Morphology

Philipp Koehn

24 March 2016

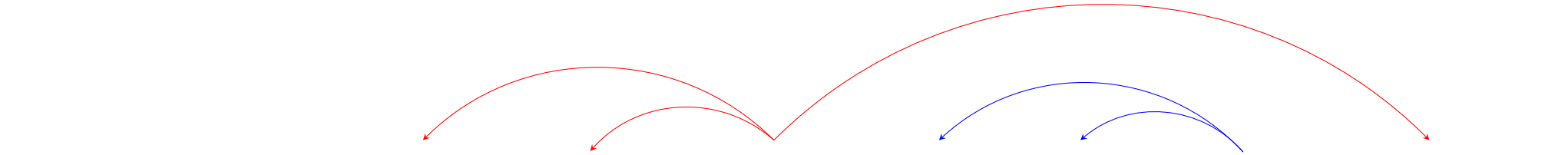
A Naive View of Language

- Language needs to name
  - nouns: objects in the world (dog)
  - verbs: actions (jump)
  - adjectives and adverbs: properties of objects and actions (brown, quickly)

- Relationship between these have to specified
  - word order
  - morphology
  - function words

Marking of Relationships: Agreement

- From Catullus, First Book, first verse (Latin):



Cui	dono	lepidum	novum	libellum	arida	modo	pumice	expolitum	?
Whom	I-present	lovely	new	little-book	dry	manner	pumice	polished	?

(To whom do I present this lovely new little book now polished with a dry pumice?)

- Gender (and case) agreement links adjectives to nouns

Marking of Relationships to Verb: Case

- German:

Die Frau	gibt	dem Mann	den Apfel
The woman	gives	the man	the apple
subject		indirect object	object

- Case inflection indicates role of noun phrases

Case Morphology vs. Prepositions

- Two different word orderings for English:
  - The woman gives the man the apple
  - The woman gives the apple **to** the man

- Japanese:

女性は	男性に	アップルの	を与えます
woman SUBJ	man OBJ	apple OBJ2	gives

- Is there a real difference between prepositions and noun phrase case inflection?

Writingwordstogether

- Definition of word boundaries purely an artifact of writing system
- Differences between languages
  - Agglutinative compounding
    Informatikseminar vs. computer science seminar
  - Function word vs. affix
- Border cases
  - Joe's — one token or two?
  - Morphology of affixes often depends on phonetics / spelling conventions
    dog+s → dogs vs. pony → ponies

    ... but note the English function word a:
    a donkey vs. an aardvark

Relationship between Noun Phrases

- In English handled with possessive case, prepositions, or word order
- Possessive case somewhat interchangeable with of preposition

  the dog's bone vs. the bone of the dog

- Mulitiple modifiers

  the instructions by the teacher to the student about the assignment

  (teacher) student assignment instructions

Changing Part-of-Speech

- Derivational morphology allows changing part of speech of words

- Example:

  – base: nation, noun
  → national, adjective
  → nationally, adverb
  → nationalist, noun
  → nationalism, noun
  → nationalize, verb

- Sometimes distinctions between POS quite fluid (enabled by morphology)

  – I want to integrate morphology
  – I want the integration of morphology

Meaning Altering Affixes

- English

undo

redo

hypergraph

- German: zer- implies action causes destruction

Er **zer**redet das Thema → He talks the topic **to death**

- Spanish: -ito means object is small

burro → burrito

Adding Subtle Meaning

- Morphology allows adding subtle meaning
  - verb tenses: time action is occurring, if still ongoing, etc.
  - count (singular, plural): how many instances of an object are involved
  - definiteness (the cat vs. a cat): relation to previously mentioned objects
  - grammatical gender: helps with co-reference and other disambiguation

- Sometimes redundant: same information repeated many times

how does morphology impact machine translation?

Unknown Source Words

- Ratio of unknown words in WMT 2013 test set:

Source language	Ratio unknown
Russian	2.0%
Czech	1.5%
German	1.2%
French	0.5%
English (to French)	0.5%

- Caveats:
  - corpus sizes differ
  - not clear which unknown words have known morphological variants

Unknown Target Words

- Same problem, different flavor

- Harder to quantify

  (unknown words in reference?)

- Enforcing morphological constraints may have unintended consequences

  – correct morphological variant unknown (or too rare)

  → different lemma is chosen by system

Differently Encoded Information

- Languages with different sentence structure

das	behaupten	sie	wenigstens
this	claim	they	at least
the		she	

- Convert from inflected language into configuration language

  (and vice versa)

- Ambiguities can be resolved through syntactic analysis
  - the meaning the of das not possible (not a noun phrase)
  - the meaning she of sie not possible (subject-verb agreement)

Non-Local Information

- Pronominal anaphora

I saw the movie and it is good.

- How to translate it into German (or French)?
  - it refers to movie
  - movie translates to Film
  - Film has masculine gender
  - ergo: it must be translated into masculine pronoun er

- We are not handling pronouns very well

Complex Semantic Inference

- Example

Whenever I visit my uncle and his daughters,
I can't decide who is my favorite cousin.

- How to translate cousin into German? Male or female?

compound splitting

Compounds

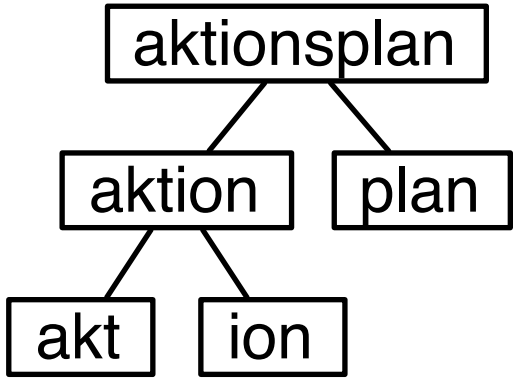
- Compounding = merging words into new bigger words
- Prevalent in German, Dutch, and Finnish
- Rare in English: homework, website

$\Rightarrow$ Compounds in source need to be split up in pre-processing

- Note related problem: word segmentation in Chinese

Compound Splitting

- Break up complex word into smaller words found in vocabulary



- Frequency-based method: geometric average of word counts
  - aktionsplan (652) → 652
  - aktion (960) / plan → **825.6**
  - aktions (5) / plan → 59.6
  - akt (224) / ion (1) / plan (710) → 54.2

Compound Merging

- When translating into a compounding language, compounds need to be created
- Original sentence (tokenized)

  der Polizeibeamte gibt dem Autofahrer einen Alkoholtest .

- Split compounds in preprocessing, build translation model with split data

  der Polizei Beamte gibt dem Auto Fahrer einen Alkohol Test .

- Detect merge points (somehow....)

  der Auto @~@ Fahrer verweigert den Polizei @~@ Alkohol @~@ Test .

- Merge compounds

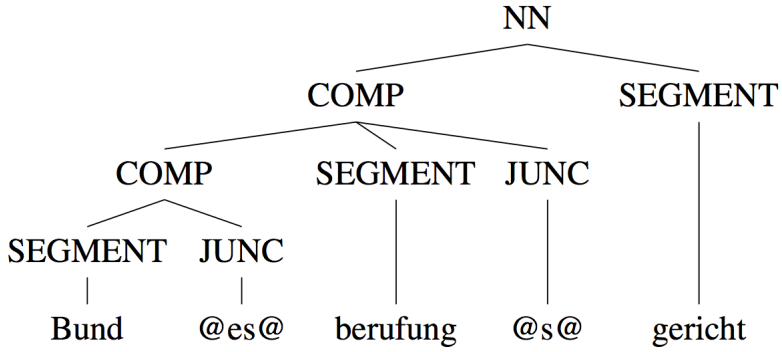
  der Autofahrer verweigert den Polizeialkoholtest .

Detecting Merge Points

- Mark compounding
  (special token @∼@ in the translation model or mark part words with Auto#)

- Classifier approach (Weller et al., 2014)
  - handle compound merging in post-processing
  - train classifier to predict for each word that it should be merged with the next
  - features:
    - part-of-speech tag
    - frequency or ratio that it occurs in compound
    - are aligned source words part of same base noun phrase etc.?

- Part of syntactic annotation in syntax-based models (Williams et al., 2014)



rich morphology in the source

German

- German sentence with morphological analysis

Er	wohnt	in	einem	großen	Haus
Er	wohnen -en+t	in	ein +em	groß +en	Haus +ϵ
He	lives	in	a	big	house

- Four inflected words in German, but English...

  also inflected both English verb live and German verb wohnen
  inflected for tense, person, count
  not inflected corresponding English words not inflected (a and big)
  $\rightarrow$ easier to translate if inflection is stripped
  less inflected English word house inflected for count
  German word Haus inflected for count and case
  $\rightarrow$ reduce morphology to singular/plural indicator

- Reduce German morphology to match English

  Er | wohnen+3P-SGL | in | ein | groß | Haus+SGL

Turkish

- Example
  - Turkish: *Sonuçlarına*$_1$ *dayanılarak*$_2$ *bir*$_3$ *ortakliği*$_4$ *oluşturulacaktır*$_5$.
  - English: **a**$_3$ **partnership**$_4$ will be **drawn-up**$_5$ on the **basis**$_2$ of **conclusions**$_1$.

- Turkish morphology → English function words (will, be, on, the, of)

- Morphological analysis

Sonuç +lar +sh +na daya +hnhl +yarak bir ortaklık +sh oluş +dhr +hl +yacak +dhr

- Alignment with morphemes

sonuç	+lar	+sh	+na		daya+hnhl	+yarak	bir	ortaklık	+sh	oluş	+dhr	+hl	+yacak	+dhr
conclusion	+s		of	the	basis	on	a	partnership		draw up	+ed		will	be

⇒ Split Turkish into morphemes, drop some

Arabic

- Basic structure of Arabic morphology

[CONJ+ [PART+ [*al+* BASE +PRON]]]

- Examples for clitics (prefixes or suffixes)
  - definite determiner al+ (English the)
  - pronominal morpheme +hm (English their/them)
  - particle l+ (English to/for)
  - conjunctive pro-clitic w+ (English and)

- Same basic strategies as for German and Turkish
  - morphemes akin to English words → separated out as tokens
  - properties (e.g., tense) also expressed in English → keep attached to word
  - morphemes without equivalence in English → drop

Arabic Preprocessing Schemes

ST Simple tokenization (punctuations, numbers, remove diacritics)

wsynhY Alr}ys jwlth bzyArp AlY trkyA .

D1 Decliticization: split off conjunction clitics

w+ synhy Alr}ys jwlth bzyArp <lY trkyA .

D2 Decliticization: split off the class of particles

w+ s+ ynhy Alr}ys jwlth b+ zyArp <lY trkyA .

D3 Decliticization: split off definite article (Al+) and pronominal clitics

w+ s+ ynhy Al+ r}ys jwlp +P_{3MS} b+ zyArp <lY trkyA .

MR Morphemes: split off any remaining morphemes

w+ s+ y+ nhy Al+ r}ys jwl +p +h b+ zyAr +p <lY trkyA .

EN English-like: use lexeme and English-like POS tags, indicates pro-dropped verb subject as a separate token

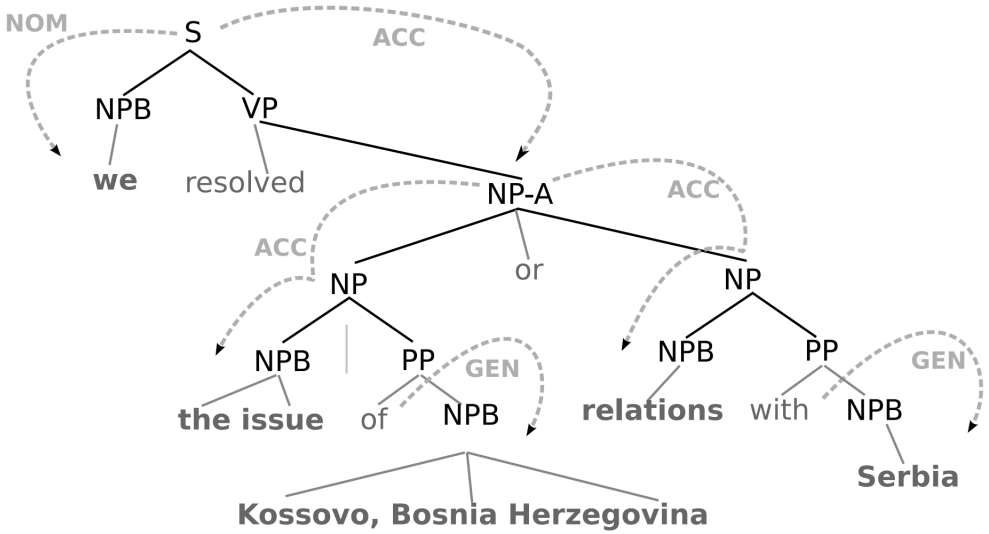
w+ s+ >nhY$_{VBP}$ +S_{3MS} Al+ r}ys$_{NN}$ jwlp$_{NN}$ +P_{3MS} b+ zyArp$_{NN}$ <lY trky$_{NNP}$

missing information in the source

Enriching the Source

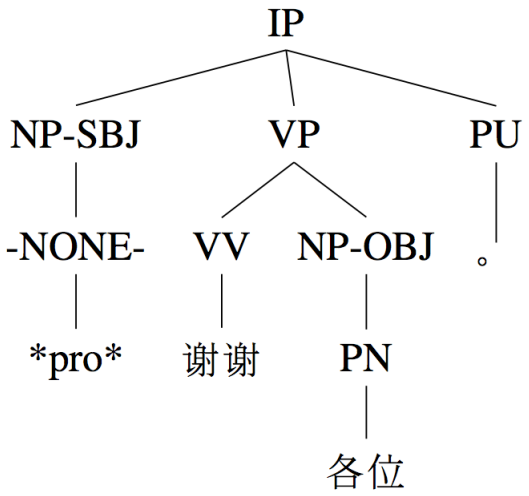
- Translating from morphologically poor to rich language
- Idea: Add annotation to source
  - morphological analysis
  - syntactic parsing (phrase structure and dependencies)
  - semantic analysis
  - prediction models that consider context
- Surprisingly little work in this area

Adding Case Information



- Translating
  - from language with word order marking of noun phrases (e.g., English)
  - to language with morphological case marking (e.g., Greek, German)
- Case information needed when generating target, but it is not local
- Method (Avramidis and Koehn, 2008)
  - parse English source sentence
  - detect "case" of each noun phrase
  - annotate words that map to inflected forms (nouns, adjectives, determiners)

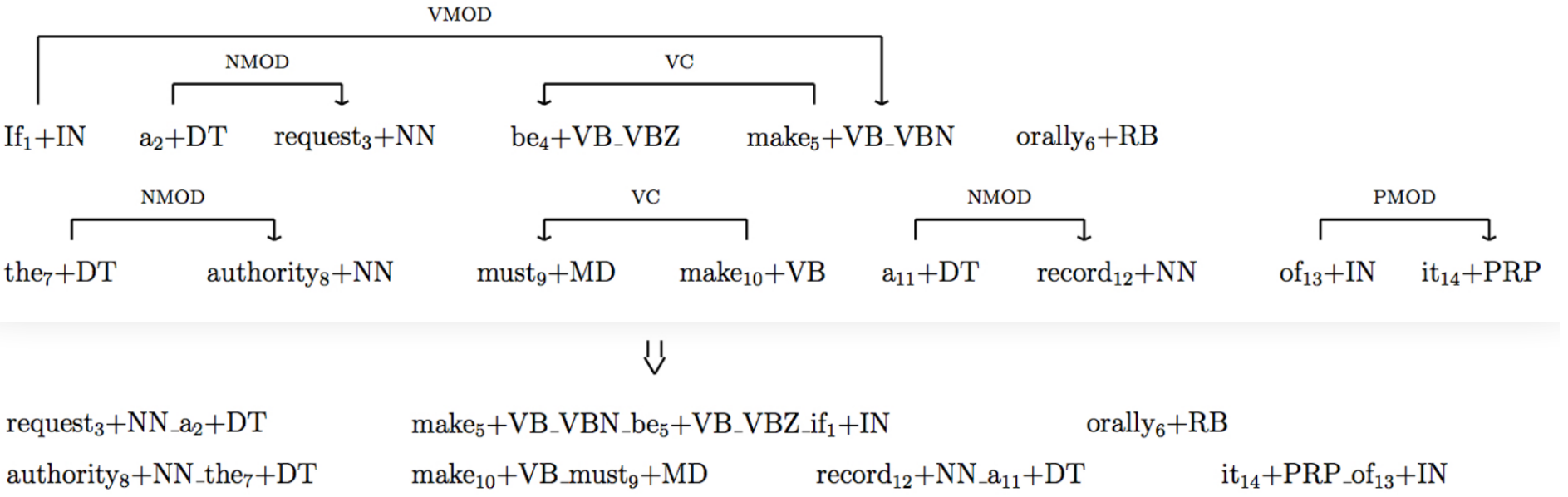
Special Tokens for Empty Categories



(IP (NP-SBJ (-NONE- *pro*)) VP PU)

- Linguistic analysis of some languages suggests the existence of empty categories
- Most commonly known: pro-drop, omission of pronouns
- Method (Chung and Gildea, 2010) for Chinese–English
  - detect empty categories with parser and structured maximum entropy model
  - insert special token in source side of parallel corpus

Transforming into Complex Morphology



- English-Turkish: generation of complex morphology
- Method (Yeniterzi and Oflazer, 2010)
  - parse English sentence
  - annotate each word with part-of-speech tag
  - attach function words that will be part of Turkish morphology

generating target side morphology

Problem

- Example: Case inflection in German

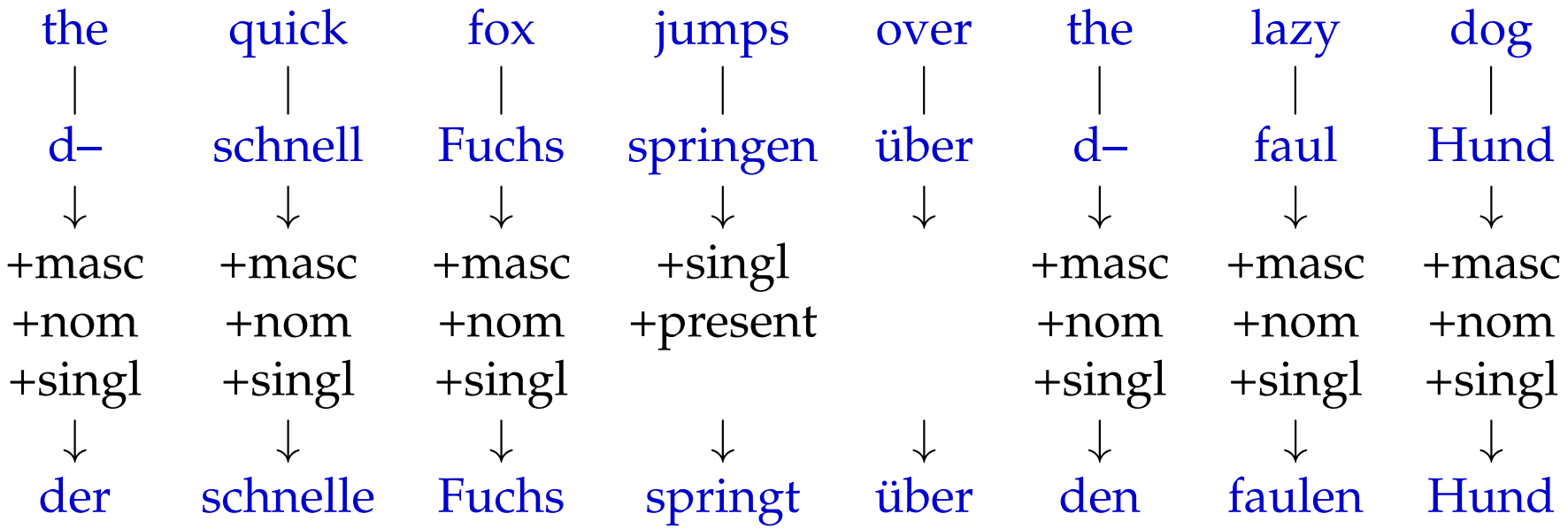
$$\text{quick fox} \rightarrow \begin{cases} \text{schnelle Fuchs} \\ \text{schnellen Fuchses} \\ \text{schnellem Fuchs} \\ \text{schnellen Fuchs} \end{cases}$$

- Relevant information is not local to the phrase rule

- Sparse data
  - differentiating between inflected forms splits statistical evidence
  - some cases of correct inflection may be missing

$\Rightarrow$ Translation into lemma, inflection as post-processing

Inflection Prediction



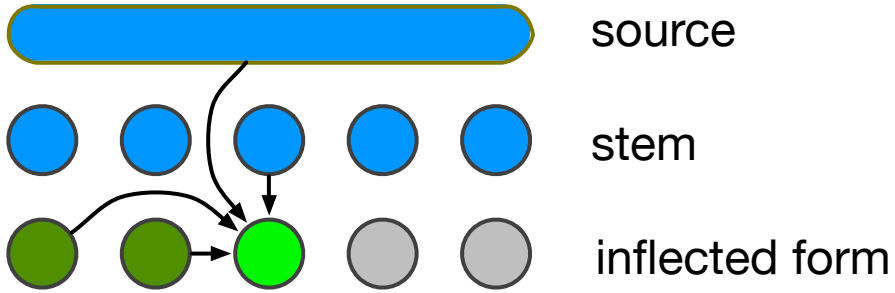
- Inflection as classification task
- Morphological properties typically come from morphological analyzer, but can also be learned unsupervised

Model

- Given
  - string of source words
  - string of target words
  - word alignments
  - morphological and syntactic properties of source words
- Predict
  - morphological properties of target words
- Sequence prediction:

  prediction of morphological properties of earlier words
  affect prediction for subsequent words

Maximum Entropy Markov Models



- Predicting one inflected form at a time (Toutanova et al., 2008)

$$p(\text{form}|\text{stem}, \text{src}) = \prod_{t=1}^{n} p(\text{form}_t|\text{form}_{t-2}, \text{form}_{t-1}, \text{stem}_t, \text{source})$$

- Log-linear model with features

$$p(\text{form}_t|\text{form}_{t-2}, \text{form}_{t-1}, \text{stem}_t, \text{source})$$
$$= \exp \tfrac{1}{Z} \textstyle\sum_i \lambda_i h_i(\text{form}_t, \text{form}_{t-2}, \text{form}_{t-1}, \text{stem}_t, \text{source})$$

- Could also use conditional random fields (Fraser et al., 2012)

Synthetic Phrase Pairs

- Inflection by post-processing is pipelining (bad!)
  - decisions made by translation model cannot be changed
  - but, say, surface form language model may have important evidence

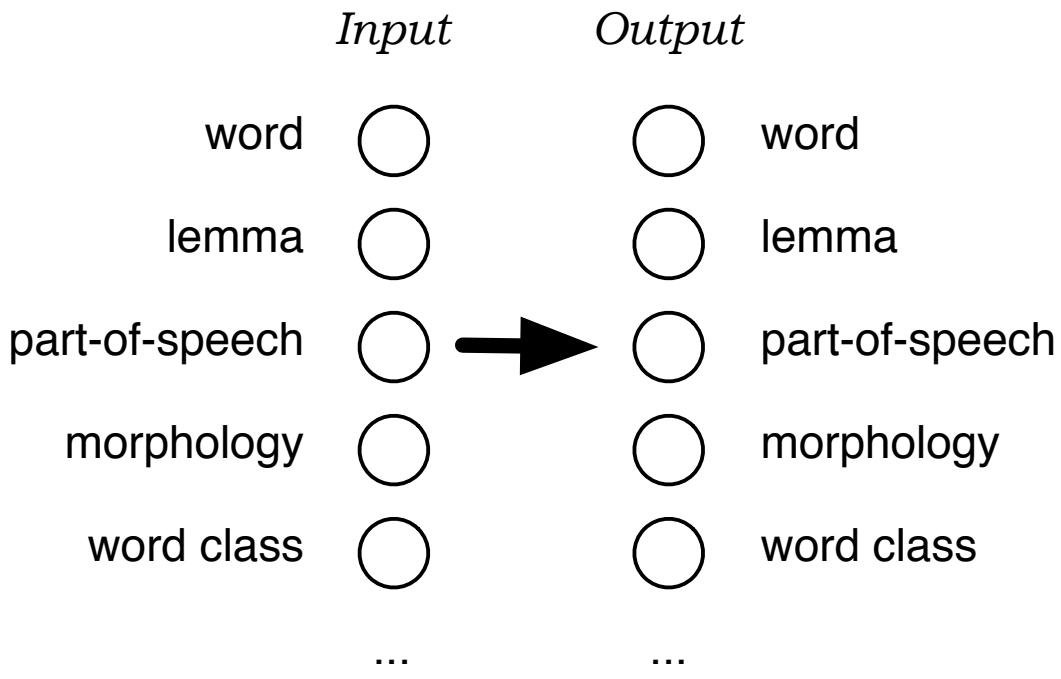
$\Rightarrow$ Extend phrase table (Chahuneau et al., 2013)

  - build inflection model to predict target side inflection
  - use model to predict target side inflection in parallel data
  - add predicted variants as additional phrase pairs

factored models

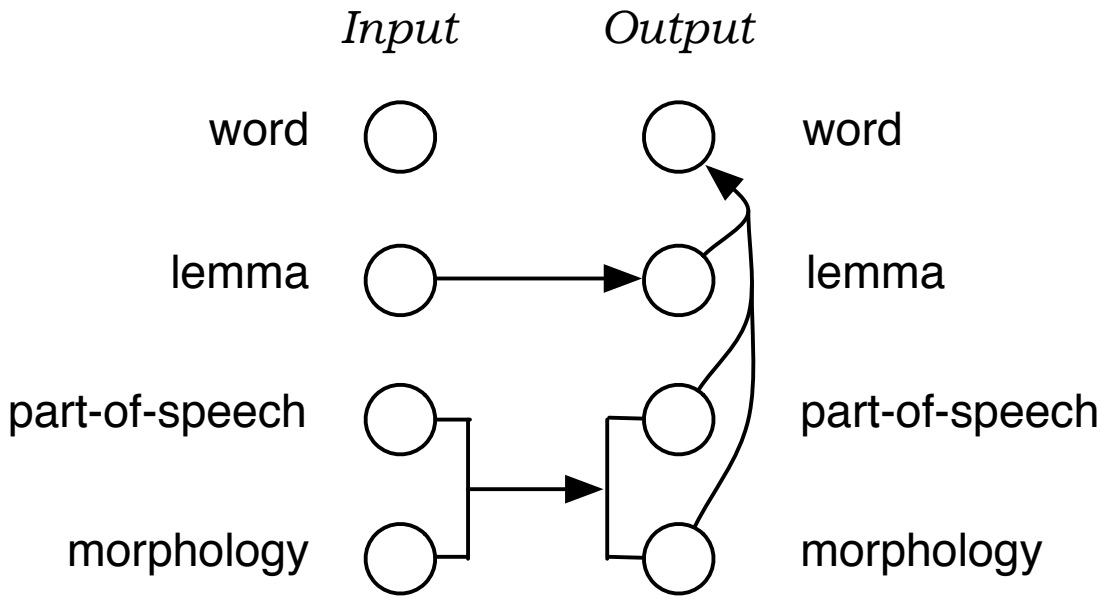
Factored Representation

- Factored representation of words



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

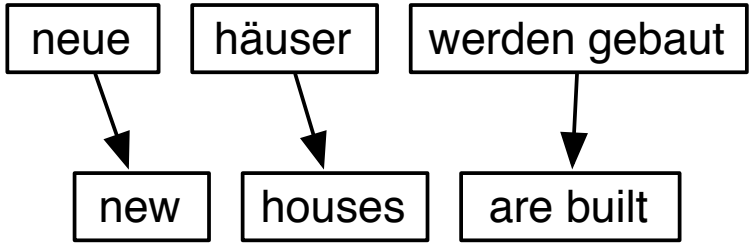
Morphological Analysis and Generation



- Three steps
  - translation of lemmas
  - translation of part-of-speech and morphological information
  - generation of surface forms

Decomposition of Factored Translation

- Traditional phrase-based translation



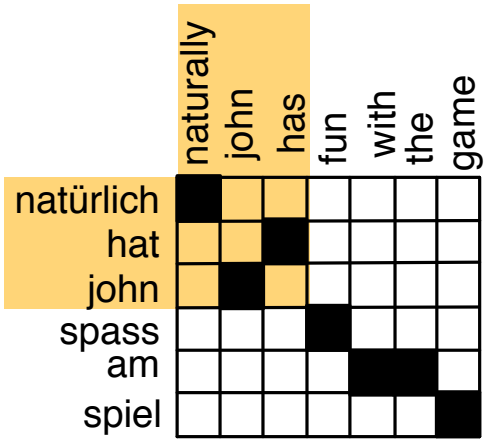
- Decomposition of phrase translation häuser into English
  1. **Translation:** Mapping lemmas
     - haus → house, home, building, shell
  2. **Translation:** Mapping morphology
     - NN|plural-nominative-neutral → NN|plural, NN|singular
  3. **Generation:** Generating surface forms
     - house|NN|plural → houses
     - house|NN|singular → house
     - home|NN|plural → homes
     - ...

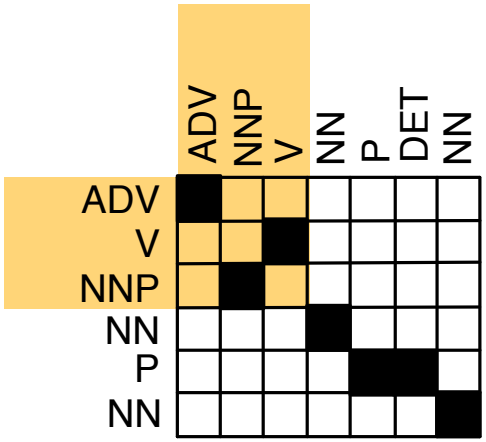
Expansion

Translation Mapping lemmas		**Translation** Mapping morphology		**Generation** Generating surface forms
?|house|?|?		?|house|NN|plural		houses|house|NN|plural
		?|house|NN|singular		house|house|NN|singular
?|home|?|?		?|home|NN|plural		homes|home|NN|plural
	$\Rightarrow$	?|home|NN|singular	$\Rightarrow$	home|home|NN|singular
?|building|?|?		?|building|NN|plural		buildings|building|NN|plural
		?|building|NN|singular		building|building|NN|singular
?|shell|?|?		?|shell|NN|plural		shells|shell|NN|plural
		?|shell|NN|singular		shell|shell|NN|singular

Learning Phrase Translations

- Learning translation step models follows phrase-based model training





natürlich hat john — naturally john has ADV V NNP — ADV NNP V

- Generation models aestimated on the output side only

  → only monolingual data needed

- Features: conditional probabilities

Efficient Decoding

- Factored models create translation options
  - independent of application context

  → pre-compute before decoding

- Expansion may create too many translation options

  → intermediate pruning required

- Fundamental search algorithm does not change

Final Comments

- Need to balance rich surface form translation vs. decomposition
- Parameterization difficult
- Pre- / post-processing schemes $\rightarrow$ pipelining
- Supervised vs. unsupervised morphological analysis

$\Rightarrow$ some general principles learned, but no comprehensive solution yet