Statistical machine translation today

- Best performing methods based on *phrases*
  - short sequences of words
  - no use of explicit syntactic information
  - no use of morphological information
  - currently best performing method

- Progress in *syntax-based* translation
  - tree transfer models using syntactic annotation
  - still no use of morphological information
  - slower, more complex, and lower translation quality
  - active research, closing the performance gap?
Morphology for machine translation

- Models treat *car* and *cars* as completely different words
  - training occurrences of *car* have no effect on learning translation of *cars*
  - if we only see *car*, we do not know how to translate *cars*
  - rich morphology (German, Arabic, Finnish, Czech, ...) → many word forms

- Better approach
  - analyze surface word forms into *lemma* and *morphology*, e.g.: *car + plural*
  - translate lemma and morphology separately
  - generate target surface form

Factored translation models

- **Factored representation** of words

  ![Factored Representation](chart)

- Goals
  - **Generalization**, e.g. by translating lemmas, not surface forms
  - **Richer model**, e.g. using syntax for reordering, language modeling)
Decomposing translation: example

- *Translate* lemma and syntactic information *separately*

  - `lemma` ⇒ `lemma`
  - `part-of-speech` ⇒ `part-of-speech`
  - `morphology` ⇒ `morphology`

Decomposing translation: example

- *Generate surface* form on target side

  - `surface`
  - `↑`
  - `lemma`
  - `part-of-speech`
  - `morphology`
Translation process

- Extension of phrase model
  - translation step is one-to-one mapping of word sequences

- Mapping of foreign words into English words broken up into steps
  - translation step: maps foreign factors into English factors
  - generation step: maps English factors into English factors

- Order of mapping steps is chosen to optimize search

Translation process: example

Input: (Autos, Auto, NNS)

1. Translation step: lemma $\Rightarrow$ lemma
   $(?, \text{car}, ?), (?, \text{auto}, ?)$

2. Generation step: lemma $\Rightarrow$ part-of-speech
   $(?, \text{car, NN}), (?, \text{car, NNS}), (?, \text{auto, NN}), (?, \text{auto, NNS})$

3. Translation step: part-of-speech $\Rightarrow$ part-of-speech
   $(?, \text{car, NN}), (?, \text{car, NNS}), (?, \text{auto, NNP}), (?, \text{auto, NNS})$

4. Generation step: lemma,part-of-speech $\Rightarrow$ surface
   (car, car, NN), (cars, car, NNS), (auto, auto, NN), (autos, auto, NNS)
Integration with factored language models

- **Factored language models**: back-off to factors with richer statistics
  - if preceding word is rare, current word hard to predict
  → back-off to part-of-speech tags

- Example
  - \( \text{count}(\text{scotland is}) = \text{count}(\text{scotland fish}) = \text{count}(\text{scotland yellow}) = 0 \)
  - \( \text{count}(\text{NNP is}) > \text{count}(\text{NNP fish}) > \text{count}(\text{NNP yellow}) \)

- Gains shown for speech recognition and translation

Richer models for machine translation

- **Reordering** is often due to syntactic reasons
  - French-English: \( \text{NN ADJ} \rightarrow \text{ADJ NN} \)
  - Chinese-English: \( \text{NN1 F NN2} \rightarrow \text{NN1 NN2} \)
  - Arabic-English: \( \text{VB NN} \rightarrow \text{NN VB} \)

- **Syntactic coherence** may be modeled using syntactic tags
  - n-gram models of *part-of-speech tags* may aid grammaticality of output
  - sequence models over *morphological tags* may aid agreement (e.g., case, number, and gender agreement in noun phrases)
Adding linguistic markup to output

- High order language models over POS
- Motivation: syntactic tags should enforce syntactic sentence structure
- Results: No major impact with 7-gram POS model
- Analysis: local grammatical coherence already fairly good, POS sequence LM model not strong enough to support major restructuring

Local agreement (esp. within noun phrases)

- High order language models over POS and morphology
- Motivation
  - *DET-sgl* NOUN-*sgl* good sequence
  - *DET-sgl* NOUN-*plural* bad sequence
Agreement within noun phrases

- Experiment: 7-gram POS, morph LM in addition to 3-gram word LM

- Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Agreement errors in NP</th>
<th>devtest</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>15% in NP ≥ 3 words</td>
<td>18.22 BLEU</td>
<td>18.04 BLEU</td>
</tr>
<tr>
<td>factored model</td>
<td>4% in NP ≥ 3 words</td>
<td>18.25 BLEU</td>
<td>18.22 BLEU</td>
</tr>
</tbody>
</table>

- Example
  - baseline: ... zur zwischenstaatlichen methoden ...
  - factored model: ... zu zwischenstaatlichen methoden ...

- Example
  - baseline: ... das zweite wichtige änderung ...
  - factored model: ... die zweite wichtige änderung ...

Morphological generation model

- Our motivating example

- Translating lemma and morphological information more robust
Initial results

- Results on 1 million word News Commentary corpus (German–English)

<table>
<thead>
<tr>
<th>System</th>
<th>In-domain</th>
<th>Out-of-domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>18.19</td>
<td>15.01</td>
</tr>
<tr>
<td>With POS LM</td>
<td>19.05</td>
<td>15.03</td>
</tr>
<tr>
<td>Morphgen model</td>
<td>14.38</td>
<td>11.65</td>
</tr>
</tbody>
</table>

- What went wrong?
  - why back-off to lemma, when we know how to translate surface forms?
  → loss of information

Solution: alternative decoding paths

- Allow both surface form translation and morphgen model
  - prefer surface model for known words
  - morphgen model acts as back-off
Results

• Model now beats the baseline:

<table>
<thead>
<tr>
<th>System</th>
<th>In-domain</th>
<th>Out-of-domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>18.19</td>
<td>15.01</td>
</tr>
<tr>
<td>With POS LM</td>
<td>19.05</td>
<td>15.03</td>
</tr>
<tr>
<td>Morphgen model</td>
<td>14.38</td>
<td>11.65</td>
</tr>
<tr>
<td>Both model paths</td>
<td>19.47</td>
<td>15.23</td>
</tr>
</tbody>
</table>

Adding annotation to the source

• Source words may contain insufficient information to map phrases
  – English-German: what case for noun phrases?
  – Chinese-English: plural or singular
  – pronoun translation: what do they refer to?

• Idea: add additional information to the source that makes the required information available locally (where it is needed)
Case information for English–German

Input \hspace{1cm} Output

\begin{itemize}
\item Detect in English, if noun phrase is subject/object (using parse tree)
\item Map information into case morphology of German
\item Use case morphology to generate correct word form
\end{itemize}

Factored models: open questions

\begin{itemize}
\item What is the\textit{ best decomposition} into translation and generation steps?
\item Same segmentation for all translation steps?
\item \textit{What information} is useful?
  \begin{itemize}
  \item translation: mostly lexical, or lemmas for richer statistics
  \item reordering: syntactic information useful
  \item language model: syntactic information for overall grammatical coherence
  \end{itemize}
\item Use of annotation tools vs. \textit{automatically discovered} word classes
\item \textit{Other decoding steps} besides phrase translation and word generation?
\end{itemize}