Alignment in Machine Translation

CMSC 723 / LING 723 / INST 725

MARINE CARPUAT
marine@cs.umd.edu
<table>
<thead>
<tr>
<th>Centauri/Arcturan [Knight, 1997]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp</td>
</tr>
<tr>
<td><strong>1a. ok-voon ororok sprok .</strong></td>
</tr>
<tr>
<td><strong>1b. at-voon bichat dat .</strong></td>
</tr>
<tr>
<td><strong>2a. ok-drubel ok-voon anok plok sprok .</strong></td>
</tr>
<tr>
<td><strong>2b. at-drubel at-voon pippat rrat dat .</strong></td>
</tr>
<tr>
<td><strong>3a. erok sprok izok hihok ghirok .</strong></td>
</tr>
<tr>
<td><strong>3b. totat dat arrat vat hilat .</strong></td>
</tr>
<tr>
<td><strong>4a. ok-voon anok drok brok jok .</strong></td>
</tr>
<tr>
<td><strong>4b. at-voon krat pippat sat lat .</strong></td>
</tr>
<tr>
<td><strong>5a. wiwok farok izok stok .</strong></td>
</tr>
<tr>
<td><strong>5b. totat jjat quat cat .</strong></td>
</tr>
<tr>
<td><strong>6a. lalok sprok izok jok stok .</strong></td>
</tr>
<tr>
<td><strong>6b. wat dat kratquat cat .</strong></td>
</tr>
<tr>
<td>English</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>1a. Garcia and associates.</td>
</tr>
<tr>
<td>1b. Garcia y asociados.</td>
</tr>
<tr>
<td>2a. Carlos Garcia has three associates.</td>
</tr>
<tr>
<td>2b. Carlos Garcia tiene tres asociados.</td>
</tr>
<tr>
<td>3a. his associates are not strong.</td>
</tr>
<tr>
<td>3b. sus asociados no son fuertes.</td>
</tr>
<tr>
<td>4a. Garcia has a company also.</td>
</tr>
<tr>
<td>4b. Garcia también tiene una empresa.</td>
</tr>
<tr>
<td>5a. its clients are angry.</td>
</tr>
<tr>
<td>5b. sus clientes están enfadados.</td>
</tr>
<tr>
<td>6a. the associates are also angry.</td>
</tr>
<tr>
<td>6b. los asociados también están enfadados.</td>
</tr>
</tbody>
</table>
When I look at an article in Russian, I say to myself: This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.
A Statistical Approach to Machine Translation

IBM
Thomas J. Watson Research Center
Yorktown Heights, NY

In this paper, we present a statistical approach to machine translation. We describe the application of our approach to translation from French to English and give preliminary results.

The COLING Paper Review

The validity of statistical (information theoretic) approach to MT has indeed been recognized, as the authors mention, by Weaver as early as 1949. And was universally recognized as mistaken by 1950. (cf. Hutchins, MT: Past, Present, Future, Ellis Horwood, 1986, pp. 30ff. and references therein) The crude force of computers is not science. The paper is simply beyond the scope of COLING.

More about the IBM story: 20 years of bitext workshop
The noisy channel model decomposes machine translation into two independent subproblems:

- Language modeling
- Translation modeling / Alignment
WORD ALIGNMENT
How can we model $p(f|e)$?

• We’ll describe the word alignment models introduced in early 90s at IBM

• Assumption: each French word $f$ is aligned to exactly one English word $e$
  – Including NULL

```
NULL      And      the      program      has      been      implemented
Le       programme    a      ete      mis      en      application
```
Word Alignment
Vector Representation

• Alignment vector $a = [2, 3, 4, 5, 6, 6, 6, 6, 6, 6, 6]$
  – length of $a = \text{length of sentence f}$
  – $a_i = j$ if French position $i$ is aligned to English position $j$
Formalizing the connection between word alignments & the translation model

\[ p(f_1, f_2, \ldots, f_m \mid e_1, e_2, \ldots, e_l, m) = \sum_{a \in A} p(f_1, \ldots, f_m, a_1, \ldots, a_m \mid e_1, \ldots, e_l, m) \]

• We define a conditional model
  – Projecting word translations
  – Through alignment links
How many possible alignments in $A$?

- How many possible alignments for $(f,e)$ where
  - $f$ is French sentence with $m$ words
  - $e$ is an English sentence with $l$ words

- For each of $m$ French words, we choose an alignment link among $(l+1)$ English words

- Answer: $(l + 1)^m$
IBM Model 1: generative story

• Input
  – an English sentence of length $l$
  – a length $m$

• For each French position $i$ in $1..m$
  – Pick an English source index $j$
    
  
    \[ q(j \mid i, l, m) = \frac{1}{l + 1} \]

  – Choose a translation

  \[ t(f_i \mid e_{a_i}) \]
IBM Model 1: generative story

• Input
  – an English sentence of length $l$
  – a length $m$

• For each French position $i$ in 1..m
  – Pick an English source index $j$
  – Choose a translation

Alignment is based on word positions, not word identities. Alignment probabilities are UNIFORM.

\[
q(j \mid i, l, m) = \frac{1}{l + 1}
\]

Words are translated independently.
IBM Model 1: Parameters

- $t(f|e)$
  - Word translation probability table
  - for all words in French & English vocab

| f       | e     | $p(f | e)$ |
|---------|-------|-----------|
| le      | the   | 0.42      |
| la      | the   | 0.4       |
| programme | the  | 0.001     |
| a       | has   | 0.78      |
| ...     | ...   | ...       |
IBM Model 1: generative story

• Input
  – an English sentence of length l
  – a length m

• For each French position \( i \) in \( 1..m \)
  – Pick an English source index \( j \)
  – Choose a translation

\[
q(j \mid i, l, m) = \frac{1}{l + 1}
\]

\[
t(f_i \mid e_{a_i})
\]

\[
p(f_1 \ldots f_m, a_1 \ldots a_m \mid e_1 \ldots e_l, m) = \prod_{i=1}^{m} q(a_i \mid i, l, m) t(f_i \mid e_{a_i})
\]
Improving on IBM Model 1: IBM Model 2

• Input
  – an English sentence of length l
  – a length m

• For each French position $i$ in 1..m
  – Pick an English source index $j$ $q(j \mid i, l, m)$
  – Choose a translation $t(f_i \mid e_{a_i})$

Remove assumption that $q$ is uniform
IBM Model 2: Parameters

• $q(j|i,l,m)$
  – now a table
  – not uniform as in IBM1

• How many parameters are there?

<table>
<thead>
<tr>
<th>j</th>
<th>$q(j \mid 1, 6, 7)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.27</td>
</tr>
<tr>
<td>2</td>
<td>0.14</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>48</td>
<td>1E-75</td>
</tr>
</tbody>
</table>
2 Remaining Tasks

Inference
• Given
  – a sentence pair (e,f)
  – an alignment model with parameters $t(e|f)$ and $q(j|i,l,m)$
• What is the most probable alignment $a$?

Parameter Estimation
• Given
  – training data (lots of sentence pairs)
  – a model definition
• how do we learn the parameters $t(e|f)$ and $q(j|i,l,m)$?
Inference

• Inputs
  – Model parameter tables for $t$ and $q$
  – A sentence pair

<table>
<thead>
<tr>
<th>NULL</th>
<th>And</th>
<th>the program has been implemented</th>
</tr>
</thead>
<tbody>
<tr>
<td>Le</td>
<td>programme a ete mis en application</td>
<td></td>
</tr>
</tbody>
</table>

• How do we find the alignment $a$ that maximizes $P(e,a|f)$?
  – Hint: recall independence assumptions!
Inference

• Inputs
  – Model parameter tables for t and q
  – A sentence pair

  ![Sentence Pair Diagram]

• How do we find the alignment $a$ that maximizes $P(e,a|f)$?
  – Hint: recall independence assumptions!
Inference

• Inputs
  – Model parameter tables for t and q
  – A sentence pair


• How do we find the alignment \( a \) that maximizes \( P(e,a|f) \)?
  – Hint: recall independence assumptions!
Inference

• Inputs
  – Model parameter tables for t and q
  – A sentence pair

  ![Sentence Pair Diagram]

• How do we find the alignment $a$ that maximizes $P(e,a|f)$?
  – Hint: recall independence assumptions!
Inference

• Inputs
  – Model parameter tables for t and q
  – A sentence pair

\[
\text{NULL} \quad \text{And} \quad \text{the} \quad \text{program} \quad \text{has} \quad \text{been} \quad \text{implemented}
\]

\[
\text{Le} \quad \text{programme} \quad \text{a} \quad \text{ete} \quad \text{mis} \quad \text{en} \quad \text{application}
\]

• How do we find the alignment \( a \) that maximizes \( P(e,a|f) \)?
  – Hint: recall independence assumptions!
Inference

• **Inputs**
  – Model parameter tables for t and q
  – A sentence pair

![Sentence alignment diagram]

• How do we find the alignment \( a \) that maximizes \( P(e, a | f) \)?
  – Hint: recall independence assumptions!
1 Remaining Task

**Inference**
- Given a sentence pair \((e,f)\), what is the most probable alignment \(a\)?

**Parameter Estimation**
- How do we learn the parameters \(t(e|f)\) and \(q(j|i,l,m)\) from data?
Parameter Estimation

• Problem
  – Parallel corpus gives us \((e, f)\) pairs only, \(a\) is hidden

• We know how to
  – estimate \(t\) and \(q\), given \((e, a, f)\)
  – compute \(p(e, a|f)\), given \(t\) and \(q\)

• Solution: Expectation-Maximization algorithm (EM)
  – E-step: given hidden variable, estimate parameters
  – M-step: given parameters, update hidden variable
Parameter Estimation: EM

Algorithm 1 (soft EM)

initialize parameters $t$ and $q$ to something
repeat until convergence
for every sentence
    for every target position $j$
        for every source position $i$
            count($f_j, e_i$) += $P(a_i = j | e_i, f_j)$
            count($e_i$) += $P(a_i = j | e_i, f_j)$
            count($j, i, l, m$) += $P(a_i = j | e_i, f_j)$
            count($j, i, l, m$) += $P(a_i = j | e_i, f_j)$

$$t(f | e) = \frac{\text{count}(f, e)}{\text{count}(e)}$$
$$q(j | i, l, m) = \frac{\text{count}(j, i, l, m)}{\text{count}(i, l, m)}$$

Use “Soft” values instead of binary counts
Parameter Estimation: soft EM

- Soft EM considers all possible alignment links
- Each alignment link now has a weight

\[
P(a_i = j \mid e_i, f_j) = \frac{q(j \mid i, l, m) \cdot t(f_i \mid e_j)}{\sum_{j'=1}^{l} q(j' \mid i, l, m) \cdot t(f_i \mid e_{j'})}
\]
EM for IBM Model 1

• Expectation (E)-step:
  – Compute expected counts for parameters \( (t) \) based on summing over hidden variable

• Maximization (M)-step:
  – Compute the maximum likelihood estimate of \( t \) from the expected counts
EM example: initialization

For the rest of this talk, French = Spanish

green house  the house

casa verde  la casa

|                 | $t(casa|green)$ | $t(verte|green)$ | $t(la|green)$ |
|-----------------|-----------------|------------------|---------------|
| $t(casa|house)$ | $\frac{1}{3}$   | $\frac{1}{3}$   | $\frac{1}{3}$ |
| $t(casa|the)$   | $\frac{1}{3}$   | $\frac{1}{3}$   | $\frac{1}{3}$ |
EM example: E-step
(a) compute probability of each alignment \( p(a|f,e) \)

Note: we’re making simplification assumptions in this example
• No NULL word
• We only consider alignments were each French and English word is aligned to something
• We ignore \( q! \)
EM example: E-step
(b) normalize to get \( p(a|f,e) \)

\[
\begin{array}{c|c|c}
\text{green} & \text{house} & \text{green} \\
\hline
\text{casa} & \text{verde} & \text{casa} \\
\end{array}
\]

\[
\frac{P(a|f,e) = \frac{1}{9}}{2/9} = \frac{1}{2}
\]

\[
\begin{array}{c|c|c}
\text{the} & \text{house} & \text{the} \\
\hline
\text{la casa} & \text{casa} & \text{la casa} \\
\end{array}
\]

\[
\frac{P(a|f,e) = \frac{1}{9}}{2/9} = \frac{1}{2}
\]

\[
\frac{P(a|f,e) = \frac{1}{9}}{2/9} = \frac{1}{2}
\]

\[
\frac{P(a|f,e) = \frac{1}{9}}{2/9} = \frac{1}{2}
\]
EM example: E-step
(c) compute expected counts
(weighting each count by p(a|e,f)

\[ P(a|f,e) = \frac{1/9}{2/9} = \frac{1}{2} \]

<table>
<thead>
<tr>
<th>green</th>
<th>house</th>
<th>green</th>
<th>house</th>
<th>the</th>
<th>house</th>
<th>the</th>
<th>house</th>
</tr>
</thead>
<tbody>
<tr>
<td>casa</td>
<td>verde</td>
<td>casa</td>
<td>verde</td>
<td>la</td>
<td>casa</td>
<td>la</td>
<td>casa</td>
</tr>
</tbody>
</table>

\[ P(a|f,e) = \frac{1/9}{2/9} = \frac{1}{2} \]

| tcount(casa|green) = \frac{1}{2} | tcount(verte|green) = \frac{1}{2} | tcount(la|green) = 0 | total(green) = 1 |
| tcount(casa|house) = \frac{1}{2} + \frac{1}{2} | tcount(verte|house) = \frac{1}{2} | tcount(la|house) = \frac{1}{2} | total(house) = 2 |
| tcount(casa|the) = \frac{1}{2} | tcount(verte|the) = 0 | tcount(la|the) = \frac{1}{2} | total(the) = 1 |
EM example: M-step
Compute probability estimate by normalizing expected counts

|                | $t(casa|green)$ | $t(vero|green)$ | $t(la|green)$ |
|----------------|-----------------|-----------------|---------------|
| $t(casa|house)$ | $\frac{1/2}{1} = \frac{1}{2}$ | $\frac{1/2}{1} = \frac{1}{2}$ | $\frac{0}{1} = 0$ |
| $t(casa|the)$  | $\frac{1/2}{1} = \frac{1}{2}$ | $\frac{0}{1} = 0$ | $\frac{1/2}{1} = \frac{1}{2}$ |
| $t(vero|house)$| $\frac{1/2}{2} = \frac{1}{4}$ | $\frac{1/2}{2} = \frac{1}{4}$ | $\frac{1/2}{2} = \frac{1}{4}$ |
| $t(vero|the)$  | $\frac{0}{1} = 0$ | $\frac{0}{1} = 0$ | $\frac{1/2}{1} = \frac{1}{2}$ |
EM example: next iteration

\[
\begin{align*}
\text{green} & \quad \text{house} \\
\text{casa} & \quad \text{verde} \\
P(a, f | e) &= t(\text{casa}, \text{green}) \\
& \quad \times t(\text{verde}, \text{house}) \\
&= \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}
\end{align*}
\]

\[
\begin{align*}
\text{green} & \quad \text{house} \\
\text{casa} & \quad \text{verde} \\
P(a, f | e) &= t(\text{verde}, \text{green}) \\
& \quad \times t(\text{casa}, \text{house}) \\
&= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}
\end{align*}
\]

\[
\begin{align*}
\text{the} & \quad \text{house} \\
\text{la} & \quad \text{casa} \\
P(a, f | e) &= t(\text{la}, \text{the}) \\
& \quad \times t(\text{casa}, \text{house}) \\
&= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}
\end{align*}
\]

\[
\begin{align*}
\text{the} & \quad \text{house} \\
\text{la} & \quad \text{casa} \\
P(a, f | e) &= t(\text{casa}, \text{the}) \\
& \quad \times t(\text{la}, \text{house}) \\
&= \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}
\end{align*}
\]
Parameter Estimation with EM

- EM guarantees that data likelihood does not decrease across iterations

\[
\log \mathcal{L}(t, q \mid E, F) = \log \prod_{n=1}^{N} \sum_{f^{(n)} \mid e^{(n)}} p(f^{(n)} \mid e^{(n)})
\]

\[
= \sum_{n=1}^{N} \log \sum_{a \in A} p(f^{(n)}, a \mid e^{(n)})
\]

- EM can get stuck in a local optimum
  - Initialization matters
Word Alignment with IBM Models 1, 2

• Probabilistic models with strong independence assumptions
  – Results in linguistically naïve models
    • asymmetric, 1-to-many alignments
  – But allows efficient parameter estimation and inference

• Alignments are hidden variables
  – unlike words which are observed
  – require unsupervised learning (EM algorithm)
PHRASE-BASED MODELS
Phrase-based models

- Most common way to model $P(F|E)$ nowadays (instead of IBM models)

\[
P(F|E) = \prod_{i=1}^{I} \phi(f_i, \bar{e}_i)d(a_i - b_{i-1})
\]

- Probability of two consecutive English phrases being separated by a particular span in French

- Start position of $f_i$

- End position of $f_{i-1}$
Phrase alignments are derived from word alignments.

This means that the IBM model represents \( P(\text{Spanish} \mid \text{English}) \).

Get high confidence alignment links by intersecting IBM word alignments from both directions.
Phrase alignments are derived from word alignments.

<table>
<thead>
<tr>
<th>Mary</th>
<th>did</th>
<th>not</th>
<th>slap</th>
<th>the</th>
<th>green</th>
<th>witch</th>
</tr>
</thead>
<tbody>
<tr>
<td>María</td>
<td>no</td>
<td>dió</td>
<td>una</td>
<td>a</td>
<td>la</td>
<td>verde</td>
</tr>
</tbody>
</table>

 Improve recall by adding some links from the union of alignments.
Phrase alignments are derived from word alignments

Extract phrases that are **consistent** with word alignment

(Maria, Mary), (no, did not),
(slap, dió una bofetada), (verde, green),
(a la, the), (bruja, witch),
(Maria no, Mary did not),
(no dió una bofetada, did not slap),
(dió una bofetada a la, slap the),
bruja verde, green witch),
(a la bruja verde, the green witch), …
Phrase Translation Probabilities

• Given such phrases we can get the required statistics for the model from

\[ \phi(\bar{f}, \bar{e}) = \frac{\text{count}(\bar{f}, \bar{e})}{\sum_{\bar{f}} \text{count}(\bar{f}, \bar{e})} \]
Phrase-based Machine Translation

\[ \hat{E} = \arg \max_{E \in \text{English}} \left( P(F|E) \cdot P(E) \right) \]

\[ \prod_{i \in S} \phi(\bar{f}_i, \bar{e}_i) d(a_i - b_{i-1}) P(E) \]
RECAP
Noisy Channel Model for Machine Translation

\[ \hat{E} = \arg\max_{E \in \text{English}} \frac{P(F|E)}{P(E)} \]

- The **noisy channel model** decomposes machine translation into two independent subproblems
  - Language modeling
  - Translation modeling / Alignment
Word Alignment with IBM Models 1, 2

• Probabilistic models with strong independence assumptions

• Alignments are hidden variables
  – unlike words which are observed
  – require unsupervised learning (EM algorithm)

• Word alignments often used as building blocks for more complex translation models
  – E.g., phrase-based machine translation