statistical machine translation

PART 3: DECODING & EVALUATION
Statistical Machine Translation

- Challenges to statistical machine translation
- Sentence alignment
- IBM model
- Phrase-based translation
- Decoding
- Evaluation
How to use the noisy channel

• How does this work?

\[ E^* = \arg \max_{E} P(F|E)P(E) \]

• \( P(E) \) is a language model (e.g., N-gram) and encodes knowledge of word order.

• \( P(F|E) \) is a word-level translation model that encodes only knowledge on an unordered word-by-word basis.

• Combining these models can give us naturalness and fidelity, respectively.
How to use the noisy channel

• Example from Koehn and Knight using only conditional likelihoods of Spanish words given English words.

• *Que hambre tengo yo*

  →

  *What hunger have I*  \[ P(S|E) = 1.4E^{-5} \]

  *Hungry I am so*  \[ P(S|E) = 1.0E^{-6} \]

  *I am so hungry*  \[ P(S|E) = 1.0E^{-6} \]

  *Have I that hunger*  \[ P(S|E) = 2.0E^{-5} \]

  ...
How to use the noisy channel

• ... and with the English language model

• Que hambre tengo yo
  →
  What hunger have I  \[ P(S|E)P(E) = 1.4E^{-5} \times 1.0E^{-6} \]
  Hungry I am so        \[ P(S|E)P(E) = 1.0E^{-6} \times 1.4E^{-6} \]
  I am so hungry         \[ P(S|E)P(E) = 1.0E^{-6} \times 1.0E^{-4} \]
  Have I that hunger     \[ P(S|E)P(E) = 2.0E^{-5} \times 9.8E^{-7} \]
  ...

Sentence alignment

• We often need to align sentences before we can align words.

• We’ll look at two broad classes of methods:
  1. Methods that only look at sentence length,
  2. Methods based on lexical matches, or “cognates”.
**Word alignment**

- **Word alignments** can be 1:1, N:1, 1:N, 0:1, 1:0, ... E.g.,

  - "zero fertility" word: not translated (1:0)
  - "spurious" words: generated from 'nothing' (0:1)
  - One word translated as several words (1:N)

**Diagram:**

```
Canada ˈs program has been implemented
```

```
Le programme du Canada a été mis en application
```

**Note:** that this is only one possible alignment.
## IBM models

<table>
<thead>
<tr>
<th>IBM Model 1</th>
<th>lexical translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM Model 2</td>
<td>adds absolute <em>re-ordering</em> model</td>
</tr>
<tr>
<td>IBM Model 3</td>
<td>adds <em>fertility</em> model</td>
</tr>
</tbody>
</table>
IBM Model 1: alignments

• An **alignment**, $a$, identifies the **English word** that ‘produced’ the given French word at each index.

  • $a = \{a_1, ..., a_{LF}\}$ where $a_j \in \{0, ..., L_E\}$
  
  • E.g., $a = \{0, 3, 0, 1, 4, 5, 6, 6, 6\}$

---

**Diagram:**

- $e_0$ to $e_6$: English words: $\emptyset$, Canada, ‘s, program, has, been, implemented
- $f_1$ to $f_9$: French words: Le, programme, du, Canada, a, été, mis, en, application

- $a_1 = 0$ from $e_0$ to $f_1$
- $a_9 = 6$ from $e_6$ to $f_9$
IBM-1: expectation-maximization

1. Initialize translation parameters $P(f | e)$ (e.g., randomly).

2. Expectation: Given the current $\theta_k = P(f | e)$, compute the expected value of $Count(f, e)$ for all words in training data $\mathcal{O}$.

3. Maximization: Given the expected value of $Count(f, e)$, compute the maximum likelihood estimate of $\theta_k = P(f | e)$.
IBM-1: expectation-maximization

• First, we **initialize** our parameters, \( \theta = P(f|e) \).

• In the **Expectation** step, we compute **expected** counts:
  - \( TCount(f, e) \): the total number of times \( e \) and \( f \) are aligned.
  - \( Total(e) \): the total number of \( e \).

  *This has to be done in steps by first computing \( P(F, a|E) \) then \( P(a|F, E) \)*

• In the **Maximization** step, we perform MLE with the expected counts.
IBM-1 EM

1. Initialize $P(f \mid e)$
2. Make grid of all possible alignments
3. Compute $P(F \mid a, E) \rightarrow$ Products of $P(f \mid e)$
4. Compute $P(a \mid E, F) \rightarrow$ Divide by sum of rows from step 3
5. Compute $TCount \rightarrow$ Sum relevant probabilities from step 4
6. Compute $Total \rightarrow$ Sum over rows from step 5
7. Compute $P(f \mid e) = \frac{TCount(f,e)}{Total(e)}$
phrases

words
Phrase-based statistical MT

- **Phrase-based** statistical MT involves segmenting sentences into contiguous blocks or segments.
  - Each phrase is probabilistically **translated**.
  - Example: $P(\text{zu Hausse} | \text{at home})$
  - Each phrase is probabilistically **re-ordered**.

**Example**:

- English: *The green witch is at home this week*
- German: *Diese Woche ist die grüne Hexe zu Hausse*
Phrase-based statistical MT

• Phrase-based SMT allows many-to-many word mappings.

• Larger context allows for some disambiguation that is not possible in word-based alignment.
  • E.g.,

  \[ P(\text{coup} \mid \text{stroke}) \]
  vs.

  \[
  P(\text{coup de poing} \mid \text{punch}) > \\
  P(\text{coup de poing} \mid \text{stroke of fist}) > \\
  P(\text{coup d’œil} \mid \text{glance}) > \\
  P(\text{coup d’œil} \mid \text{stroke of eye})
  \]
Learning phrase-translations

• Typically, we use alignment templates (Och et al., 1999).
• Start with a word-alignment, then build phrases.

Maria no dió una bofetada a la bruja verde

This word-alignment is produced by a model like IBM-3
Learning phrase-translations

- A phrase alignment **must** contain **all** word alignments for each of its rows and columns.
- Collect **all** phrase alignments that are **consistent** with the **word alignment**, e.g.

```
<table>
<thead>
<tr>
<th></th>
<th>Maria</th>
<th>no</th>
<th>dió</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>did</td>
<td>not</td>
<td>slap</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Consistent

```
<table>
<thead>
<tr>
<th></th>
<th>Maria</th>
<th>no</th>
<th>dió</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>did</td>
<td>not</td>
<td>slap</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Inconsistent

```
<table>
<thead>
<tr>
<th></th>
<th>Maria</th>
<th>no</th>
<th>dió</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>did</td>
<td>not</td>
<td>slap</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Inconsistent
Learning phrase-translations

• **Given word-alignments** (produced automatically or otherwise), we do *not* need to do EM training. E.g.,

\[
P(f_1 f_2 | e_1 e_2 e_3) = \frac{\text{Count}(f_1 f_2, e_1 e_2 e_3)}{\text{Count}(e_1 e_2 e_3)}
\]
Phrase-based translation in practice

What is the legal drinking age in Quebec?

Quel est l'âge légal pour boire au Québec?
No more decoding models
Decoding

- **Decoding** is the act of translating another language into your native language.
  - Decoding is an NP-complete problem (Knight, 1999).

- IBM Models often decoded with **stack decoding** or **A* search**.

  - Introduces **greedy decoding** – start with a solution and incrementally try to **improve** it.
First stage of greedy method

- For each French word $f_j$, pick the English word $e^*$ such that

$$e^* = \arg\max_e P(f_j | e)$$

- This gives an initial alignment, e.g.,

<table>
<thead>
<tr>
<th>Bien</th>
<th>entendu</th>
<th>,</th>
<th>il</th>
<th>parole</th>
<th>d’</th>
<th>une</th>
<th>belle</th>
<th>victoire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well</td>
<td>heard</td>
<td>,</td>
<td>it</td>
<td>talking</td>
<td>Ø</td>
<td>a</td>
<td>beautiful</td>
<td>victory</td>
</tr>
</tbody>
</table>

(Better: *quite naturally, he talks about a great victory*)
Some transformations

- **Change**$(j, e)$: sets translation of $f_j$ to $e$
  - Usually we only consider English words $e$ that are in the top $N$ ranked translations for $f_j$.

- **Change$_2$**$(j_1, e_1, j_2, e_2)$: sets translation of $f_{j_1}$ to $e_1$ and translation of $f_{j_2}$ to $e_2$
  - Like performing two Change transformations in sequence, but **without** evaluating the intermediate string.

- **ChangeAndInsert**$(j, e_1, e_2)$: sets translation of $f_j$ to $e_1$ and inserts $e_2$ at its most likely position.
Some more transformations

- **RemoveInfertile**$(i)$: Removes $e_i$ if $e_i$ is aligned with no French words.

- **SwapSeg**$(i_1, i_2, j_1, j_2)$: Swaps segment $e_{i_1:i_2}$ with segment $e_{j_1:j_2}$ such that segments do not overlap.

- **JoinWords**$(i_1, i_2)$: Removes $e_{i_1}$ and aligns all French words that were aligned to $e_{i_1}$ to $e_{i_2}$. 
Iterating greedily

• We have an initial pair \((E^{(0)}, a^{(0)})\).

• Use local transformations to map \((E, a)\) to new pairs, \((E', a')\).

• At each iteration, \(k\), take the highest probability pair from all possible transformations
  
  • i.e., if \(\mathcal{R}(E^{(k)}, a^{(k)})\) is the set of all \((E, a)\) ‘reachable’ from \((E^{(k)}, a^{(k)})\), then at each iteration:

\[
(E^{(k+1)}, a^{(k+1)}) = \arg\max_{(E, a) \in \mathcal{R}(E^{(k)}, a^{(k)})} P(E)P(F, a | E)
\]
Example of greedy search

<table>
<thead>
<tr>
<th>Bien</th>
<th>entendu</th>
<th>,</th>
<th>il</th>
<th>parole</th>
<th>d’</th>
<th>une</th>
<th>belle</th>
<th>victoire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well</td>
<td>heard</td>
<td>,</td>
<td>it</td>
<td>talking</td>
<td>Ø</td>
<td>a</td>
<td>beautiful</td>
<td>victory</td>
</tr>
</tbody>
</table>

$\textbf{Change2}(5, \textit{talks}, 8, \textit{great})$
Example of greedy search

<table>
<thead>
<tr>
<th>Bien</th>
<th>entendu</th>
<th>,</th>
<th>il</th>
<th>parole</th>
<th>d’</th>
<th>une</th>
<th>belle</th>
<th>victoire</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Well</strong></td>
<td><strong>heard</strong></td>
<td>,</td>
<td>it</td>
<td>talks</td>
<td>Ø</td>
<td>a</td>
<td>great</td>
<td>victory</td>
</tr>
</tbody>
</table>

*Change2*(2, *understood*, 6, *about*)

<table>
<thead>
<tr>
<th>Bien</th>
<th>entendu</th>
<th>,</th>
<th>il</th>
<th>parole</th>
<th>d’</th>
<th>une</th>
<th>belle</th>
<th>victoire</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Well</strong></td>
<td><strong>understood</strong></td>
<td>,</td>
<td>it</td>
<td>talks</td>
<td><em>about</em></td>
<td>a</td>
<td>great</td>
<td>victory</td>
</tr>
</tbody>
</table>
Example of greedy search

<table>
<thead>
<tr>
<th>Bien</th>
<th>entendu</th>
<th>,</th>
<th>il</th>
<th>parole</th>
<th>d’</th>
<th>une</th>
<th>belle</th>
<th>victoire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well</td>
<td>understood</td>
<td>,</td>
<td><em>it</em></td>
<td>talks</td>
<td><em>about</em></td>
<td>a</td>
<td>great</td>
<td>victory</td>
</tr>
</tbody>
</table>

Change(4, *he*)
Example of greedy search

<table>
<thead>
<tr>
<th>Bien</th>
<th>entendu</th>
<th>,</th>
<th>il</th>
<th>parole</th>
<th>d’</th>
<th>une</th>
<th>belle</th>
<th>victoire</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Well</strong></td>
<td><strong>understood</strong></td>
<td>,</td>
<td>he</td>
<td>talks</td>
<td>about</td>
<td>a</td>
<td>great</td>
<td>victory</td>
</tr>
</tbody>
</table>

Change2(1, *quite*, 2, *naturally*)

<table>
<thead>
<tr>
<th>Bien</th>
<th>entendu</th>
<th>,</th>
<th>il</th>
<th>parole</th>
<th>d’</th>
<th>une</th>
<th>belle</th>
<th>victoire</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quite</strong></td>
<td><strong>naturally</strong></td>
<td>,</td>
<td>he</td>
<td>talks</td>
<td>about</td>
<td>a</td>
<td>great</td>
<td>victory</td>
</tr>
</tbody>
</table>
Greedy transformations

• At each iteration, we try *each possible* transformation.

• For each possible transformation, we evaluate

\[ P(E)P(F, a|E) \]

• We choose the transformation that gives the highest probability, and iterate until some stopping condition.
Evaluation of MT systems

<table>
<thead>
<tr>
<th>Human</th>
<th>According to the data provided today by the Ministry of Foreign Trade and Economic Cooperation, as of November this year, China has actually utilized 46.959B US dollars of foreign capital, including 40.007B US dollars of direct investment from foreign businessmen.</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM4</td>
<td>The Ministry of Foreign Trade and Economic Cooperation, including foreign direct investment 40.007B US dollars today provide data include that year to November China actually using foreign 46.959B US dollars and</td>
</tr>
<tr>
<td>Yamada/Knight</td>
<td>Today’s available data of the Ministry of Foreign Trade and Economic Cooperation shows that China’s actual utilization of November this year will include 40.007B US dollars for the foreign direct investment among 46.959B US dollars in foreign capital.</td>
</tr>
</tbody>
</table>

How can we objectively compare the quality of two translations?
Automatic evaluation

• We want an automatic and effective method to objectively rank competing translations.
  • Word Error Rate (WER) measures the number of erroneous word insertions, deletions, substitutions in a translation.
    • E.g., Reference: how to recognize speech
      Translation: how understand a speech

• Problem: There are many possible valid translations.
  (There’s no need for an exact match)
Challenges of evaluation

• Human judges: expensive, slow, non-reproducible (different judges – different biases).

• Multiple valid translations, e.g.:
  • **Source:** *Il s’agit d’un guide qui assure que l’armée sera toujours fidèle au Parti*
  • **T1:** *It is a guide to action that ensures that the military will forever heed Party commands*
  • **T2:** *It is the guiding principle which guarantees the military forces always being under command of the Party*
BLEU evaluation

• **BLEU (BiLingual Evaluation Understudy)** is an automatic and popular method for evaluating MT.
  • It uses **multiple** human **reference** translations, and looks for local matches, allowing for phrase movement.

• **Candidate**: *n.* a translation produced by a machine.

• There are a few parts to a **BLEU score**...
Example of BLEU evaluation

- **Reference 1**: It is a guide to action that ensures that the military will forever heed Party commands
- **Reference 2**: It is the guiding principle which guarantees the military forces always being under command of the Party
- **Reference 3**: It is the practical guide for the army always to heed the directions of the party

- **Candidate 1**: It is a guide to action which ensures that the military always obeys the commands of the party
- **Candidate 2**: It is to insure the troops forever hearing the activity guidebook that party direct
**BLEU: Unigram precision**

- The **unigram precision** of a candidate is $$\frac{C}{N}$$ where $$N$$ is the number of words in the candidate and $$C$$ is the number of words in the candidate which are in at least one reference.

- e.g., **Candidate 1**: *It is a guide to action which ensures that the military always obeys the commands of the party*
  - Unigram precision = $$\frac{17}{18}$$

  *(obey appears in none of the three references).*
BLEU: Modified unigram precision

• **Reference 1:** *The lunatic is on the grass*
• **Reference 2:** *There is a lunatic upon the grass*
• **Candidate:** *The the the the the the the the*
  • Unigram precision $= \frac{7}{7} = 1$

• **Capped unigram precision:**
  A candidate word type $w$ can only be correct a maximum of $\text{cap}(w)$ times.
  • e.g., with $\text{cap}(\text{the}) = 2$, the above gives
    $$p_1 = \frac{2}{7}$$
BLEU: Generalizing to $N$-grams

• Generalizes to higher-order $N$-grams.
  
  • Reference 1: *It is* a guide to action that ensures that the military will forever heed Party commands
  
  • Reference 2: *It is* the guiding principle which guarantees the military forces always being under command of the Party
  
  • Reference 3: *It is* the practical guide for the army always to heed the directions of the party
  
  • Candidate 1: *It is* a guide to action which ensures that the military always obeys the commands of the party
  
  • Candidate 2: *It is* to insure the troops forever hearing the activity guidebook that party direct

Bigram precision, $p_2$

\[ p_2 = 10/17 \]

\[ p_2 = 1/13 \]
BLEU: Precision is not enough

- **Reference 1**: It is a guide to action that ensures that the military will forever heed Party commands
- **Reference 2**: It is the guiding principle which guarantees the military forces always being under command of the Party
- **Reference 3**: It is the practical guide for the army always to heed the directions of the party

- **Candidate 1**: of the

Unigram precision, $p_1 = \frac{2}{2} = 1$  Bigram precision, $p_2 = \frac{1}{1} = 1$
BLEU: Brevity

• Solution: Penalize brevity.

• **Step 1:** for each candidate, find the reference **most similar in length.**

• **Step 2:** $c_i$ is the length of the $i^{th}$ candidate, and $r_i$ is the nearest length among the references,

$$brevity_i = \frac{r_i}{c_i}$$

• **Step 3:** multiply precision by the (0..1) **brevity penalty:**

$$BP = \begin{cases} 
1 & \text{if } brevity < 1 \\
 e^{1-brevity} & \text{if } brevity \geq 1
\end{cases}$$

Bigger = too brief
BLEU: Final score

• On slide 39, \( r_1 = 16, r_2 = 17, r_3 = 16, \) and \( c_1 = 18 \) and \( c_2 = 14, \)

\[
brevity_1 = \frac{17}{18} \quad BP_1 = 1
\]

\[
brevity_2 = \frac{16}{14} \quad BP_2 = e^{1 - \left(\frac{8}{7}\right)} = 0.8669
\]

• **Final score** of candidate \( C: \)

\[
BLEU = BP_C \times (p_1 p_2 \ldots p_n)^{1/n}
\]

where \( p_n \) is the \( n \)-gram precision. (You can set \( n \) empirically)
Example: Final BLEU score

- Reference 1: I am afraid Dave
- Reference 2: I am scared Dave
- Reference 3: I have fear David
- Candidate: I fear David

\[
\text{brevity} = \frac{4}{3} \geq 1 \quad \text{so} \quad BP = e^{1 - \left(\frac{4}{3}\right)}
\]

- \(p_1 = \frac{1+1+1}{3} = 1\)
- \(p_2 = \frac{1}{2}\)

\[
\text{BLEU} = BP \left(p_1 p_2\right)^{\frac{1}{2}} = e^{1 - \left(\frac{4}{3}\right)} \left(\frac{1}{2}\right)^{\frac{1}{2}} \approx 0.5067
\]

Assume \(\text{cap}() = 2\) for all N-grams

Also assume BLEU order \(n = 2\)
BLEU: summary

• BLEU is a geometric mean over $n$-gram precisions.
  • These precisions are capped to avoid strange cases.
    • E.g., the translation “the the the the” is not favoured.
  • This geometric mean is weighted so as not to favour unrealistically short translations, e.g., “the”

• Initially, evaluations showed that BLEU predicted human judgements very well, but:
  • People started optimizing MT systems to maximize BLEU. Correlations between BLEU and humans decreased.
(Aside) Bias in machine translation

- **Turkish**: o bir doktor
  - **English**: he is a doctor

- **Turkish**: o bir hemsire
  - **English**: she is a nurse
(Aside) Other evaluation methods

• **METEOR** is a weighted F-measure (combination of recall and precision)
• **Translation Error Rate** computes the string edit distance between the reference and the hypothesis.
• The **RIBES** metric looks at rank correlation to measure word order similarity between system and reference translations.
(Preview) Neural machine translation

"le chat est noir" <EOS>
[ 02 85 03 12 99 ]

Encoder

Context

Decoder

"the cat is black"
[ 00 42 82 16 04 ]

[ 42 82 16 04 99 ]

"the cat is black" <EOS>
Reading


• (optional) Gale & Church “A Program for Aligning Sentences in Bilingual Corpora” (on course website)

• **Useful reading on IBM Model-1**: Section 25.5 of the 2nd edition of the Jurafsky & Martin text.
  • 1st edition available at Robarts library.

• **Other**: Manning & Schütze Sections 13.0, 13.1.2 (Gale&Church), 13.1.3 (Church), 13.2, 13.3, 14.2.2