Morphological Knowledge in Statistical Machine Translation

Kristina Toutanova
Microsoft Research

Morphology

- **Morpheme** – the smallest unit of language that carries information about meaning or function
- **Word** – the smallest **free** form in language (that does not have to occur in a fixed position relative to neighboring elements)
- Typical MT systems model the translation process based on words

Language differences in word granularity

- Some words in one language correspond to bound affix morphemes in another

wa+li+al+maktaba+āt   وللمكتبات

and for the library+pl

Language differences in word granularity

- Languages mark different amounts of grammatical information using inflection

	Russian	English
Noun gender	3	1
Noun case	6	1
Adjective gender, number, case	3 x 2 x 6	1
Verbs person, number	3 x 2	2

Language differences in word granularity

- Languages exhibit different amount of compounding

elin-keino-tulo-vero-laki (life's means income tax law)

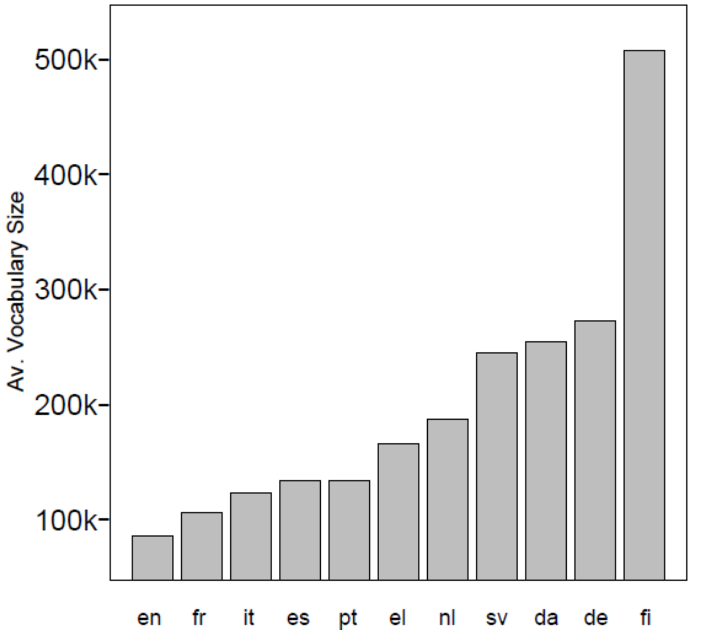
income tax law

Finnish-English example

Challenges for Machine Translation

- Standard word alignment and translation models work best when the mapping between words in largely one-to-one
  - Breaks for languages with different word granularity
- Rich morphology leads to sparsity
  - Translation rules with less coverage
  - Poor estimation of translation probabilities
- Rich systems of grammatical agreement lead to insufficiency of standard language models
  - Need longer context from source and target to predict correct target forms

Impact of Morphology on Vocabulary Size



tietä+isi+mme

know+would+we

Creutz et al. 2005

Slide from Costa-jussà
& Quirk NAACL 2013 tutorial

Opportunities and Challenges in Modeling Morphology for MT

- Achieve better source-target alignment
- Expand translation rule coverage
- Generalize statistics by parameter sharing among morphologically related words

- Morphological analysis is not observed
- Morphological analyzers are hard to obtain for many languages
- Can make incorrect predictions based on less specific evidence

Use of Morphogical Knoweldge

- Alignment – basic units and correspondence among them

- Translation rules
  - Defining the set of options

- Modeling
  - Morphology-related models for scoring candidates

Outline

- Unsupervised Induction of Morphology
- Pre-processing to reduce language divergence [Alignment, Rules, Modeling]
- Factored translation models [Rules, Modeling]
- Models for generation of complex morphology [Rules, Modeling]
- Scoring models for rich target morphology [Modeling]

Unsupervised Induction of Morphology

Unsupervised Morphology

- For many languages, no high-quality analyzers available.
- Even when we have supervised analysis, it is not clear what is the optimal segmentation for a given language pair and data size [Goldwater & McKlosky, Habash & Sadat 2006].
- Can we have an unsupervised morphological analyzer determining the optimal units?

Unsupervised Morphological Segmentation

- Monolingual morphological segmentation
- Bilingual morphological segmentation
- Supervised versus unsupervised morphology for translation performance

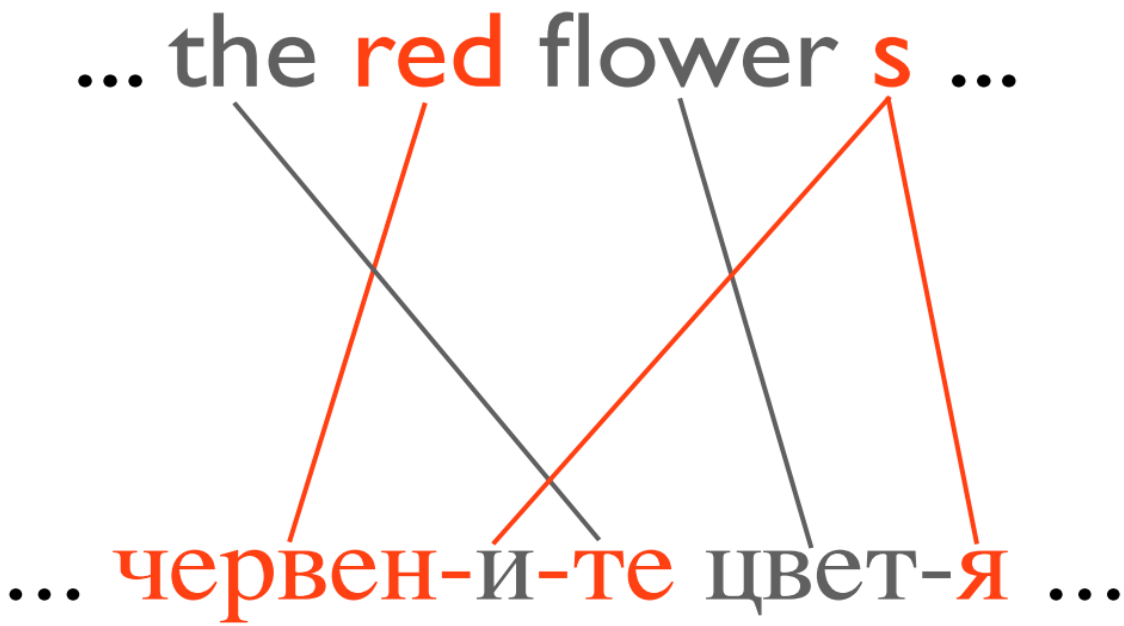
Monolingual morphological segmentation

walialmaktabaāt →wa+li+al+maktaba+āt

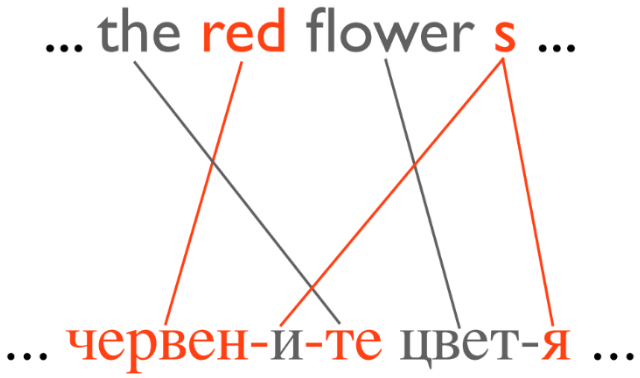
- Morfessor [Creutz et al, 2005]
  - Categories-MAP uses an HMM statistical model with prefix,stem, and suffix states
  - Publicly available
- [Poon et al 2009],[Naradowsky & Toutanova 2011]
  - Feature-rich models, higher accuracy on Arabic and Hebrew
- Active area of research

Bilingual Morphological Segmentation

- Given source segmentation into words or morphemes, segment and align the target to the source
- Target segmentation may vary to match source units



Models using standard IBM-1 and HMM alignment modes [Chung & Gildea 09]



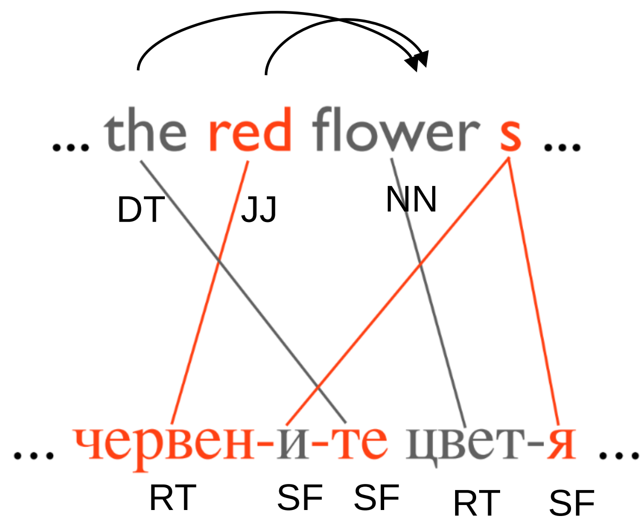
$$P(m_1, m_2, m_3, m_4, m_5, A \mid e) = \prod_{i=1}^{5} \tfrac{1}{5}\, p(m_i \mid e_{a_i})\varphi(|\, m_i\,|)$$

Use our standard alignment models except now the target segmentation is a hidden variable.

Inference is fast using a dynamic program like the one for semi-markov CRFs.

Improvement in MT over monolingual segmentation.

Model using richer morpho-syntactic information [Naradowsky & Toutanova 2011]



- Model based on HMM word alignment model
- Leverage source morpho-syntactic information
- Generate latent morpheme state – prefix, root, stem
- Distortion model aware of source and target morpho-syntactic context

Supervised versus Unsupervised Morphology

- [Chung & Gildea 09] on Korean-English
  - supervised vs unsupervised BLEU 7.27 vs 7.46
- [Chahuneau et al 13] on English-Russian
  - word baseline 15.7
  - supervised vs unsupervised BLEU 16.7 vs 16.2
- [Stallard et al 12] on Arabic-English 35mln train
  - word baseline 43.45
  - supervised vs unsupervised BLEU 45.64 vs 45.84

Outline

- Unsupervised Induction of Morphology
- **Pre-processing to reduce language divergence** [Alignment, Rules, Modeling]
- Factored translation models [Rules, Modeling]
- Models for generation of complex morphology [Rules, Modeling]
- Scoring models for rich target morphology [Modeling]

Preprocessing to Reduce Language Divergence

[Alignment, Rules, Modeling]

Preprocessing to Reduce Language Divergence

- Transform source tokens but leave target tokens alone (or enrich target words)
- From highly inflect to less inflected language
  - Remove some information from source
  - Convert bound morphemes to free
- From less inflected to more inflected language
  - Enrich the source words using syntactic information
  - Covert free morphemes to bound

Preprocessing for high → low [Goldwater & McClosky 2005]

- For several morphological features, try splitting them off as pseudowords, dropping them, or appending to the lemma

Words:	`Pro někoho by její provedení mělo smysl .`
Lemmas:	`pro někdo být jeho provedení mít smysl .`
Lemmas+Pseudowords:	`pro někdo být PER_3 jeho provedení mít PER_X smysl .`
Modified Lemmas:	`pro někdo být+PER_3 jeho provedení mít+PER_X smysl .`

It would make sense for somebody to do it

- Optimal scheme: lemmatize words, treat person and negation as pseudo-words, append number and tense
- Gain 6 BLEU points using 20K sent training data

Preprocessing for high → low [Habash & Sadat 2006]

Arabic Morphology

[CONJ+ [PART+ [Al+ BASE +PRON]]]

TOK	
ST	Splitting off punctuation and numbers
D1	Declitization (w+, f+)
D2	Declitization (D1+ l+, k+, b+, s+)
D3	Declitization (D1,D2, Al+)
MR	Stem + affixival morphemes
EN	English-like

Slide from Costa-jussà
& Quirk NAACL 2013 tutorial

Preprocessing for high → low
[Habash & Sadat 2006]

Input	wsynhY	Alr}ys	jwlth	bzyArp	AlY	trkyA.	
Gloss	and will fi nish	the president	tour his	with visit	to	Turkey	.
English	The president will fi nish his tour with a visit to Turkey.						
ST	wsynhY	Alr}ys	jwlth	bzyArp	AlY	trkyA	.
D1	w+ synhy	Alr}ys	jwlth	bzyArp	<lY	trkyA	.
D2	w+ s+ ynhy	Alr}ys	jwlth	b+ zyArp	<lY	trkyA	.
D3	w+ s+ ynhy	Al+ r}ys	jwlp +P$_{3MS}$	b+ zyArp	<lY	trkyA	.
MR	w+ s+ y+ nhy	Al+ r}ys	jwl +p +h	b+ zyAr +p	<lY	trkyA	.
EN	w+ s+ >nhY$_{VBP}$ +S$_{3MS}$	Al+ r}ys$_{NN}$	jwlp$_{NN}$ +P$_{3MS}$	b+ zyArp$_{NN}$	<lY$_{IN}$	trkyA$_{NNP}$	.

The optimal segmentation dependent on training set size.

For a training set of 50,000 words: EN best,
gaining 7 to 8 BLEU points.

For training set of 5 million words: D2 best, gaining 1 to 2 BLEU points.

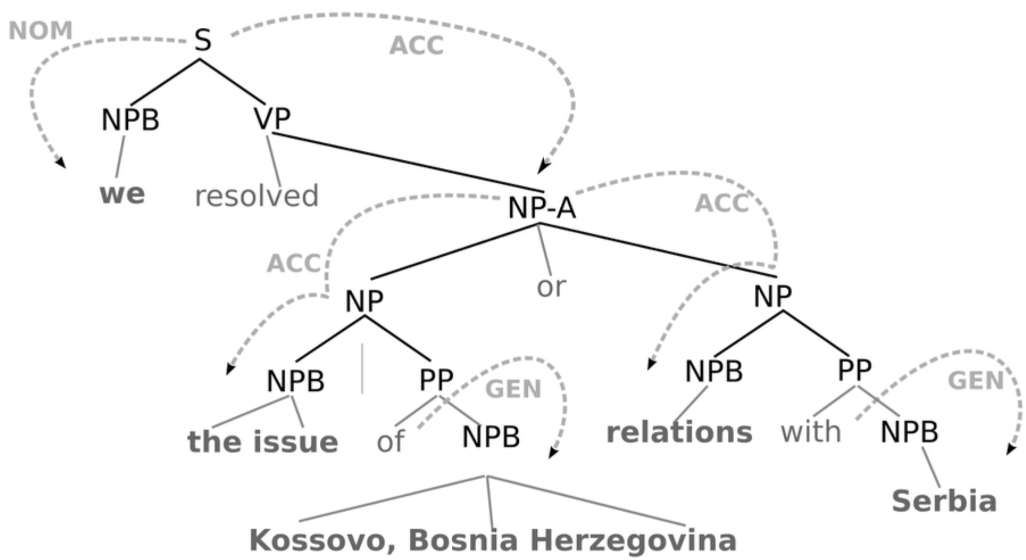
Preprocessing for low → high [Avramidis & Koehn 08]

In English-to-Greek translation we need to predict case for nouns and person for verbs.

- EN: The president, after reading the press review and the announcements, left his office
- GR: The president[nominative], after reading[3S] the press review[Accusative,S] and the announcements[***<u>Accusative</u>***,p], left[3S] his office[Accusative,S]

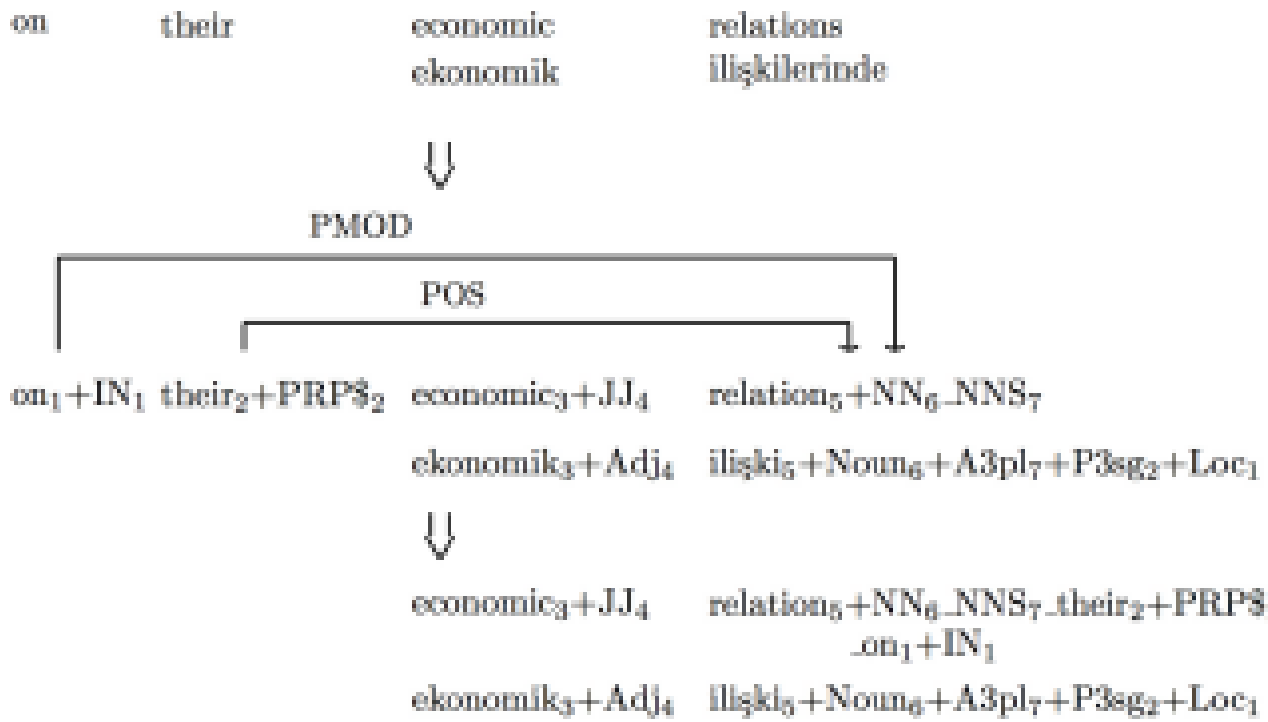
Preprocessing for low → high [Avramidis & Koehn 08]

- Annotate English source with rules looking at syntactic tree for noun case and verb person



- Results: small improvement in BLEU but large error reduction in noun and verb inflection errors

Preprocessing for low → high
[Yeniterzi & Oflazer 2010]



25 rules specifying how to convert function words in English into Turkish morphemes

5 BLEU points improvement for a 50K training corpus (in combination with factored translation models)

Outline

- Unsupervised Induction of Morphology
- Pre-processing to reduce language divergence [Alignment, Rules, Modeling]
- **Factored translation models** [Rules, Modeling]
- Models for generation of complex morphology [Rules, Modeling]
- Scoring models for rich target morphology [Modeling]

Factored Translation Models

[Rules, Modeling]

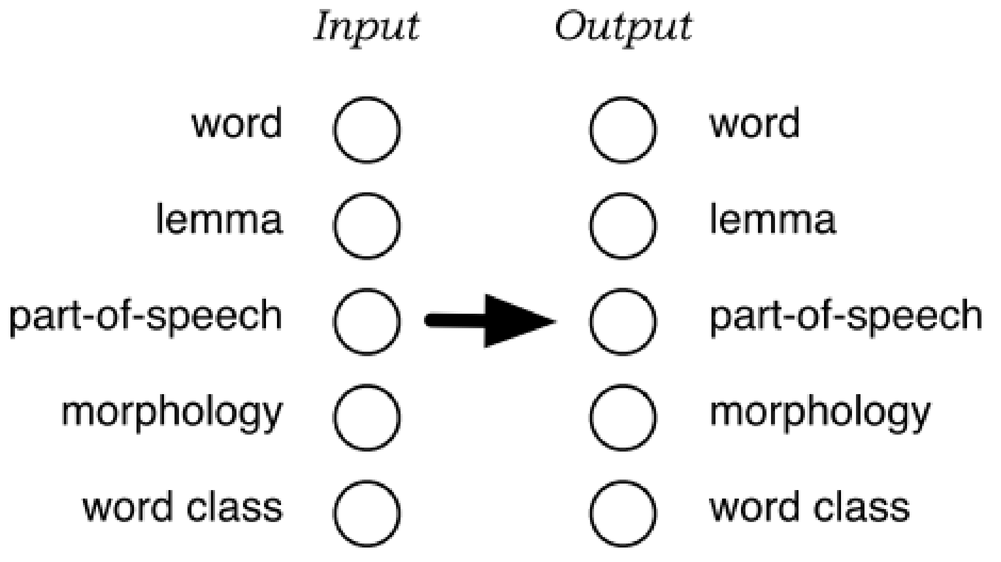
Factored Translation Models [Koehn & Hoang 2007]

- The phrase-based model sees every word as a sequence of factors (indicating morphological, syntactic, or semantic information)

(word) =⇒ (word, lemma, PoS, morphology, ...)

- The system can now generalize over factors in addition to words

Factored Translation Models



Can define target phrase generation in a factored way

Can use richer information for modeling

Example: Decomposing translation

- Translate the lemmas and syntactic features separately

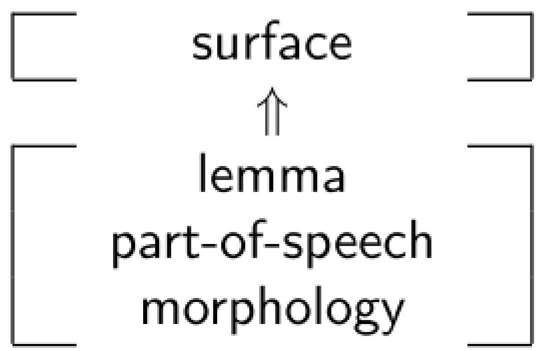
$$\begin{bmatrix}\text{lemma}\end{bmatrix} \Rightarrow \begin{bmatrix}\text{lemma}\end{bmatrix}$$

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

Slide by Philipp Koehn

Example: Decomposing translation

- Generate surface forms on target side



Slide by Philipp Koehn

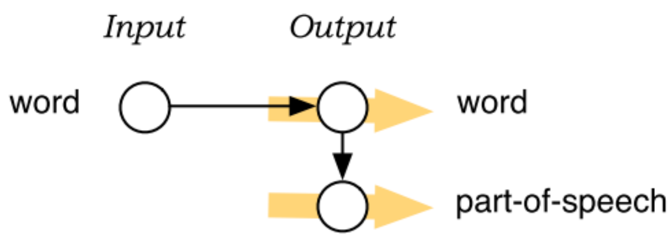
Example: Decomposing Translation

Input: *(Autos, Auto, NNS)*

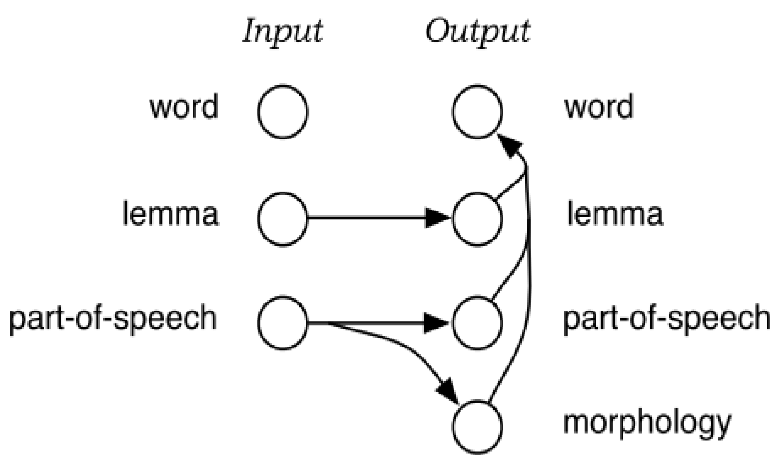
1. Translation step: lemma ⇒ lemma
   (?, car, ?), (?, auto, ?)
2. Generation step: lemma ⇒ part-of-speech
   (?, car, NN), (?, car, NNS), (?, auto, NN), (?, auto, NNS)
3. Translation step: part-of-speech ⇒ part-of-speech
   (?, car, NN), (?, car, NNS), (?, auto, NNP), (?, auto, NNS)
4. Generation step: lemma,part-of-speech ⇒ surface
   (car, car, NN), (cars, car, NNS), (auto, auto, NN), (autos, auto, NNS)

Slide by Philipp Koehn

Results with Factored Translation Models



Enriching output and using high-order LM over POS: gains 1 to 2 BLEU points using small training set [Koehn & Hoang 07]



Generation through lemma and morphology: gain 19.05 →19.47 when using alternative decoding for a small German-English system

Outline

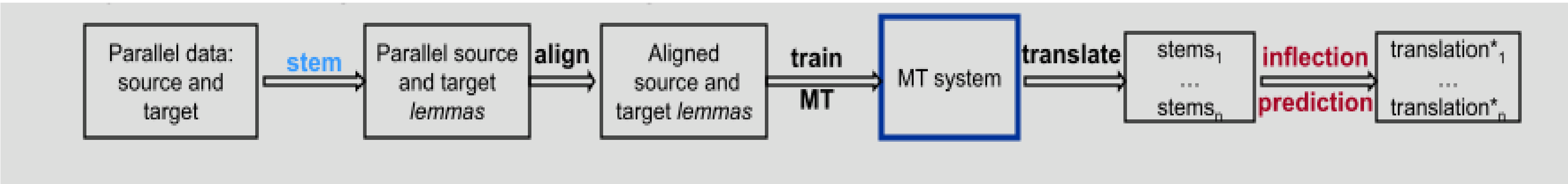
- Unsupervised Induction of Morphology
- Pre-processing to reduce language divergence [Alignment, Rules, Modeling]
- Factored translation models [Rules, Modeling]
- **Models for generation of complex morphology** [Rules, Modeling]
- Scoring models for rich target morphology [Modeling]

Models for Generation of Complex Morphology

[Rules, Modeling]

Models for generation of complex morphology

- Factors the translation process into translation from source to target stem sequence and a separate inflection prediction component



The Problem(s)

Please select one of the values

The Problem(s)

Please select one of the values

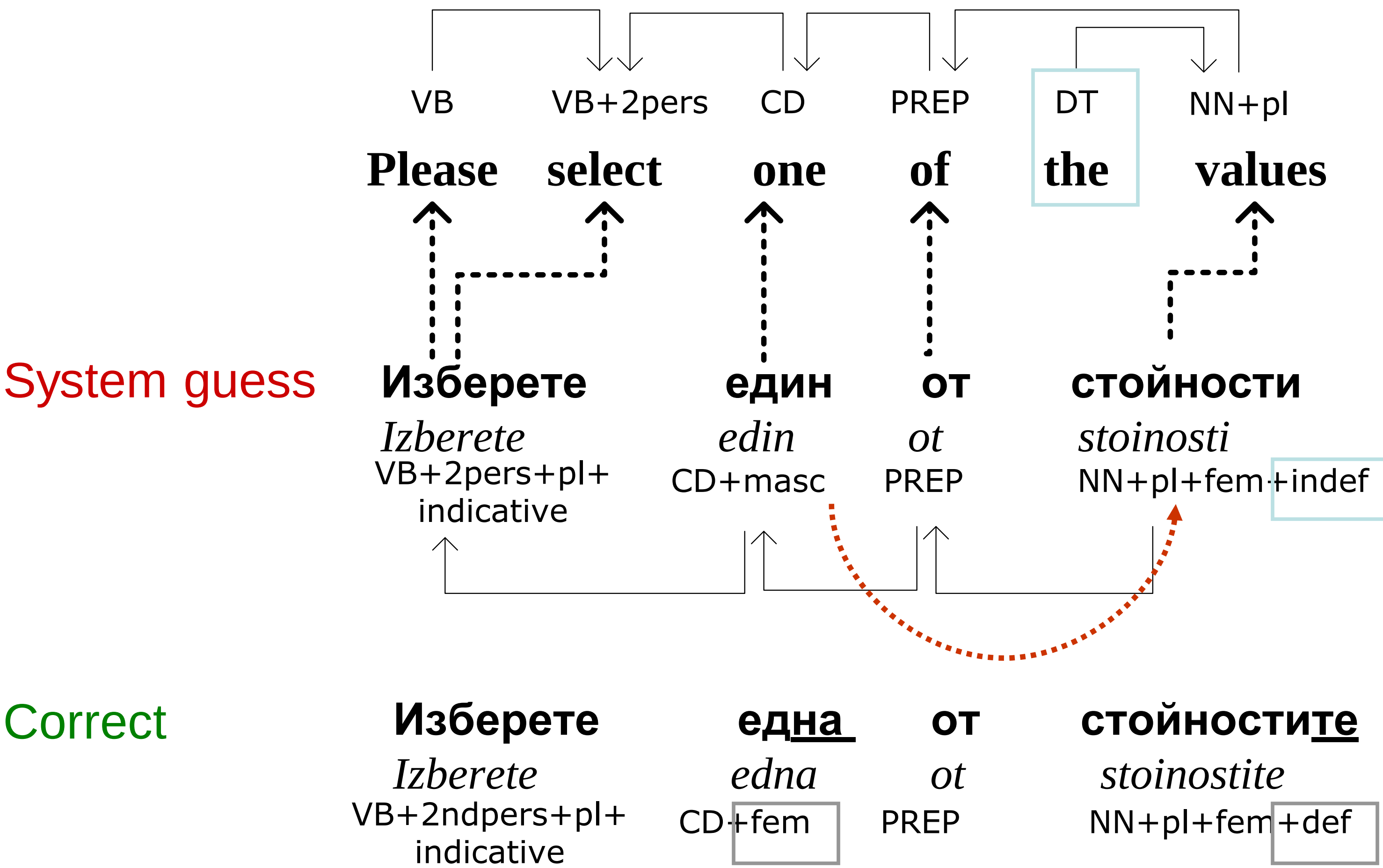
System guess **Изберете един от стойности**

Izberete edin ot stoinosti

The Problem(s)

	Please	**select**	**one**	**of**	**the**	**values**
System guess	**Изберете** *Izberete*		**един** *edin*	**от** *ot*		**стойности** *stoinosti*
Correct	**Изберете** *Izberete*		**една** *edna*	**от** *ot*		**стойностите** *stoinostite*

The Problem(s)



Morphology Prediction

- Morphology generation as classification: Classify each stem into an inflected form

System guess	**eliminare**	**un**	**vincolo**	**di**	**chiave**	**primario**
Possible inflections	eliminare elimino elimini eliminiamo …	un una	vincolo vincoli	di del dei della …	chiave chiavi	primario primaria primari primarie

Morphology Prediction

- Morphology generation as classification: Classify each stem into an inflected form

System guess	**eliminare**	**un**	**vincolo**	**di**	**chiave**	**primario**
Possible inflections	**eliminare**	**un**	**vincolo**	**di**	**chiave**	primario
	elimino	una	vincoli	del	chiavi	**primaria**
	elimini			dei		primari
	eliminiamo			della		primarie
	…			…		

Generation of Complex Morphology
[Minkov et al 07, Toutanova et al 08]

- Morphology
  - Russian, Arabic
  - Lexicon operations
- The task of inflection prediction
- A log-linear model
- Features
  - Lexical, Syntax and Morphology
- Evaluation

Russian Morphology

- 3 genders, 2 numbers, 6 cases
- Nouns have gender, and inflect for number and case
- Adjectives agree with nouns in number, gender, and case; have short and long forms;
- Verbs agree with Subject person and number (past tense agrees with gender and number) –not many variations though in our domain

Я	люблю	мой	синий	карандаш.
I	*love*	*my*	*blue*	*pencil*
Pers1 Sing	Pers1 Sing	Acc Masc Sing	Acc Masc Sing	Acc Masc Sing

Arabic morphology

- Arabic: inflection + clitics
  - Prefixes: Conj/Prep/Compl/Def (in strict order)
  - Suffixes: Object/Possessive pronouns

(from Bar-Haim et al)

وللمكتبات

/walilmaktabāt/

و+ل+ال+مكتبة+ات

wa+li+al+maktaba+āt

and+for+the+library+plural

and for the libraries

فقلناها

/faqulnāhā/

ف+ قال+ نا+ ها

fa+qul+na+hā

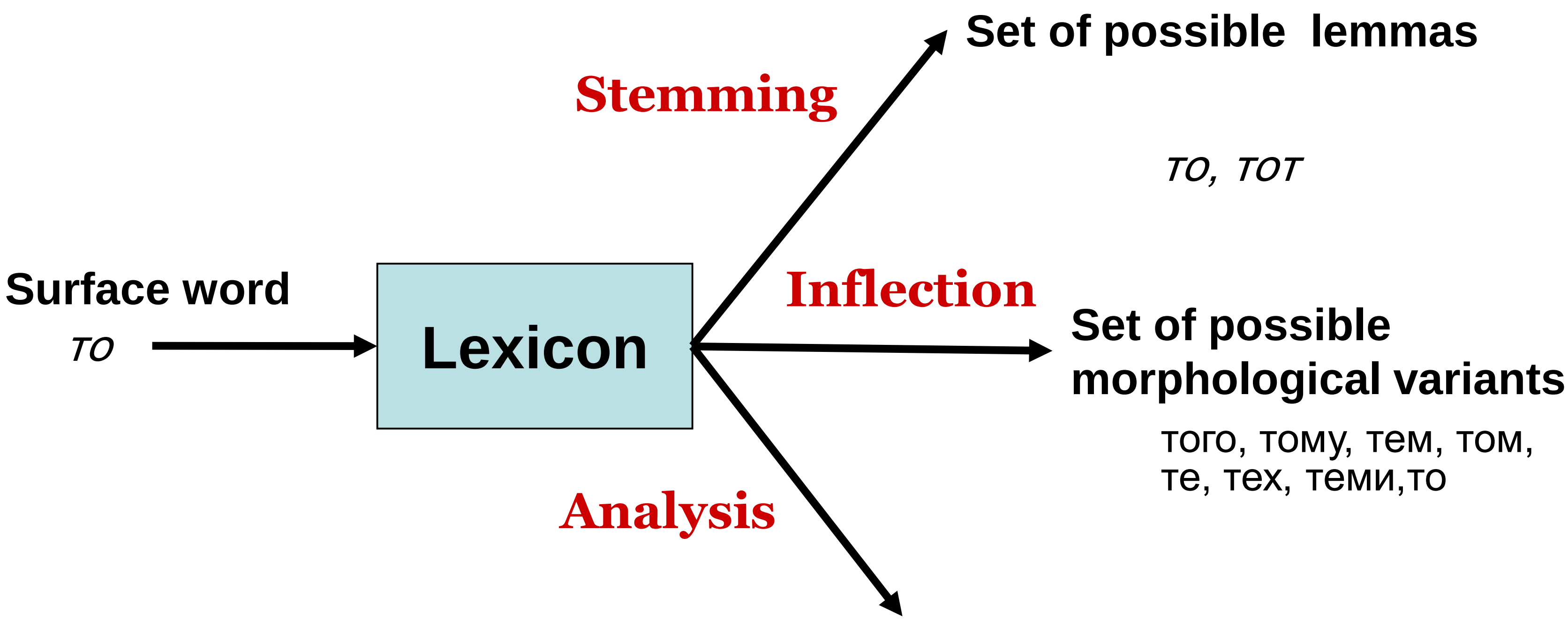
so+said+we+it

so we said it

(from Nizar Habash)

- Agreement:
  - In person, number, gender and definiteness

Lexicon Operations



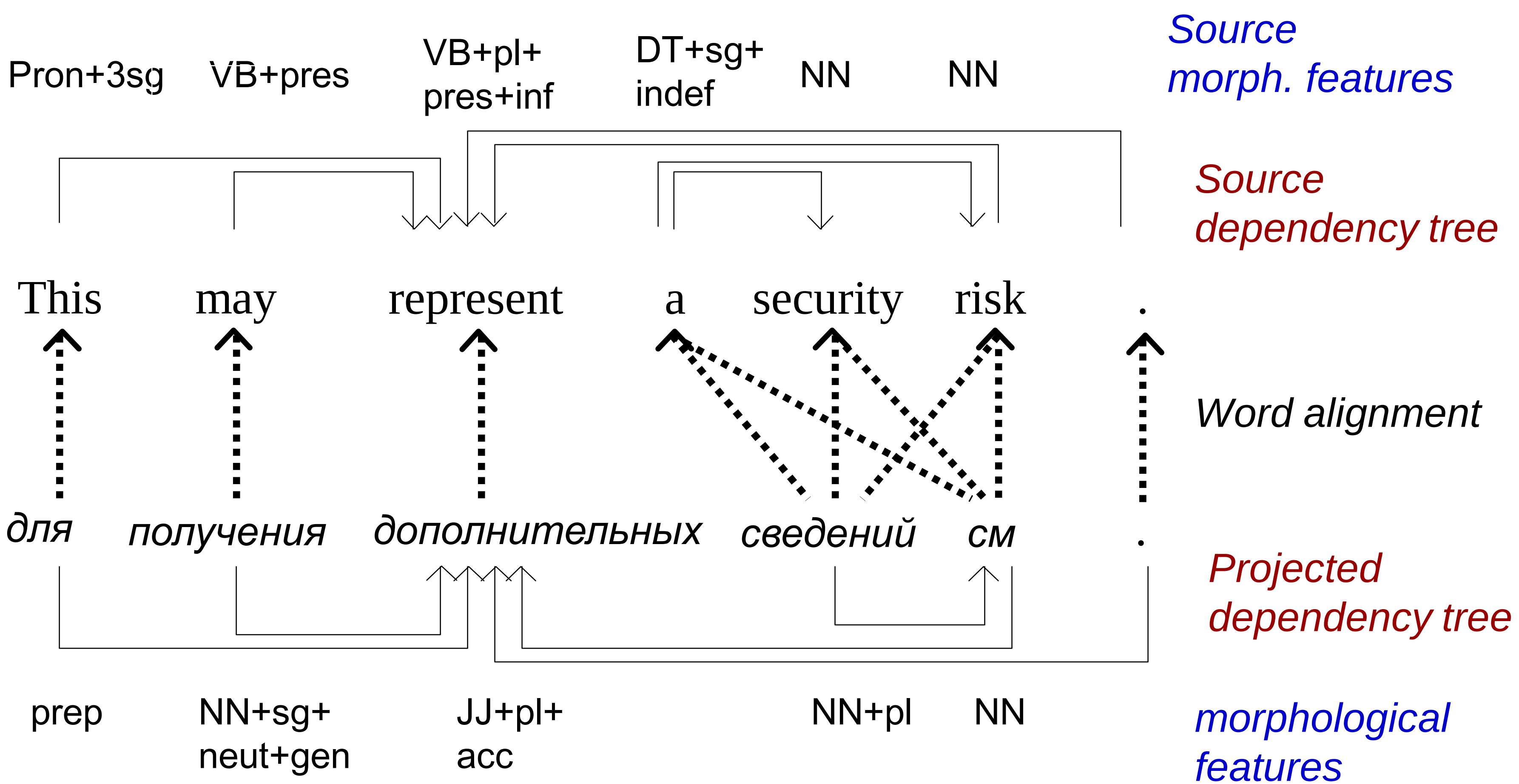
Set of possible morphological analyses

тот+PronAdj+DemPron+Neut+Sg+NomAcc (that)

то то+Pron+Neut+Inanim+Sg+NomAcc (it)

то то+Conj (then)

Linguistic Annotation & Features

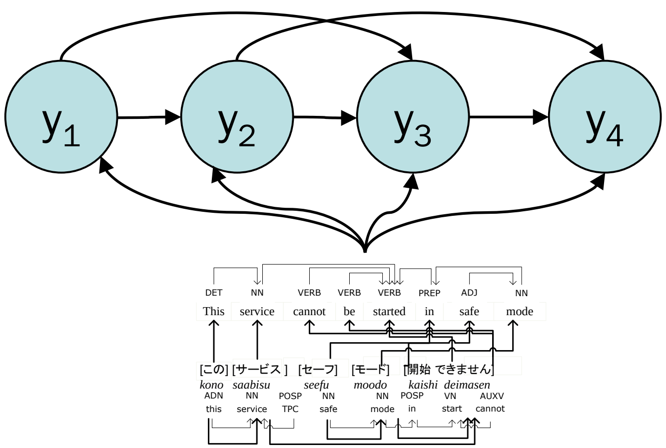


Inflection Prediction

- Given lemmatized text, predict the inflection of each word.

$$y_i \in \mathit{Inflections}(stem_i)$$

- A sequence Conditional Markov Model
  - globally conditioned on the source sentence, the target sentence content words or stems, and the linguistic annotations of the context
  - local probability distributions are estimated with log-linear (maximum entropy) models



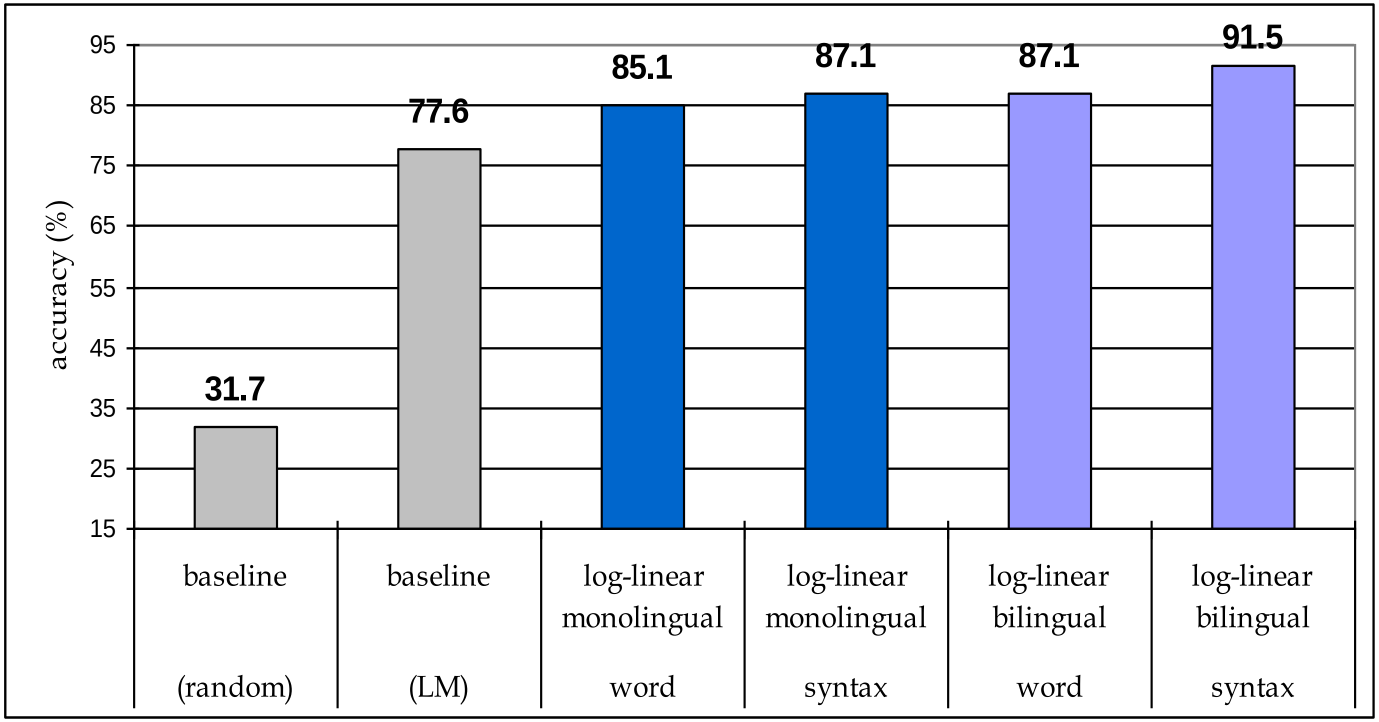
$$p(\overline{y} \mid \overline{x}) = \prod_{t=1}^{n} p(y_t \mid y_{t-1}, y_{t-2}, x_t)$$

Reference Experiments

Data	Eng-Russian	Eng-Arabic
Training	1M	~0.5M
Dev	1K	1K
Test	1K	1K

- Baselines
  - Random baseline (pick a label at random)
  - Word-trigram language model baseline
    - Trained using the CMU toolkit on the same training dataset
- Models: log-linear models
  - *Monolingual, Bilingual, Word, Syntax*
- Lexicons:
  - Russian…, Arabic: Buckwalter
  - Evaluated only on words in the lexicon

Russian inflection prediction: accuracy



The log-linear monolingual word model significantly outperforms the language model 77.6 → 85.1

Using syntactic and morphological information reduces the error by 35% for the best bilingual models 87.1 → 91.5

Integrating inflection models with an SMT system

- a chaining (factoring) approach

Baseline SMT

$$\Pr(f_1, f_2, \ldots, f_n \mid e_1, e_2, \ldots, e_m) =$$

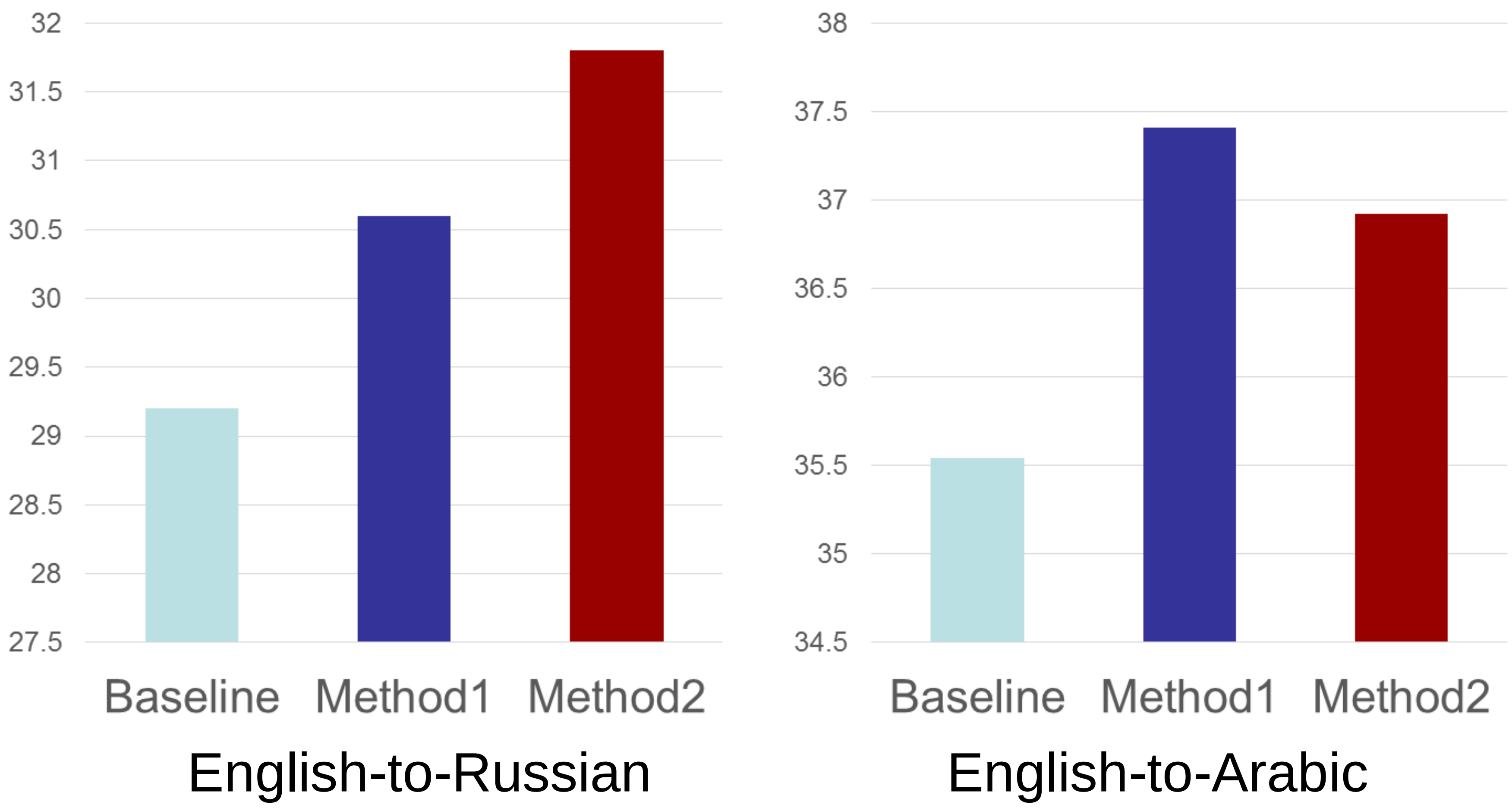
$$\Pr_{SMT}(stem(f_1), stem(f_2), \ldots, stem(f_n) \mid e_1, e_2, \ldots, e_m) \times$$

$$\Pr_{InflM}(f_1, \ldots, f_n \mid stem(f_1), \ldots, stem(f_n), e_1, \ldots, e_m)$$

Inflection Prediction model

- Method 1 – train baseline system to predict full target forms and ignore the produced inflections [Rules, Modeling]
- Method 2 – train baseline system to predict target sequences of stems (pre-process parallel data by stemming) [Alignment, Rules, Modeling]
-

Results for Integration with tree-to-string MT system



Other advancements in translating to morphologically rich languages

- Predict word formation (compound merging) in addition to inflection with a feature-rich generation model [Fraser at al 2013]
- Study which morphological features are best predicted by the MT system, and which ones are best predicted through a separate generation model [Kholy & Habash 2012]
- Using a feature-rich model, extend the translation rules for an MT system on a sentence basis to generate possible inflections for target words [Chahuneau et al 13]
  - Use phrase-based decoder with additional feature
- Reverse self-training – adding automatically translated data from Czech-to-English to improve English-Czech translation [Bojar & Tamchyna 2013]

Outline

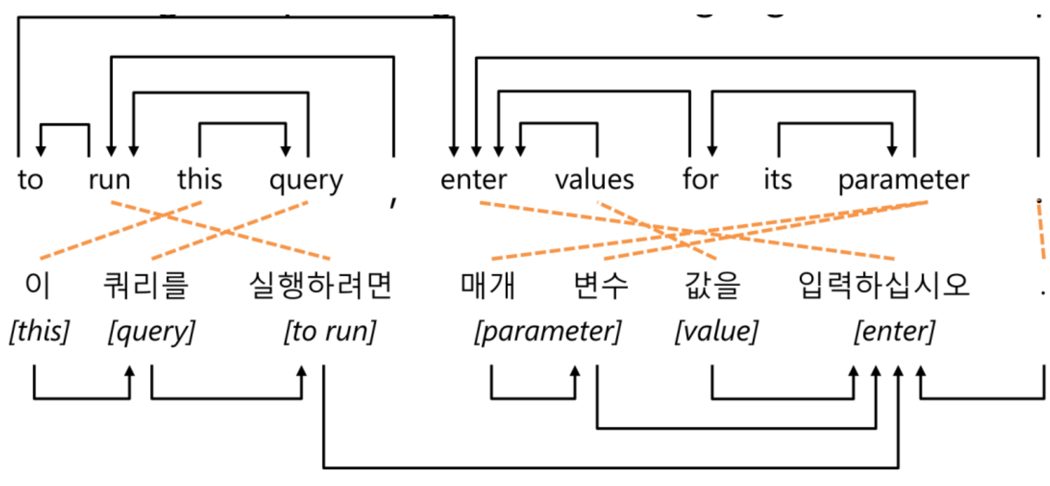
- Unsupervised Induction of Morphology
- Pre-processing to reduce language divergence [Alignment, Rules, Modeling]
- Factored translation models [Rules, Modeling]
- Models for generation of complex morphology [Rules, Modeling]
- **Scoring models for rich target morphology** [Modeling]

Scoring Models for Rich Target Morphology

[Modeling]

Toward tighter integration of feature-rich models for morphology [Jeong et al 2010]

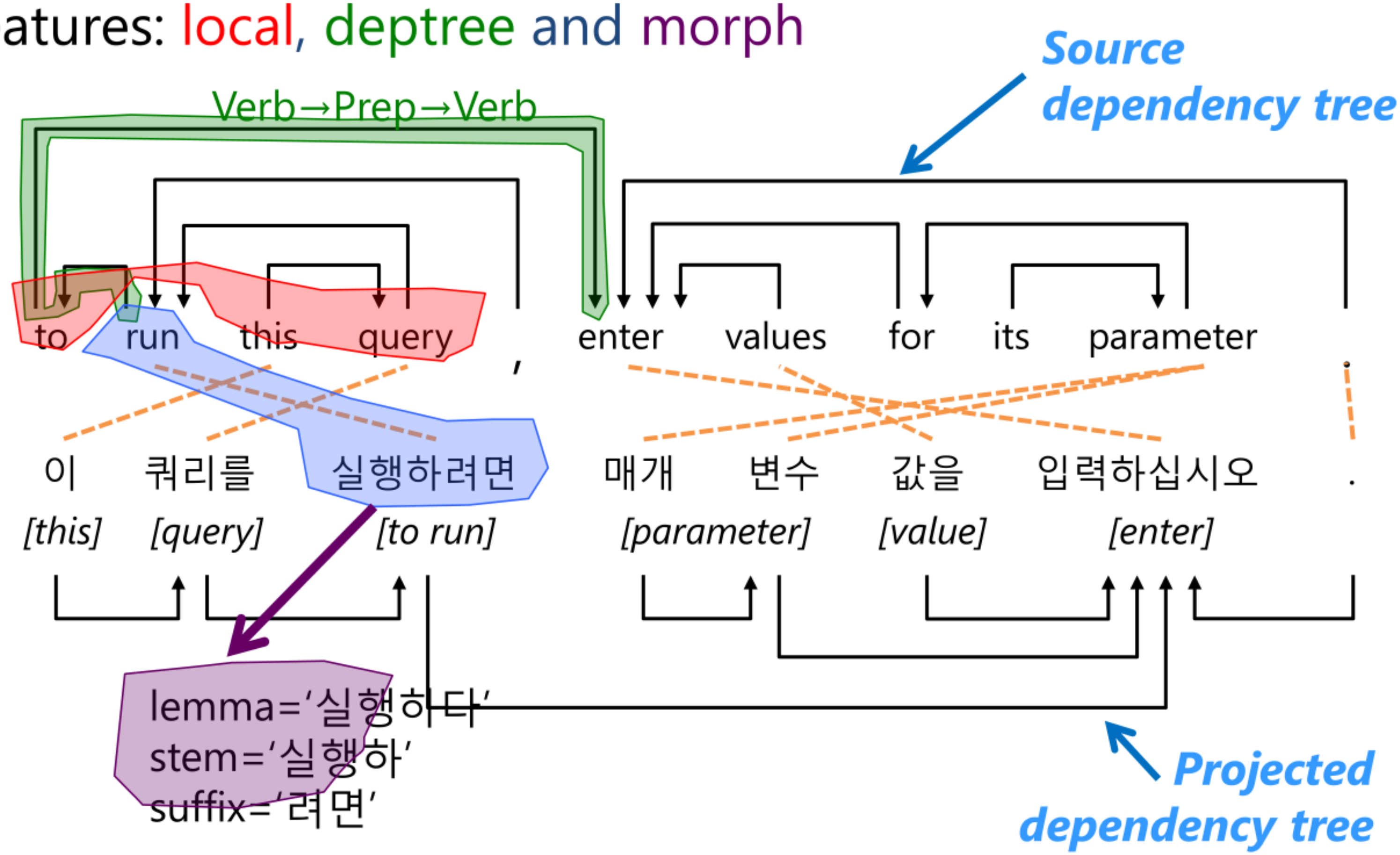
- Given: (a) group of source words + (b) context from whole source sentence
- Predict the target translations
- Parallel data provides training pairs



- Integrated as a feature in tree-to-string decoder

Model features

Features: local, deptree and morph



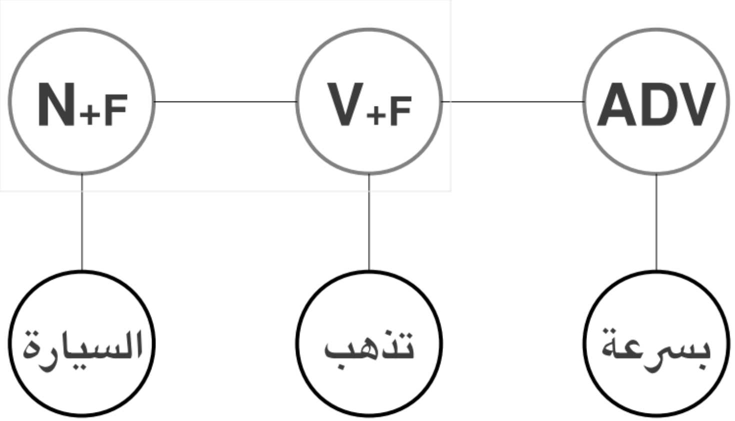
MT results

	MERT Dev		Test	
	Baseline	+DL	Baseline	+DL
Bulgarian	21.78	22.44	19.00	19.63
Czech	11.87	12.45	11.90	12.38
Korean	61.23	62.04	59.04	59.52

Table 8: Results (BLEU) on MT task

- Pros:
  - Tightly integrated with decoder
- Cons:
  - Only impacts modeling, not Alignment or Rules
  - No target context used

A Class-Based Agreement Model for Generating Accurately Inflected Translations
[Green & DeNero 2012]



- Keep baseline hypothesis space, define new feature: class-based agreement model.
- Compute best morphological segmentation and tagging of target hypotheses during decoding.
- Efficient decoder integration.
- Gains of 1 BLEU on average for train size 500 million words.

Summary

- Unsupervised morphology is useful in MT.
- Pre-processing and re-defining the basic units can be very effective.
- Factored Models generalize translation rules and incorporate more information locally.
- Feature-rich models for generation into morphologically rich languages improve quality.
- New features in standard decoders targeted at agreement and sparsely reduction are effective.