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

$$\begin{bmatrix} \text{surface} \\ \text{lemma} \\ \text{part of speech} \\ \text{morphology} \\ \text{word class} \\ \ldots \end{bmatrix} \Rightarrow \begin{bmatrix} \text{surface} \\ \text{lemma} \\ \text{part of speech} \\ \text{morphology} \\ \text{word class} \\ \ldots \end{bmatrix}$$

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

Decomposing translation: example

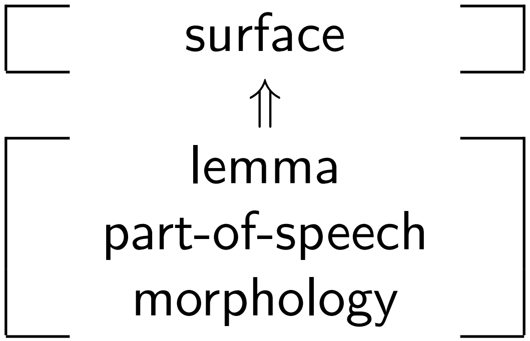
- *Translate* lemma and syntactic information *separately*

$$\begin{bmatrix}\text{lemma}\end{bmatrix} \Rightarrow \begin{bmatrix}\text{lemma}\end{bmatrix}$$

$$\begin{bmatrix}\text{part-of-speech}\\ \text{morphology}\end{bmatrix} \Rightarrow \begin{bmatrix}\text{part-of-speech}\\ \text{morphology}\end{bmatrix}$$

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 ⇒ lemma
   (?, car, ?), (?, auto, ?)

2. Generation step: lemma ⇒ part-of-speech
   (?, car, NN), (?, car, NNS), (?, auto, NN), (?, auto, NNS)

3. Translation step: part-of-speech ⇒ part-of-speech
   (?, car, NN), (?, car, NNS), (?, auto, NNP), (?, auto, NNS)

4. Generation step: lemma,part-of-speech ⇒ 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

  – count(*scotland is*) $=$ count(*scotland fish*) $=$ count(*scotland yellow*) $= 0$

  – count(*NNP is*) $>$ count(*NNP fish*) $>$ count(*NNP yellow*)

- Gains shown for speech recognition and translation

Richer models for machine translation

- **Reordering** is often due to syntactic reasons
  - French-English: *NN ADJ* → *ADJ NN*
  - Chinese-English: *NN1 F NN2* → *NN1 NN2*
  - Arabic-English: *VB NN* → *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)