Factored Models for Morphology

Philipp Koehn, University of Edinburgh

26 January 2011

Translating between all EU-27 languages

Target Language

	en	bg	de	cs	da	el	es	et	fi	fr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
en	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
bg	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
de	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
cs	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
da	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
el	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
es	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
et	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
fi	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
fr	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
hu	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
it	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
lt	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
lv	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
mt	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
nl	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
pl	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
pt	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
ro	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
sk	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
sl	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
sv	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

(using the Acquis corpus)

[from Koehn et al., 2009]

What Makes Machine Translation Hard?

- Finding explanatory factors for diverging performance of Europarl systems

Explanatory Factor	R^2
Target vocabulary size	**0.388**
Reordering amount	0.384
Language similarity	0.366
Source vocabulary size	**0.045**

[from Birch et al., 2008]

- These factors together explain 75% of the differences in performance

- Similar results in study of Acquis systems [Koehn et al., 2009]

Why Morphology?

Content words:

bite

man

dog

How can we encode their relation?

Word Order (English)

Content words:

bite

man

dog

Defined word order: subject, verb, object

dog bite man

Function Words (Japanese)

Content words:

bite

man

dog

Place marker word after (or before) content word to indicate its role

bite dog subject man object

(a lot like prepositional phrases in English)

Affixes (German, Hebrew, ...)

Content words:

bite

man

dog

Add affix to content word to indicate its role

bite dog-subject man-object

(prepositions may become affixes)

Advantage of Affixes: Freer Word Order

- The following German sentences mean the same:

  Der Mann gibt der Frau das Buch.
  Das Buch gibt der Mann der Frau.
  Der Frau gibt der Mann das Buch.
  Der Mann gibt das Buch der Frau.
  Das Buch gibt der Frau der Mann.
  Der Frau gibt das Buch der Mann.

- Placing of content words allows for nuanced emphasis

Additional Information

- Count (singular/dual/plural)

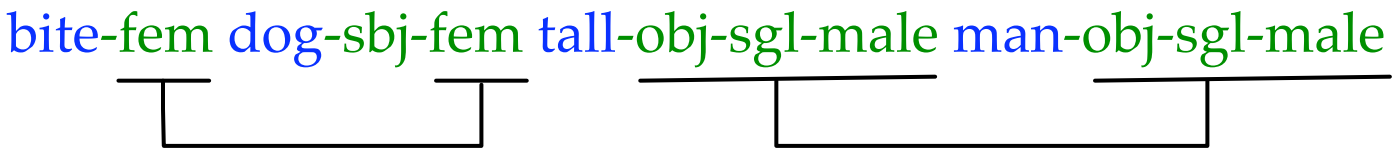
- Gender
  - in English, you more likely refer to your brother or sister than your sibling, but a cousin is gender-neutral
  - in other languages, words like scientist are always gender-specific

- Definiteness
  - indicating reference to a prior mentioned or well-established object
  - in English only in singular determiners the vs. a

$\Rightarrow$ subtly adding additional information

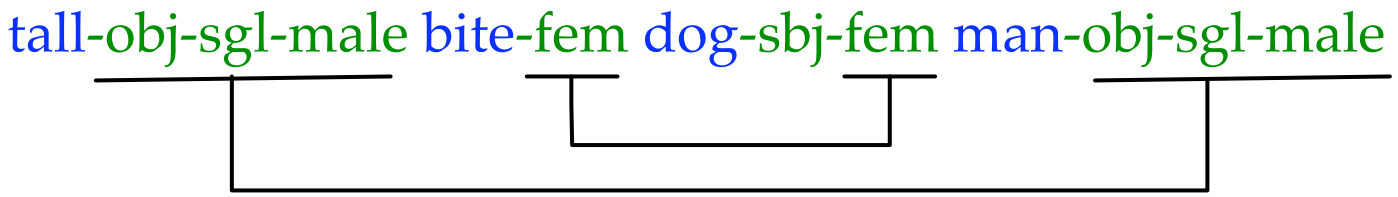
Agreement

- More than one word may contain additional information



related words have to agree
(subject-verb, within noun phrase)

- Even more free word order possible



Derivational Morphology

- Changing part of speech
  - organize → organization, organizer
  - systematic and highly productive
- Generic change of meaning
  - German -chen makes objects small
  - English verb prefixes re- (doing it again) co- (doing it together)
- Compounds (homework, website)

Productivity of Derivational Morphology

- word (614,000,000 hits on Google)
- wordify (8,840 hits on Google)
- wordification (2,350 hits on Google)
- wordificator (8 hits on Google)
- wordifier (2,820 hits on Google)
- wordificationism (1 hit on Google)

> I think you're confusing the term "Democracy" with "Capitalism"; I think you mean "Has Capitalism failed"?
>
> No. It hasn't.
>
> I agree, Hambone; I'm just trying to correct the wordificationism.
>
> Where in the world did you get the word "wordificationism"? Not in the Merriam-Webster dictionary, not in the Thesaurus...

- wordificationist (0 hit on Google, Fall 2010)

Problems for Machine Translation

- Increased vocabulary size → sparse data
- Often added ambiguity (many interpretations per surface form)
- Lack of information in source
- Enforcing long distance agreement
- Transfer between different annotation schemes (free to fixed word order)

Ambiguity: Forms of the German the

Case	Singular			Plural		
	male	fem.	n.	male	fem.	n.
nominative (subject)	der	die	das	die	die	die
genitive (possessive)	des	der	des	der	der	der
dative (indirect object)	dem	der	dem	den	den	den
accusative (direct object)	den	die	das	die	die	die

Not only many different forms,
but each form is highly ambiguous

Major Approaches

- Splitting approach
- Factored approach
- Enriching approach

Splitting

- Source side
  e.g., Arabic–English: split off w- (and) and al- (the) prefixes

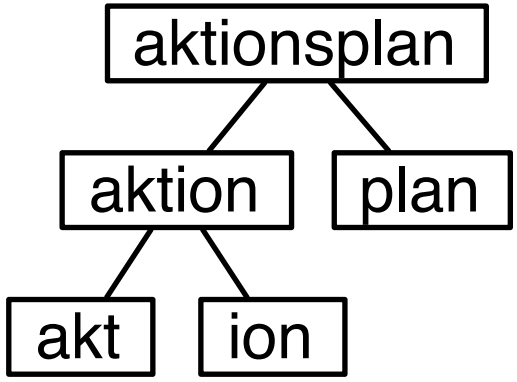
- Target side
  e.g., English–Turkish: see work by Kemal Oflazer's group

- May also drop irrelevant morphemes

- Compound splitting

Compound Splitting

- Compounding common in German, Finnish, Greek, ...
  - increased vocabulary size
  - leads to sparse data problems and unknown words

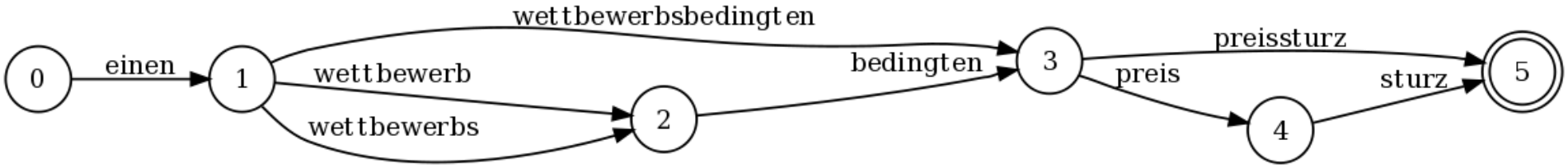


- Frequency based method for compound splitting [Koehn and Knight, 2003]
  - break up, if parts are more frequent than whole
  - geometric mean: $S_{\text{best}} = \text{argmax}_S(\prod_{p_i \in S} \text{count}(p_i))^{\frac{1}{n}}$

Preserving Ambiguity

- Many possible splits

⇒ Encode them in an input lattice [Dyer, 2009]



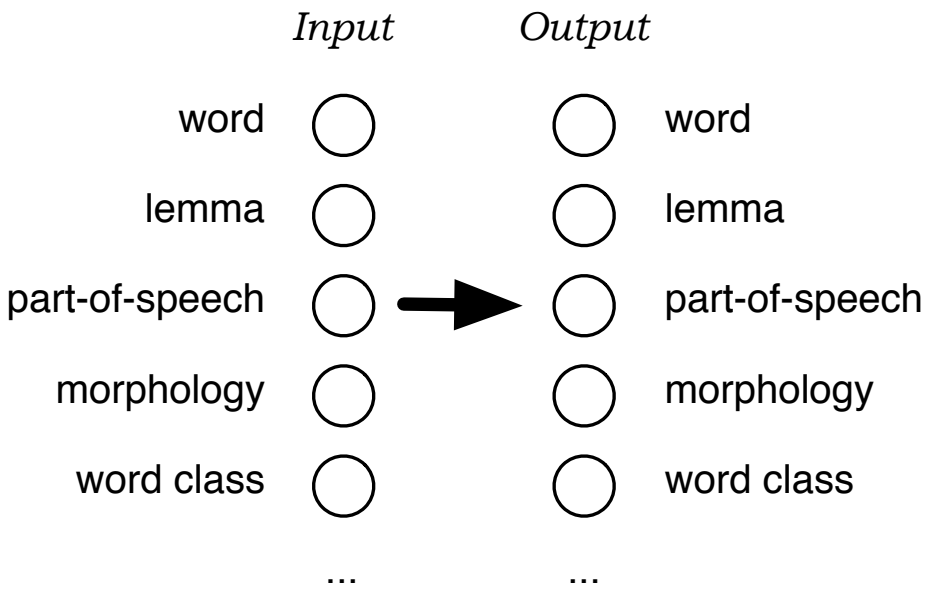
- Decoder chooses optimal source path

Compound Merging

- Work by ... (cite)
- Split compounds on target side of training data
- Indicate splits
  - split token aktion @∼@ plan
  - annotate one part aktion∼ plan
- Merging as deterministic post-processing step

Factored Translation Models

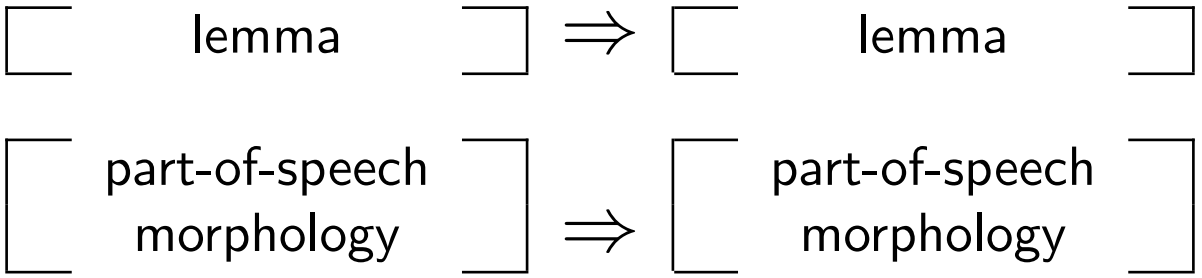
- Factored represention of words [Koehn and Hoang, 2007]



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

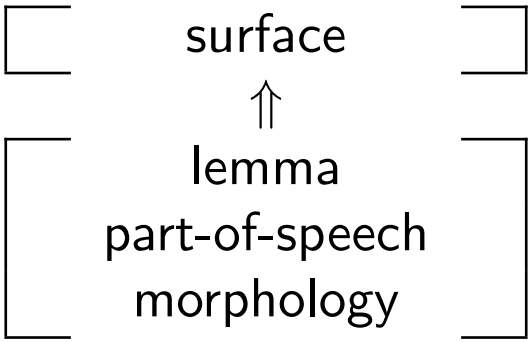
Decomposing Translation: Example

- Translate lemma and syntactic information separately



Decomposing Translation: Example

- Generate surface form on target side



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)

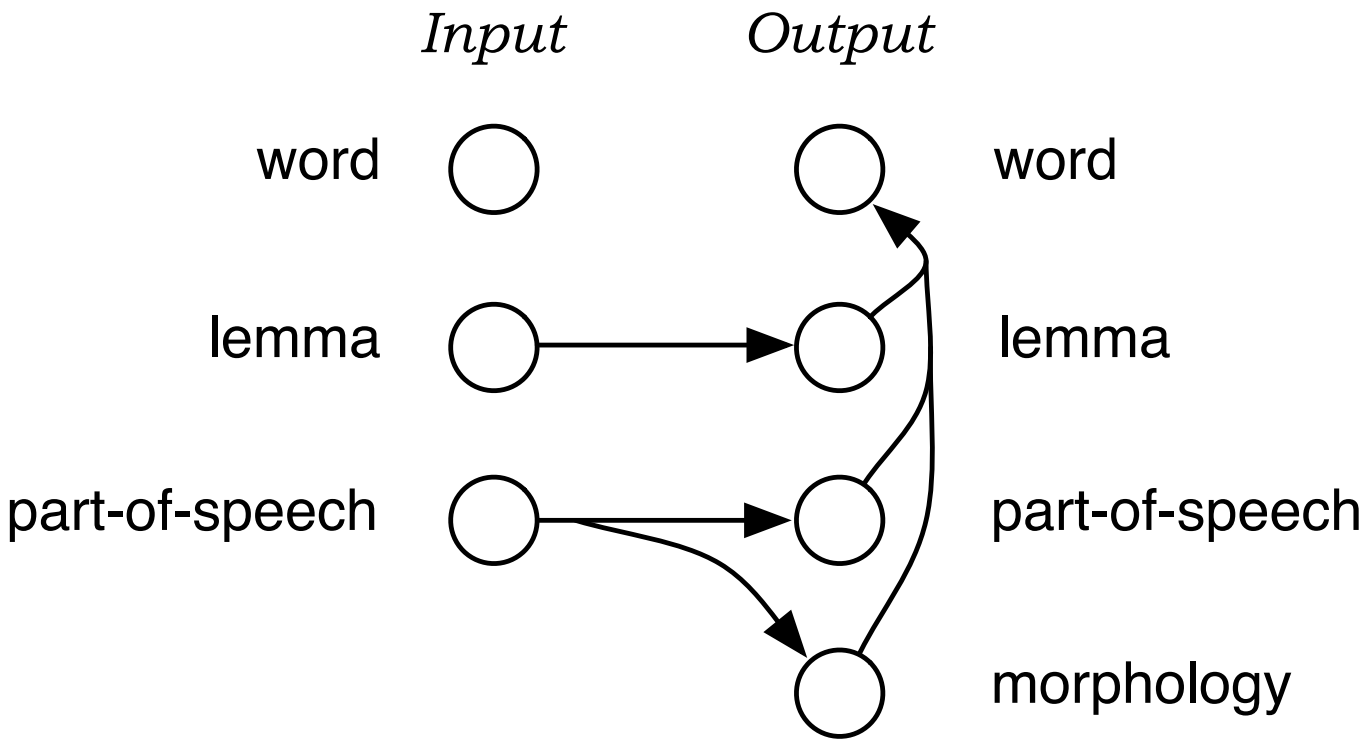
4. Generation step: lemma,part-of-speech ⇒ surface
   (car, car, NN), (cars, car, NNS), (auto, auto, NN), (autos, auto, NNS)

Efficient Factored Model Decoding

- Problem: Explosion of number of translation options
  - originally limited to 20 per input phrase
  - even with simple model, now 1000s of mapping expansions possible

- Solution: Additional pruning of translation options
  - keep only the best expanded translation options
  - current default 50 per input phrase
  - decoding only about 2-3 times slower than with surface model

Morphological generation model



Initial Results

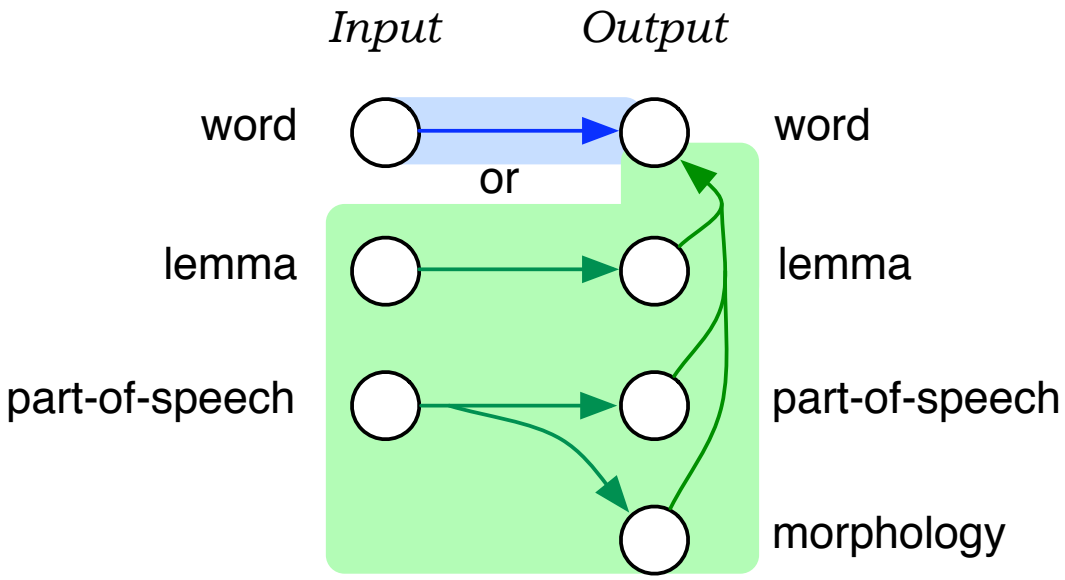
- Results on 1 million word News Commentary corpus (German–English)

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**

Open Issues

- Factored decoding for complex models such as the morph-gen model is broken
- Bad exploration of search space
  (see next slide)
- No proper back-off
  - decomposed model should only be used for unknown and short phrases
  - translation rare phrase could be interpolated (offline)
- Should be addressed — any volunteers?

Bad Exploration of Search Space

- Search for translation options is exhaustive with panic pruning

- Example for unusual part-of-speech patterns:

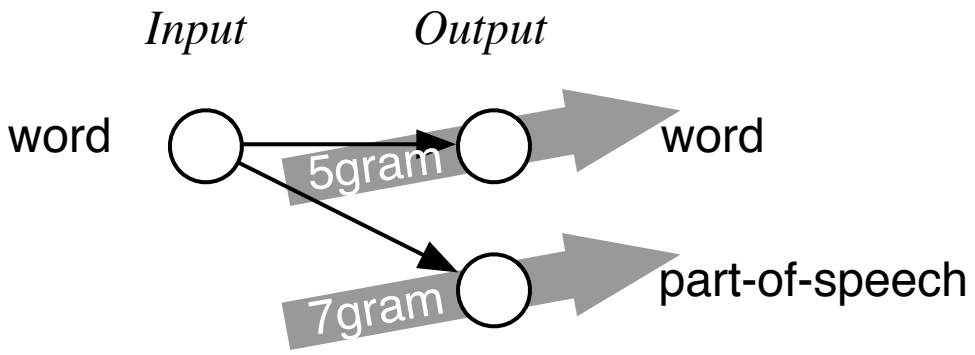
  Staatsanwalt → attorney general , prosecutor , ...
  ↓ ↓ ↓
  NN ADJ NN

  NN → NN , NNS , PN , ADJ , ... , ... , NN ADJ

  - preferred translation attorney/NN general/ADJ
  - unusual part-of-speech mapping NN → NN ADJ may be pruned

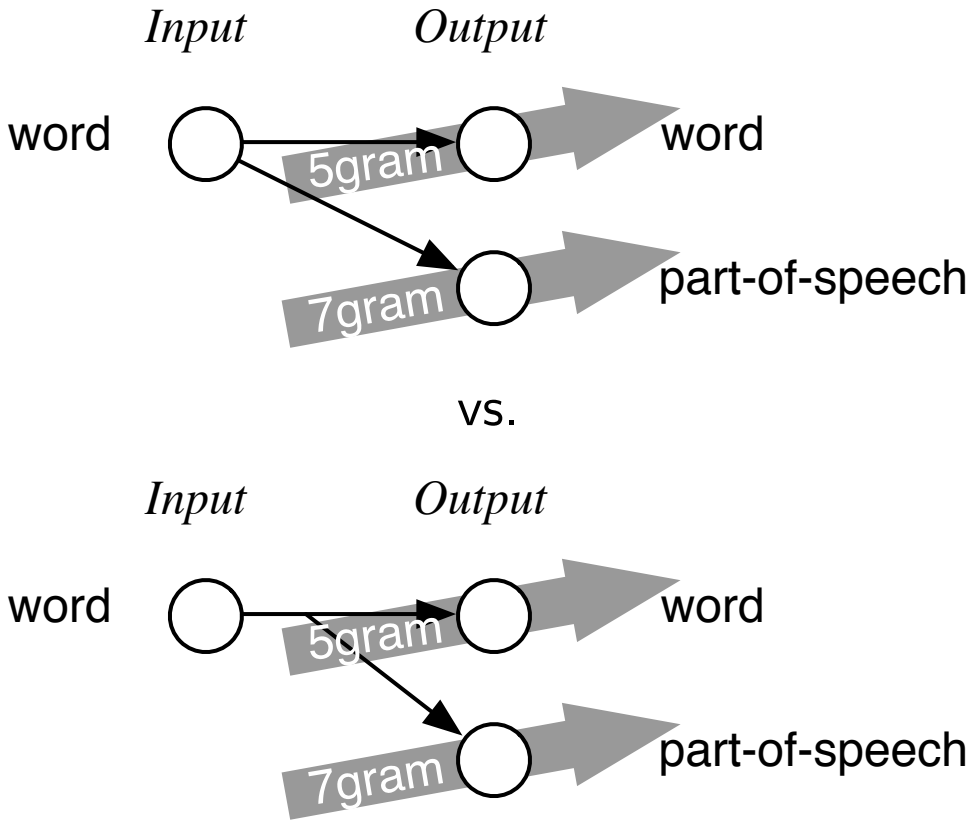
- Also: for long phrases and words with many associated part-of-speech tags, computing all possibilities computationally too expensive

Enriching Output



- Generation of POS tags on the target side
- Use of high order language models over POS (7-gram, 9-gram)
- Motivation: syntactic tags should enforce syntactic sentence structure model not strong enough to support major restructuring

Decomposed vs. Joint



Better: generating both factors in same translation step

Morphological Tags

die	hellen	Sterne	erleuchten	das	schwarze	Himmel
(the)	(bright)	(stars)	(illuminate)	(the)	(black)	(sky)
fem	fem	fem	-	neutral	neutral	male
plural	plural	plural	plural	sgl.	sgl.	sgl
nom.	nom.	nom.	-	acc.	acc.	acc.

- Violation of noun phrase agreement in gender
  - das schwarze and schwarze Himmel are perfectly fine bigrams
  - but: das schwarze Himmel is not
- If relevant n-grams does not occur in the corpus, a lexical n-gram model would *fail to detect* this mistake
- Morphological sequence model: $p(\text{N-male}|\text{J-male}) > p(\text{N-male}|\text{J-neutral})$

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 ...

BLEU Results

Systems for WMT10

Language Pair	Baseline	Factored
Spanish-English	26.03	26.20 (+0.17)
French-English	25.92	26.13 (+0.21)
German-English	19.51	21.09 (+0.24)
Czech-English	21.19	21.33 (+0.14)
English-Spanish	24.65	24.37 (−0.28)
English-French	24.70	24.74 (+0.04)
English-German (POS)	14.81	15.03 (+0.22)
English-German (morph)	14.81	15.28 (+0.47)

Insufficient Input

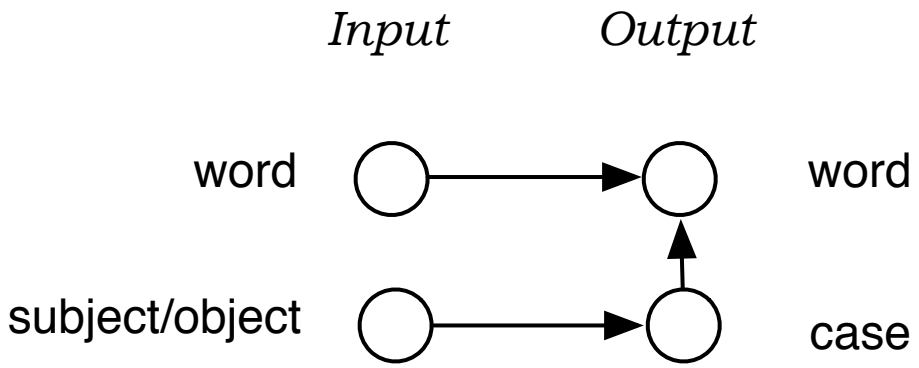
- Examples

$$\text{Habla español.} = \text{He/she speaks Spanish.}$$

$$\text{His cousin is friendly.} = \begin{cases} \text{Sein Vetter} & \text{(male)} \\ \text{Seine Base} & \text{(female)} \end{cases} \text{ ist freundlich.}$$

- Occurs frequently when output language is morphologically richer
- May require document context for resolution

Case Information for English–Greek



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

Results English-Greek

System	devtest	test07
baseline	18.13	18.05
enriched	18.21 (+0.08)	18.20 (+0.15)

- Improvement in verb inflection

System	Verb count	Errors	Missing
baseline	311	19.0%	7.4%
enriched	294	5.4%	2.7%

- Improvement in noun phrase inflection

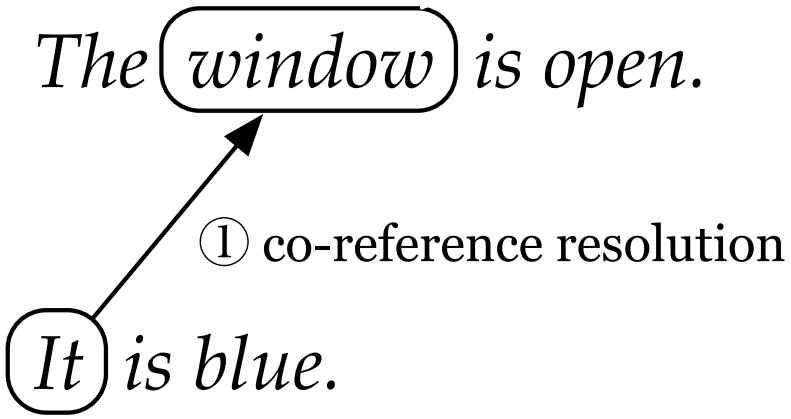
System	NPs	Errors	Missing
baseline	247	8.1%	3.2%
enriched	239	5.0%	5.0%

Pronoun Translation

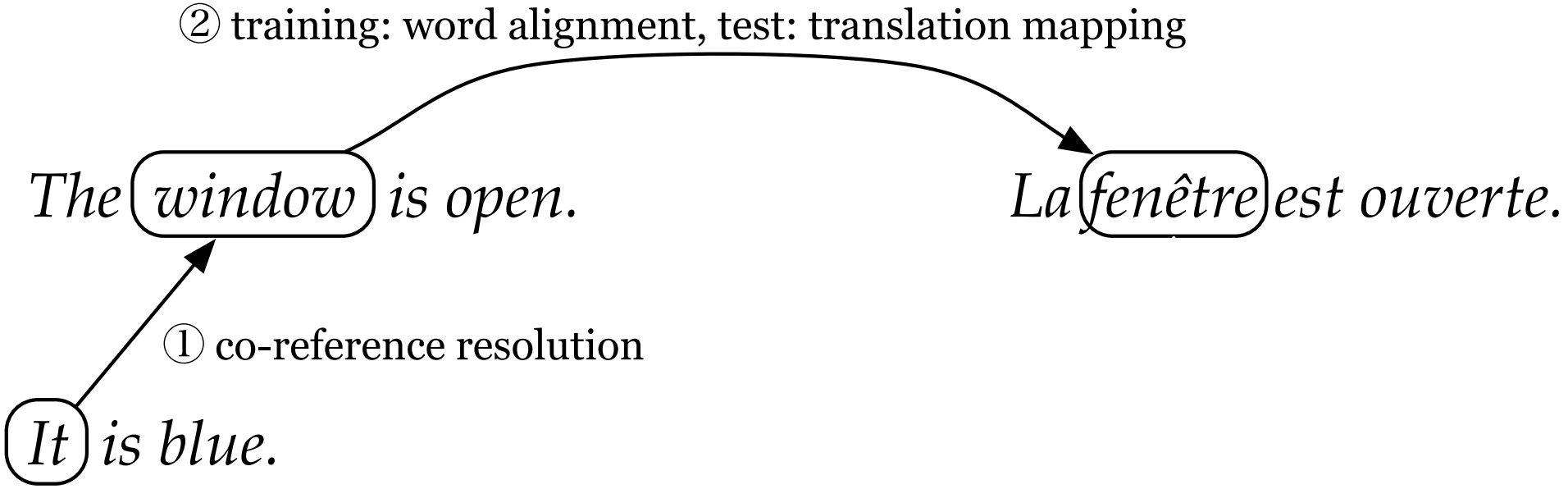
The English it receives a grammatical gender in translation.

The window is open.	La fenêtre est ouverte.	
It is blue.	**Elle** est bleue.	CORRECT
The window is open.	La fenêtre est ouverte.	
It is black.	**Il** est noir.	WRONG
The oven is open.	Le four est ouverte.	
It is new.	**Elle** est neuve.	WRONG
The door is open.	La porte est ouverte. .	
It is new.	**Elle** est neuve.	CORRECT

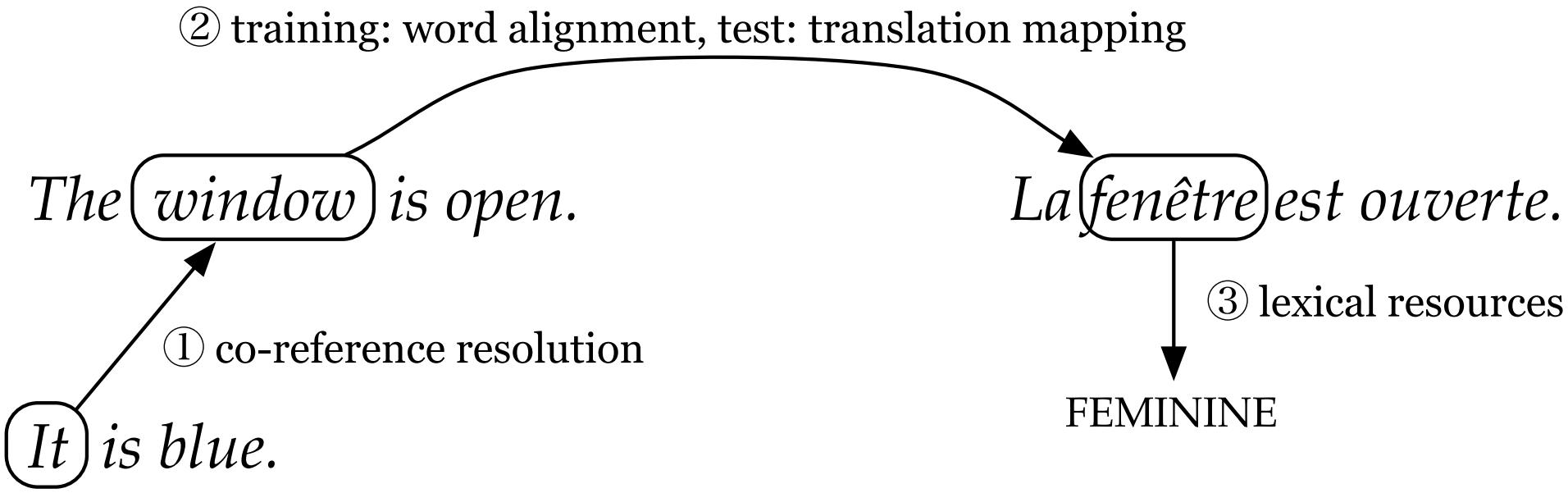
Co-Reference Resolution



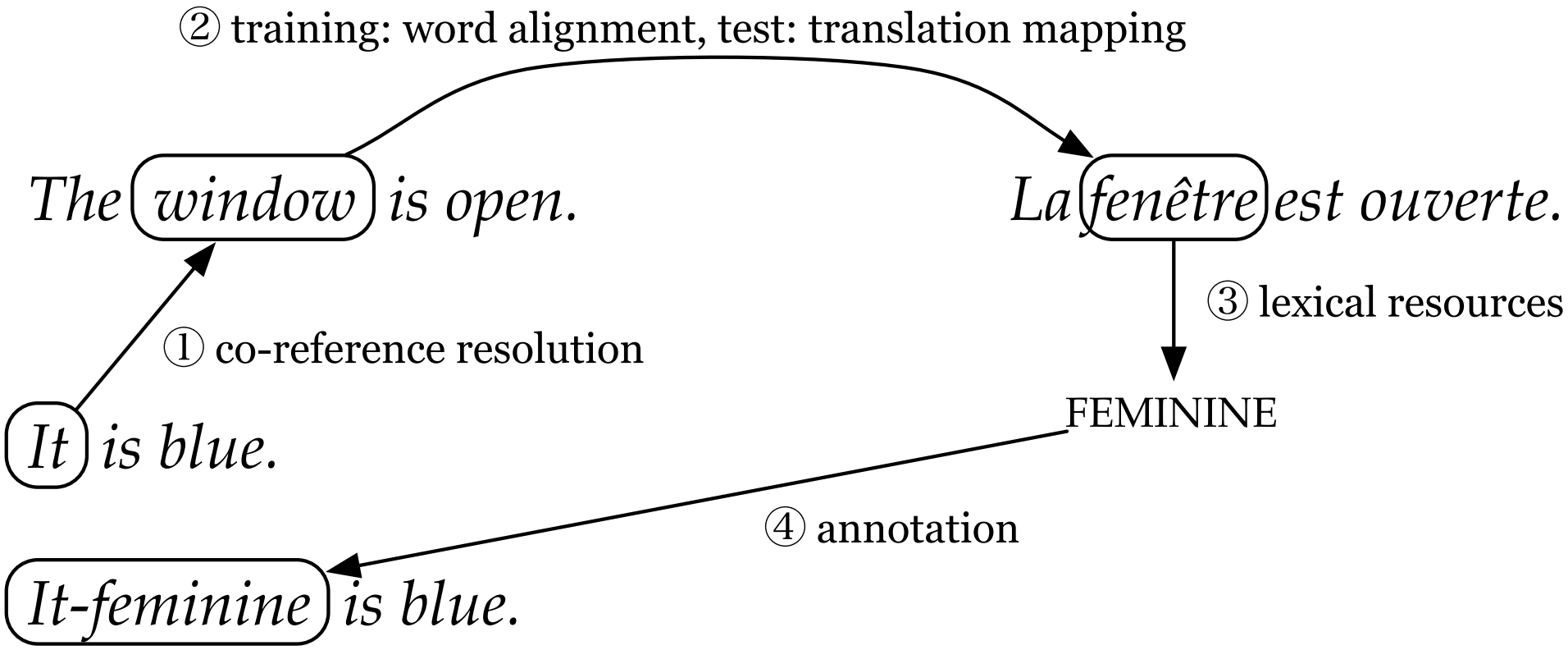
Word Alignment



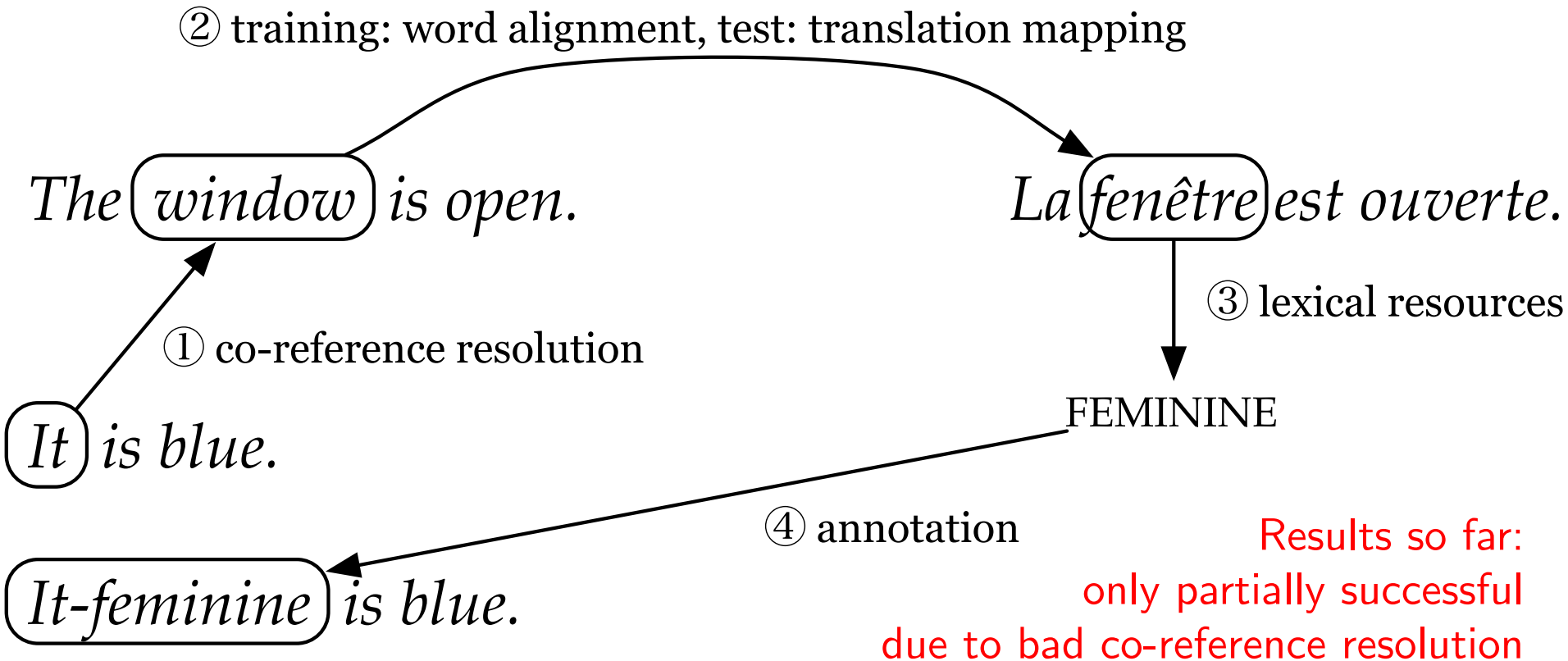
Gender Detection



Enriching Source



Enriching Source



Results so far:
only partially successful
due to bad co-reference resolution

Problems for Machine Translation

- Increased vocabulary size → sparse data
  splitting / factored approach
- Often added ambiguity (many interpretations per surface form)
  factored approach
- Lack of information in source
  enriching approach
- Enforcing long distance agreement
  unsolved
- Transfer between different annotation schemes (free to fixed word order)
  unsolved

Syntax to the Rescue

- Syntactic structure better at enforcing agreement
- Reordering driven by morphology (tree-to-string)
  - S → NP-acc$_1$ frißt NP-nom$_2$; X$_2$ eats X$_1$
- Local agreement within noun phrases
  - NP-dat → the X man ; dem ADJ-male-dat-sgl-def Manne
- Long-range agreement within clauses
  - S → X$_1$ eats X$_2$; NP-nom$_1$ frißt NP-acc$_2$

Problems

- Adding ambiguity
  - DET-male-nom-sgl → the ; der
  - DET-fem-gen-sgl → the ; der
  - DET-neutral-gen-pl → the ; der

  → spurious ambiguity during decoding

- Increasing number of non-terminals and rules
  - bigger models
  - more complex decoding
  - overly specific rules are less applicable

Synchronous Unification Grammar

- Ongoing work...

- Principles
  - separate translation rules and constraints
  - overcome interpretation ambiguity by maintaining sets in hypotheses
  - overcome sparsity of forms by generation step

Questions?