6.881: Natural Language Processing
Machine Translation II

Philipp Koehn
CS and AI Lab
MIT

November 23, 2004
Outline

● Lecture I
  – Introduction to Machine Translation
  – Principles of Statistical MT
  – Word-Based Models
  – Phrase-Based Models

● Lecture II
  – Beam Search Decoding
  – Evaluation
  – The Challenge of Syntax
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
Decoding Algorithm

- Goal of the decoding algorithm:
  Put models to work, perform the actual translation
Greedy Decoder

- Greedy Hill-climbing [Germann, 2003]
  - start with gloss
  - improve probability with actions
  - use 2-step look-ahead to avoid some local minima
Beam-Search Decoding Process

- Build translation left to right
  - select foreign words to be translated
Beam-Search Decoding Process

- Build translation left to right
  - select foreign words to be translated
  - find English phrase translation
  - add English phrase to end of partial translation
Beam-Search Decoding Process

- Build translation left to right
  - select foreign words to be translated
  - find English phrase translation
  - add English phrase to end of partial translation
  - mark foreign words as translated
Beam-Search Decoding Process

- One to many translation

Maria no dio una bofetada a la bruja verde

Mary did not
Beam-Search Decoding Process

- Many to one translation
Beam-Search Decoding Process

- Many to one translation

Maria no dio una bofetada a la bruja verde

Mary did not slap the

Philipp Koehn, CSAIL, MIT
Beam-Search Decoding Process

- Reordering
Beam-Search Decoding Process

- Translation finished
## Translation Options

<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</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></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></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>the</td>
<td>the witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Look up possible phrase translations
  - many different ways to segment words into phrases
  - many different ways to translate each phrase
Hypothesis Expansion

<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</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></td>
<td></td>
<td></td>
<td>by</td>
<td>green witch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td></td>
<td>slap</td>
<td></td>
<td></td>
<td>to</td>
<td>the</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>
</tbody>
</table>

- e: no English words
- f: no foreign words covered
- p: probability 1

Start with null hypothesis
Hypothesis Expansion

<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</td>
<td>slap</td>
<td>by</td>
<td>green</td>
<td>witch</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>slap</td>
<td>to</td>
<td>the</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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>e:</th>
<th>f:</th>
<th>p:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>*</td>
<td>.534</td>
</tr>
</tbody>
</table>

- **Pick translation option**
- **Create hypothesis**
  - e: add English phrase Mary
  - f: first foreign word covered
  - p: probability 0.534
### Hypothesis Expansion

<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</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></td>
<td></td>
<td></td>
<td>by</td>
<td></td>
<td>green witch</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td></td>
<td></td>
<td></td>
<td>slap</td>
<td>to</td>
<td>the</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></td>
<td></td>
<td>slap</td>
<td></td>
<td></td>
<td>the</td>
<td>the witch</td>
</tr>
</tbody>
</table>

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

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

- **Add another hypothesis**
Hypothesis Expansion

- Further hypothesis expansion
Hypothesis Expansion

- ... until all foreign words covered
  - find best hypothesis that covers all foreign words
  - backtrack to read off translation
Hypothesis Expansion

- Adding more hypothesis

⇒ Explosion of search space
Explosion of Search Space

- Number of hypotheses is exponential with respect to sentence length

$\Rightarrow$ Decoding is NP-complete [Knight, 1999]

$\Rightarrow$ Need to reduce search space

- risk free: hypothesis recombination
- risky: histogram/threshold pruning
Hypothesis Recombination

- Different paths to the same partial translation
Hypothesis Recombination

- Different paths to the same partial translation

⇒ Combine paths
  - drop weaker hypothesis
  - keep pointer from worse path
Hypothesis Recombination

- Recombined hypotheses do not have to match completely
- No matter what is added, weaker path can be dropped, if:
  - last two English words match (matters for language model)
  - foreign word coverage vectors match (effects future path)
Hypothesis Recombination

- Recombined hypotheses do not have to match completely
- No matter what is added, weaker path can be dropped, if:
  - last two English words match (matters for language model)
  - foreign word coverage vectors match (effects future path)

⇒ Combine paths
Pruning

- Hypothesis recombination is not sufficient

⇒ Heuristically discard weak hypotheses

- Organize Hypothesis in stacks, e.g. by
  - same foreign words covered
  - same number of foreign words covered (Pharaoh does this)
  - same number of English words produced

- Compare hypotheses in stacks, discard bad ones
  - histogram pruning: keep top $n$ hypotheses in each stack (e.g., $n=100$)
  - threshold pruning: keep hypotheses that are at most $\alpha$ times the cost of best hypothesis in stack (e.g., $\alpha = 0.001$)
Comparing Hypotheses

- Comparing hypotheses with same number of foreign words covered

Maria no dio una bofetada a la bruja verde

<table>
<thead>
<tr>
<th>e: Mary did not</th>
<th>f: *-------</th>
<th>p: 0.154</th>
</tr>
</thead>
<tbody>
<tr>
<td>better partial translation</td>
<td>covers easier part --&gt; lower cost</td>
<td></td>
</tr>
</tbody>
</table>

- Hypothesis that covers easy part of sentence is preferred

⇒ Need to consider future cost
Future Cost Estimation

- Estimate cost to translate remaining part of input

- **Step 1: find cheapest translation options**
  - find cheapest translation option for each input span
  - compute translation model cost
  - estimate language model cost (no prior context)
  - ignore reordering model cost

- **Step 2: compute cheapest cost**
  - for each contiguous span:
    - find cheapest sequence of translation options

- **Precompute and lookup**
  - precompute future cost for each contiguous span
  - future cost for any coverage vector:
    - sum of cost of each contiguous span of uncovered words
  → no expensive computation during run time
Word Lattice Generation

- Search graph can be easily converted into a word lattice
  - can be further mined for n-best lists
  → enables reranking approaches
  → enables discriminative training
Evaluation

- Manual Evaluation
  - human judge output
  - expensive

- Automatic Evaluation
  - machines judge output
  - fast
  - reliable?

- Task-Oriented Evaluation
  - humans do task with MT
  - tests usefulness of MT
Manual Evaluation

- Correct yes/no
  - simple
  - longer sentences almost always have at least one error

- Correct on scale
  - 0=bad, 5=perfect
  - disagreement between judges

- More detailed judgments
  - adequacy: how well is meaning preserved?
  - fluency: is it good English?
  - ...
Manual Evaluation

- Give grade from 0=bad to 5=perfect
  - In the First Two Months Guangdong's Export of High-Tech Products 3.76 Billion US Dollars
  - The Guangdong provincial foreign trade and economic growth has made important contributions.
  - Suicide explosion in Jerusalem

- Agreement
Automatic Evaluation

- Why automatic evaluation metrics?
  - manual evaluation is too slow
  - evaluation on large test sets reveals minor improvements
  - automatic tuning to improve machine translation performance

- History
  - Word Error Rate
  - BLEU since 2002
  - BLEU in short: overlap with reference translations
Bi-Lingual Evaluation Understudy (BLEU)

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 office all receives 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.

BLEU4 formula

\[ bp \times \exp (\log p_1 + \log p_2 + \log p_3 + \log p_4) \]

\[ p_1 = 1\text{-gram precision} \]
\[ p_2 = 2\text{-gram precision} \]
\[ p_3 = 3\text{-gram precision} \]
\[ p_4 = 4\text{-gram precision} \]

\[ bp = \text{brevity penalty:} \quad \min(1, \exp(\text{words-in-reference / words-in-} \]

Philipp Koehn, CSAIL, MIT
Correlation with Manual Evaluation

- Correlates with human evaluation (adequacy, fluency)

\[ R^2 = 88.0\% \]
\[ R^2 = 90.2\% \]
The Challenge of Syntax

- Remember the pyramid
Advantages of Syntax-Based Translation

- Reordering for syntactic reasons
  - e.g., move German object to end of sentence

- Better explanation for function words
  - e.g., prepositions, determiners

- Conditioning to syntactically related words
  - translation of verb may depend on subject or object

- Use of syntactic language models
Inversion Transduction Grammars

- Generation of both English and foreign trees [Wu, 1997]
- Rules (binary and unary)
  - \( A \rightarrow A_1 A_2 | A_1 A_2 \)
  - \( A \rightarrow A_1 A_2 | A_2 A_1 \)
  - \( A \rightarrow e | f \)
  - \( A \rightarrow e | \ast \)
  - \( A \rightarrow \ast | f \)

⇒ Common binary tree required
  - limits the complexity of reorderings
Syntax Trees

Mary did not slap the green witch

- English binary tree
Syntax Trees (2)

Spanish binary tree

Maria no daba una bofetada a la bruja verde
Syntax Trees (3)

- Combined tree with reordering of Spanish
- Can such trees be learned from data?
- Do common tree exist with real syntax on both sides?
Dependency Structure

- Common dependency tree
- Interest in dependency-based translation models
String to Tree Translation

- Use of English syntax trees [Yamada and Knight, 2001]
  - exploit rich resources on the English side
  - obtained with statistical parser [Collins, 1997]
  - flattened tree to allow more reorderings
  - works well with syntactic language model
Yamada and Knight [2001]

 lider

# VB

PRP VB1 VB2

he adores VB TO

listening TO MN
to music

# reorder

PRP VB2 VB1

he TO VB adores

MN TO listening

mus  to

# insert

PRP VB2 VB1

he ha TO VB ga adores desu

MN TO listening no

music to

# translate

PRP VB2 VB1

kare ha TO VB ga daisuki desu

MN TO kiku no

ongaku wo

# take leaves

Kare ha ongaku wo kiku no ga daisuki desu
Syntactic Language Model

- Good syntax tree $\rightarrow$ good English
- Allows for long distance constraints

- Left translation preferred by syntactic LM
String to Tree Transfer and Syntactic LM

- Work presented at this MT Summit by Charniak, Knight, Yamada
  - more grammatical correct output
  - more perfectly translated sentences
  - ... but no improvement in BLEU

- Syntactic transfer and LM on top of phrase translation
  - parse a lattice generated by phrase-based MT
  - no results yet
Augment Models with Syntactic Features

- Intuition: other models work fine, syntax provides additional clues

- Define syntactic properties that should hold
  - preservation of plural
  - output should have verb
  - no dropping of content words
  - ...

- 2003 summer workshop at John Hopkins: little improvement
Clause Structure

- Syntax tree from German parser
  - statistical parser by Amit Dubay, trained on TIGER treebank
Reordering When Translating

Reordering when translating into English

- tree is flattened
- clause level constituents line up

Philipp Koehn, CSAIL, MIT
Clause Level Reordering

Clause level reordering is a well defined task

- label German constituents with their English order
- done this for 300 sentences, two annotators, high agreement

Philipp Koehn, CSAIL, MIT
Systematic Reordering German → English

- Many types of reorderings are systematic
  - move verb group together
  - subject - verb - object
  - move negation in front of verb

⇒ Write rules by hand
  - apply rules to test and training data
  - train standard phrase-based SMT system

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline system</td>
<td>25.2%</td>
</tr>
<tr>
<td>with manual rules</td>
<td>26.8%</td>
</tr>
</tbody>
</table>
Integration

- Transform f into f’ with our methods
- Translate n-best restructurings with phrase-based MT
  - uses both transformation score and translation/language model score
  - if no restructuring → baseline performance
- Transformation does not need to be perfect
  - phrase-based model may still reorder

Philipp Koehn, CSAIL, MIT
Improved Translations

- we must also this criticism should be taken seriously.
  → we must also take this criticism seriously.

- i am with him that it is necessary, the institutional balance by means of a political revaluation of both the commission and the council to maintain.
  → i agree with him in this, that it is necessary to maintain the institutional balance by means of a political revaluation of both the commission and the council.

- thirdly, we believe that the principle of differentiation of negotiations note.
  → thirdly, we maintain the principle of differentiation of negotiations.

- perhaps it would be a constructive dialog between the government and opposition parties, social representative a positive impetus in the right direction.
  → perhaps a constructive dialog between government and opposition parties and social representative could give a positive impetus in the right direction.