Data Intensive Linguistics — Lecture 19
Machine translation (VI): Advanced Topics

Philipp Koehn

16 March 2006
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

  \[
  \begin{align*}
  &\text{surface} \\
  &\text{lemma} \\
  &\text{part of speech} \\
  &\text{morphology} \\
  &\text{word class} \\
  &\ldots
  \end{align*}
  \quad \rightarrow \quad
  \begin{align*}
  &\text{surface} \\
  &\text{lemma} \\
  &\text{part of speech} \\
  &\text{morphology} \\
  &\text{word class} \\
  &\ldots
  \end{align*}
  \]

- **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*

\[
\begin{array}{c}
\text{lemma} \quad \Rightarrow \quad \text{lemma} \\
\text{part-of-speech} \quad \Rightarrow \quad \text{part-of-speech} \\
\text{morphology} \quad \Rightarrow \quad \text{morphology}
\end{array}
\]
Decomposing translation: example

- *Generate surface* form on target side

```
  surface
  \uparrow
  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: \((\text{Autos}, \text{Auto}, \text{NNS})\)

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

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

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

4. Generation step: lemma,part-of-speech \(\Rightarrow\) surface
   \((\text{car}, \text{car}, \text{NN}), (\text{cars}, \text{car}, \text{NNS}), (\text{auto}, \text{auto}, \text{NN}), (\text{autos}, \text{auto}, \text{NNS})\)
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: $NN \text{ ADJ} \rightarrow \text{ADJ} \ NN$
  – Chinese-English: $NN1 \ F \ NN2 \rightarrow NN1 \ NN2$
  – Arabic-English: $VB \ NN \rightarrow \ NN \ VB$

• **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)
Factored models: open questions

- What is the *best decomposition* into translation and generation steps?

- Same segmentation for all translation steps?

- *What information* is useful?
  - translation: mostly lexical, or lemmas for richer statistics
  - reordering: syntactic information useful
  - language model: syntactic information for overall grammatical coherence

- Use of annotation tools vs. *automatically discovered* word classes

- *Back-off* models (use complex mappings, if available)