Empirical Methods in Natural Language Processing
Lecture 19
Machine translation (VI):
Factored Translation Models

Philipp Koehn

10 March 2008

School of informatics

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

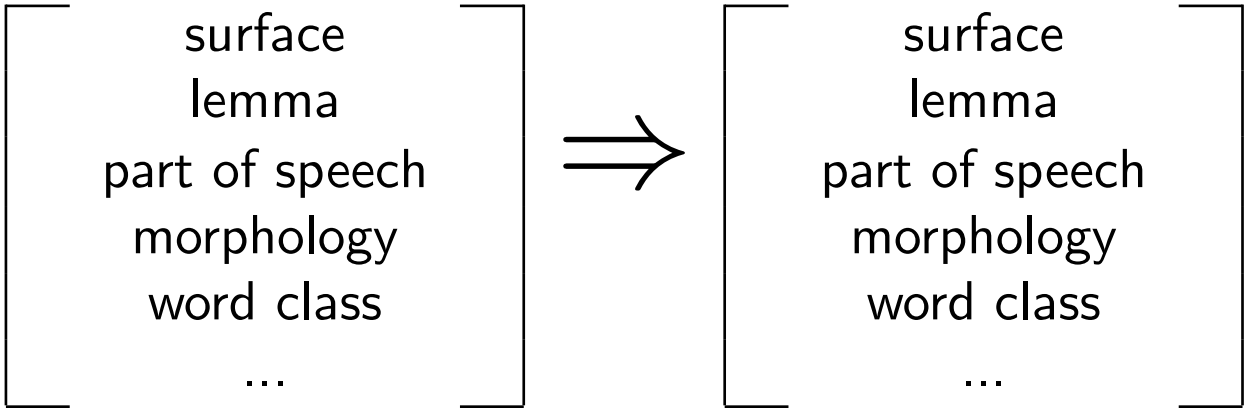
- Progress in *syntax-based* translation
  - tree transfer models using syntactic annotation
  - still no use of morphological information
  - slower, more complex, and lower translation quality
  - active research, closing the performance gap?

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, ...) $\rightarrow$ 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

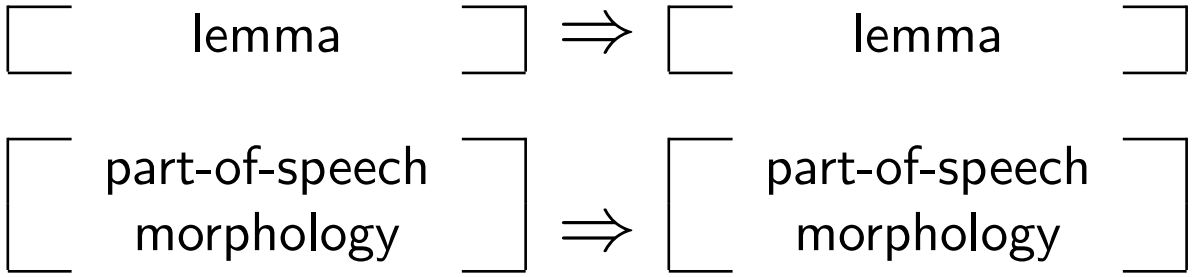
- **Factored represention** of words



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

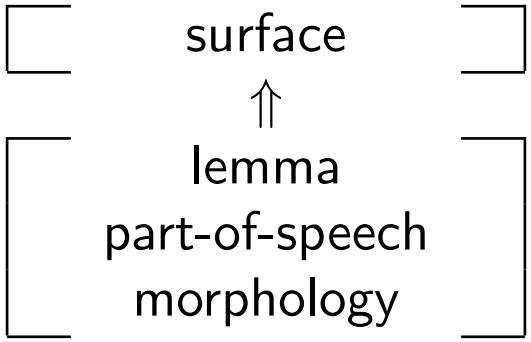
Decomposing translation: example

- *Translate* lemma and syntactic information *separately*



Decomposing translation: example

- *Generate surface* form on target side



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

  $\rightarrow$ 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*)

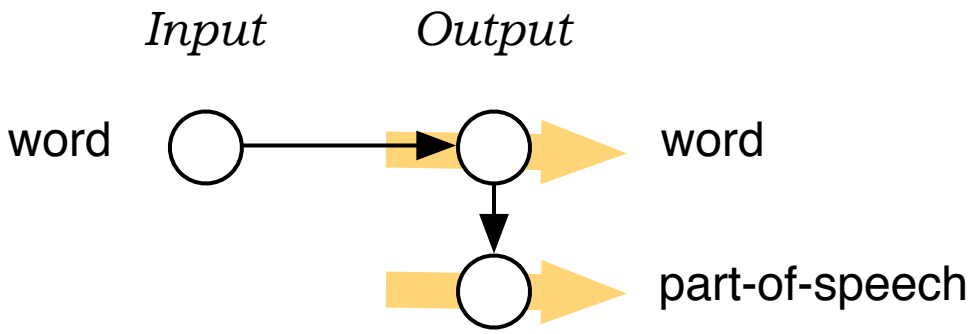
- Gains shown for speech recognition and translation

Richer models for machine translation

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

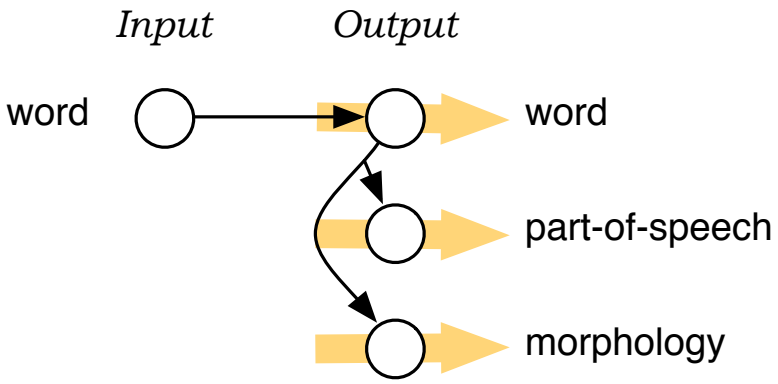
- **Syntactic coherence** may be modeled using syntactic tags
  - n-gram models of *part-of-speech tags* may aid grammaticality of output
  - sequence models over *morphological tags* may aid agreement (e.g., case, number, and gender agreement in noun phrases)

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

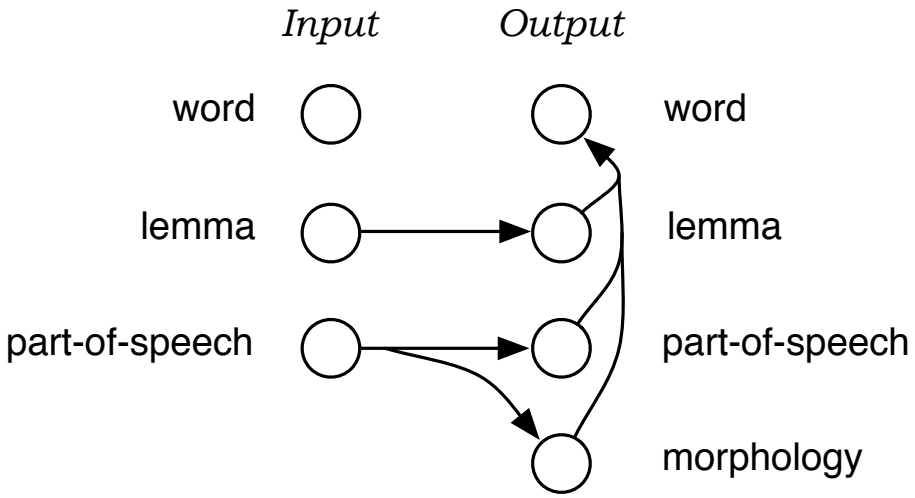
Agreement within noun phrases

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

Method	Agreement errors in NP	devtest	test
baseline	15% in NP $\geq$ 3 words	18.22 BLEU	18.04 BLEU
factored model	4% in NP $\geq$ 3 words	18.25 BLEU	18.22 BLEU

- 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

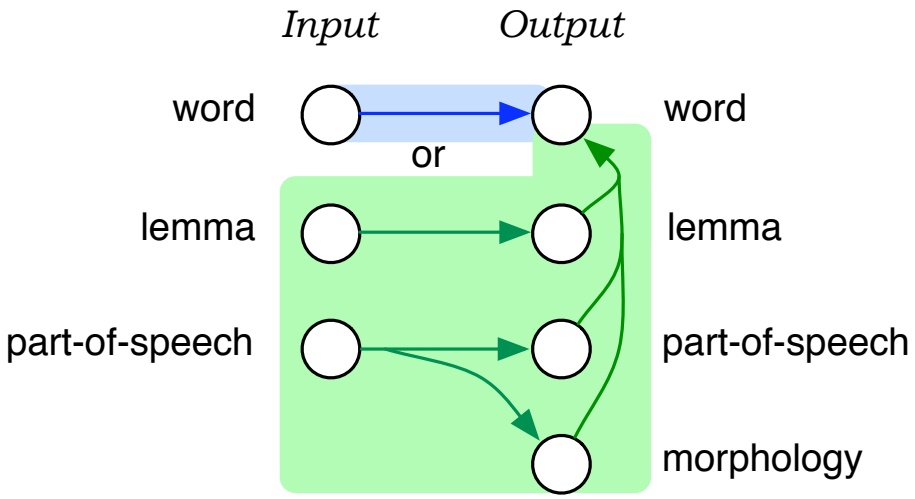
- Results on 1 million word News Commentary corpus (German–English)

System	In-doman	Out-of-domain
Baseline	18.19	15.01
With POS LM	19.05	15.03
Morphgen model	14.38	11.65

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

  $\rightarrow$ 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:

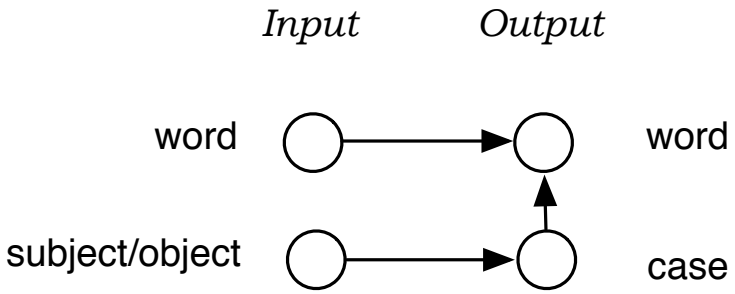
System	In-doman	Out-of-domain
Baseline	18.19	15.01
With POS LM	19.05	15.03
Morphgen model	14.38	11.65
Both model paths	19.47	15.23

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



- Detect in English, if noun phrase is subject/object (using parse tree)
- Map information into case morphology of German
- Use case morphology to generate correct word form

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
- *Other decoding steps* besides phrase translation and word generation?