Machine Translation Concluded

David Kauchak
CS159 – Fall 2020

Language translation

¡Hola!

Word models: IBM Model 1

Mary did not slap the green witch

Maria no dio una botefada a la bruja verde

Each foreign word is aligned to exactly one English word

This is the ONLY thing we model!

\[
p(f_1 f_2 \ldots f_F | a_1 a_2 \ldots a_F | e_1 e_2 \ldots e_E) = \prod_{i=1}^{F} p(f_i | e_{a_i})
\]
Training without alignments

Initially assume a $p(f|e)$ are equally probable

Repeat:
- Enumerate all possible alignments
- Calculate how probable the alignments are under the current model (i.e. $p(f|e)$)
- Recalculate $p(f|e)$ using counts from all alignments, weighted by how probable they are

(Note: theoretical algorithm)

EM alignment

E-step
- Enumerate all possible alignments
- Calculate how probable the alignments are under the current model (i.e. $p(f|e)$)

M-step
- Recalculate $p(f|e)$ using counts from all alignments, weighted by how probable they are

(Note: theoretical algorithm)
### E-step: Given \( p(F|E) \), calculate \( p(A,F|E) \)

Normalize by the sum

\[
E \rightarrow \text{step: } \text{Given } p(F|E), \text{ calculate } p(A,F|E)
\]

### M-step: calculate unnormalized counts for \( p(f|e) \) given the alignments

Then, calculate the probabilities by normalizing the counts

---

The page contains a mathematical derivation involving conditional probabilities and the calculation of alignments between two sets of variables, \( A \) and \( F \), given an event \( E \). The text describes the steps in estimating the conditional probabilities and normalizing the counts to obtain the final probabilities.

- **E-step**: Given \( P(F|E) \), calculate \( P(A,F|E) \)
  - Normalize by the sum

- **M-step**: Calculate unnormalized counts for \( p(f|e) \) given the alignments
  - Then, calculate the probabilities by normalizing the counts

The page includes detailed mathematical expressions and diagrams illustrating the alignment process and the calculation steps.
Implementation details

For |E| English words and |F| foreign words, how many alignments are there?

Repeat:

E-step

- Enumerate all possible alignments
- Calculate how probable the alignments are under the current model (i.e. \( p(f|e) \))

M-step

- Recalculate \( p(f|e) \) using counts from all alignments, weighted by how probable they are

Thought experiment

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

His wife talks to him.

Su mujer habla con él.

The sharks await.

Los tiburones esperan.

\[
p(f_i | e_j) = \frac{\text{count}(f \text{ aligned-to } e)}{\text{count}(e)}
\]

\[
p(\text{el} | \text{the}) = 0.5
\]

\[
p(\text{Los} | \text{the}) = 0.5
\]

If we had the alignments

Input: corpus of English/Foreign sentence pairs along with alignment

for (E, F) in corpus:

for aligned words (e, f) in pair (E,F):

\[\text{count}(e, f) += 1\]

\[\text{count}(e) += 1\]

for all (e, f) in count:

\[p(f | e) = \frac{\text{count}(e, f)}{\text{count}(e)}\]
If we had the alignments

Input: corpus of English/Foreign sentence pairs along with alignment

for (E, F) in corpus:
    for e in E:
        for f in F:
            if f aligned-to e:
                count(e, f) += 1
                count(e) += 1

for all (e, f) in count:
    p(f | e) = count(e, f) / count(e)

Are these equivalent?

Thought experiment #2

The old man is happy. He has fished many times.
El viejo está feliz porque ha pescado muchos veces.

Use partial counts:
- count(viejo | man) 0.8
- count(viejo | old) 0.2

Without the alignments

Key: use expected counts (i.e., how likely based on the current model), rather than actual counts
Without alignments

\[ p(f \rightarrow e) : \text{probability that } f \text{ is aligned to } e \text{ in this pair} \]

a b c

y z

What is \( p(y \rightarrow a) \)?

Put another way, of all things that \( y \) could align to in this sentence, how likely is it to be \( a \)?

Of all things that \( y \) could align to, how likely is it to be \( a \): \[ p(y | a) \]

Does that do it?

No! \( p(y | a) \) is how likely \( y \) is to align to \( a \) over the whole data set.

21

Without alignments

\[ p(f \rightarrow e) : \text{probability that } f \text{ is aligned to } e \text{ in this pair} \]

a b c

y z

\[ p(y | a) \]

Of all things that \( y \) could align to, how likely is it to be \( a \):

\[ p(y | a) + p(y | b) + p(y | c) \]

22

Without the alignments

Input: corpus of English/Foreign sentence pairs along with alignment

\[
\text{for } (E, F) \text{ in corpus:} \\
\text{for } e \text{ in } E; \\
\text{for } f \text{ in } F; \\
p(f \rightarrow e) = p(f(e)) / \sum_{e \text{ in } E} p(f | e) \\
\text{count}(e,f) += p(f \rightarrow e) \\
\text{count}(e) += p(f \rightarrow e) \\
\text{for all } (e,f) \text{ in count:} \\
p(f | e) = \text{count}(e,f) / \text{count}(e)
\]

23

Without alignments

\[ p(f \rightarrow e) : \text{probability that } f \text{ is aligned to } e \text{ in this pair} \]

a b c

y z

\[ p(y | a) \]

Of all things that \( y \) could align to, how likely is it to be \( a \):

\[ p(y | a) \]

24
EM: without the alignments

Input: corpus of English/Foreign sentence pairs along with alignment

for some number of iterations:
   for (E, F) in corpus:
      for e in E:
         for f in F:
            $p(f \rightarrow e) = \frac{p(f|e) \cdot \sum_{\hat{e}} p(f|\hat{e})}{\sum_{\hat{e}} p(f|\hat{e})}$
            count(e, f) += $p(f \rightarrow e)$
            count(e) += $p(f \rightarrow e)$
      for all (e, f) in count:
         $p(f|e) = \frac{\text{count}(e, f)}{\text{count}(e)}$

Where are the E and M steps?

Calculate how probable the alignments are under the current model (i.e. $p(f|e)$)
EM: without the alignments

Input: corpus of English/Foreign sentence pairs along with alignment

for some number of iterations:
  for (E, F) in corpus:
    for e in E:
      for f in F:
        \( p(f \rightarrow e) = \frac{\text{count}(e, f)}{\sum \text{count}(e)} \)
        count(e, f) += \( p(f \rightarrow e) \)
        count(e) += \( p(f \rightarrow e) \)

for all \((e, f)\) in count:
  \( p(f | e) = \text{count}(e, f) / \text{count}(e) \)

Recalculate \( p(f | e) \) using counts from all alignments, weighted by how probable they are

NULL

Sometimes foreign words don’t have a direct correspondence to an English word

Adding a NULL word allows for \( p(f \mid \text{NULL}) \), i.e. words that appear, but are not associated explicitly with an English word

Implementation: add “NULL” (or some unique string representing NULL) to each of the English sentences, often at the beginning of the sentence

- \( p(\text{casa} \mid \text{NULL}) = 1/3 \)
- \( p(\text{verde} \mid \text{NULL}) = 1/3 \)
- \( p(\text{la} \mid \text{NULL}) = 1/3 \)

Benefits of word-level model

Rarely used in practice for modern MT system

Two key side effects of training a word-level model:
- Word-level alignment
- \( p(f \mid e) \): translation dictionary

Word alignment

100 iterations

How should these be aligned?

- green house
- casa verde
- the house
- la casa
### Word alignment

|          | p(casa | green) | p(vertede | green) | p(la | green) |
|----------|-------------|------------|------------|-------------|
|          | 0.005       | 0.995      | 0          |
|          | p(casa | house) | p(vertede | house) | p(la | house) |
|          | ~1.0        | ~0.0       | ~0.0       |
|          | p(casa | the) | p(vertede | the) | p(la | the) |
|          | 0.005       | 0          | 0.995      |

100 iterations

100 iterations

```
Why?
```

```
green house  casa verde
```

```
the house  la  casa
```

### Word-level alignment

\[
\text{alignment}(E, F) = \arg \max_A p(A, F \mid E)
\]

Which for IBM model 1 is:

\[
\text{alignment}(E, F) = \arg \max \prod_{i=1}^{n} p(f_i \mid e_i)
\]

Given a model (i.e. trained \( p(f \mid e) \)), how do we find this?

Align each foreign word \( f \) in \( F \) to the English word \( e \) in \( E \) with highest \( p(f \mid e) \)

\[
\alpha_i = \arg \max f_i \mid e_i p(f_i \mid e_i)
\]

### Word-alignment Evaluation

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

How good of an alignment is this?
How can we quantify this?

System:

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

Human:

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

How can we quantify this?
Word-alignment Evaluation

System:
The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

Human:
The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

Precision and recall:

\[
\text{Precision: } \frac{6}{7}, \quad \text{Recall: } \frac{6}{10}
\]

37

Word-alignment Evaluation

System:
The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

Human:
The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

38

Problems for Statistical MT

Preprocessing
Language modeling
Translation modeling
Decoding
Parameter optimization
Evaluation

39

What kind of Translation Model?

Mary did not slap the green witch

Word-level models
Phrasal models
Syntactic models
Semantic models

Maria no dio una bofetada a la bruja verde

40
**Phrasal translation model**

The models define probabilities over inputs

\[ p(f \mid e) \]

1. Sentence is divided into phrases

---

**Phrase table**

<table>
<thead>
<tr>
<th>Translation</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>of course</td>
<td>0.5</td>
</tr>
<tr>
<td>naturally</td>
<td>0.3</td>
</tr>
<tr>
<td>of course,</td>
<td>0.15</td>
</tr>
<tr>
<td>, of course,</td>
<td>0.05</td>
</tr>
</tbody>
</table>
**Phrasal translation model**

The models define probabilities over inputs:

\[ p(f \mid e) \]

**Advantages of Phrase-Based**

Many-to-many mappings can handle non-compositional phrases

Easy to understand

Local context is very useful for disambiguating

- "Interest rate" → ...
- "Interest in" → ...

The more data, the longer the learned phrases

- Sometimes whole sentences!
Syntax-based models

Benefits
- Can use syntax to motivate word/phrase movement
- Could ensure grammaticality

Two main types:
• $p(\text{foreign string} \mid \text{English parse tree})$
• $p(\text{foreign parse tree} \mid \text{English parse tree})$

Tree to string rule

Tree to tree example

Problems for Statistical MT
- Preprocessing
- Language modeling
- Translation modeling
- Decoding
- Parameter optimization
- Evaluation
MT Evaluation

How do we do it?

What data might be useful?

Source only

Manual:
- SSER (subjective sentence error rate)
- Correct/Incorrect
- Error categorization

Extrinsic:
Objective usage testing

Automatic:
- WER (word error rate)
- BLEU (Bilingual Evaluation Understudy)
- NIST

Automatic Evaluation

Common NLP/machine learning/AI approach

Reference (human) translation:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Machine translation:
The American [?] international airport and its office all receives one calls self the sand Arab rich business [?] and so on electronic mail, which sends out; The threat will be able after public place and so on the airport to start the biochemistry attack, [?] highly alerts after the maintenance.

Machine translation 2:
United States Office of the Guam International Airport and were received by a man claiming to be Saudi Arabian businessman Osama bin Laden, sent emails, threats to airports and other public places will launch a biological or chemical attack, remain on high alert in Guam.

Ideas?
**Multiple Reference Translations**

**Reference translation 1:**

The Saudi Arabian Osama bin Laden and his other wealthy associates who command a great deal of money and influence the world are said to be planning to launch a biological attack against the U.S. island of Guam and other public places. They said there was a high state of alert after receiving an email from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport and its offices.

**Reference translation 2:**

The U.S. island of Guam is maintaining a high state of alert after the Guam International Airport and other public places such as the airport were put on alert. The Office of the Guam International Airport and other public places received an email from someone calling himself the sand Arab millionaire which threatens to launch a biochemical attack on such public places such as the airport. Guam International Airport and its offices both received an email from a person claiming to be the Saudi Arabian Osama bin Laden.

**Reference translation 3:**

The US International Airport of Guam is maintaining a high state of alert after the Guam International Airport and other public places such as the airport were put on alert. The Office of the Guam International Airport and other public places received an email from Mr. Bin Laden and that threatened to launch a biological and chemical attack on such public places as airport. Guam International Airport and its offices are maintaining a high state of alert after receiving an email from a person claiming to be the Saudi Arabian Osama bin Laden. They said there was a high state of alert after receiving an email from someone calling himself the sand Arab rich business which sends out electronic mail, which threatens to launch a biochemical attack against public places such as the airport and so on.

**Reference translation 4:**

A high state of alert needs to be in high precaution about Airport and other public places. Guam International Airport and other public places were put on alert after receiving an email from Saudi Arabia. They said there would be biochemistry air raid to Guam International Airport and other public places. Guam International Airport and its offices are maintaining a high state of alert after the Guam International Airport and its offices both received an email from someone calling himself the sand Arab millionaire which threatens to launch a biochemical attack against public places such as the airport.

---

**N-gram precision example**

*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 which ensures that the military always obey the commands of the party.

*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.

What percentage of machine n-grams can be found in the reference translations? Do unigrams, bigrams and trigrams.

*Unigrams:* 17/18

---

**BLEU Evaluation Metric**

(Papineni et al., ACL-2002)

**Basic idea:**

Combination of n-gram precisions of varying size

What percentage of machine n-grams can be found in the reference translation?
N-gram precision example

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.

Unigrams: 17/18
Bigrams: 10/17

N-gram precision example

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.

Unigrams: 17/18
Bigrams: 10/17
Trigrams: 7/16

N-gram precision example 2

Candidate 2: It is to ensure the army forever hearing the directions guide that party commands.
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.

Unigrams: 12/14

N-gram precision example 2

Candidate 2: It is to ensure the army forever hearing the directions guide that party commands.
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.

Unigrams: 12/14
N-gram precision example 2

Candidate 2: It is to ensure the army forever hearing the directions guide that party commands.

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.

Unigrams: 12/14
Bigrams: 4/13

Any problems/concerns?
Reference (human) translation:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Machine translation:
The American [?] international airport and its[?] the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

**Amazingly hard to “game” the system (i.e. find a way to change machine output so that BLEU goes up, but quality doesn’t)**

**BLEU Evaluation Metric**
(Papineni et al, ACL-2002)

- N-gram precision (score is between 0 & 1)
  - What percentage of machine n-grams can be found in the reference translation?
  - Not allowed to use same portion of reference translation twice (can’t cheat by typing out “the the the the”)

- Brevity penalty
  - Can’t just type out single word “the” (precision 1.0)

**BLEU Tends to Predict Human Judgments**

![Slide from G. Doddington (NIST)](image)

**BLEU: Problems?**

Doesn’t care if an incorrectly translated word is a name or a preposition
- gave it to Albright (reference)
- gave it at Albright (translation #1)
- gave it to altar (translation #2)

What happens when a program reaches human level performance in BLEU but the translations are still bad?
- maybe sooner than you think ...