>
<td>f: ---------</td>
<td>e: Mary</td>
</tr>
<tr>
<td>p: 1</td>
<td>f: ---------</td>
<td>e: witch</td>
</tr>
<tr>
<td>p: .534</td>
<td>f: *--------</td>
<td>e: Mary</td>
</tr>
<tr>
<td>p: .182</td>
<td>f: -------*-</td>
<td>e: witch</td>
</tr>
<tr>
<td>p: .043</td>
<td>f: <em>-</em>**----</td>
<td>e: slap</td>
</tr>
<tr>
<td>p: .015</td>
<td>f: *****----</td>
<td>e: slap</td>
</tr>
<tr>
<td>p: .015</td>
<td>f: *******--</td>
<td>e: the</td>
</tr>
<tr>
<td>p: .004283</td>
<td>f: *******--</td>
<td>e: the</td>
</tr>
<tr>
<td>p: .000271</td>
<td>f: *********</td>
<td>e: green witch</td>
</tr>
<tr>
<td>p: .000271</td>
<td>f: *********</td>
<td>e: bruja verde</td>
</tr>
</tbody>
</table>

Connections with the Log Linear Model

Hypothesis Expansion
**Hypothesis Explosion!**

- **Q:** How much time to find the best translation?
  - Exponentially many translations, in length of source sentence
  - NP-hard, just like for word translation models
  - So, we will use approximate search techniques!
Hypothesis Lattices

Can recombine if:
• Last two English words match
• Foreign word coverage vectors match
Decoder Pseudocode

Initialization: Set beam \( Q = \{ q_0 \} \) where \( q_0 \) is initial state with no words translated

For \( i = 0 \) … \( n - 1 \) [where \( n \) in input sentence length]

• For each state \( q \in \text{beam}(Q) \) and phrase \( p \in \text{ph}(q) \)
  1. \( q' = \text{next}(q,p) \) [compute the new state]
  2. \( \text{Add}(Q,q',q,p) \) [add the new state to the beam]

Notes:
• \( \text{ph}(q) \): set of phrases that can be added to partial translation in state \( q \)
• \( \text{next}(q,p) \): updates the translation in \( q \) and records which words have been translated from input
• \( \text{Add}(Q,q',q,p) \): updates beam, \( q' \) is added to \( Q \) if it is in the top-\( n \) overall highest scoring partial translations
Decoder Pseudocode

Initialization: Set beam $Q=\{q_0\}$ where $q_0$ is initial state with no words translated

For $i=0 \ldots n-1$ where $n$ in input sentence length

• For each state $q \in \text{beam}(Q)$ and phrase $p \in \text{ph}(q)$
  1. $q' = \text{next}(q, p)$ [compute the new state]
  2. Add($Q, q', q, p$) [add the new state to the beam]

Possible State Representations:
• Full: $q = (e, b, \alpha)$, e.g. (“Joe did not give,” 11000000, 0.092)
  • $e$ is the partial English sentence
  • $b$ is a bit vector recorded which source words are translated
  • $\alpha$ is score of translation so far
Decoder Pseudocode

Initialization: Set beam $Q=\{q_0\}$ where $q_0$ is initial state with no words translated

For $i=0 \ldots n-1$ [where $n$ in input sentence length]

• For each state $q \in \text{beam}(Q)$ and phrase $p \in \text{ph}(q)$
  1. $q' = \text{next}(q,p)$ [compute the new state]
  2. Add($Q,q',q,p$) [add the new state to the beam]

Possible State Representations:
• Full: $q = (e, b, \alpha)$, e.g. (“Joe did not give,” 11000000, 0.092)
• Compact: $q = (e_1, e_2, b, r, \alpha)$,
  • e.g. (“not,” “give,” 11000000, 4, 0.092)
  • $e_1$ and $e_2$ are the last two words of partial translation
  • $r$ is the length of the partial translation
• Compact representation is more efficient, but requires back pointers to get the final translation
Pruning

Problem: easy partial analyses are cheaper
  - Solution 1: separate bean for each number of foreign words
  - Solution 2: estimate forward costs (A*-like)

María no dio una bofetada a la bruja verde

<table>
<thead>
<tr>
<th>e: Mary did not</th>
<th>f: **-------</th>
<th>p: 0.154</th>
</tr>
</thead>
<tbody>
<tr>
<td>e: the</td>
<td>f: -------**---</td>
<td>p: 0.354</td>
</tr>
</tbody>
</table>

better partial translation

covers easier part

--> lower cost
• Hypothesis expansion in a stack decoder
  – translation option is applied to hypothesis
  – new hypothesis is dropped into a stack further down
Stack Decoding

Stack Decoding Algorithm

1: place empty hypothesis into stack 0
2: for all stacks 0...n − 1 do
3:   for all hypotheses in stack do
4:     for all translation options do
5:       if applicable then
6:         create new hypothesis
7:         place in stack
8:         recombine with existing hypothesis if possible
9:         prune stack if too big
10:     end if
11:   end for
12: end for
13: end for
Decoder Pseudocode (Multibeam)

Initialization:
• set $Q_0 = \{q_0\}$, $Q_i = \{}$ for $i = 1 \ldots n$ [n is input sent length]

For $i=0 \ldots n-1$
• For each state $q \in \text{beam}(Q_i)$ and phrase $p \in \text{ph}(q)$
  1. $q' = \text{next}(q, p)$
  2. Add($Q_j, q', q, p$) where $j = \text{len}(q')$

Notes:
• $Q_i$ is a beam of all partial translations where $i$ input words have been translated
• $\text{len}(q)$ is the number of bits equal to one in $q$ (the number of words that have been translated)
Making it Work (better)

The “Fundamental Equation of Machine Translation” (Brown et al. 1993)

\[ \hat{e} = \arg \max_e P(e \mid f) \]

\[ = \arg \max_e P(e) \times P(f \mid e) / P(f) \]

\[ = \arg \max_e P(e) \times P(f \mid e) \]
Making it Work (better)

What StatMT people do in the privacy of their own homes

\[ \text{argmax } P(e \mid f) = e \]

\[ \text{argmax } P(e) \times P(f \mid e) / P(f) = e \]

\[ \text{argmax } P(e)^{1.9} \times P(f \mid e) \ldots \text{ works better!} \]

Which model are you now paying more attention to?
Making it Work (better)

What StatMT people do in the privacy of their own homes

$$\arg\max \ P(e \mid f) = e$$

$$\arg\max \ P(e) \times P(f \mid e) / P(f)$$

$$\arg\max \ P(e)^{1.9} \times P(f \mid e) \times 1.1^{\text{length}(e)}$$

Rewards longer hypotheses, since these are ‘unfairly’ punished by $P(e)$
Making it Work (better)

What StatMT people do in the privacy of their own homes

$$\text{argmax } P(e)^{1.9} \times P(f \mid e) \times 1.1^{\text{length}(e)} \times KS^{3.7} \ldots$$

Lots of knowledge sources vote on any given hypothesis. Each has a weight

“Knowledge source” = “feature function” = “score component”.
argmax_e 1.9×log P(e) + 1.0×log P(f | e) + 1.1× log length(e) + 3.7×KS + …
= argmax_e Σ_i w_i f_i

So, we have two things:

– “Features” f, such as log language model score
– A weight w for each feature that indicates how good a job it does at indicating good translations
No Data like More Data!

- Discussed for LMs, but can new understand full model!
Tuning for MT

- Features encapsulate lots of information
  - Basic MT systems have around 6 features
  - \( P(e|f) \), \( P(f|e) \), lexical weighting, language model

- How to tune feature weights?

- Idea 1: Use your favorite classifier
Why Tuning is Hard

- Problem 1: There are latent variables
  - Alignments and segmentations
  - Possibility: forced decoding (but it can go badly)

\[
\begin{align*}
x & : \quad \text{le parlement} \quad \text{adopte} \quad \text{la} \quad \text{résolution} \quad \text{législative} \\
h & : \\
y & : \quad \text{parliament} \quad \text{has} \quad \text{adopted} \quad \text{the} \quad \text{resolution}
\end{align*}
\]
Why Tuning is Hard

- Problem 2: There are many right answers
  - The reference or references are just a few options
  - No good characterization of the whole class

- BLEU isn’t perfect, but even if you trust it, it’s a corpus-level metric, not sentence-level
The perceptron algorithm

- Iteratively processes the training set, reacting to training errors
- Can be thought of as trying to drive down training error

The (online) perceptron algorithm:

- Start with zero weights
- Visit training instances \((x_i, y_i)\) one by one
  - Make a prediction
    \[
    y^* = \arg\max_y w \cdot \phi(x_i, y)
    \]
  - If correct \((y^* = y_i)\): no change, goto next example!
  - If wrong: adjust weights
    \[
    w = w + \phi(x_i, y_i) - \phi(x_i, y^*)
    \]
Perceptron training

For each training example \((x, y)\): \cite{Collins:2002}

\[
\begin{align*}
    w &\leftarrow w + \Phi(x, y_t) \quad | \quad y_t &= y \\
    w &\leftarrow w - \Phi(x, y_p) \quad | \quad y_p &= \text{DECODE}(x)
\end{align*}
\]

\[
\begin{align*}
    w &\leftarrow w + \Phi(x, y_t, h_t) \quad | \quad y_t, h_t &= ??? \\
    w &\leftarrow w - \Phi(x, y_p, h_p) \quad | \quad y_p, h_p &= \text{DECODE}(x)
\end{align*}
\]
Update strategies

\[ w \leftarrow w + \Phi(x, y_t, h_t) - \Phi(x, y_p, h_p) \]

Training example (reference)

\( x \): voté sur demande d’urgence
\( y \): vote on a request for urgent procedure
Update strategies

\[ w \leftarrow w + \Phi(x, [y_t, h_t]) - \Phi(x, y_p, h_p) \]

Training example (reference)

\( x \): voté sur demande d’urgence
\( y \): vote on a request for urgent procedure

Reachable translations

\( x \): voté sur demande d’urgence
\( h_p \):
\( y_p \): vote on emergency request

Current prediction

\( x \): voté sur demande d’urgence
\( h_t \):
\( y_t \): vote on a request for urgent procedure

Bold update
Update strategies

\[ \mathbf{w} \leftarrow \mathbf{w} + \Phi(\mathbf{x}, \mathbf{y}_t, \mathbf{h}_t) - \Phi(\mathbf{x}, \mathbf{y}_p, \mathbf{h}_p) \]

Training example (reference)
- \( \mathbf{x} \): voté sur demande d’urgence
- \( \mathbf{y} \): vote on a request for urgent procedure

**Reachable translations**

**Current prediction**
- \( \mathbf{x} \): voté sur demande d’urgence
- \( \mathbf{h}_p \):
- \( \mathbf{y}_p \): vote on emergency request

**Local update**
- \( \mathbf{x} \): voté sur demande d’urgence
- \( \mathbf{h}_t \):
- \( \mathbf{y}_t \): vote on an urgent request

**Bold update**
- \( \mathbf{x} \): voté sur demande d’urgence
- \( \mathbf{h}_t \):
- \( \mathbf{y}_t \): vote on a request for urgent procedure
Update strategies

\[ w \leftarrow w + \Phi(x, [y_t, h_t]) - \Phi(x, y_p, h_p) \]

Training example (reference)
- \( x \): voté sur demande d’urgence
- \( y \): vote on a request for urgent procedure

Reachable translations
- \( x \): voté sur demande d’urgence
- \( h_t \)
- \( y_t \): vote on an urgent request

Local update

Current prediction
- \( x \): voté sur demande d’urgence
- \( h_p \)
- \( y_p \): vote on emergency request

Bold update: skip example
Update strategies

\[ w \leftarrow w + \Phi(x, y_t, h_t) - \Phi(x, y_p, h_p) \]

Training example (reference)
- **x**: voté sur demande d’urgence
- **y**: vote on a request for urgent procedure

<table>
<thead>
<tr>
<th>Decoder</th>
<th>Bold</th>
<th>Local</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monotonic</td>
<td>34.3</td>
<td>34.6</td>
</tr>
<tr>
<td>Limited distortion</td>
<td>33.5</td>
<td>34.7</td>
</tr>
</tbody>
</table>

Current prediction

Bold update: skip example
Why Tuning is Hard

- **Problem 3: Computational constraints**
  - Discriminative training involves repeated decoding
  - Very slow! So people tune on sets much smaller than those used to build phrase tables
Minimum Error Rate Training

- Standard method: minimize BLEU directly (Och 03)
  - MERT is a discontinuous objective
  - Only works for max ~10 features, but works very well then
  - Here: k-best lists, but forest methods exist (Machery et al 08)
MERT: Convex Upper Bound of BLEU