NOVEMBER 26, 2019
ANTONIS ANASTASOPOULOS
MATERIALS LARGELY BORROWED FROM JUNJIE HU AND AUSTIN MATTHEWS

MACHINE TRANSLATION
OVERVIEW
• 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 to Norbert Wiener, March, 1947
ORDER YOUR KAWHE/COFFEE IN MĀORI

He mōwai māku I’ll have a flat white
He pango poto māku I’ll have a short black
He pango roa māku I’ll have a long black
He rate pini māku I’ll have a soy latte
He kaputino māku I’ll have a cappuccino
He rate māku I’ll have a latte
He tiakarete wera māku I’ll have a hot chocolate

Rahi Size

Kei te pēhea koe? How’s it going?
Anei taku kapu mau tonu Here is my reusable cup

(S) Paku
(M) Waenga
(L) Nui

Hei kawe atu To take away
Ki konei To have here
NEW ENGLISH TRANSLATION
NOVUM TESTAMENTUM GRÆCÆ

NEW
TESTAMENT
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CLASSIC SOUPS</strong></td>
<td><strong>Sm.</strong></td>
<td><strong>Lg.</strong></td>
</tr>
<tr>
<td>57.</td>
<td>House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)</td>
<td>1.50</td>
</tr>
<tr>
<td>58.</td>
<td>Chicken Rice Soup</td>
<td>1.85</td>
</tr>
<tr>
<td>59.</td>
<td>Chicken Noodle Soup</td>
<td>1.85</td>
</tr>
<tr>
<td>60.</td>
<td>Cantonese Wonton Soup</td>
<td>1.50</td>
</tr>
<tr>
<td>61.</td>
<td>Tomato Clear Egg Drop Soup</td>
<td>1.65</td>
</tr>
<tr>
<td>62.</td>
<td>Regular Wonton Soup</td>
<td>1.10</td>
</tr>
<tr>
<td>63.</td>
<td>Hot &amp; Sour Soup</td>
<td>1.10</td>
</tr>
<tr>
<td>64.</td>
<td>Egg Drop Soup</td>
<td>1.10</td>
</tr>
<tr>
<td>65.</td>
<td>Egg Drop Wonton Mix</td>
<td>1.10</td>
</tr>
<tr>
<td>66.</td>
<td>Tofu Vegetable Soup</td>
<td>NA</td>
</tr>
<tr>
<td>67.</td>
<td>Chicken Corn Cream Soup</td>
<td>NA</td>
</tr>
<tr>
<td>68.</td>
<td>Crab Meat Corn Cream Soup</td>
<td>NA</td>
</tr>
<tr>
<td>69.</td>
<td>Seafood Soup</td>
<td>NA</td>
</tr>
</tbody>
</table>
We want a model of $p(e | f)$.
**NOISY CHANNEL MT**

- $p(e)$
- "English"
- $e$ (encoded)
- Channel
- "Foreign"
- $f$ (decoded)
- $p(f|e)$
- **decode**
\[ \hat{e} = \arg \max_e p(e|f) \]

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

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

“Language Model”

“Translation Model”
NOISY CHANNEL DIVISION OF LABOR

• Language model – $p(e)$
  • is the translation fluent, grammatical, and idiomatic?
  • use any model of $p(e)$ – typically an $n$-gram model

• Translation model – $p(f|e)$
  • “reverse” translation probability
  • ensures adequacy of translation
My legal name is Alexander Perchov.
My legal name is Alexander Perchov. But all of my many friends dub me Alex, because that is a more flaccid-to-utter version of my legal name. Mother dubs me Alexi-stop-spleening-me!, because I am always spleening her.
My legal name is Alexander Perchov. But all of my many friends dub me Alex, because that is a more flaccid-to-utter version of my legal name. Mother dubs me Alexi-stop-spleening-me!, because I am always spleening her. If you want to know why I am always spleening her, it is because I am always elsewhere with friends, and disseminating so much currency, and performing so many things that can spleen a mother.
TRANSFORMATION MODEL

\( p(f|e) \) gives the channel probability – the probability of translating an English sentence into a foreign sentence

\( f = \text{je voudrais un peu de frommage} \)
\( e_1 = \text{I would like some cheese} \) \( p(f|e) \)
\( e_2 = \text{I would like a little of cheese} \) \( 0.4 \)
\( e_3 = \text{There is no train to Barcelona} \) \( 0.5 \)
\( e_3 = \text{There is no train to Barcelona} \) \( >0.00001 \)
• How do we parameterize $p(f|e)$?

$$p(f|e) = \frac{\text{count}(f, e)}{\text{count}(e)}$$

• There are a lot of sentences: this won’t generalize to new inputs
LEXICAL TRANSLATION

How do we translate a word? Look it up in a dictionary!

Haus: house, home, shell, household

Multiple translations
Different word senses, different registers, different inflections

house, home are common
shell is specialized (the Haus of a snail is its shell)
## HOW COMMON IS EACH?

<table>
<thead>
<tr>
<th>Translation</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>house</td>
<td>5000</td>
</tr>
<tr>
<td>home</td>
<td>2000</td>
</tr>
<tr>
<td>shell</td>
<td>100</td>
</tr>
<tr>
<td>household</td>
<td>80</td>
</tr>
</tbody>
</table>
\[ \hat{p}_{\text{MLE}}(e \mid \text{Haus}) = \begin{cases} 
0.696 & \text{if } e = \text{house} \\
0.279 & \text{if } e = \text{home} \\
0.014 & \text{if } e = \text{shell} \\
0.011 & \text{if } e = \text{household} \\
0 & \text{otherwise} 
\end{cases} \]
• Goal: a model $p(e|f, m)$

• where $e$ and $f$ are complete English and Foreign sentences

$e = \langle e_1, e_2, \ldots, e_m \rangle \quad f = \langle f_1, f_2, \ldots, f_n \rangle$
LEXICAL TRANSLATION

Goal: a model \( p(\text{e|f,m}) \)

where \( \text{e} \) and \( \text{f} \) are complete English and Foreign sentences

Lexical translation makes the following assumptions:

1. Each word \( e_i \) in \( \text{e} \) is generated from exactly one word in \( \text{f} \)
2. Thus, we have a latent alignment \( a_i \) that indicates which word \( e_i \) “came from.” Specifically it came from \( f_{a_i} \).
3. Given the alignments \( \text{a} \), translation decisions are conditionally independent of each other and depend only on the aligned source word \( f_{a_i} \).
Putting our assumptions together, we have:

\[ p(e | f, m) = \sum_{a \in [0,n]^m} p(a | f, m) \times \prod_{i=1}^{m} p(e_i | f_{a_i}) \]
• Most of the action for the first 10 years of MT was here. Words weren’t the problem. Word order was hard.
Alignments can be visualized by drawing links between two sentences, and they are represented as vectors of positions:

\[ \mathbf{a} = (1, 2, 3, 4)^\top \]
• Words may be reordered during translation

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

• A source word may not be translated at all

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

• Words may be inserted during translation

• E.g. English *just* does not have an equivalent

• But these words must be explained – we typically assume every source sentence contains a NULL token

\[ a = (1, 2, 3, 0, 4)^T \]
• A source word may translate into **more than one** target word

\[ a = (1, 2, 3, 4, 4)^T \]
MANY-TO-ONE TRANSLATION

- More than one source word may not translate as a unit in lexical translation.
IBM MODEL 1

Simplest possible lexical translation model

Additional assumptions:
- The $m$ alignment decisions are independent
- The alignment distribution for each $a_i$ is uniform over all source words and NULL

\[
\text{for each } i \in [1, 2, \ldots, m] \\
\quad a_i \sim \text{Uniform}(0, 1, 2, \ldots, n) \\
\quad e_i \sim \text{Categorical}(\theta_{f_{a_i}})
\]
TRANSLATING WITH MODEL 1

0  1  2  3  4
NULL das Haus ist klein

1  2  3  4
Language model says: 😊

TRANSLATING WITH MODEL 1

null das Haus ist klein
the house is small
TRANSLATING WITH MODEL 1

Language model says: 😞
How do we learn the parameters \( p(elf) \)?

“Chicken and egg” problem:
If we had the alignments, we could estimate the translation probabilities (MLE estimation)
If we had the translation probabilities we could find the most likely alignments (greedy)
We start with an English Sentence \( e = e_1e_2\ldots e_n \)

1. Choose the length of the Spanish sentence \( m \), with uniform probability \( \epsilon = \frac{1}{M} \), where \( M \) is the maximum allowed length of any Spanish sentence in the corpus.
2. Generate an alignment \( a_1, \ldots, a_m \) again with uniform probability.
3. Generate Spanish words \( f_1, \ldots, f_m \) each with probability \( t(f_j \mid e_{a_j}) \) or \( t(f_j \mid \text{NULL}) \)

How can we estimate the \( t(f \mid e) \) parameters?
EM ALGORITHM

Pick some random (or uniform) starting parameters

Repeat until bored (~5 iterations for lexical translation models):

1. Using the current parameters, compute “expected” alignments \( p(a_i | e, f) \) for every target word token in the training data
2. Keep track of the expected number of times \( f \) translates into \( e \) throughout the whole corpus
3. Keep track of the number of times \( f \) is used in the source of any translation
4. Use these estimates in the standard MLE equation to get a better set of parameters
EM FOR MODEL 1

... la maison ... la maison blue ... la fleur ...  

... the house ... the blue house ... the flower ...

- Initial step: all alignments equally likely

- Model learns that, e.g., la is often aligned with the
... la maison ... la maison blue ... la fleur ...

... the house ... the blue house ... the flower ...

• After one iteration

• Alignments, e.g., between la and the are more likely
EM FOR MODEL 1

... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...

- After another iteration

- It becomes apparent that alignments, e.g., between fleur and flower are more likely (pigeon hole principle)
EM FOR MODEL 1

... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...

\[
p(\text{la}|\text{the}) = 0.453 \\
p(\text{le}|\text{the}) = 0.334 \\
p(\text{maison}|\text{house}) = 0.876 \\
p(\text{bleu}|\text{blue}) = 0.563 \\
... \\
\]

- Parameter estimation from the aligned corpus
EM ALGORITHM - PSEUDOCODE

1. Initialize $t(\cdot \mid e)$ to uniform: $t(f \mid e) = \frac{1}{|V_f|}$ where $V_f$ is the Spanish vocabulary, and $e$ is any English word or NULL.

2. E-step: Calculate the expected number of times that word $e$ is translated as $f$.
   For each $i, j$ the transition that generates $f_j$ from $e_i$ “competes” with the transitions that generate $f_j$ from the other English words (or NULL). So we update our expected counts $c(f, e)$ as follows:

   $$c(f_j, e_i) \leftarrow c(f_j, e_i) + \frac{t(f_j \mid e_i)}{t(f_j \mid \text{NULL}) + \sum_i t(f_j \mid e_i)}$$

   $$c(f_j, \text{NULL}) \leftarrow c(f_j, \text{NULL}) + \frac{t(f_j \mid \text{NULL})}{t(f_j \mid \text{NULL}) + \sum_i t(f_j \mid e_i)}$$

3. M-step: Estimate the model’s parameters based on the expected counts.
   Let $t(f \mid e) \leftarrow \frac{c(f, e)}{\sum_f c(f, e)}$ where $e$ is any English word or NULL.

4. Go to step 2.
### Convergence

<table>
<thead>
<tr>
<th>e</th>
<th>f</th>
<th>initial</th>
<th>1st it.</th>
<th>2nd it.</th>
<th>3rd it.</th>
<th>...</th>
<th>final</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>das</td>
<td>0.25</td>
<td>0.5</td>
<td>0.6364</td>
<td>0.7479</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>book</td>
<td>das</td>
<td>0.25</td>
<td>0.25</td>
<td>0.1818</td>
<td>0.1208</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>house</td>
<td>das</td>
<td>0.25</td>
<td>0.25</td>
<td>0.1818</td>
<td>0.1313</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>the</td>
<td>buch</td>
<td>0.25</td>
<td>0.25</td>
<td>0.1818</td>
<td>0.1208</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>book</td>
<td>buch</td>
<td>0.25</td>
<td>0.5</td>
<td>0.6364</td>
<td>0.7479</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>a</td>
<td>buch</td>
<td>0.25</td>
<td>0.25</td>
<td>0.1818</td>
<td>0.1313</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>book</td>
<td>ein</td>
<td>0.25</td>
<td>0.5</td>
<td>0.4286</td>
<td>0.3466</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>a</td>
<td>ein</td>
<td>0.25</td>
<td>0.5</td>
<td>0.5714</td>
<td>0.6534</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>the</td>
<td>haus</td>
<td>0.25</td>
<td>0.5</td>
<td>0.4286</td>
<td>0.3466</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>house</td>
<td>haus</td>
<td>0.25</td>
<td>0.5</td>
<td>0.5714</td>
<td>0.6534</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>
EXTENSIONS

Phrase-based MT:
Allow multiple words to translate as chunks (including many-to-one)
Introduce another latent variable, the source *segmentation*

Adapted from Koehn (2006)
Alignment Priors:
Instead of assuming the alignment decisions are uniform, impose (or learn) a prior over alignment grids:

Chahuneau et al. (2013)
EXTENSIONS

Syntactic structure

Rules of the form:

\[ X \text{ 之一} \rightarrow \text{one of the } X \]

Chiang (2005), Galley et al. (2006)
How do we evaluate translation systems’ output?

Central idea: “The closer a machine translation is to a professional human translation, the better it is.”

Most commonly used metric is called BLEU
Candidate 1: It is a guide to action which ensures that the military always obey the commands of the party.

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 the command of the Party.

Reference 3: It is the practical guide for the army always to heed directions of the party.

Unigram Precision: 17/18

Adapted from slides by Arthur Chan
ISSUE OF N-GRAM PRECISION

What if some words are over-generated?
  e.g. “the”

An extreme example

Candidate: the the the the the the the.

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

N-gram Precision: 7/7

Solution: reference word should be exhausted after it is matched.

Adapted from slides by Arthur Chan
ISSUE OF N-GRAM PRECISION

What if some words are just dropped?

Another extreme example

  Candidate: *the*.

  Reference 1: *My mom likes the blue flowers*.

  Reference 2: *My mother prefers the blue flowers*.

N-gram Precision: 1/1

Solution: add a penalty if the candidate is too short.

Adapted from slides by Arthur Chan
BLEU

\[
\text{BLEU} = \left( p_1 \cdot p_2 \cdot p_3 \cdot p_4 \right)^{\frac{1}{4}} \max\left(1, e^{1-\frac{r}{c}}\right)
\]

Ranges from 0.0 to 1.0, but usually shown multiplied by 100

An increase of +1.0 BLEU is usually a conference paper

MT systems usually score in the 10s to 30s (40-50s?)

Human translators usually score in the 70s and 80s
A SHORT SEGUE

Word- and phrase-based (“symbolic”) models were cutting edge for decades (up until ~2014)

Such models are still the most widely used in commercial applications

Since 2014 most research on MT has focused on neural models
FULLY NEURAL TRANSLATION

Fully end-to-end RNN-based translation model

Encode the source sentence using one RNN

Generate the target sentence one word at a time using another RNN

Sutskever et al. (2014)
ATTENTIONAL MODEL

The encoder-decoder model struggles with long sentences

An RNN is trying to compress an arbitrarily long sentence into a finite-length worth vector

What if we only look at one (or a few) source words when we generate each output word?

Bahdanau et al. (2014)
Our large black dog bit the poor mailman.

うちの大きな黒い犬が可哀想な郵便屋に噛みついた。
THE ATTENTION MODEL

Encoder

Decoder

I am a student

Bahdanau et al. (2014)
THE ATTENTION MODEL

Bahdanau et al. (2014)
I am a student
THE ATTENTION MODEL

Encoder

Attention Model

Context Vector

Decoder

I am a student

Bahdanau et al. (2014)
THE ATTENTION MODEL

Encoder

Attention Model

Decoder

Context Vector

Bahdanau et al. (2014)
THE ATTENTION MODEL

Bahdanau et al. (2014)
THE ATTENTION MODEL

Encoder

Attention Model

Decoder

I am a student

Bahdanau et al. (2014)
THE ATTENTION MODEL

Bahdanau et al. (2014)
THE ATTENTION MODEL

Encoder

I
am
astudent
</s>

Decoder

je
suis étudiant

Context Vector

Bahdanau et al. (2014)
The Attention Model

Encoder

I am a student

Attention Model

Context Vector

Decoder

je suis étudiant</s>

Bahdanau et al. (2014)
CONVOLUTIONAL ENCODER-DECODER

Gehring et. al 2017

**CNN:**
- encodes words within a fixed size window
- Parallel computation
- Shortest path to cover a wider range of words

**RNN:**
- sequentially encode a sentence from left to right
- Hard to parallelize
• Idea: Instead of using an RNN to encode the source sentence and the partial target sentence, use self-attention!
THE TRANSFORMER

Attention Model

Encoder

Decoder

I am a student</s>

je suis étudiant</s>

Context Vector

Vaswani et al. (2017)
• Self-attention weight can detect long-term dependencies within a sentence, e.g., make ... more difficult
THE TRANSFORMER

Computation is easily parallelizable

Shorter path from each target word to each source word $\rightarrow$ stronger gradient signals

Empirically stronger translation performance

Empirically trains substantially faster than more serial models

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>Training Cost (FLOPs)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EN-DE</td>
<td>EN-FR</td>
</tr>
<tr>
<td>ByteNet [17]</td>
<td>23.75</td>
<td>39.2</td>
</tr>
<tr>
<td>Deep-Att + PosUnk [37]</td>
<td>24.6</td>
<td>39.92</td>
</tr>
<tr>
<td>GNMT + RL [36]</td>
<td>25.16</td>
<td>40.46</td>
</tr>
<tr>
<td>ConvS2S [9]</td>
<td>26.03</td>
<td>40.56</td>
</tr>
<tr>
<td>MoE [31]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deep-Att + PosUnk Ensemble [37]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GNMT + RL Ensemble [36]</td>
<td>26.30</td>
<td>41.16</td>
</tr>
<tr>
<td>ConvS2S Ensemble [9]</td>
<td>26.36</td>
<td>41.29</td>
</tr>
<tr>
<td>Transformer (base model)</td>
<td>27.3</td>
<td>38.1</td>
</tr>
<tr>
<td>Transformer (big)</td>
<td>28.4</td>
<td>41.0</td>
</tr>
</tbody>
</table>
CURRENT RESEARCH DIRECTIONS ON NEURAL MT

• Incorporation syntax into Neural MT
• Handling of morphologically rich languages
• Optimizing translation quality (instead of corpus probability)
• Multilingual models
CURRENT RESEARCH DIRECTIONS ON NEURAL MT

• Incorporation syntax into Neural MT
• Handling of morphologically rich languages
• Optimizing translation quality (instead of corpus probability)
• Multilingual models

Exploring Massively Multilingual, Massive Neural Machine Translation
Friday, October 11, 2019

Posted by Ankur Bapna, Software Engineer and Orhan Firat, Research Scientist, Google Research
CURRENT RESEARCH DIRECTIONS ON NEURAL MT

• Incorporation syntax into Neural MT
• Handling of morphologically rich languages
• Optimizing translation quality (instead of corpus probability)
• Multilingual models

Achieving Human Parity on Automatic Chinese to English News Translation

Hany Hassan,* Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou

Microsoft AI & Research
CURRENT RESEARCH DIRECTIONS ON NEURAL MT

- Incorporation syntax into Neural MT
- Handling of morphologically rich languages
- Optimizing translation quality (instead of corpus probability)
- Multilingual models
- Document-level translation


Samuel Läubli\(^1\)  Rico Sennrich\(^{1,2}\)  Martin Volk\(^1\)

\(^{1}\)Institute of Computational Linguistics, University of Zurich
{lueubli,volk}@cl.uzh.ch
CURRENT RESEARCH DIRECTIONS ON NEURAL MT

• Incorporation syntax into Neural MT
• Handling of morphologically rich languages
• Optimizing translation quality (instead of corpus probability)
• Multilingual models
• Document-level translation
• Domain adaptation and robustness