COS 484: Natural Language Processing

Machine Translation

Fall 2019
Translation

Communication is the key to solving the world's problems.

• One of the “holy grail” problems in artificial intelligence

• Practical use case: Facilitate communication between people in the world

• Extremely challenging (especially for low-resource languages)
Easy and not so easy translations

• Easy:
  
  • I like apples ↔ ich mag Äpfel (German)

• Not so easy:

  • I like apples ↔ J'aime les pommes (French)

  • I like red apples ↔ J'aime les pommes rouges (French)

  • les ↔ the but les pommes ↔ apples
MT basics

• **Goal:** Translate a sentence \( w^{(s)} \) in a **source language (input)** to a sentence in the **target language (output)**

• Can be formulated as an optimization problem:

\[
\hat{w}^{(t)} = \arg \max_{w^{(t)}} \psi(w^{(s)}, w^{(t)})
\]

• where \( \psi \) is a scoring function over source and target sentences

• Requires **two** components:

  • Learning algorithm to compute parameters of \( \psi \)
  
  • Decoding algorithm for computing the best translation \( \hat{w}^{(t)} \)
Why is MT challenging?

• Single words may be replaced with multi-word phrases
  • I like apples ↔ J'aime les pommes

• Reordering of phrases
  • I like red apples ↔ J'aime les pommes rouges

• Contextual dependence
  • les ↔ the but les pommes ↔ apples

Extremely large output space ➔ Decoding is NP-hard
Vauquois Pyramid

- Hierarchy of concepts and distances between them in different languages
- Lowest level: individual words/characters
- Higher levels: syntax, semantics
- Interlingua: Generic language-agnostic representation of meaning
Evaluating translation quality

• Two main criteria:
  
  • **Adequacy:** Translation $w^{(t)}$ should adequately reflect the linguistic content of $w^{(s)}$
  
  • **Fluency:** Translation $w^{(t)}$ should be fluent text in the target language

<table>
<thead>
<tr>
<th>Test</th>
<th>Adequate?</th>
<th>Fluent?</th>
</tr>
</thead>
<tbody>
<tr>
<td>To Vinay it like Python</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Vinay debugs memory leaks</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Vinay likes Python</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>

Different translations of *A Vinay le gusta Python*
Evaluation metrics

• Manual evaluation is most accurate, but expensive

• Automated evaluation metrics:
  • Compare system hypothesis with reference translations
  • BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002):
    • Modified n-gram precision

\[ p_n = \frac{\text{number of } n\text{-grams appearing in both reference and hypothesis translations}}{\text{number of } n\text{-grams appearing in the hypothesis translation}} \]
BLEU

$$\text{BLEU} = \exp \left( \frac{1}{N} \sum_{n=1}^{N} \log p_n \right)$$

Two modifications:

- To avoid $\log 0$, all precisions are smoothed
- Each n-gram in reference can be used at most once
- Ex. **Hypothesis**: to to to to to vs **Reference**: to be or not to be should not get a unigram precision of 1

Precision-based metrics favor short translations

- Solution: Multiply score with a brevity penalty for translations shorter than reference, $e^{1-r/h}$
BLEU

- Correlates somewhat well with human judgements
BLEU scores

<table>
<thead>
<tr>
<th>Translation</th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>$p_4$</th>
<th>BP</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vinay likes programming in Python</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sys1</td>
<td>$\frac{2}{7}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>.21</td>
</tr>
<tr>
<td>Sys2</td>
<td>$\frac{3}{3}$</td>
<td>$\frac{1}{2}$</td>
<td>0</td>
<td>0</td>
<td>.51</td>
<td>.33</td>
</tr>
<tr>
<td>Sys3</td>
<td>$\frac{4}{6}$</td>
<td>$\frac{3}{5}$</td>
<td>$\frac{2}{4}$</td>
<td>$\frac{1}{3}$</td>
<td>1</td>
<td>.76</td>
</tr>
</tbody>
</table>

Sample BLEU scores for various system outputs

- Alternatives have been proposed:
  - METEOR: weighted F-measure
  - Translation Error Rate (TER): Edit distance between hypothesis and reference
Data

- Statistical MT relies requires **parallel corpora**

<table>
<thead>
<tr>
<th>1. Chapter 4, Koch (DE)</th>
<th>de</th>
<th>es</th>
</tr>
</thead>
<tbody>
<tr>
<td>context</td>
<td>Wir möchten sicherstellen, daß hierauf bereits in den Erwägungsgründen hingewiesen wird und die uneindeutig formulierter Frist, innerhalb der der Rat eine Entscheidung treffen muß, auf maximal drei Monate fixiert wird.</td>
<td>Quisíramos asegurar que se aluda ya a esto en los considerandos y que el plazo, imprecisamente formulado, dentro del cual el Consejo ha de adoptar una decisión, se fije en tres meses como máximo.</td>
</tr>
<tr>
<td>as early as the recitals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>and that the period within which the Council has to make a decision - which is not clearly worded - is set at a maximum of three months.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. Chapter 3, FÅRM (SV)</th>
<th>de</th>
<th>es</th>
</tr>
</thead>
<tbody>
<tr>
<td>context</td>
<td>Unsere Erfahrungen mit moderner Verwaltung besagen, daß Transparenz, Dezentralisation der Verantwortlichkeiten und eine qualifizierte Auswertung oft ebenso effektiv sind wie bürokratische Detailkontrolle.</td>
<td>Nuestras experiencias en materia de administración moderna nos señalan que la apertura, la descentralización de las responsabilidades y las evaluaciones bien hechas son a menudo tan eficaces como los controles burocráticos detallados.</td>
</tr>
<tr>
<td>as effective as detailed bureaucratic supervision</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Europarl, Koehn, 2005)

- And lots of it!

- Not available for many low-resource languages in the world
Statistical MT

\[ \hat{w}^{(t)} = \arg \max_{w^{(t)}} \psi(w^{(s)}, w^{(t)}) \]

- Scoring function \( \psi \) can be broken down as follows:
  \[
  \psi(w^{(s)}, w^{(t)}) = \psi_A(w^{(s)}, w^{(t)}) + \psi_F(w^{(t)})
  \]
  \( (\text{adequacy}) \quad (\text{fluency}) \)

- Allows us to estimate parameters of \( \psi \) on separate data
  - \( \psi_A \) from aligned corpora
  - \( \psi_F \) from monolingual corpora
Noisy channel model

Generative process for source sentence

Use Bayes rule to recover $w^{(t)}$ that is maximally likely under the conditional distribution $p_{T|S}$ (which is what we want)
Noisy channel model

\[ p_T \xrightarrow{\text{Target sentence}} \Psi_A(w^{(s)}, w^{(t)}) \triangleq \log p_{S|T}(w^{(s)} | w^{(t)}) \]

\[ \Psi_F(w^{(t)}) \triangleq \log p_T(w^{(t)}) \]

\[ \Psi(w^{(s)}, w^{(t)}) = \log p_{S|T}(w^{(s)} | w^{(t)}) + \log p_T(w^{(t)}) = \log p_{S,T}(w^{(s)}, w^{(t)}). \]

- Allows us to use a language model \( p_T \) to improve fluency

- Use Bayes rule to recover \( w^{(t)} \) that is maximally likely under the conditional distribution \( p_{T|S} \) (which is what we want)
IBM Models

• Early approaches to statistical MT

• How can we define the translation model $p_{S|T}$?

• How can we estimate the parameters of the translation model from parallel training examples?

• Make use of the idea of alignments
Alignments

- **Key question:** How should we align words in source to words in target?

\[
A(w^{(s)}, w^{(t)}) = \{(A, \emptyset), (Vinay, Vinay), (le, likes), (gusta, likes), (Python, Python)\}.
\]

**good**

\[
A(w^{(s)}, w^{(t)}) = \{(A, Vinay), (Vinay, likes), (le, Python), (gusta, \emptyset), (Python, \emptyset)\}.
\]

**bad**
Incorporating alignments

• Joint probability of alignment and translation can be defined as:

\[
p(w^{(s)}, A \mid w^{(t)}) = \prod_{m=1}^{M^{(s)}} p(w_m^{(s)}, a_m \mid w_{a_m}^{(t)}, m, M^{(s)}, M^{(t)})
\]

\[
= \prod_{m=1}^{M^{(s)}} p(a_m \mid m, M^{(s)}, M^{(t)}) \times p(w_m^{(s)} \mid w_{a_m}^{(t)}).
\]

• \(M^{(s)}, M^{(t)}\) are the number of words in source and target sentences

• \(a_m\) is the alignment of the \(m^{th}\) word in the source sentence, i.e. it specifies that the \(m^{th}\) word is aligned to the \(a_m^{th}\) word in target

Is this sufficient?
Incorporating alignments

Multiple source words may align to the same target word!

\[ a_1 = 2, \quad a_2 = 3, \quad a_3 = 4, \ldots \]
Reordering and word insertion

\[
a = (3, 4, 2, 1)^\top
\]

\[
a = (1, 2, 3, 0, 4)^\top
\]

Assume extra NULL token

(Slide credit: Brendan O’Connor)
Independence assumptions

\[
p(w^{(s)}, A \mid w^{(t)}) = \prod_{m=1}^{M^{(s)}} p(w_m^{(s)}, a_m \mid w_{a_m}^{(t)}, m, M^{(s)}, M^{(t)})
\]

\[
= \prod_{m=1}^{M^{(s)}} p(a_m \mid m, M^{(s)}, M^{(t)}) \times p(w_m^{(s)} \mid w_{a_m}^{(t)}).
\]

- Two independence assumptions:
  - Alignment probability factors across tokens:
    \[
p(A \mid w^{(s)}, w^{(t)}) = \prod_{m=1}^{M^{(s)}} p(a_m \mid m, M^{(s)}, M^{(t)}).
    \]
  - Translation probability factors across tokens:
    \[
p(w^{(s)} \mid w^{(t)}, A) = \prod_{m=1}^{M^{(s)}} p(w_m^{(s)} \mid w_{a_m}^{(t)}).
    \]
How do we translate?

- We want: \( \arg \max_{w(t)} p(w(t) \mid w(s)) = \arg \max_{w(t)} \frac{p(w(s), w(t))}{p(w(s))} \)

- Sum over all possible alignments:

  \[
  p(w(s), w(t)) = \sum_{A} p(w(s), w(t), A)
  \]

  \[
  = p(w(t)) \sum_{A} p(A) \times p(w(s) \mid w(t), A)
  \]

- Alternatively, take the max over alignments

- Decoding: Greedy/beam search
• Assume \( p(a_m | m, M^{(s)}, M^{(t)}) = \frac{1}{M^{(t)}} \)

• Is this a good assumption?

Every alignment is equally likely!
Each source word is aligned to at most one target word.

Further, assume $p(a_m \mid m, M^{(s)}, M^{(t)}) = \frac{1}{M^{(t)}}$.

We then have:

$$p(w^{(s)}, w^{(t)}) = p(w^{(t)}) \sum_{A} \left( \frac{1}{M^{(t)}} \right)^{M^{(s)}} p(w^{(s)} \mid w^{(t)})$$

How do we estimate $p(w^{(s)} = v \mid w^{(t)} = u)$?
IBM Model 1

- If we had word-to-word alignments, we could compute the probabilities using the MLE:

\[ p(v | u) = \frac{\text{count}(u, v)}{\text{count}(u)} \]

- where \( \text{count}(u, v) = \) #instances where word \( u \) was aligned to word \( v \) in the training set

- However, word-to-word alignments are often hard to come by

What can we do?
EM for Model 1

• (E-Step) If we had an accurate translation model, we can estimate likelihood of each alignment as:

\[ q_m(a_m \mid w^{(s)}, w^{(t)}) \propto p(a_m \mid m, M^{(s)}, M^{(t)}) \times p(w_m^{(s)} \mid w_{a_m}^{(t)}) , \]

• (M Step) Use expected count to re-estimate translation parameters:

\[ p(v \mid u) = \frac{E_q[\text{count}(u, v)]}{\text{count}(u)} \]

\[ E_q[\text{count}(u, v)] = \sum_m q_m(a_m \mid w^{(s)}, w^{(t)}) \times \delta(w_m^{(s)} = v) \times \delta(w_{a_m}^{(t)} = u) . \]
IBM Model 2

- Slightly relaxed assumption:
  
  - $p(a_m | m, M^{(s)}, M^{(t)})$ is also estimated, not set to constant

- Original independence assumptions still required:
  
  - Alignment probability factors across tokens:
    
    $$p(A | w^{(s)}, w^{(t)}) = \prod_{m=1}^{M^{(s)}} p(a_m | m, M^{(s)}, M^{(t)}).$$

  - Translation probability factors across tokens:
    
    $$p(w^{(s)} | w^{(t)}, A) = \prod_{m=1}^{M^{(s)}} p(w^{(s)}_m | w^{(t)}_{a_m}).$$
Other IBM models

Model 1: lexical translation
Model 2: additional absolute alignment model
Model 3: extra fertility model
Model 4: added relative alignment model
Model 5: fixed deficiency problem.
Model 6: Model 4 combined with a HMM alignment model in a log linear way

• Models 3 - 6 make successively weaker assumptions
  • But get progressively harder to optimize
• Simpler models are often used to ‘initialize’ complex ones
  • e.g train Model 1 and use it to initialize Model 2 parameters
Phrase-based MT

• Word-by-word translation is not sufficient in many cases

   Nous allons prendre un verre

   (literal)  We will take a glass

   (actual)  We’ll have a drink

• Solution: build alignments and translation tables between multiword spans or “phrases”
Phrase-based MT

- Solution: build alignments and translation tables between multiword spans or “phrases”
- Translations condition on multi-word units and assign probabilities to multi-word units
- Alignments map from spans to spans
Vauquois Pyramid

- Hierarchy of concepts and distances between them in different languages
- Lowest level: individual words/characters
- Higher levels: syntax, semantics
- Interlingua: Generic language-agnostic representation of meaning
Syntactic MT

- Rather than use phrases, use a *synchronous context-free grammar*: constructs “parallel” trees in two languages simultaneously

\[
\begin{align*}
\text{NP} & \rightarrow [\text{DT}_1 \text{ JJ}_2 \text{ NN}_3; \text{DT}_1 \text{ NN}_3 \text{ JJ}_2] \\
\text{DT} & \rightarrow [\text{the, la}] \\
\text{DT} & \rightarrow [\text{the, le}] \\
\text{NN} & \rightarrow [\text{car, voiture}] \\
\text{JJ} & \rightarrow [\text{yellow, jaune}]
\end{align*}
\]

- Assumes parallel syntax up to reordering
- Translation = parse the input with “half” the grammar, read off other half

(Slide credit: Greg Durrett)
Syntactic MT

- Relax this by using lexicalized rules, like "syntactic phrases"
- Leads to HUGE grammars, parsing is slow

Grammar:

\[
S \rightarrow \{\text{VP . ; I VP .}\} \text{ OR } S \rightarrow \{\text{VP . ; you VP .}\}
\]

\[
\text{VP} \rightarrow \{\text{lo haré ADV ; will do it ADV}\}
\]

\[
S \rightarrow \{\text{lo haré ADV . ; I will do it ADV .}\}
\]

\[
\text{ADV} \rightarrow \{\text{de muy buen grado ; gladly}\}
\]

Slide credit: Dan Klein

Next time: Neural machine translation