Machine Translation Contd

Prof. Sameer Singh

CS 295: STATISTICAL NLP
WINTER 2017

March 7, 2017

Based on slides from Richard Socher, Chris Manning, Philipp Koehn, and everyone else they copied from.
Upcoming...

- **Project**
  - Status report due tonight: **March 7, 2017**
  - Almost final report, only **5 pages**

- **Summaries**
  - Paper summaries: **March 14**
  - Summary 2 graded

- **Homework**
  - Homework 4 is due on **March 13**
  - Write-up, code, and data released.
  - **Lowest grade of the homeworks will be dropped**
Outline

- Decoding Algorithms
- Syntax-Based MT
- Neural MT Models
Outline

Decoding Algorithms

Syntax-Based MT

Neural MT Models
Phrase Decoding: Stacks

**function** STACK DECODING(source sentence)

initialize stack with a null hypothesis

**loop do**

pop best hypothesis $h$ off of stack

if $h$ is a complete sentence, **return** $h$

**for each** possible expansion $h'$ of $h$

assign a score to $h'$

push $h'$ onto stack

maximum size, $B$
Monotonic Decoding

\[
\text{function} \quad \text{STACK DECODING}(\text{source sentence}) \\
\text{initialize stack with a null hypothesis} \\
\text{loop do} \\
\quad \text{pop best hypothesis } h \text{ off of stack} \\
\quad \text{if } h \text{ is a complete sentence, return } h \\
\quad \text{for each possible expansion } h' \text{ of } h \\
\quad \quad \text{assign a score to } h' \\
\quad \quad \text{push } h' \text{ onto stack}
\]
Monotonic Phrase 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>Mary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<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>to</td>
<td>the</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>slap</td>
<td>to</td>
<td>the the witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>did not give</td>
<td>slap</td>
<td>to</td>
<td>the witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Monotonic Phrase Decoding

(Mary)

"not"

"did not"

"no"

"did not give"
Monotonic Phrase Decoding

(Mary) (did not)
Monotonic Phrase Decoding

(Mary) (did not) (slap)
Monotonic Phrase Decoding

(Mary) (did not) (slap) (the)
Monotonic Phrase Decoding

(Mary) (did not) (slap) (the) (green witch)
Non-Monotonic Decoding

\[ h \in P_1 \cup P_2 \]

\[ F = \text{ any P that doesn't overlap with } h. \]

**function** STACK DECODING(source sentence)

initialize stack with a null hypothesis

**loop do**

pop best hypothesis \( h \) off of stack

**if** \( h \) is a complete sentence, **return** \( h \)

for each possible expansion \( h' \) of \( h \)

assign a score to \( h' \)

push \( h' \) onto stack
Non-Monotonic 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>did not</td>
<td></td>
<td>a slap</td>
<td>by</td>
<td>green witch</td>
<td>to</td>
<td>the</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td></td>
<td>slap</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>did not give</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>slap</td>
<td></td>
<td>the witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

e: 
f: ---------
p: 1
Non-Monotonic Decoding

Maria
no
dio
una
bofetada
a
la
bruja
verde

Mary
not
give
a
slap
to
the
witch
green
did not
a slap
by
to the

Mary
did not give

slap
the witch

e: Mary
f: *--------
p: .534
Non-Monotonic Decoding

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

Mary did not give a slap to the green witch.

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

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

e: witch
f: *-----------*

e: witch slap
f: _ _ _ _ _ _
Non-Monotonic Decoding
Non-Monotonic Decoding

Maria

no
dio una bofetada
ala
bruja verde

Mary

not
give
a slap
to
the
witch
green

did not

a slap
by
to the

do not
give

slap
to the witch

---

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

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

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

e: did not
f: *-------*
p: .154

e: slap
f: ***-----*
p: .015

e: the
f: *-----***
p: .004283

e: green witch
f: *****
p: .000271
Non-Monotonic Decoding

Maria
no
dio
una
bofetada
a
la
bruja
verde

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

did not give

slap the witch

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

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

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

e: did not
f: *-------
p: .154

e: the
f: *-------
p: .015

e: green witch
f: *-------
p: .000271
Comparing Partial Translations

$h_1 \quad E: \text{“Mary”}$

$F: \star \ldots$

$c: \quad \quad$

$h_2 \quad E: \text{“Mary did not slap the”}$

$F: \star \star \star \star \star \star \ldots$

$c: \quad \quad$
Hypothesis Recombination

\[ p=1 \quad \text{Mary} \quad p=0.534 \quad \text{did not give} \quad p=0.092 \]
\[ p=0.164 \quad \text{give} \quad p=0.044 \]

\[ p=1 \quad \text{Mary} \quad p=0.534 \quad \text{did not give} \quad p=0.092 \]
\[ p=0.164 \quad \text{give} \]
Hypothesis Recombination
Multi-Stack Decoding

function STACK DECODING(source sentence)

initialize stack with a null hypothesis
loop do
    pop best hypothesis $h$ off of stack
    if $h$ is a complete sentence, return $h$
    for each possible expansion $h'$ of $h$
        assign a score to $h'$
        push $h'$ onto stack

return max(stack[n])
Outline

- Decoding Algorithms
- Syntax-Based MT
- Neural MT Models
Limits of the Phrase Model

Non-Contiguous Phrases

Ich habe das Auto gekauft
I bought the car

Syntactic Transformations

Den Antrag verabschiedet das Parlament
The draft approves the Parliament
Syntax leads to Good English
The Vauquios Triangle

"Vauquois Triangle"

- **INTERLINGUA**
- **SEMANTICS**
- **SYNTAX**
- **PHRASES**
- **WORDS**

**SOURCE**  
**TARGET**

```
Yo lo haré mañana  
I will do it tomorrow
```

```
P( lo haré mañana | VP ) = 0.8
```

| English (E) | P( E | lo haré ) |
|-------------|---------------|
| will do it  | 0.8           |
| will do so  | 0.2           |

| English (E) | P( E | mañana ) |
|-------------|------------|
| tomorrow    | 0.7        |
| morning     | 0.3        |
String to Tree Translation

[Yamada and Knight, 2001]
String to Tree Translation

[Diagram of string to tree translation process]

Kare ha ongaku wo kiku no ga daisuki desu  
[from Yamada and Knight, 2001]
Synchronous CFGs

\[ A \rightarrow A_1 A_2 \parallel A_1 A_2 \]
\[ A \rightarrow A_1 A_2 \parallel A_2 A_1 \]
\[ A \rightarrow e \parallel f \]
\[ A \rightarrow e \parallel * \]
\[ A \rightarrow * \parallel f \]

Mary did not slap the green witch

Maria no daba una bofetada a la bruja verde

Mary did not slap the green witch

Maria * no daba una bofetada a la bruja verde
Outline

Decoding Algorithms

Syntax-Based MT

Neural MT Models
Neural MT Models

Input text → Encoder → Decoder → Translated text
Recurrent Neural Networks

\[ y_t = \text{softmax} \left( W_0 h_t \right) \]
\[ h_t = (W_1 x_t + W'_1 h_{t-1}) \]
\[ \tanh h \]
Recurrent Neural Networks

Encoder

rodillas  de  abejas

Bee’s  knees

\[ h_t = \tanh(W_i x_t + W_h h_{t-1}) \]
\[ y_t = \sigma(W_y h_t) \]
Different Weights

\[ h_t = \tanh \left( W_1 x_t + W_1 h_{t-1} \right) \]

rodillas \rightarrow de \rightarrow abejas

Bee’s \rightarrow knees

\[ h_t = \tanh \left( W_2 h_{t-1} \right) \]
More Connections

rodillas → de → abejas

Bee’s

knees

<->

Bee’s
\[ f = (\text{La, croissance, économique, s'est, ralentie, ces, dernières, années, .}) \]

\[ e = (\text{Economic, growth, has, slowed, down, in, recent, years, .}) \]
Other Extensions

Reverse

rodillas de abejas
de de abejas

Bee’s knees

CS 295: STATISTICAL NLP (WINTER 2017)
Stacking

rodillas → de → abejas → Bee’s → knees

Other Extensions
Regular Recurrent Units

\[ h_t = \tanh \left( W_1 x_t + W'_1 h_{t-1} \right) \]

\[ h_t = \tanh \left( W_1 x_t + r_t \odot h_{t-1} \right) \]

\[ r_t = \sigma \left( W_2 x_t + W'_2 h_{t-1} \right) \]
Gated Recurrent Units

\[ h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \tilde{h}_t \]

\[ \tilde{h}_t = \tanh (W x_t + r_t \circ U h_{t-1}) \]

\[ r_t = \sigma \left( W^{(r)} x_t + U^{(r)} h_{t-1} \right) \]

\[ z_t = \sigma \left( W^{(z)} x_t + U^{(z)} h_{t-1} \right) \]
Long Short-Term Memory

\[ i_t = \sigma \left( W^{(i)} x_t + U^{(i)} h_{t-1} \right) \]
\[ f_t = \sigma \left( W^{(f)} x_t + U^{(f)} h_{t-1} \right) \]
\[ o_t = \sigma \left( W^{(o)} x_t + U^{(o)} h_{t-1} \right) \]
\[ \tilde{c}_t = \tanh \left( W^{(c)} x_t + U^{(c)} h_{t-1} \right) \]
\[ c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \]
\[ h_t = o_t \circ \tanh(c_t) \]
Neural MT Results

<table>
<thead>
<tr>
<th>Method</th>
<th>test BLEU score (ntst14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline System [29]</td>
<td>33.30</td>
</tr>
<tr>
<td>Cho et al. [5]</td>
<td>34.54</td>
</tr>
<tr>
<td>Best WMT’14 result [9]</td>
<td><strong>37.0</strong></td>
</tr>
<tr>
<td>Rescoring the baseline 1000-best with a single forward LSTM</td>
<td>35.61</td>
</tr>
<tr>
<td>Rescoring the baseline 1000-best with a single reversed LSTM</td>
<td>35.85</td>
</tr>
<tr>
<td>Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs</td>
<td><strong>36.5</strong></td>
</tr>
<tr>
<td>Oracle Rescoring of the Baseline 1000-best lists</td>
<td>~45</td>
</tr>
</tbody>
</table>

Sequence to Sequence Learning by Sutskever et al. 2014
Trend in Machine Translation

Multilingual Neural MT: Naïve

- **Spanish**
  - English-Spanish Encoder
  - English-Spanish Decoder

- **Hindi**
  - English-Hindi Encoder
  - English-Hindi Decoder

- **Chinese**
  - English-Chinese Encoder
  - English-Chinese Decoder
Multilingual NMT: Decoder

- Spanish
  - English-Spanish Encoder

- Hindi
  - English-Hindi Encoder

- Chinese
  - English-Chinese Encoder

Shared Decoder

English

CS 295: STATISTICAL NLP (WINTER 2017)
Multilingual NMT: Encoder

- English Encoder
  - English-Spanish Decoder
  - English-Hindi Decoder
  - English-Chinese Decoder
  - Spanish
  - Hindi
  - Chinese
Google Neural MT

- Spanish Encoder
- Spanish Decoder
- Hindi Encoder
- Hindi Decoder
- Chinese Encoder
- Chinese Decoder
- English Encoder
- English Decoder

GNMT
Out of Control!

unwatchable
 CSI Miami

@shrdlu_
Out of Control!

Microsoft Translator

Translate

Russian German English Detect language  

Deutschland

Deutsch, deutsch, deutsch, deutsch, deutsch, deutsch, deutsch

Naturally has a German "Wetten, dass ... ?" invented
Vielen Dank für die schönen Stunden!
Wir sind die freundlichsten Kunden auf dieser Welt
Wir sind bescheiden, wir haben Geld
Die Allerbesten in jedem Sport
Die Steuern hier sind Weltrekord
Bereisen Sie Deutschland und bleiben Sie hier!
Auf diese Art von Besuchern warten wir
Es kann jeder hier wohnen, dem es gefällt
Wir sind das freundlichste Volk auf dieser Welt

Germany

German, English, German, German, German, German, and English

Of course a German has "betting that ... ?" invented
Thanks for the nice hours!
We are the friendliest customers in this world
We are modest, we have money
The very best in any sport
The taxes here are a world record
Travel to Germany and stay here!
We are waiting for this kind of visitors
Anyone who likes it can live here
We are the friendliest people in this world

English, German, German, German, and German
Out of Control!

knife, fork, knife,
(The trailing comma messes this one up.)

Messer, Messer, Messer,

Messer, Gabel, Messer, Messer, Messer, Messer, Messer, Messer, Messer, Messer

Screen monitor styling Projector styling Print styling — back to. 2010-01-20 with adjustable interlinear. Knife, fork; knife, knife, knife, knife;