CS395T: Structured Models for NLP
Lecture 12: Machine Translation

Greg Durrett

Adapted from Dan Klein – UC Berkeley
Administrivia

Project 2 due one week from today!

P1 test set results: top 3

Su Wang: 84.03 F1 (86.10 P / 82.05 R)
Larger window size and Wikipedia gazetteer

Prateek Shrishail Kolhar: 82.32 F1 (82.61 P / 82.07 R)
Conjunctions of words, POS, and shapes in neighborhood
Very fast vectorized implementation (15s per epoch)

Yasumasa Onoe: 78.55 F1 (78.27 P / 78.83 R)
Used transition probabilities from HMM, character
5-grams and other feature tuning
Machine Translation
Atlanta, preso il killer del palazzo di Giustizia

ATLANTA - La grande paura che per 26 ore ha attanagliato Atlanta è finita: Brian Nichols, l'uomo che aveva ucciso tre persone a palazzo di Giustizia e che ha poi ucciso un agente di dogana, s'è consegnato alla polizia, dopo avere cercato rifugio nell'alloggio di una donna in un complesso d'appartamenti alla periferia della città. Per tutto il giorno, il centro della città, sede della Coca Cola e dei Giochi 1996, cuore di una popolosa area metropolitana, era rimasto paralizzato.

Atlanta, taken the killer of the palace of Justice

ATLANTA - The great fear that for 26 hours has gripped Atlanta is ended: Brian Nichols, the man who had killed three persons to palace of Justice and that a customs agent has then killed, s' is delivered to the police, after to have tried shelter in the lodging of one woman in a complex of apartments to the periphery of the city. For all the day, the center of the city, center of the Coke Strains and of Giochi 1996, heart of one popolosa metropolitan area, was remained paralyzed.
Levels of Transfer

- **Interlingua**
- **Semantics**
- **Syntax**
- **Phrases**
- **Words**

**Source** to **Target** with language translation examples:

**English (E)**

<table>
<thead>
<tr>
<th>English (E)</th>
<th>( P(E \mid \text{lo haré} ) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>will do it</td>
<td>0.8</td>
</tr>
<tr>
<td>will do so</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**Translated Examples**

- \( \text{Yo lo haré mañana} \) → \( \text{I will do it tomorrow} \)
- \( \text{Yo lo haré mañana} \) → \( \text{I will do it tomorrow} \)

**Table Example**

<table>
<thead>
<tr>
<th>English (E)</th>
<th>( P(E \mid \text{mañana} ) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>tomorrow</td>
<td>0.7</td>
</tr>
<tr>
<td>morning</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Word-Level MT: Examples

*la politique de la haine.*
politics of hate.
the policy of the hatred.

*nous avons signé le protocole.*
we did sign the memorandum of agreement.
we have signed the protocol.

*où était le plan solide ?*
buts where was the solid plan?
where was the economic base?
Phrasal MT: Examples

Le président américain Barack Obama doit annoncer lundi de nouvelles mesures en faveur des constructeurs automobile. General motors et Chrysler avaient déjà bénéficié fin 2008 d'un prêt d'urgence cumulé de 17,4 milliards de dollars, et ont soumis en février au Trésor un plan de restructuration basé sur un total de 22 milliards de dollars d'aides publiques supplémentaires.

U.S. President Barack Obama to announce Monday new measures to help automakers. General Motors and Chrysler had already received late in 2008 a cumulative emergency loan of 17.4 billion dollars, and submitted to the Treasury in February in a restructuring plan based on a total of 22 billion dollars in additional aid.
Metrics
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)
  - Lots more: TER, HTER, METEOR, ...

- BLEU:
  - $P_1 = \text{unigram precision}$
  - $P_2, P_3, P_4 = \text{bi-, tri-, 4-gram precision}$
  - Weighted geometric mean of $P_1-4$
  - Brevity penalty (why?)
  - Somewhat hard to game...
  - Magnitude only meaningful on same language, corpus, number of references, probably only within system types...

---

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
received 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.
Automatic Metrics Work (?)

slide from G. Doddington (NIST)
Systems Overview
Corpus-Based MT

Modeling correspondences between languages

Sentence-aligned parallel corpus:

<table>
<thead>
<tr>
<th>Spanish</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yo lo haré mañana</td>
<td>I will do it tomorrow</td>
</tr>
<tr>
<td>Hasta pronto</td>
<td>See you soon</td>
</tr>
<tr>
<td>Hasta pronto</td>
<td>See you around</td>
</tr>
</tbody>
</table>

Machine translation system:

Novel Sentence

<table>
<thead>
<tr>
<th>Spanish</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yo lo haré pronto</td>
<td>I will do it soon</td>
</tr>
<tr>
<td></td>
<td>I will do it around</td>
</tr>
<tr>
<td></td>
<td>See you tomorrow</td>
</tr>
</tbody>
</table>
Phrase-Based System Overview

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)

Many slides and examples from Philipp Koehn or John DeNero
Phrase-Based System Overview

Phrase table $P(f|e)$

Unlabeled English data

Language model $P(e)$

Noisy channel model: combine scores from translation model + language model to translate foreign to English

$P(e|f) \propto P(f|e)P(e)$

"Translate faithfully but make fluent English"

Many slides and examples from Philipp Koehn or John DeNero
Word Alignment
Word Alignment

1. **Align words with a probabilistic model**

2. **Infer presence of larger structures from this alignment**

3. **Translate with the larger structures**
What is the anticipated cost of collecting fees under the new proposal?

En vertu des nouvelles propositions, quel est le coût prévu de perception des droits?
Unsupervised Word Alignment

- Input: a *bitext*: pairs of translated sentences
  - nous acceptons votre opinion .
  - we accept your view .

- Output: *alignments*: pairs of translated words
  - Not always one-to-one!
1-to-Many Alignments

And₁ the₂ program₃ has₄ been₅ implemented₆

Le₁ programme₂ a₃ été₄ mis₅ en₆ application₇
Evaluating Models

How do we measure quality of a word-to-word model?

- **Method 1**: use in an end-to-end translation system
  - Slow development cycle
  - Misleading if your MT system was “tuned” for certain aspects of bad alignments

- **Method 2**: measure quality of the alignments produced
  - Easy to measure
  - Hard to know what the gold alignments should be
  - Often does not correlate well with translation quality
Alignment Error Rate

- **Alignment Error Rate**

- **Sure**: □
- **Possible**: ◯
- **Predicted**: ■

\[
AER(A, S, P) = \left(1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}\right)
\]

\[
= \left(1 - \frac{3 + 3}{3 + 4}\right) = \frac{1}{7}
\]

In 1978, on a enregistré 1,122,000 divorces sur le continent.

In 1978, Americans divorced 1,122,000 times.
IBM Model 1
IBM Model 1 (Brown 93)

- Alignments: a hidden vector called an alignment specifies which English source (or a special null token) is responsible for each French target word.

\[
a = a_1 \ldots a_J
\]

\[
\begin{align*}
\text{And}_1 & \quad \text{the}_2 & \quad \text{program}_3 & \quad \text{has}_4 & \quad \text{been}_5 & \quad \text{implemented}_6 \\
| & \quad | & \quad | & \quad | & \quad | & \quad |
\text{Le}_1 & \quad \text{programme}_2 & \quad \text{a}_3 & \quad \text{été}_4 & \quad \text{mis}_5 & \quad \text{en}_6 & \quad \text{application}_7
\end{align*}
\]

\[
P(f, a|e) = \prod_j P(a_j = i) P(f_j|e_i)
\]

\[
= \prod_j \frac{1}{I+1} P(f_j|e_i)
\]

\[
P(f|e) = \sum_a P(f, a|e)
\]
Thank you, I shall do so gladly.

Gracias, lo haré de muy buen grado.

**Model Parameters**

\[
P(A_1 = 1) = \frac{1}{10}, \text{ nothing to learn}
\]

\[
P(F_1 = \text{Gracias} | A_1 = 1) = P(\text{Gracias} | \text{Thank}) \leftarrow \text{learn these translation probs}
\]
EM for Model 1

- Model 1 Parameters: Translation probabilities \( P(f_j|e_i) \)
- Start with \( P(f_j|e_i) \) uniform, including \( P(f_j|\text{null}) \)
- For each sentence, for each foreign position \( j \):
  - Calculate posterior over English positions

\[
P(a_j = i|f, e) = \frac{P(f_j|e_i)}{\sum_{i'} P(f_j|e_{i'})}
\]

- Increment count of word \( f_j \) with word \( e_i \) by these amounts
- Do for whole corpus, re-estimate \( P(f|e) \) with M-step
There’s a reason they designed models 2-5!

Problems: alignments jump around, align everything to rare words

Experimental setup:
- Training data: 1.1M sentences of French-English text, Canadian Hansards
- Evaluation metric: alignment error Rate (AER)
- Evaluation data: 447 hand-aligned sentences
- Post-intersection: standard practice to train models in each direction then intersect their predictions [Och and Ney, 03]

- Second model is basically a filter on the first
  - Precision jumps, recall drops
  - End up not guessing hard alignments

<table>
<thead>
<tr>
<th>Model</th>
<th>P/R</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 E→F</td>
<td>82/58</td>
<td>30.6</td>
</tr>
<tr>
<td>Model 1 F→E</td>
<td>85/58</td>
<td>28.7</td>
</tr>
<tr>
<td>Model 1 AND</td>
<td>96/46</td>
<td>34.8</td>
</tr>
</tbody>
</table>
HMM Model: Local Monotonicity
Monotonic Translation

Japan shaken by two new quakes

Le Japon secoué par deux nouveaux séismes
Local Order Change

Japan is at the junction of four tectonic plates

Le Japon est au confluent de quatre plaques tectoniques
The HMM Model

- Want local monotonicity: most jumps are small
- HMM model (Vogel 96)

\[ P(f, a | e) = \prod_{j} P(a_j | a_{j-1}) P(f_j | e_i) \]

- Re-estimate using the forward-backward algorithm

| \( f \)     | \( t(f | e) \) |
|------------|--------------|
| nationale  | 0.469        |
| national   | 0.418        |
| nationaux  | 0.054        |
| nationales | 0.029        |
HMM Examples

- nous
- ne
- avons
- pas
- cru
- bon
- de
- assister
- la
- reunion
- et
- en
- avons
- informe
- le
- cojo
- en
- consquence
- we
- deemed
- it
- inadvisable
- to
- attend
- the
- meeting
- and
- so
- informed
- cojo
- nous
- ne
- avons
- pas
- cru
- bon
- de
- assister
- la
- reunion
- et
- en
- avons
- informe
- le
- cojo
- en
- consquence
### AER for HMMs

<table>
<thead>
<tr>
<th>Model</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 INT</td>
<td>19.5</td>
</tr>
<tr>
<td>HMM E→F</td>
<td>11.4</td>
</tr>
<tr>
<td>HMM F→E</td>
<td>10.8</td>
</tr>
<tr>
<td>HMM AND</td>
<td>7.1</td>
</tr>
<tr>
<td>HMM INT</td>
<td>4.7</td>
</tr>
<tr>
<td>GIZA M4 AND</td>
<td>6.9</td>
</tr>
</tbody>
</table>
Language Modeling
Phrase-Based System Overview

Phrase table $P(f|e)$

Unlabeled English data

Language model $P(e)$

Noisy channel model: combine scores from translation model + language model to translate foreign to English

"Translate faithfully but make fluent English"

Many slides and examples from Philipp Koehn or John DeNero
N-gram Language Modeling

- Could give several lectures on this!

- Estimate \( P(\mathbf{w}_n | \mathbf{w}_{n-k}, \mathbf{w}_{n-k+1}, \ldots, \mathbf{w}_{n-1}) \)

- Generative model: read off counts and normalize
  - \( P(\text{fox} \mid \text{the quick brown}) = 0.9 \), etc.

- Very complex distributions, need to smooth
  - Interpolate with lower-order models
  - Lots of complex techniques
Phrase-Based MT
Phrase-Based System Overview

- We have a phrase table now (ran aligner, extracted phrases and counted them to get scores) – phrase extraction and counting are tricky, but we’ll ignore this...
Phrase-Based System Overview

Unlabeled English data

Phrase table $P(f|e)$

Language model $P(e)$

$P(e|f) \propto P(f|e)P(e)$

Noisy channel model: combine scores from translation model + language model to translate foreign to English

“Translate faithfully but make fluent English”

Many slides and examples from Philipp Koehn or John DeNero
Phrase-Based Translation Overview

**Input:** lo haré rápidamente.

**Translations:** I’ll do it quickly.

quickly I’ll do it.

**Objective:**

\[
\arg \max_e [P(f|e) \cdot P(e)]
\]

\[
\arg \max_e \left[ \prod_{\langle e, f \rangle} P(f|e) \cdot \prod_{i=1}^{\mid e \mid} P(e_i|e_{i-1}, e_{i-2}) \right]
\]
Decoder design is important: [Koehn et al. 03]
### Phrase-Based Decoding

<table>
<thead>
<tr>
<th>Maria</th>
<th>no</th>
<th>dio</th>
<th>una</th>
<th>bofetada</th>
<th>a</th>
<th>la</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>not</td>
<td>give</td>
<td>a</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td>witch</td>
<td>green</td>
</tr>
<tr>
<td></td>
<td>did not</td>
<td>a slap</td>
<td>by</td>
<td>green witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>slap</td>
<td>to the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>did not give</td>
<td>to</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>slap</td>
<td>the witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Cost is \( \text{LM} \times \text{TM} \)

It’s an HMM?
- \( \Pr(e|e_1, e_2) \)
- \( \Pr(f|e) \)

State includes
- Exposed English
- Position in foreign

Dynamic program loop?

\[
\begin{align*}
\text{for } (\text{fPosition in } 1 \ldots [f]) \\
&\text{for } (\text{eContext in allEContexts}) \\
&\quad \text{for } (\text{eOption in translations[fPosition]}) \\
&\quad \quad \text{score} = \text{scores[fPosition-1][eContext]} \times \text{LM(eContext+eOption)} \times \text{TM(eOption, fWord[fPosition])} \\
&\quad \quad \text{scores[fPosition][eContext[2]+eOption]} = \max \text{ score}
\end{align*}
\]
Beam Decoding

- For real MT models, this kind of dynamic program is a disaster (why?)
- Standard solution is beam search: for each position, keep track of only the best k hypotheses

```plaintext
for (fPosition in 1…|f|) 
  for (eContext in bestEContexts[fPosition]) 
    for (eOption in translations[fPosition]) 
      score = scores[fPosition-1][eContext] * LM(eContext+eOption) * TM(eOption, fWord[fPosition])
      bestEContexts.maybeAdd(eContext[2]+eOption, score)
```

Example from David Chiang
### Phrase Translation

<table>
<thead>
<tr>
<th>Maria</th>
<th>no</th>
<th>dio</th>
<th>una</th>
<th>bofetada</th>
<th>a</th>
<th>la</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
</table>

- Mary did not give a slap to the witch green
did not a slap by green witch
no slap to the

did not give to

- slab the witch

- If monotonic, almost an HMM; technically a semi-HMM

  for (fPosition in 1…|f|)
  for (lastPosition < fPosition)
  for (eContext in eContexts)
  for (eOption in translations[fPosition])
    ... combine hypothesis for (lastPosition ending in eContext) with eOption

- If distortion... now what?
Non-Monotonic Phrasal MT

e: Mary
f: *-------
p: .534

e: Mary did not
f: **--------
p: .122

e: Mary slap
f: *-----***
p: .043

e: witch
f: ---------*
p: .182

e: 
f: 
p: 1
Problem: easy partial analyses are cheaper
- Solution 1: use beams per foreign subset
- Solution 2: estimate forward costs (A*-like)
The Pharaoh Decoder

Maria | no | dio | una | bofetada | a | la | bruja | verde

Mary did not give a slap to the witch green

Maria did not give a slap by the green witch

Mary did not give a slap to the witch

Maria | no | dio una bofetada | a la | bruja | verde

Mary | did not | slap | the | green | witch
Hypothesis Lattices

<table>
<thead>
<tr>
<th>Maria</th>
<th>no</th>
<th>dio</th>
<th>una</th>
<th>bofetada</th>
<th>a</th>
<th>la</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>not give</td>
<td>a</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td>witch</td>
<td>green</td>
<td></td>
</tr>
<tr>
<td>did not</td>
<td>a slap</td>
<td>by</td>
<td>green witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>slap</td>
<td>to the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>did not give</td>
<td></td>
<td>to</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>slap</td>
<td>the</td>
<td>witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Syntactic Models
<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>lo haré de muy buen grado .</td>
<td></td>
</tr>
</tbody>
</table>

Grammar
## Translating with Tree Transducers

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>lo haré de muy buen grado .</td>
<td></td>
</tr>
</tbody>
</table>

**Grammar**

\[
ADV \rightarrow (\text{de muy buen grado} ; \text{gladly})
\]
Translating with Tree Transducers

Grammar

\[
\begin{align*}
\text{ADV} & \rightarrow \langle \text{de muy buen grado} ; \text{gladly} \rangle \\
\end{align*}
\]
Translating with Tree Transducers

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>lo haré</td>
<td>I gladly</td>
</tr>
<tr>
<td>de muy buen grado</td>
<td></td>
</tr>
</tbody>
</table>

**Grammar**

\[
S \rightarrow \langle \text{lo haré } \text{ADV . ; I will do it } \text{ADV . } \rangle \\
\text{ADV } \rightarrow \langle \text{de muy buen grado ; gladly } \rangle
\]
Translating with Tree Transducers

Input

S

<table>
<thead>
<tr>
<th>lo haré</th>
<th>ADV</th>
</tr>
</thead>
<tbody>
<tr>
<td>de muy buen grado</td>
<td>.</td>
</tr>
</tbody>
</table>

Output

S

I will do it gladly .

Grammar

S → ⟨ lo haré ADV . ; I will do it ADV . ⟩

ADV → ⟨ de muy buen grado ; gladly ⟩
Translating with Tree Transducers

**Input**

<table>
<thead>
<tr>
<th>S</th>
<th>ADV</th>
</tr>
</thead>
<tbody>
<tr>
<td>lo haré</td>
<td>de muy buen grado</td>
</tr>
</tbody>
</table>

**Output**

I will do it gladly

**Grammar**

\[
S \rightarrow \langle \text{lo haré } \text{ADV} . ; \text{I will do it } \text{ADV} . \rangle \\
\text{ADV} \rightarrow \langle \text{de muy buen grado} ; \text{gladly} \rangle
Translating with Tree Transducers

Input

\[ \text{lo haré} \quad \text{de muy buen grado} \quad . \]

Output

\[ \text{ADV} \quad \text{I} \quad \text{gladly} \]

Grammar

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

\[
\text{ADV} \rightarrow \langle \text{de muy buen grado} \quad ; \quad \text{gladly} \quad \rangle
\]
# Translating with Tree Transducers

**Input**

| lo haré     | de muy buen grado | ... |

**Output**

<table>
<thead>
<tr>
<th>ADV</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>gladly</td>
</tr>
</tbody>
</table>

## Grammar

<table>
<thead>
<tr>
<th>Rule</th>
<th>Nonterminals</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP</td>
<td>lo haré ADV ; will do it ADV</td>
</tr>
<tr>
<td>S</td>
<td>lo haré ADV . ; I will do it ADV .</td>
</tr>
<tr>
<td>ADV</td>
<td>de muy buen grado ; gladly</td>
</tr>
</tbody>
</table>
Translating with Tree Transducers

**Input**

<table>
<thead>
<tr>
<th>VP</th>
<th>ADV</th>
</tr>
</thead>
<tbody>
<tr>
<td>lo haré</td>
<td>de muy buen grado .</td>
</tr>
</tbody>
</table>

**Output**

VP

ADV

will do it gladly

**Grammar**

\[
\begin{align*}
VP & \rightarrow \langle \lo haré \ ADV ; \ will \ do \ it \ ADV \rangle \\
S & \rightarrow \langle \lo haré \ ADV . ; \ I \ will \ do \ it \ ADV . \rangle \\
ADV & \rightarrow \langle \ de \ muy \ buen \ grado ; \ gladly \rangle
\end{align*}
\]
Translating with Tree Transducers

Input

<table>
<thead>
<tr>
<th>lo haré</th>
<th>de muy buen grado</th>
</tr>
</thead>
</table>

Output

will do it gladly

Grammar

\[
S \rightarrow \langle \ \mathrm{VP} \ . \ ; \ I \ \mathrm{VP} \ . \ \rangle \\
\mathrm{VP} \rightarrow \langle \ \mathrm{lo} \ \mathrm{har}é \ \mathrm{ADV} \ ; \ \mathrm{will} \ \mathrm{do} \ \mathrm{it} \ \mathrm{ADV} \ \rangle \\
\mathrm{S} \rightarrow \langle \ \mathrm{lo} \ \mathrm{har}é \ \mathrm{ADV} \ . \ ; \ I \ \mathrm{will} \ \mathrm{do} \ \mathrm{it} \ \mathrm{ADV} \ . \ \rangle \\
\mathrm{ADV} \rightarrow \langle \ \mathrm{de} \ \mathrm{muy} \ \mathrm{buen} \ \mathrm{grado} \ ; \ \mathrm{gladly} \ \rangle 
\]
Translating with Tree Transducers

Input

```
S
  VP
    ADV
    lo haré de muy buen grado .
```

Output

```
S
  VP
    ADV
    I will do it gladly .
```

Grammar

```
S → ⟨ VP ; I VP . ⟩

VP → ⟨ lo haré ADV ; will do it ADV ⟩

S → ⟨ lo haré ADV . ; I will do it ADV . ⟩

ADV → ⟨ de muy buen grado ; gladly ⟩
```
Translating with Tree Transducers

**Input**

```
S
  VP
  lo haré de muy buen grado .
```

**Output**

```
S
  VP
  ADV
  I will do it gladly .
```

**Grammar**

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

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

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

\[
\text{ADV} \rightarrow \langle \text{de muy buen grado ; gladly } \rangle
\]
Syntactic Translation

- Lots of complexity: large phrase tables, errors introduced by parsers, parses don’t agree, inference is harder, ...

- Good for some languages (Japanese->English), but generally more trouble than it’s worth

- Easier method: syntactic “pre-reordering”
MT: Takeaways

- Word alignments: unsupervised process for finding word-level correspondences. Turn these into phrase level correspondences -> phrase table

- Language model: estimate n-gram model on a very large corpus

- Translation process: use beam search to find the best translation $\arg\max_e P(f|e)P(e)$