Lecture 8: 
Phrase-Based Translation 
Ankur Parikh – Google

Slides based on notes from Michael Collins and also material from Slav Petrov, Philipp Koehn, and Chris Dyer
Overview of Rest of Class

- **Today** - Phrase Based Machine Translation
- **Nov 13, 20** - Parsing (Guest Lecturer: Slav Petrov)
- **Nov 27** - Compositional Semantics
- **Dec 4** - Neural Machine Translation
- **Dec 11** - Summarization / Text Generation
Outline

- Machine Translation Overview
- Evaluation Metrics
- IBM Model Decoding
- A Phrase Based MT System:
  - Constructing a Phrase Table
  - Finding the Best Derivation
- Scoring the Derivation
Outline

- Machine Translation Overview
- Evaluation Metrics
- IBM Model Decoding
- A Phrase Based MT System:
  - Constructing a Phrase Table
  - Finding the Best Derivation
- Scoring the Derivation
Machine Translation

- Translate text from one language to another
- Recombines fragments of example translations
- Challenges:
  - What fragments? [learning to translate]
  - How to make efficient? [fast translation search]
  - Fluency vs fidelity
Corpus-Based MT

Modeling correspondences between languages

Sentence-aligned parallel corpus:

Yo lo haré mañana
I will do it tomorrow

Hasta pronto
See you soon

Hasta pronto
See you around

Machine translation system:

Yo lo haré pronto
I will do it soon

Model of translation
I will do it around

See you tomorrow
Levels of Transfer

- **Source**: Words → Phrases → Syntax → Semantics → Interlingua
- **Target**: Words → Phrases → Syntax → Semantics

Example:

- **Spanish**: Yo lo haré mañana
- **English**: I will do it tomorrow

- **Probability Calculation**:
  
  $P(\text{E} | \text{lo haré }) = 0.8$

- **Conditional Probabilities**:
  
  | English (E)          | $P(\text{E} | \text{lo haré })$ |
  |----------------------|-------------------------------|
  | will do it           | 0.8                           |
  | will do so           | 0.2                           |

- **Spanish-French Translation**:
  
  | English (E)          | $P(\text{E} | \text{mañana }$ |
  |----------------------|-------------------------------|
  | tomorrow             | 0.7                           |
  | morning              | 0.3                           |
MT System Components

Language Model

source

P(e)

Translation Model

channel

P(f|e)

decoder

e

best

e

observed

f

\[ e^* = \arg\max_e P(e|f) = \arg\max_e P(f|e)P(e) \]
MT System Components

Language Model

source
P(e)

best

decoder

e

Translation Model

channel
P(f|e)

observed

f

\[ e^* = \arg\max_e P(e|f) = \arg\max_e P(f|e) P(e) \]
Outline

- Machine Translation Overview
- Evaluation Metrics
- IBM Model Decoding
- A Phrase Based MT System:
  - Constructing a Phrase Table
  - Finding the Best Derivation
- Scoring the Derivation
MT: Evaluation

- Human evaluations: subject measures, fluency/adequacy
- Automatic measures: n-gram match to references
  - NIST measure: n-gram recall (worked poorly)
  - BLEU: n-gram precision (no one really likes it, but everyone uses it)
- BLEU:
  - P1 = unigram precision
  - P2, P3, P4 = bi-, tri-, 4-gram precision
  - Weighted geometric mean of P1-4
  - Brevity penalty (why?)
  - Somewhat hard to game…
Why Automatic Evaluation is Hard

Exact reconstruction (24 of 38)

Please give me your response as soon as possible.
⇒ Please give me your response as soon as possible.

Reconstruction preserving meaning (8 of 38)

Now let me mention some of the disadvantages.
⇒ Let me mention some of the disadvantages now.

Garbage reconstruction (6 of 38)

In our organization research has two missions.
⇒ In our missions research organization has two.
Candidate: the the the the the the the the
Reference 1: The cat is on the mat.
Reference 2: There is a cat on the mat.

vanilla unigram precision = 1.0  (yikes!!)
Use a modified unigram precision instead by clipping counts. Let $x_w$ = maximum times number of times a word appears in any given translation

**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.

$x_{\text{the}} = 2$

modified unigram precision $= \frac{\min(7, 2)}{7} = \frac{2}{7}$
Can then combine $n$-grams and weighted precisions

$$\exp\left(\sum_{n} w_n \rho_n \right)$$

weight

(modified) precision for given $n$-gram size

What is the problem with this?
So far we only have a precision-based measure e.g. consider the problem case below:

**Candidate:** of the

**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 the directions of the party.*

Modified unigram/bigram precisions are both 1.0
Typically, recall is used to solve this problem, but recall doesn’t work well with multi-reference.

Candidate 1: 

I always do.

Candidate 2: 

I always invariably perpetually do.

Reference 1: 

I always do.

Reference 2: 

I invariably do.

Reference 3: 

I perpetually do.

Candidate 2 will get a high recall, even if it is incorrect.
The proposed solution in BLEU is to use a brevity penalty (BP), encouraging the candidate translation to have length similar to one of the reference translations.

\[ c = \text{candidate translation length} \]
\[ r = \text{effective reference translation length} \]

\[
BP = \begin{cases} 
1 & c > r \\
\exp(1 - \frac{r}{c}) & c \leq r 
\end{cases}
\]

\[
\text{BLEU} = BP \times \exp\left(\sum_n w_n p_n \right)
\]
Automatic Metrics Work (?)
Outline

- Machine Translation Overview
- Evaluation Metrics
- IBM Model Decoding
- A Phrase Based MT System:
  - Constructing a Phrase Table
  - Finding the Best Derivation
- Scoring the Derivation
Translation with IBM Models

\[ e^* = \arg\max_e P(e|f) = \arg\max_e P(f|e)P(e) \]

Let us first pretend the translation model \( P(f|e) \) is IBM Model 2.

Two challenges:
- Computing the marginal probability \( P(f | e) \)
- Taking the argmax over all possible \( e \)
Translation with IBM Models

\[ e^* = \text{argmax}_e P(e|f) = \text{argmax}_e P(f|e)P(e) \]

Let us first pretend the translation model \( P(f|e) \) is IBM Model 2.

Two challenges:
- Computing the marginal probability \( P(f | e) \)
- Taking the argmax over all possible \( e \)
IBM Models: Computing $P(f|e)$

$$P(f|e) = \sum_{a_1, \ldots, a_m} P(f_1, \ldots, f_m, a_1, \ldots, a_m | e_1, \ldots, e_l, m)$$

$$= \sum_{a_1, \ldots, a_m} \prod_{i=1}^{m} q(a_i | i, l, m) t(f_i | e_{a_i})$$

*Intractable in general*
IBM Model 1: Computing $P(f|e)$

IBM Model 1 is special and it is possible to exactly compute the marginal probability exactly

$$P(f|e) = \sum_{a_1, \ldots, a_m} \prod_{i=1}^{m} \frac{1}{\ell + 1} t(f_i | e_{a_i})$$

$$\prod_{i=1}^{m} \frac{1}{\ell + 1} \sum_{a_i} t(f_i | e_{a_i})$$
IBM Model 2: Computing $P(f|e)$

One typical heuristic is to replace the sum over all possible alignments with just the top $Z$ most likely alignments.

$$P(f|e) = \sum_{a_1, \ldots, a_m} P(f_1, \ldots, f_m, a_1, \ldots, a_m | e_1, \ldots, e_l, m)$$

$$= \sum_{a_1, \ldots, a_m} \prod_{i=1}^{m} q(a_i | i, l, m) t(f_i | e_{a_i})$$

$$\approx \sum_{z=1}^{Z} \prod_{i=1}^{m} q(a_{i}^{z} | i, l, m) t(f_i | e_{a_i}^{z})$$

$(a_{i}^{1}, \ldots, a_{i}^{Z})$ is the $z^{th}$ highest scoring alignment

In which situations would we expect this to work well?
Computing the most likely alignment is tractable for IBM Model 2

\[ a_1^*, \ldots, a_m^* = \arg\max_{a_1, \ldots, a_m} P(f_1, \ldots, f_m, a_1, \ldots, a_m | e_1, \ldots, e_L, m) \]
\[ = \arg\max_{a_1, \ldots, a_m} \prod_{i=1}^{m} q(a_i | i, l, m) t(f_i | e_{a_i}) \]

\[ a_i^* = \arg\max_{a_i} (q(a_i | i, l, m) t(f_i | e_{a_i})) \]
Decoding

\[ e^* = \arg \max_e P(e|f) = \arg \max_e P(f|e)P(e) \]

Let us first pretend the translation model \( P(f|e) \) is IBM Model 2.

Two challenges:

- Computing the marginal probability \( P(f | e) \)
- Taking the \( \text{argmax} \) over all possible \( e \)

Number of possibilities is huge!!

- NP-hard, reduces to Traveling Salesman Problem (if \( P(e) \) is a bigram language model)
- Even if we knew the set of words in the optimal \( e \), still \( m! \) number of permutations.

We will discuss strategies for this problem later in the lecture.
No one actually uses IBM models as the translation model.

Instead we use the alignments extracted from the IBM models to discover phrases and do translation at the phrase level.
Phrase-Based Systems

Morgen \rightarrow Tomorrow
fliege \rightarrow I
ich \rightarrow will fly
nach Kanada \rightarrow to the conference
zur Konferenz \rightarrow in Canada

cat \mid|\mid chat \mid|\mid 0.9
the cat \mid|\mid le chat \mid|\mid 0.8
dog \mid|\mid chien \mid|\mid 0.8
house \mid|\mid maison \mid|\mid 0.6
my house \mid|\mid ma maison \mid|\mid 0.9
language \mid|\mid langue \mid|\mid 0.9
...

Sentence-aligned corpus

Word alignments

Phrase table (translation model)
Outline

- Machine Translation Overview
- Evaluation Metrics
- IBM Model Decoding

- A Phrase Based MT System:
  - Constructing a Phrase Table
  - Finding the Best Derivation

- Scoring the Derivation
A phrase lexicon is a set of tuples. Phrases do not have to match in length.

\[(f, e, g)\]

- (au, to the, 0.5)
- (au banque, to the bank, 0.01)
- (allez au banque, go to the bank, -2.5)
Constructing a Phrase Lexicon

Assume an alignment matrix $A^{(k)}$ is given for each example $k$.

$$A^{(k)}_{i,j} = 1 \text{ if Foreign word } i \text{ is aligned to English word } j, \ 0 \text{ otherwise}$$

We can use IBM models (in combination with some heuristics) for example, to construct such a matrix.
Bidirectional Alignment

**English to Spanish**

- Mary
- did
- not
- slap

**Spanish to English**

- Maria no daba una bofetada a la bruja verde

**Intersection**

- Mary did not slap the green witch
Constructing a Phrase Lexicon

- Once we have an alignment matrix, the next step is to use it to extract phrases / scores.

- We iterate over all training examples $k$
  - For each phrase pairs $(s, t)$ from the source and $(s', t')$ from the target add it to our phrase lexicon if it is consistent with the alignment matrix.
Constructing a Phrase Lexicon

nous devons aussi prendre ces critiques au sérieux

we must also take these criticisms seriously

\( O(m^4) \) possible phrase pairs (but we will enumerate sparsely)

\((s,t) = (1, 2), (s', t') = (2, 5)\) corresponds to:

(nous devons, must also take these)

Add these phrases to the phrase lexicon if they are consistent with A
Consistent Phrases

- All words of the phrase have to align to each other (or to nothing)
Extracting Phrases

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch),
(vero, green), (Maria no, Mary did not), (no daba una bofetada, did not slap),
(daba una bofetada a la, slap the), (bruja verde, green witch),
(Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),
(Maria no daba una bofetada a la, Mary did not slap the),
(daba una bofetada a la bruja verde, slap the green witch),
(no daba una bofetada a la bruja verde, did not slap the green witch),
(Maria no daba una bofetada a la bruja verde, Mary did not slap the green witch)
Phrase Weights

How the MT community estimates $P(\bar{f} | \bar{e})$

Parallel training sentences provide phrase pair counts.

Gracias, lo haré de muy buen grado. Thank you, I shall do so gladly.

lo haré $\leftrightarrow$ I shall do so
44 times in the corpus

All phrase pairs are counted, and counts are normalized.

Gracias, lo haré de muy buen grado.
Thank you, I shall do so gladly.

$P(\bar{f} | \bar{e}) = \frac{\text{count}(\bar{f}, \bar{e})}{\text{count}(\bar{e})}$
Phrase Scoring

\[
\phi(\bar{e}_j | \bar{f}_i) = \frac{c(\bar{f}_i, \bar{e}_j)}{c(\bar{f}_i)} \\
g(\bar{e}_j, \bar{f}_i) = \log \phi(\bar{e}_j | \bar{f}_i)
\]

- Learning weights has been tried, several times:
  - [Marcu and Wong, 02]
  - [DeNero et al, 06]
  - … and others

- Seems not to work well, for a variety of partially understood reasons

- Main issue: big chunks get all the weight, obvious priors don’t help
  - Though, [DeNero et al 08]
Phrase Table Example

- Phrase translations for ‘der Vorschlag’:

| English              | $\phi(\bar{e}|f)$ | English              | $\phi(\bar{e}|f)$ |
|----------------------|-------------------|----------------------|-------------------|
| the proposal         | 0.6227            | the suggestions      | 0.0114            |
| ’s proposal          | 0.1068            | the proposed         | 0.0114            |
| a proposal           | 0.0341            | the motion           | 0.0091            |
| the idea             | 0.0250            | the idea of          | 0.0091            |
| this proposal        | 0.0227            | the proposal ,       | 0.0068            |
| proposal             | 0.0205            | its proposal         | 0.0068            |
| of the proposal      | 0.0159            | it                   | 0.0068            |
| the proposals        | 0.0159            | ...                  | ...               |

- Lexical variation, morphology, function words
Outline

- Machine Translation Overview
- Evaluation Metrics
- IBM Model Decoding
- A Phrase Based MT System:
  - Constructing a Phrase Table
  - Finding the Best Derivation
- Experiments
Derivations

- A derivation \( y \) is a finite sequence of phrases from the phrase table.

- Denote each phrase with the short-hand:
  
  \[
  p = (s, t, e)
  \]
  
  start index of source phrase - denote as \( s(p) \)
  
  end index of source phrase - denote as \( t(p) \)
  
  target phrase
Derivations

Example phrases in this notation:

\[ p_1 = (1, 3, \text{we must also}) \]
\[ p_2 = (7, 7, \text{take}) \]
\[ p_3 = (4, 5, \text{this criticism}) \]
\[ p_4 = (6, 6, \text{seriously}) \]

Derivation:

\[ y = p_1 \ p_2 \ p_3 \ p_4 \]

\[ y = (1, 3, \text{we must also}) \ (7, 7, \text{take}) \ (4, 5, \text{this criticism}) \ (6, 6, \text{seriously}) \]
Constraints on the Derivation

- Each phrase in the derivation must come from the phrase table.

- Each source word is translated exactly one.

- A distortion limit that limits how far consecutive phrases can be found from each other.
Distortion Limit

- Distortion limit $d$, which limits how far consecutive phrases can be found from each other.

\[
|1 - s(p_1)| \leq d
\]

\[
|t(p_2) + 1 - s(p_3)| \leq d
\]

- In addition to improving performance, helps make decoding more efficient.
- Similar intuition to HMM alignment model
Scoring the Derivation

\[ y^* = \arg\max_y f(y) \]

\[ f(y) = h(e(y)) + \sum_{k=1}^{L} g(p_k) + \sum_{k=1}^{L-1} \eta \times |t(p_k) + 1 - s(p_{k+1})| \]

target language model \( \log p(e) \)

translation model \( \log p(f \mid e) \)

extra distortion penalty (in addition to hard constraint)
Scoring the Derivation

Log form

\[ h(e(y)) + \sum_{k=1}^{L} g(p_k) + \sum_{k=1}^{L-1} \eta \times |t(p_k) + 1 - s(p_{k+1})| \]

Exponential form

\[ \exp(h(e(y))) \prod_{k=1}^{L} \exp(g(p_k)) \prod_{k=1}^{L-1} \left( \exp \left( |t(p_k) + 1 - s(p_{k+1})| \right) \right)^{\eta} \]

\[ P_{lm}(e(y)) \prod_{k=1}^{L} \phi(f_k | \bar{e}_k) \prod_{k=1}^{L-1} \left( \exp \left( |t(p_k) + 1 - s(p_{k+1})| \right) \right)^{\eta} \]
Decoding

- Define a state:

\[ (e_{inc}, b, r, \alpha) \]

- Example:

\[ y = (1, 3, \text{we must also}) \ (7, 7, \text{take}) \ (4, 5, \text{this criticism}) \]

(we must also take this criticism, 111101, 5, \alpha)
Decoding

- Initialize the state to:

  \((\cdot, 0000000, 0, 0)\)

- When considering the next step of the translation consider all possible phrases that could append to the above state.
  - Subject to constraints

- Keep the K-best incremental translations so far (beam search)
Decoding
Outline

- Machine Translation Overview
- Evaluation Metrics
- IBM Model Decoding
- A Phrase Based MT System:
  - Constructing a Phrase Table
  - Finding the Best Derivation
- Experiments
Results

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Model4</th>
<th>Phrase</th>
<th>Lex</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-German</td>
<td>0.2040</td>
<td>0.2361</td>
<td>0.2449</td>
</tr>
<tr>
<td>French-English</td>
<td>0.2787</td>
<td>0.3294</td>
<td>0.3389</td>
</tr>
<tr>
<td>English-French</td>
<td>0.2555</td>
<td>0.3145</td>
<td>0.3247</td>
</tr>
<tr>
<td>Finnish-English</td>
<td>0.2178</td>
<td>0.2742</td>
<td>0.2806</td>
</tr>
<tr>
<td>Swedish-English</td>
<td>0.3137</td>
<td>0.3459</td>
<td>0.3554</td>
</tr>
<tr>
<td>Chinese-English</td>
<td>0.1190</td>
<td>0.1395</td>
<td>0.1418</td>
</tr>
</tbody>
</table>

Koehn et al. 2003
Phrase Translation Tables

<table>
<thead>
<tr>
<th>Max. Length</th>
<th>10k</th>
<th>20k</th>
<th>40k</th>
<th>80k</th>
<th>160k</th>
<th>320k</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>37k</td>
<td>70k</td>
<td>135k</td>
<td>250k</td>
<td>474k</td>
<td>882k</td>
</tr>
<tr>
<td>3</td>
<td>63k</td>
<td>128k</td>
<td>261k</td>
<td>509k</td>
<td>1028k</td>
<td>1996k</td>
</tr>
<tr>
<td>4</td>
<td>84k</td>
<td>176k</td>
<td>370k</td>
<td>736k</td>
<td>1536k</td>
<td>3152k</td>
</tr>
<tr>
<td>5</td>
<td>101k</td>
<td>215k</td>
<td>459k</td>
<td>925k</td>
<td>1968k</td>
<td>4119k</td>
</tr>
<tr>
<td>7</td>
<td>130k</td>
<td>278k</td>
<td>605k</td>
<td>1217k</td>
<td>2657k</td>
<td>5663k</td>
</tr>
</tbody>
</table>

Table 2: Size of the phrase translation table with varying maximum phrase length limits
Phrases do help

- But they don’t need to be long
- Why should this be?
Sources of Alignments

<table>
<thead>
<tr>
<th>Method</th>
<th>Training corpus size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10k</td>
</tr>
<tr>
<td>AP</td>
<td>84k</td>
</tr>
<tr>
<td>Joint</td>
<td>125k</td>
</tr>
<tr>
<td>Syn</td>
<td>19k</td>
</tr>
</tbody>
</table>
Alignment Heuristics
Lexical Weighting

$$
\phi(\bar{f}_i|\bar{e}_i) = \frac{\text{count}(\bar{f}_i, \bar{e}_i)}{\text{count}(\bar{e}_i)} p_w(\bar{f}_i|\bar{e}_i)
$$

<table>
<thead>
<tr>
<th>f1</th>
<th>f2</th>
<th>f3</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>e1</td>
<td>##</td>
<td>--</td>
</tr>
<tr>
<td>e2</td>
<td>--</td>
<td>##</td>
</tr>
<tr>
<td>e3</td>
<td>--</td>
<td>##</td>
</tr>
</tbody>
</table>

$$
p_w(\bar{f}|\bar{e}, a) = p_w(f_1 f_2 f_3|e_1 e_2 e_3, a) = w(f_1|e_1) \times \frac{1}{2}(w(f_2|e_2) + w(f_2|e_3)) \times w(f_3|\text{NULL})
$$

Graph shows BLEU scores for different values of k.
Summary: Phrase-Based Systems

Morgen ||| fliege ||| ich ||| nach Kanada ||| zur Konferenz

Tomorrow ||| I will fly ||| to the conference ||| in Canada

Sentence-aligned corpus

Word alignments

Phrase table (translation model)

cat ||| chat ||| 0.9
the cat ||| le chat ||| 0.8
dog ||| chien ||| 0.8
house ||| maison ||| 0.6
my house ||| ma maison ||| 0.9
language ||| langue ||| 0.9
...

Phrase table (translation model)
More generally, define feature functions $h$: 

$$p(e|f) \propto \exp \sum_{k=1}^{m} \lambda_k \cdot h_k(e, f)$$

- $P(e)$ and $P(f|e)$ are just two of many possible feature functions:
  - $P(e|f)$, word counts, phrase counts, etc.
- Need to learn how to set weights $\lambda$
  - What function to maximize?
  - How to maximize? Is it differentiable?
Minimum Error Rate Training [Och ’03]

- Non-convex, non-differentiable objective:
  - Generate n-best list
  - Line search 1 dir. at a time
  - Use random restarts

- Each hypothesis in n-best list contributes a line:
Machine Translation Preordering

- English has Subject-Verb-Object word order, while Japanese has Subject-Object-Verb order.
- Use hand-written or automatically learned rules to change word order prior to translation [Collins et al. ’05]
- Dependency-based reordering for English-Japanese [Xu et al. ’09]
Machine Translation Reordering

- Source-side reordering for machine translation
  Use source-side syntax to guide reordering decisions
- Dependency-based reordering for English-Japanese

[Collins et al. ’05]
[Xu et al. ’09]
Classifier Reordering Results

[Reference: Lerner & Petrov ‘13]

<table>
<thead>
<tr>
<th>Language</th>
<th>BLEU Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>0.9</td>
</tr>
<tr>
<td>Hebrew</td>
<td>1.4</td>
</tr>
<tr>
<td>Indonesian</td>
<td>3.0</td>
</tr>
<tr>
<td>Irish</td>
<td>1.5</td>
</tr>
<tr>
<td>Malay</td>
<td>2.1</td>
</tr>
<tr>
<td>Welsh</td>
<td>2.6</td>
</tr>
<tr>
<td>Japanese</td>
<td>6.0</td>
</tr>
<tr>
<td>Japanese*</td>
<td>3.1</td>
</tr>
<tr>
<td>Korean</td>
<td>7.77</td>
</tr>
</tbody>
</table>

**Notes:**
- BLEU Improvement chart shows the improvement in BLEU scores for different languages.
- The languages listed are Arabic, Hebrew, Indonesian, Irish, Malay, and Welsh.
- The bars represent different reordering methods: rule-based, 1-step classifier, and 2-step classifier.
Dutch Parser in Machine Translation

- No preordering (Lerner and Petrov, 2013)
- Arc-eager preordering (L&P, 2013)
- Arc-eager preordering (updated baseline)
- Two-Registers preordering (updated baseline)

Dutch-English BLEU:
- No preordering: 34.9
- Arc-eager preordering: 35.1
- Arc-eager preordering (updated baseline): 35.4
- Two-Registers preordering (updated baseline): 35.9