Advanced Natural Language Processing
Lecture 27
Statistical Machine Translation

Philipp Koehn

23 November 2012
Machine translation

- Task: make sense of foreign text like

- One of the oldest problems in Artificial Intelligence

- AI-hard: reasoning and world knowledge required
The Rosetta stone

- Egyptian language was a mystery for centuries
- 1799 a stone with Egyptian text and its translation into Greek was found
  ⇒ Humans *could learn* how to translated Egyptian
Parallel data

- Lots of translated text available: 100s of million words of translated text for some language pairs
  - a book has a few 100,000s words
  - an educated person may read 10,000 words a day
  → 3.5 million words a year
  → 300 million a lifetime
  → soon computers will be able to see more translated text than humans read in a lifetime

⇒ Machines can learn how to translated foreign languages
Vauquois Triangle

interlingua

foreign semantics

foreign syntax

foreign words

english semantics

english syntax

english words
Euromatrix

• Proceedings of the European Parliament
  – translated into 11 official languages
  – entry of new members in May 2004: more to come...

• Europarl corpus
  – collected 20-30 million words per language
  \[\rightarrow 110 \text{ language pairs}\]

• 110 Translation systems
  – 3 weeks on 16-node cluster computer
  \[\rightarrow 110 \text{ translation systems}\]
Quality of translation systems

- **Scores** for all 110 systems

<table>
<thead>
<tr>
<th></th>
<th>da</th>
<th>de</th>
<th>el</th>
<th>en</th>
<th>es</th>
<th>fr</th>
<th>fi</th>
<th>it</th>
<th>nl</th>
<th>pt</th>
<th>sv</th>
</tr>
</thead>
<tbody>
<tr>
<td>da</td>
<td>-</td>
<td>18.4</td>
<td>21.1</td>
<td>28.5</td>
<td>26.4</td>
<td>28.7</td>
<td>14.2</td>
<td>22.2</td>
<td>21.4</td>
<td>24.3</td>
<td>28.3</td>
</tr>
<tr>
<td>de</td>
<td>22.3</td>
<td>-</td>
<td>20.7</td>
<td>25.3</td>
<td>25.4</td>
<td>27.7</td>
<td>11.8</td>
<td>21.3</td>
<td>23.4</td>
<td>23.2</td>
<td>20.5</td>
</tr>
<tr>
<td>el</td>
<td>22.7</td>
<td>17.4</td>
<td>-</td>
<td>27.2</td>
<td>31.2</td>
<td>32.1</td>
<td>11.4</td>
<td>26.8</td>
<td>20.0</td>
<td>27.6</td>
<td>21.2</td>
</tr>
<tr>
<td>en</td>
<td>25.2</td>
<td>17.6</td>
<td>23.2</td>
<td>-</td>
<td>30.1</td>
<td>31.1</td>
<td>13.0</td>
<td>25.3</td>
<td>21.0</td>
<td>27.1</td>
<td>24.8</td>
</tr>
<tr>
<td>es</td>
<td>24.1</td>
<td>18.2</td>
<td>28.3</td>
<td>30.5</td>
<td>-</td>
<td>40.2</td>
<td>12.5</td>
<td>32.3</td>
<td>21.4</td>
<td>35.9</td>
<td>23.9</td>
</tr>
<tr>
<td>fr</td>
<td>23.7</td>
<td>18.5</td>
<td>26.1</td>
<td>30.0</td>
<td>38.4</td>
<td>-</td>
<td>12.6</td>
<td>32.4</td>
<td>21.1</td>
<td>35.3</td>
<td>22.6</td>
</tr>
<tr>
<td>fi</td>
<td>20.0</td>
<td>14.5</td>
<td>18.2</td>
<td>21.8</td>
<td>21.1</td>
<td>22.4</td>
<td>-</td>
<td>18.3</td>
<td>17.0</td>
<td>19.1</td>
<td>18.8</td>
</tr>
<tr>
<td>it</td>
<td>21.4</td>
<td>16.9</td>
<td>24.8</td>
<td>27.8</td>
<td>34.0</td>
<td>36.0</td>
<td>11.0</td>
<td>-</td>
<td>20.0</td>
<td>31.2</td>
<td>20.2</td>
</tr>
<tr>
<td>nl</td>
<td>20.5</td>
<td>18.3</td>
<td>17.4</td>
<td>23.0</td>
<td>22.9</td>
<td>24.6</td>
<td>10.3</td>
<td>20.0</td>
<td>-</td>
<td>20.7</td>
<td>19.0</td>
</tr>
<tr>
<td>pt</td>
<td>23.2</td>
<td>18.2</td>
<td>26.4</td>
<td>30.1</td>
<td>37.9</td>
<td>39.0</td>
<td>11.9</td>
<td>32.0</td>
<td>20.2</td>
<td>-</td>
<td>21.9</td>
</tr>
<tr>
<td>sv</td>
<td>30.3</td>
<td>18.9</td>
<td>22.8</td>
<td>30.2</td>
<td>28.6</td>
<td>29.7</td>
<td>15.3</td>
<td>23.9</td>
<td>21.9</td>
<td>25.9</td>
<td>-</td>
</tr>
</tbody>
</table>

[from Koehn, 2005: Europarl]
Available data

- Available parallel text
  - Europarl: 50 million words in 11 languages http://www.statmt.org/europarl/
  - Acquis Communitaire: 8-50 million words in 22 EU languages
  - Canadian Hansards: 20 million words from Ulrich Germann, ISI
  - United Nations: 200 million words in five languages
  - $10^9$: 1000 million words in French–English
  - Chinese/Arabic to English: over 100 million words from LDC

- Available monolingual text (for language modeling)
  - 2.8 billion words of English from LDC
  - trillions of words on the web
More data, better translations

- **Log-scale improvements** on BLEU:
  - Doubling the training data gives constant improvement (+1 %BLEU)
- More language model data alone also helps
Lexical translation

• How to translate a word → look up in dictionary

   **Haus** — *house, building, home, household, shell.*

• *Multiple translations*
  
  – some more frequent than others
  – for instance: *house*, and *building* most common
  – special cases: *Haus* of a *snail* is its *shell*

• Note: During all the lectures, we will translate from a foreign language into English
Collect statistics

- Look at a parallel corpus (German text along with English translation)

<table>
<thead>
<tr>
<th>Translation of Haus</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>house</td>
<td>8,000</td>
</tr>
<tr>
<td>building</td>
<td>1,600</td>
</tr>
<tr>
<td>home</td>
<td>200</td>
</tr>
<tr>
<td>household</td>
<td>150</td>
</tr>
<tr>
<td>shell</td>
<td>50</td>
</tr>
</tbody>
</table>
Estimate translation probabilities

- Maximum likelihood estimation

\[
p_f(e) = \begin{cases} 
0.8 & \text{if } e = \text{house}, \\
0.16 & \text{if } e = \text{building}, \\
0.02 & \text{if } e = \text{home}, \\
0.015 & \text{if } e = \text{household}, \\
0.005 & \text{if } e = \text{shell}.
\end{cases}
\]
Alignment

- In a parallel text (or when we translate), we align words in one language with the words in the other

\[
\begin{array}{cccc}
1 & 2 & 3 & 4 \\
\text{das} & \text{Haus} & \text{ist} & \text{klein} \\
\text{the} & \text{house} & \text{is} & \text{small} \\
1 & 2 & 3 & 4 \\
\end{array}
\]

- Word positions are numbered 1–4
Alignment function

- Formalizing alignment with an alignment function

- Mapping an English target word at position $i$ to a German source word at position $j$ with a function $a : i \rightarrow j$

- Example

  $$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4\}$$
Reordering

- Words may be reordered during translation

\[ a : \{1 \rightarrow 3, 2 \rightarrow 4, 3 \rightarrow 2, 4 \rightarrow 1\} \]
One-to-many translation

- A source word may translate into multiple target words

\[
a : \{ 1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4, 5 \rightarrow 4 \}\]
Dropping words

- Words may be **dropped** when translated
  - The German article *das* is dropped

\[
a : \{1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 4\}
\]
Inserting words

- Words may be **added** during translation
  - The English *just* does not have an equivalent in German
  - We still need to map it to something: special **NULL** token

```
NULL  das  Haus  ist  klein
  the  house  is  just  small
```

```
a : {1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 0, 5 \rightarrow 4}
```
IBM Model 1

- **Generative model**: break up translation process into smaller steps
  - **IBM Model 1** only uses *lexical translation*

- Translation probability
  - for a foreign sentence \( f = (f_1, ..., f_{l_f}) \) of length \( l_f \)
  - to an English sentence \( e = (e_1, ..., e_{l_e}) \) of length \( l_e \)
  - with an alignment of each English word \( e_j \) to a foreign word \( f_i \) according to the alignment function \( a : j \rightarrow i \)

\[
p(e, a|f) = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})
\]

- parameter \( \epsilon \) is a *normalization constant*
### Example

<table>
<thead>
<tr>
<th>das</th>
<th>Haus</th>
<th>ist</th>
<th>klein</th>
</tr>
</thead>
<tbody>
<tr>
<td>e</td>
<td>t(e</td>
<td>f)</td>
<td>e</td>
</tr>
<tr>
<td>the</td>
<td>0.7</td>
<td>house</td>
<td>0.8</td>
</tr>
<tr>
<td>that</td>
<td>0.15</td>
<td>building</td>
<td>0.16</td>
</tr>
<tr>
<td>which</td>
<td>0.075</td>
<td>home</td>
<td>0.02</td>
</tr>
<tr>
<td>who</td>
<td>0.05</td>
<td>household</td>
<td>0.015</td>
</tr>
<tr>
<td>this</td>
<td>0.025</td>
<td>shell</td>
<td>0.005</td>
</tr>
</tbody>
</table>

\[
p(e, a|f) = \frac{\epsilon}{4^3} \times t(\text{the}|\text{das}) \times t(\text{house}|\text{Haus}) \times t(\text{is}|\text{ist}) \times t(\text{small}|\text{klein})
\]

\[
= \frac{\epsilon}{4^3} \times 0.7 \times 0.8 \times 0.8 \times 0.4
\]

\[
= 0.0028\epsilon
\]
Learning lexical translation models

• We would like to \textit{estimate} the lexical translation probabilities \( t(e|f) \) from a parallel corpus

• ... but we do not have the alignments

• \textbf{Chicken and egg problem}
  
  – if we had the \textit{alignments},
    \( \rightarrow \) we could estimate the \textit{parameters} of our generative model
  
  – if we had the \textit{parameters},
    \( \rightarrow \) we could estimate the \textit{alignments}
EM algorithm

• **Incomplete data**
  - if we had *complete data*, would could estimate *model*
  - if we had *model*, we could fill in the *gaps in the data*

• **Expectation Maximization (EM)** in a nutshell
  - initialize model parameters (e.g. uniform)
  - assign probabilities to the missing data
  - estimate model parameters from completed data
  - iterate
EM algorithm

... la maison ... la maison blue ... la fleur ...

... the house ... the blue house ... the flower ...

- Initial step: all alignments equally likely

- Model learns that, e.g., *la* is often aligned with *the*
EM algorithm

... la maison ... la maison blue ... la fleur ...

... the house ... the blue house ... the flower ...

- After one iteration

- Alignments, e.g., between la and the are more likely
EM algorithm

... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...

- After another iteration

- It becomes apparent that alignments, e.g., between *fleur* and *flower* are more likely (pigeon hole principle)
EM algorithm

... la maison ... la maison bleu ... la fleur ...

/ / / X / / /

... the house ... the blue house ... the flower ...

- Convergence

- Inherent hidden structure revealed by EM
EM algorithm

... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...

p(la | the) = 0.453
p(le | the) = 0.334
p(maison | house) = 0.876
p(bleu | blue) = 0.563

- Parameter estimation from the aligned corpus
IBM Model 1 and EM

- **Probabilities**
  
  \[ p(\text{the}|\text{la}) = 0.7 \quad p(\text{house}|\text{la}) = 0.05 \]
  
  \[ p(\text{the}|\text{maison}) = 0.1 \quad p(\text{house}|\text{maison}) = 0.8 \]

- **Alignments**
  
  \[
  \begin{align*}
  \text{la} & \quad \Rightarrow \text{the} & \quad \text{maison} & \quad \Rightarrow \text{house} \\
  \text{la} & \quad \Rightarrow \text{the} & \quad \text{maison} & \quad \Rightarrow \text{house} \\
  \text{la} & \quad \Rightarrow \text{the} & \quad \text{maison} & \quad \Rightarrow \text{house} \\
  \text{la} & \quad \Rightarrow \text{the} & \quad \text{maison} & \quad \Rightarrow \text{house}
  \end{align*}
  \]

  \[ p(e, a|f) = 0.56 \quad p(e, a|f) = 0.035 \quad p(e, a|f) = 0.08 \quad p(e, a|f) = 0.005 \]

  \[ p(a|e, f) = 0.824 \quad p(a|e, f) = 0.052 \quad p(a|e, f) = 0.118 \quad p(a|e, f) = 0.007 \]

- **Counts**
  
  \[ c(\text{the}|\text{la}) = 0.824 + 0.052 \quad c(\text{house}|\text{la}) = 0.052 + 0.007 \]

  \[ c(\text{the}|\text{maison}) = 0.118 + 0.007 \quad c(\text{house}|\text{maison}) = 0.824 + 0.118 \]
Higher IBM Models

<table>
<thead>
<tr>
<th>IBM Model 1</th>
<th>lexical translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM Model 2</td>
<td>adds absolute reordering model</td>
</tr>
<tr>
<td>IBM Model 3</td>
<td>adds fertility model</td>
</tr>
<tr>
<td>IBM Model 4</td>
<td>relative reordering model</td>
</tr>
<tr>
<td>IBM Model 5</td>
<td>fixes deficiency</td>
</tr>
</tbody>
</table>

- Only IBM Model 1 has *global maximum*
  - training of a higher IBM model builds on previous model

- Computationally biggest change in Model 3
  - trick to simplify estimation does not work anymore
  → *exhaustive* count collection becomes computationally too expensive
  - *sampling* over high probability alignments is used instead
IBM Model 3

Mary did not slap the green witch

Mary not slap slap slap the green witch

Mary not slap slap slap NULL the green witch

Maria no daba una botefada a la verde bruja

Maria no daba una bofetada a la bruja verde

n(3|slap)
p-null
t(la|the)
d(4|4)
Phrase-based translation

- Foreign input is segmented in phrases
  - any sequence of words, not necessarily linguistically motivated

- Each phrase is translated into English

- Phrases are reordered
Phrase translation table

- Phrase translations for *den Vorschlag*

| English                  | $\phi(e|f)$ | English                  | $\phi(e|f)$ |
|--------------------------|------------|--------------------------|------------|
| 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     | ...                      | ...        |
Translation

- Task: translate this sentence from German into English

er geht ja nicht nach hause
Translation step 1

- Task: translate this sentence from German into English

> er geht ja nicht nach hause

- Pick phrase in input, translate

> he
Translation step 2

• Task: translate this sentence from German into English

er geht ja nicht nach hause

he does not

• Pick phrase in input, translate
  – it is allowed to pick words *out of sequence* *(reordering)*
  – phrases may have multiple words: *many-to-many* translation
Translation step 3

• Task: translate this sentence from German into English

```
her geht ja nicht nach hause
he does not go
```

• Pick phrase in input, translate
Translation step 4

- Task: translate this sentence from German into English

he does not go home

- Pick phrase in input, translate
**Many translation options** to choose from
- in Europarl phrase table: **2727 matching phrase pairs** for this sentence
- by pruning to the top 20 per phrase, **202 translation options** remain
Translation options

<table>
<thead>
<tr>
<th>Er</th>
<th>Geht</th>
<th>Ja</th>
<th>Nicht</th>
<th>Nach</th>
<th>Hause</th>
</tr>
</thead>
<tbody>
<tr>
<td>He</td>
<td>Is</td>
<td>Yes</td>
<td>Not</td>
<td>After</td>
<td>House</td>
</tr>
<tr>
<td>It</td>
<td>Are</td>
<td>Is</td>
<td>Do not</td>
<td>To</td>
<td>Home</td>
</tr>
<tr>
<td>, it</td>
<td>Goes</td>
<td>, of course</td>
<td>Does not</td>
<td>According to</td>
<td>Chamber</td>
</tr>
<tr>
<td>, he</td>
<td>Go</td>
<td>Not</td>
<td>Is not</td>
<td>In</td>
<td>At home</td>
</tr>
</tbody>
</table>

- The machine translation decoder does not know the right answer
  → Search problem solved by heuristic beam search
Decoding process: find best path

er geht ja nicht nach hause
Syntax-Based Models

• Traditional statistical models operate on sequences of words

• Many translation problems can be best explained by pointing to syntax
  – reordering, e.g., verb movement in German–English translation
  – long distance agreement (e.g., subject-verb) in output

⇒ Translation models based on tree representation of language
  – significant ongoing research
  – state-of-the art for some language pairs
Synchronous Phrase Structure Grammar

- English rule

\[ \text{NP} \rightarrow \text{DET} \ JJ \ \text{NN} \]

- French rule

\[ \text{NP} \rightarrow \text{DET} \ \text{NN} \ JJ \]

- Synchronous rule (indices indicate alignment):

\[ \text{NP} \rightarrow \text{DET}_1 \ \text{NN}_2 \ JJ_3 \mid \text{DET}_1 \ JJ_3 \ \text{NN}_2 \]
Synchronous Grammar Rules

- Nonterminal rules

\[ NP \rightarrow \text{DET}_1 \text{NN}_2 \text{JJ}_3 \mid \text{DET}_1 \text{JJ}_3 \text{NN}_2 \]

- Terminal rules

\[ N \rightarrow \text{maison} \mid \text{house} \]
\[ NP \rightarrow \text{la maison bleue} \mid \text{the blue house} \]

- Mixed rules

\[ NP \rightarrow \text{la maison JJ}_1 \mid \text{the JJ}_1 \text{house} \]
Aligned Tree Pair

Phrase structure grammar trees with word alignment
(German–English sentence pair.)
Types of Syntax Models

String to String

John misses Mary
⇒ Marie manque à Jean

String to Tree

John misses Mary
⇒ S
  NP
  Marie
  VP
  manque
  PP
  à
  NP
  Jean

Tree to String

⇒ Marie manque à Jean

Tree to Tree

⇒ Marie manque à Jean
Syntax Decoding
Syntax Decoding
Syntax Decoding
Syntax Decoding

1. PRO she
2. will VAFIN
3. eine ART
4. Tasse NN
5. Kaffee NN
6. trinken VVINF
Syntax Decoding

Sie will eine Tasse Kaffee trinken.

PRO she will a cup of coffee drink

NP NP PP
DET NN IN NN
a cup of

NP
NN
Kaffee

NP
NN
coffee

VB
drink

S
VP

PRO she

will
eine

Tasse

Kaffee

trinken

NP

VP

Philipp Koehn
ANLP Lecture 27
23 November 2012
Syntax Decoding

Sie
PPER
will
VAFIN
eine
ART
Tasse
NN
Kaffee
NN
trinken
VVINF
NP
VP
S
PRO
she
VB
drink
NN
|
cup
IN
|
of
NP
PP
NP
DET
|
a
VBZ
|
wants
VB
VP
VP
TO
|
to
VP
NP
NP
DET
|
a
NP
|
cup
NN
|
of
IN

1. PRO
   she
2. Sie
   will
   VAFIN
eine
   ART
   Tasse
   NN
3. Kaffee
   NN
   trinken
   VVINF
4. NP
   NP
   DET
   a
   cup
   IN
   of
   NP
   PP
   NP
   NN
   coffee
   VB
   drinks
Sie will eine Tasse Kaffee trinken.

PRO she

S

VP

VP

wants

TO
to

NP

a cup

NP

of

NP

NN coffee

NP

VB drink

Philipp Koehn ANLP Lecture 27 23 November 2012
Computer Aided Translation

If you can’t beat them, join them..

- How can machine translation help human translators?.
Unassisted Novice Translators

L1a, L1b, L1c, L1d, L1e, L2a, L2b, L2c, L2d, L2e

L1 = native French, L2 = native English, average time per input word only typing
Unassisted Novice Translators

L1a  
L1b  
L1c  
L1d  
L1e  
L2a  
L2b  
L2c  
L2d  
L2e

L1 = native French, L2 = native English, average time per input word typing, initial and final pauses
Unassisted Novice Translators

L1a
L1b
L1c
L1d
L1e
L2a
L2b
L2c
L2d
L2e

L1 = native French, L2 = native English, average time per input word typing, initial and final pauses, short, medium, and long pauses
most time difference on intermediate pauses
Predicting Sentence Completion

- Tool makes a suggestion how to continue (in red).
- User can accept it (by pressing \textsc{tab}), or type in her own translation.
- Same idea as TransType, with minor modifications
  - show only short text chunks, not full sentence completion
  - show only one suggestion, not alternatives
Translation Options

- For each word and phrases: suggested translations
- Ranked (and color-highlighted) by probability
- User may click on suggestion → appended to text box
Das erste schwarz-grüne Bündnis auf Landesebene rückt näher: Die Spitzen von CDU und Grünen in Hamburg halten ihre Differenzen für überwindbar.

[1] Spitzen von Hamburger CDU und Grünen öffnen Weg zu Koalitionsverhandlungen

[1] Leaders of the Hamburger CDU and Greens open path to coalition negotiations.
[2] Then the CDU-leader Michael Freytag and Green party leader Anja Hajduk the division between the parties is bridgable.
Postediting Machine Translation

- Textbox is initially filled with machine translation
- User edits translation
- String edit distance to machine translation is shown (blue background)
Keystroke Log

black: keystroke, **purple:** deletion, grey: cursor move

**red:** sentence completion accept

**orange:** click on translation option

**Analysis:** Segment into periods of activity: typing, **tabbing**, **clicking**, pauses

one second before and after a keystroke is part of typing interval
# Faster and Better

<table>
<thead>
<tr>
<th>Assistance</th>
<th>Speed</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unassisted</td>
<td>4.4s/word</td>
<td>47% correct</td>
</tr>
<tr>
<td>Postedit</td>
<td>2.7s (-1.7s)</td>
<td>55% (+8%)</td>
</tr>
<tr>
<td>Options</td>
<td>3.7s (-0.7s)</td>
<td>51% (+4%)</td>
</tr>
<tr>
<td>Prediction</td>
<td>3.2s (-1.2s)</td>
<td>54% (+7%)</td>
</tr>
<tr>
<td>Prediction+Options</td>
<td>3.3s (-1.1s)</td>
<td>53% (+6%)</td>
</tr>
</tbody>
</table>
Moses: Open Source Toolkit

- **Open source** statistical machine translation system
  - reference implementation of most methods discussed today
  - phrase-based, syntax-based, factored models
  - support for very large models through memory-efficient data structures


- Main source of funding EU **EUROMATRIXPLUS** project, also supported by
  - EU-funded TC-STAR and **EUROMATRIX** project
  - US funding agencies DARPA, NSF
  - universities (Edinburgh, Maryland, MIT, ITC-irst, RWTH Aachen, ...)

Philipp Koehn  ANLP Lecture 27  23 November 2012