Text Analysis without Traditional Supervision

Noah Smith

Carnegie Mellon University

Joint work with: Jon Clark, Shay Cohen, Dipanjan Das, Chris Dyer, Kevin Gimpel, and Alon Lavie

Overview

Theme: "Marrying linguistic experts and statistical learners."

- Make linguistic knowledge more *declarative.*
- Make statistical models more *flexible*.

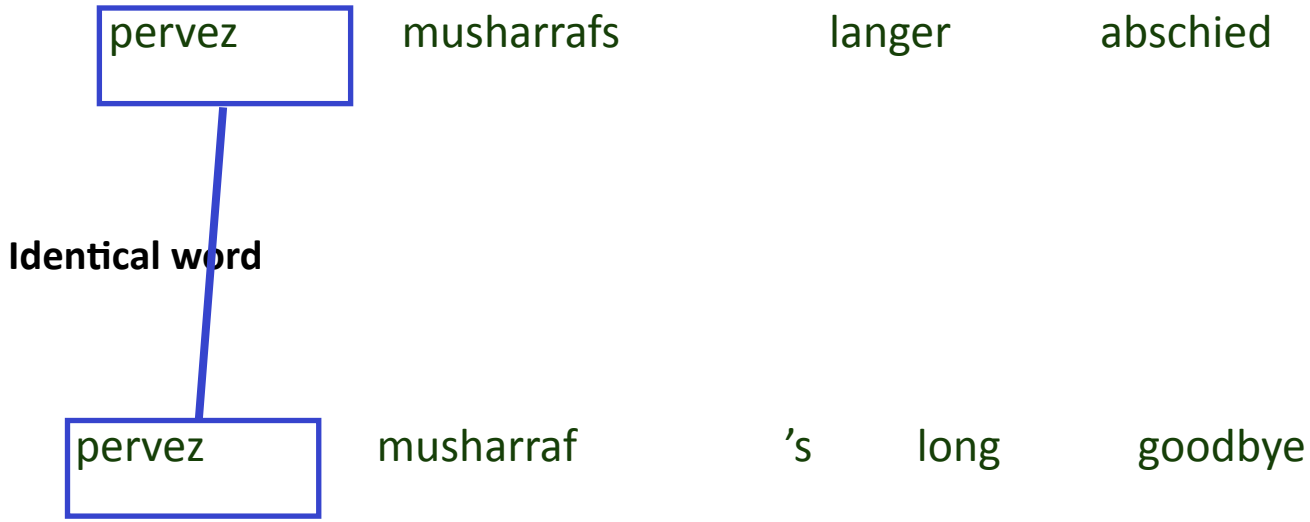
	Task(s)	Statistical Model	Linguistic Expertise
1.	word alignment	latent-variable CRF with sparsity	richer lexical translation preferences
2.	POS tagging, dependency parsing	probabilistic grammars (HMMs, PCFGs)	type-level preferences in helper languages
3.	multiword expression finding	nonparametric, Bayesian prior	non-compositional expressions are rife

1. Word Alignment

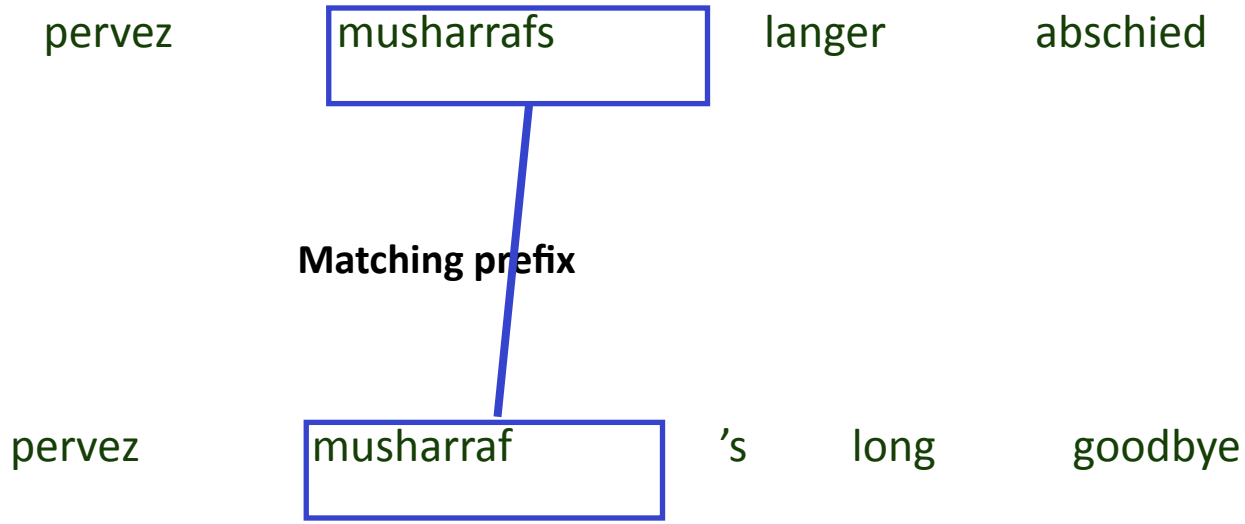
Chris Dyer
Jonathan Clark
Alon Lavie

pervez musharrafs langer abschied

pervez musharraf ’s long goodbye

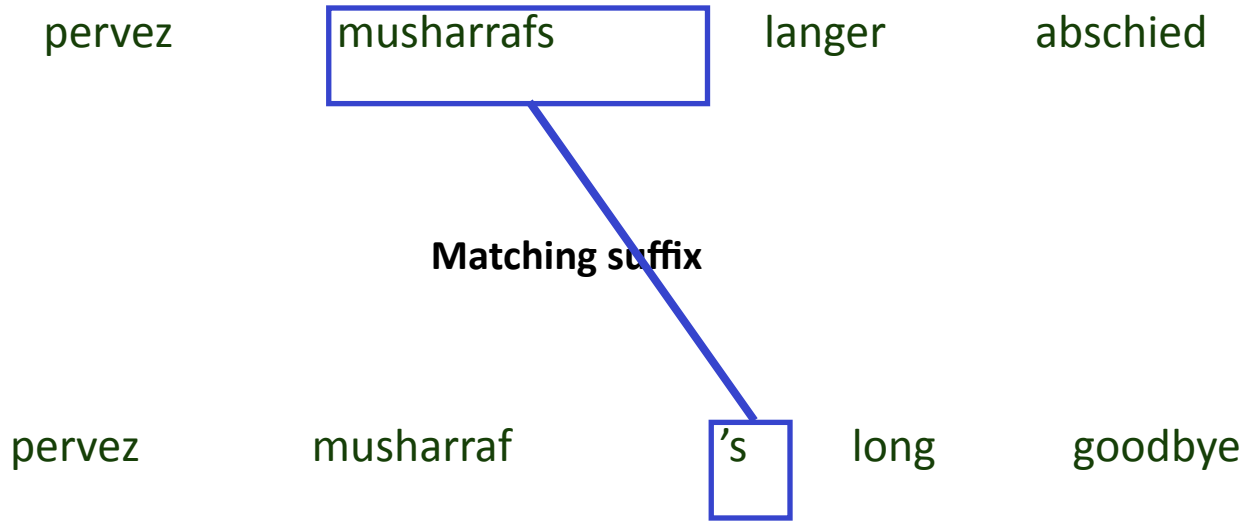


Identical word



Identical word

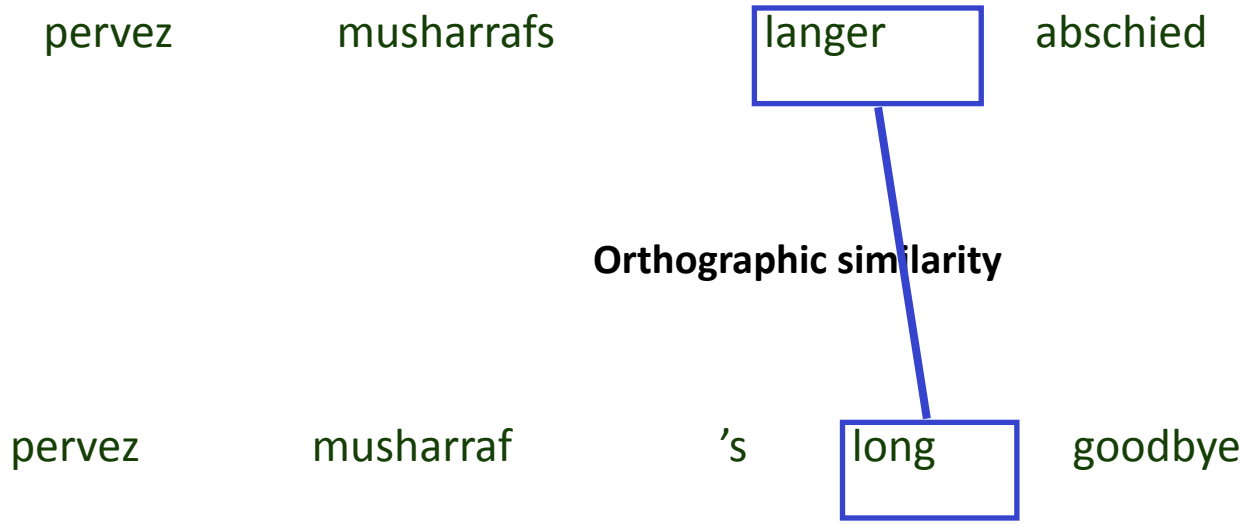
Matching prefix



Identical word

Matching prefix

Matching suffix

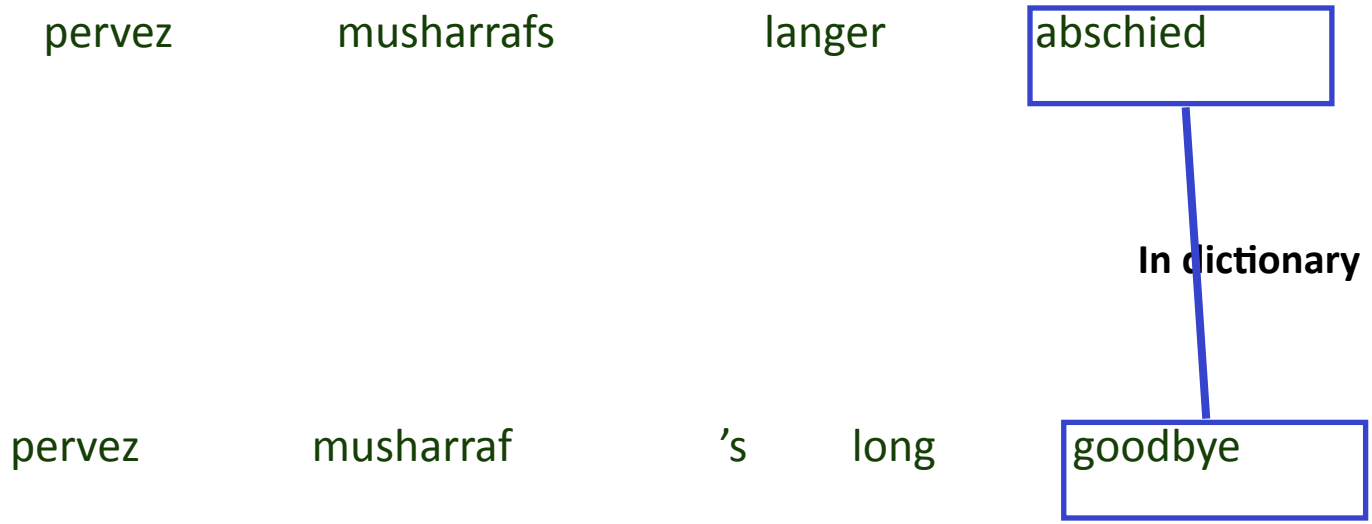


Identical word

Matching prefix

Matching suffix

Orthographic similarity



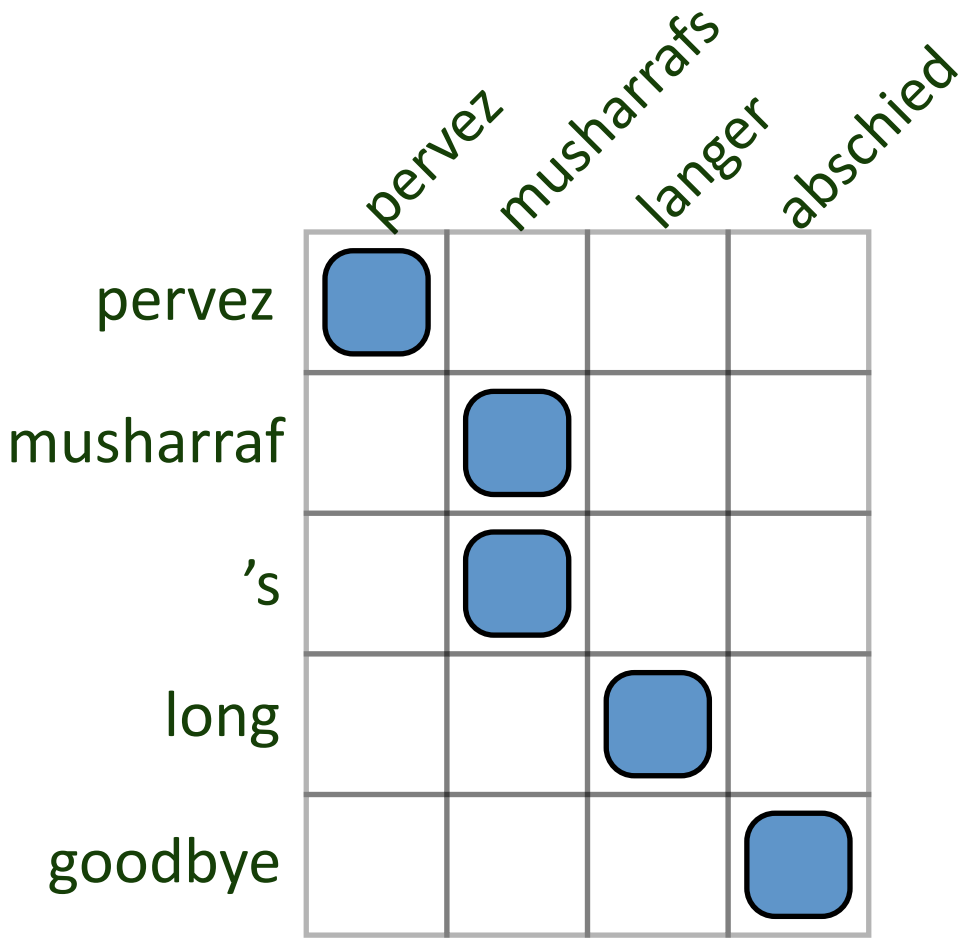
Identical word

Matching prefix

Matching suffix

Orthographic similarity

In dictionary

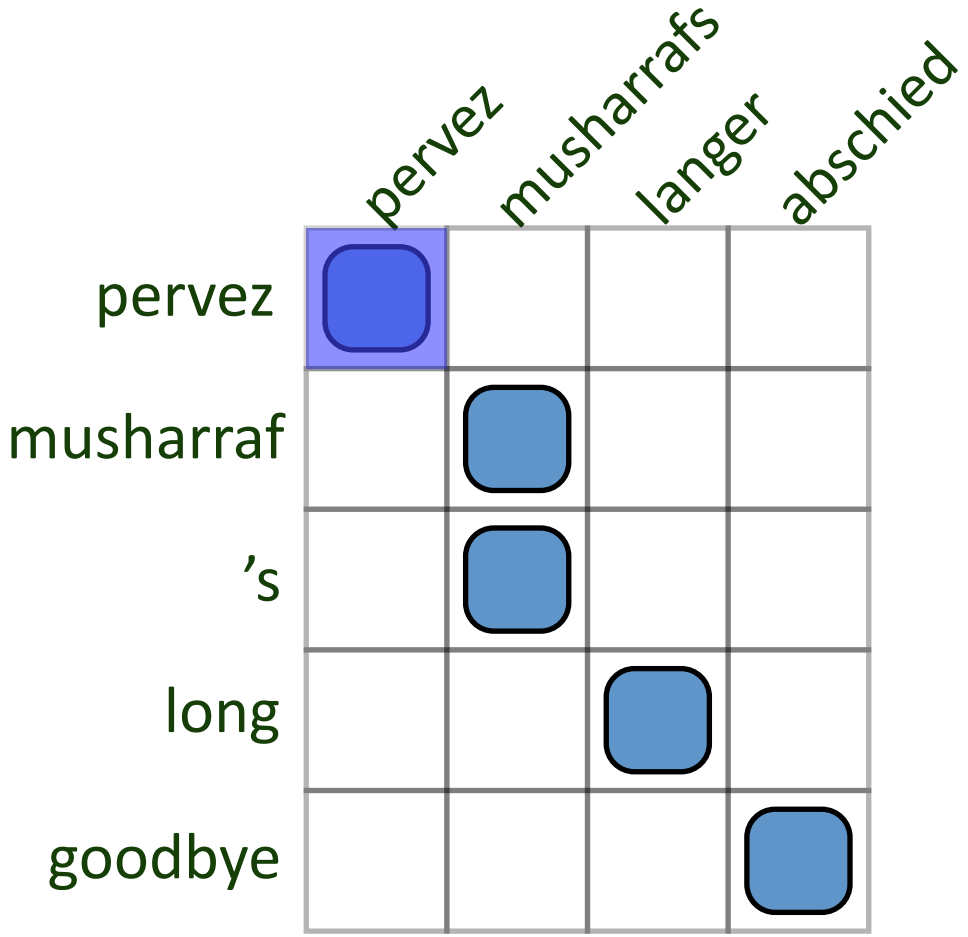


Identical word **In dictionary**

Matching prefix

Matching suffix

Orthographic similarity



Identical word

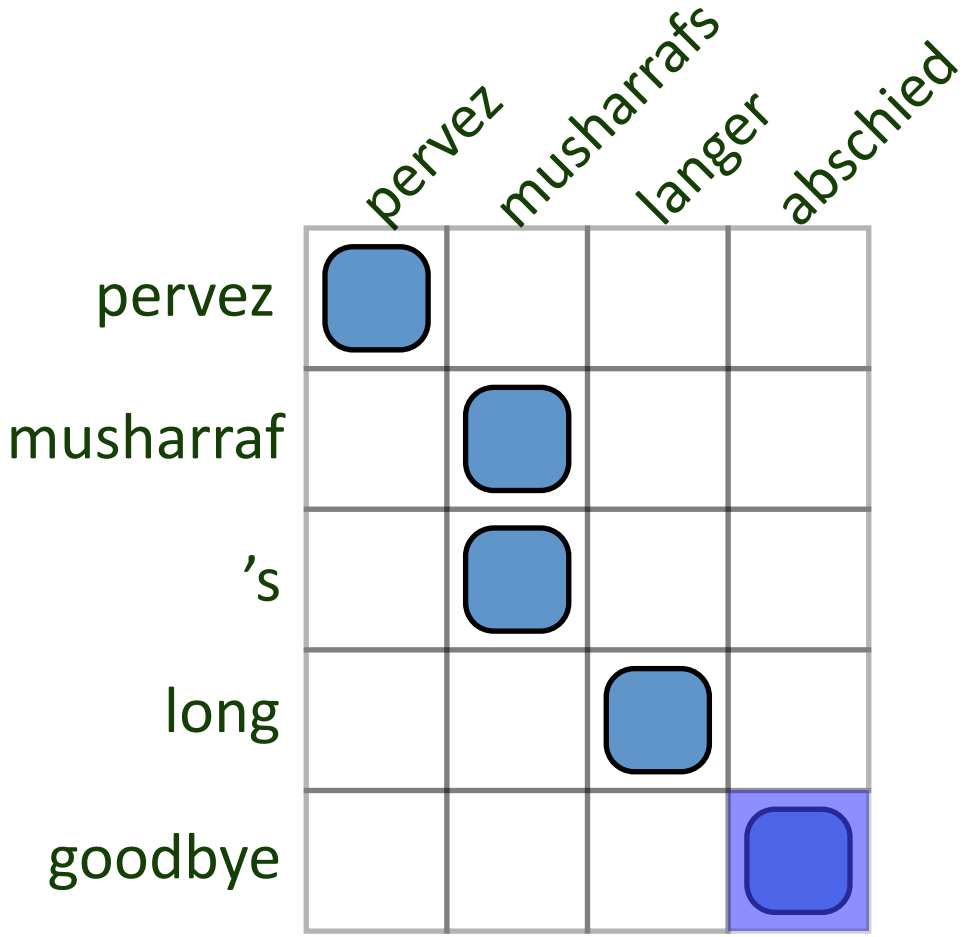
Matching prefix

Matching suffix

Orthographic similarity

In dictionary

First ↔ first

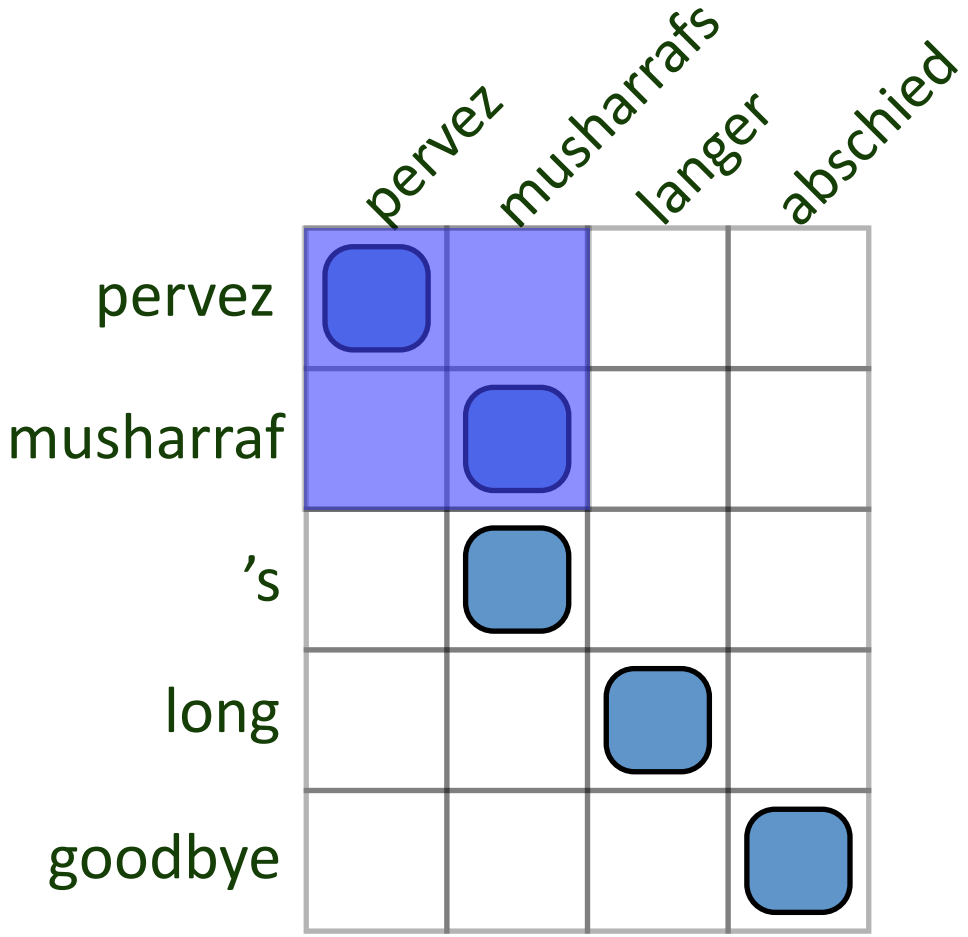


Identical word — **In dictionary**

Matching prefix — **First ↔ first**

Matching suffix — **Last ↔ last**

Orthographic similarity



Identical word

Matching prefix

Matching suffix

Orthographic similarity

In dictionary

First ↔ first

Last ↔ last

Monotonic order

Big Idea: Sparse Feature-Based Model

- Declaratively state the features that *might* help us detect word alignments.
- Combine the features into a discriminative latent-variable model (a.k.a. latent-variable conditional random field).
  - IBM models (Brown et al., 1993) and HMM (Vogel et al., 1996) recast in a feature-based framework and extended.
  - "Latent-variable" means we don't need gold-standard alignments for this language (cf. Lacoste-Julien et al., 2006).
  - We tune the hyperparameter on English-French.
- Sparse online parameter learning.
  - I.e., the data tell us which features to use.
  - Converges in just a few passes over the data.

Model and Learning

$$
\begin{aligned}
p_{\mathbf{w}}(output \mid input) &= \sum_{hidden} p_{\mathbf{w}}(output, hidden \mid input) \\
&= \frac{\sum_{hidden} \exp \mathbf{g}(output, hidden, input)^{\top} \mathbf{w}}{\sum_{output', hidden} \exp \mathbf{g}(output', hidden, input)^{\top} \mathbf{w}}
\end{aligned}
$$

$$
\max_{\mathbf{w}} p(\mathbf{w}) \times \prod_{i} p_{\mathbf{w}}(output \mid input)
$$

sparsity-inducing

Alignments Improve

	AER	Average fertility of singletons	Number of extracted translation rules matching test data
Czech-English			
Model 4	23.4	2.7	993,953
This work	**20.5**	**1.6**	**1,146,677**
Chinese-English			
Model 4		3.6	52,323
This work		**3.1**	**54,077**
Urdu-English			
Model 4		3.2	244,570
This work		**2.3**	**260,953**

Additional Notes

- Inference with dynamic programming, but prune with simpler model.
- BLEU gains as well (see Chris's talk next).
- Source code available:
  `www.cdec-decoder.org`

	Task(s)	Statistical Model	Linguistic Expertise
1.	word alignment	latent-variable CRF with sparsity	richer lexical translation preferences
2.	POS tagging, dependency parsing	probabilistic grammars (HMMs, PCFGs)	type-level preferences in helper languages
3.	multiword expression finding	nonparametric, Bayesian prior	non-compositional expressions are rife

2. Multilingual Guidance for Tagging and Parsing

Shay Cohen
Dipanjan Das

“Universals”

- There are syntactic regularities across languages (cf. Naseem et al., 2010), but we aren’t sure what they are.
  - Part of speech tag sequences
  - Syntactic attachment preferences
- Models trained on annotated data in a few languages should help us fill in the details for languages without big treebanks.

Big Idea:
A New Kind of Comparative Analysis

- Successful unsupervised learning hinges on smart *initialization*.
- Use a few supervised models in "helper" languages to design an initializer for a new language.
- **Interpolate** these models into a coarse initializer with relatively few parameters to fit to new data.
  - **No parallel data required** (cf. Yarowsky et al., 2001; Smith and Smith, 2004; Hwa et al., 2005).
- Use the coarse model to initialize standard training of a standard unsupervised model.

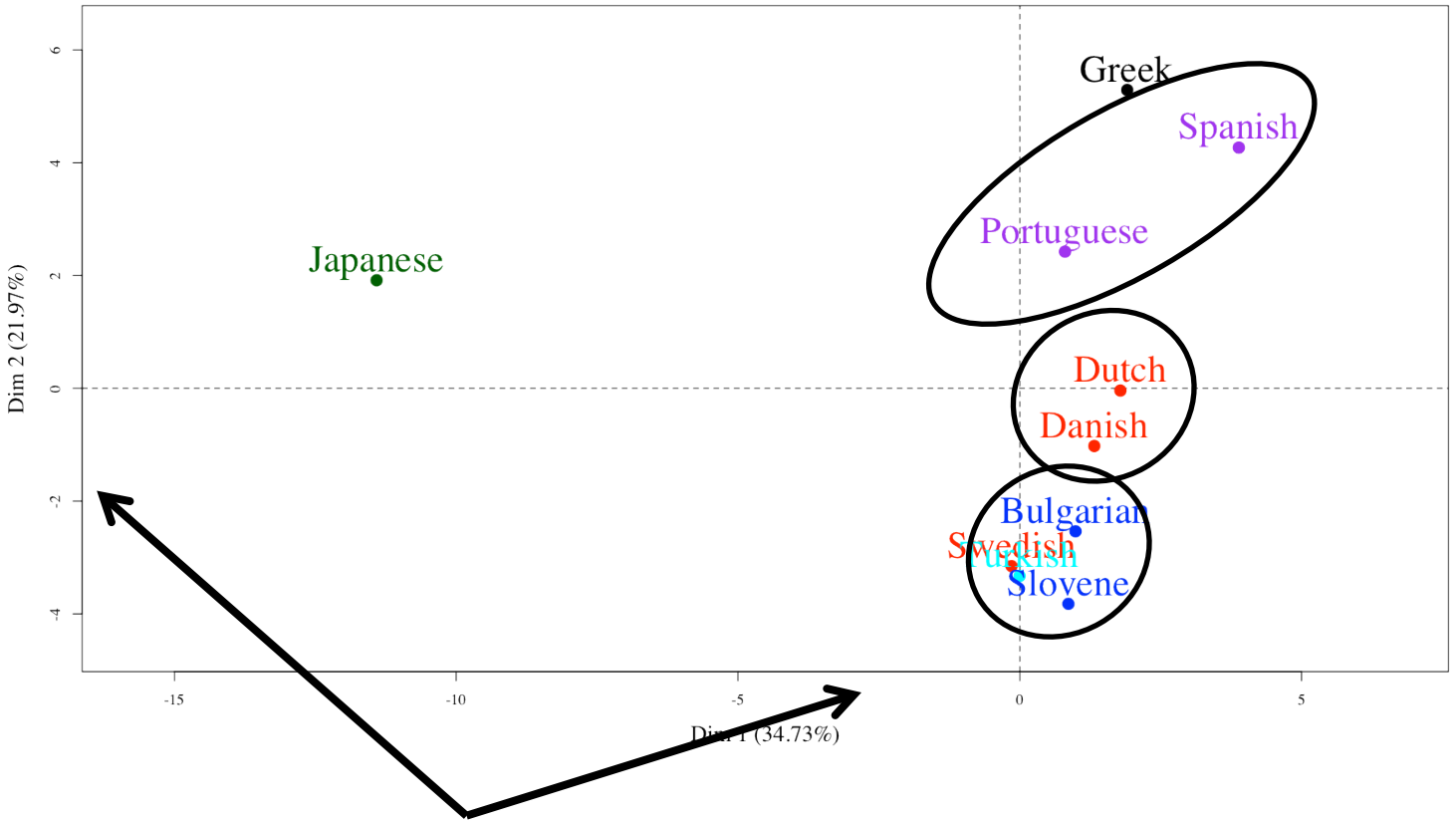
Languages

- *Helpers:* English, German, Italian, Czech
- *Test languages:* Bulgarian, Danish, Dutch, Greek, Japanese, Portuguese, Slovene, Spanish, Swedish, and Turkish

	POS tagging		Dependency parsing	
	Acc.	wins	UAS	wins
Monolingual (Berg-Kirkpatrick et al., 2010)	40.6	2		
EM (Klein and Manning, 2004)			41.4	0
Posterior regularization (Gillenwater et al., 2010)			50.2	2
Phylogenetic model (Berg-Kirkpatrick and Klein, 2010)			53.6	0
Uniform mixture initializer (without EM)	41.0	2	61.5 (61.6)	1 (3)
Full model (without EM)	**43.3**	**3**	**62.1 (62.2)**	**3 (1)**

baselines

Principal Component Analysis on Learned Coefficients

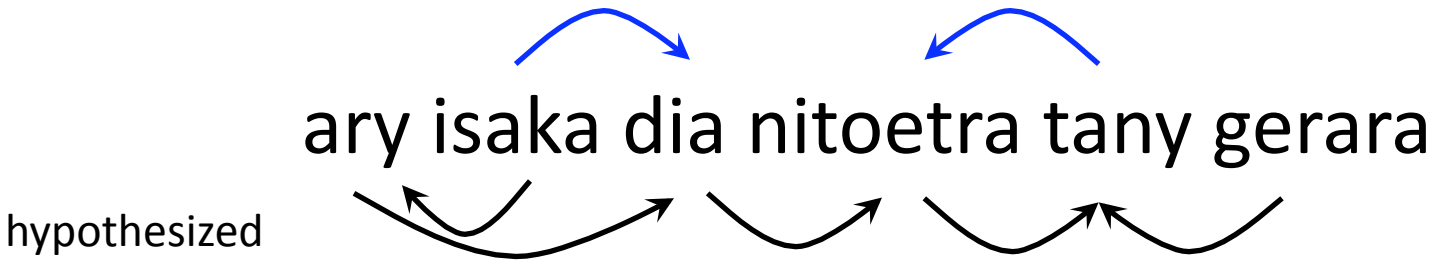


Two principal components

Additional Notes

- Lots more experimental detail in the paper.
- We've applied the model to Kinyarwanda and to Malagasy, but haven't evaluated yet.
  - Next data release will allow this!
- Unsupervised models to be provided to the data team to speed up manual annotation.

Example



so Isaac voyage dwell country Gerar

“so Isaac settled in Gerar”

	Task(s)	Statistical Model	Linguistic Expertise
1.	word alignment	latent-variable CRF with sparsity	richer lexical translation preferences
2.	POS tagging, dependency parsing	probabilistic grammars (HMMs, PCFGs)	type-level preferences in helper languages
3.	multiword expression finding	nonparametric, Bayesian prior	non-compositional expressions are rife

3. Discovering Gappy Phrases

Kevin Gimpel

Gappy Phrases

- Language models, taggers, parsers, etc. usually treat **words** as the atomic units of text.
  - Smaller than words: morphemes
  - Bigger than words: noncompositional phrases
- Phrases are not always *contiguous*!

  either ___ or; neither ___ nor; not only ___ but

  prevent ___ from

  (___); " ___ "

Big Idea: Holistic Gappy Phrase Model

- A nonparametric model for gappy phrases in monolingual or parallel text.
- Past work (Simard et al., 2005; Chiang, 2005; Galley and Manning, 2010; Kim, 2011; Xiong et al., 2011) has relied on **heuristics** or **mutual information**, not a holistic explanation of the data.
- Inspiration: Rosenfeld's (1994) **trigger** models, but patterns more complex than w_1 ___ w_2

nato must either say " yes " or " no " to the baltic states .

nato must either say " yes " or " no " to the baltic states .

- Generative story:
  - Generate number of word positions
  - Generate number of colors
  - Assign word positions to colors
  - Generate a lexical pattern for each color

nato must either say " yes " or " no " to the baltic states .

- Generative story:
  - Generate number of word positions (n = 16)
  - Generate number of colors
  - Assign word positions to colors
  - Generate a lexical pattern for each color

nato must either say " yes " or " no " to the baltic states .

- Generative story:
  - Generate number of word positions (n = 16)
  - Generate number of colors (m = 10)
  - Assign word positions to colors
  - Generate a lexical pattern for each color

nato must either say " yes " or " no " to the baltic states .

- Generative story:
  - Generate number of word positions (n = 16)
  - Generate number of colors (m = 10)
  - Assign word positions to colors
  - Generate a lexical pattern for each color

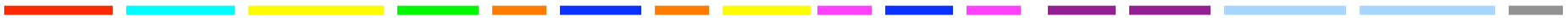
nato must either say " yes " or " no " to the baltic states .

- Generative story:
  - Generate number of word positions (n = 16)
  - Generate number of colors (m = 10)
  - Assign word positions to colors
  - Generate a lexical pattern for each color

nato must either say " yes " or " no " to the baltic states .

- Generative story:
  - Generate number of word positions (n = 16)
  - Generate number of colors (m = 10)
  - Assign word positions to colors
  - Generate a lexical pattern for each color

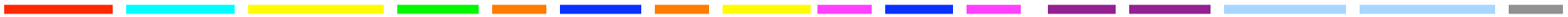
nato



nato must either say " yes " or " no " to the baltic states .

- Generative story:
  - Generate number of word positions (n = 16)
  - Generate number of colors (m = 10)
  - Assign word positions to colors
  - Generate a lexical pattern for each color

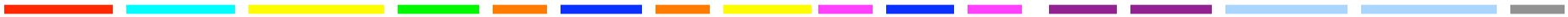
nato must



nato must either say " yes " or " no " to the baltic states .

- Generative story:
  - Generate number of word positions (n = 16)
  - Generate number of colors (m = 10)
  - Assign word positions to colors
  - Generate a lexical pattern for each color

nato must either or



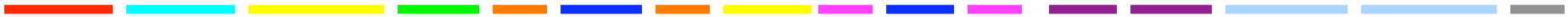
What is a **pattern**?

A sequence of symbols, possibly including the special symbol "__" which is used to indicate a gap of nonzero length

Examples:

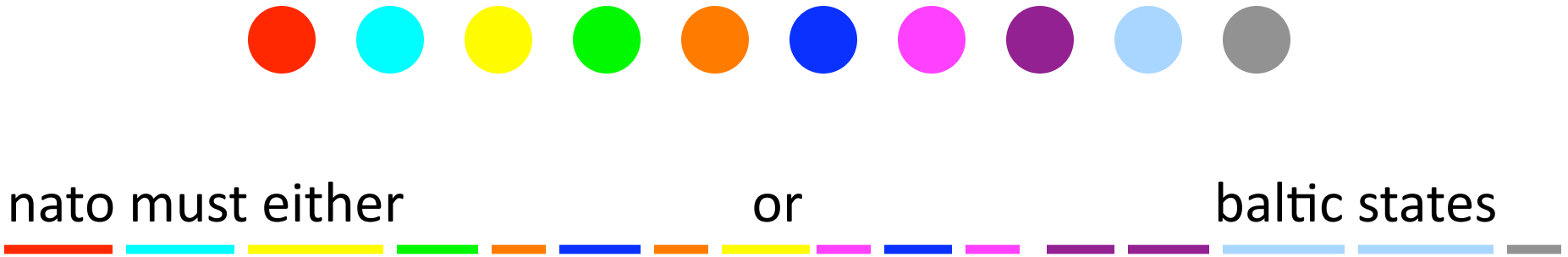
nato	the united states
must	according to the __ ,
either __ or	countries __ their __ the united states

nato must either or



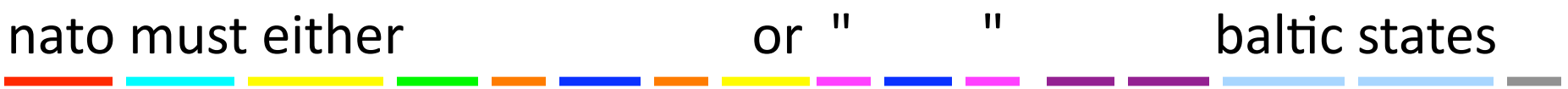
nato must either say " yes " or " no " to the baltic states .

- Generative story:
  - Generate number of word positions (n = 16)
  - Generate number of colors (m = 10)
  - Assign word positions to colors
  - Generate a lexical pattern for each color



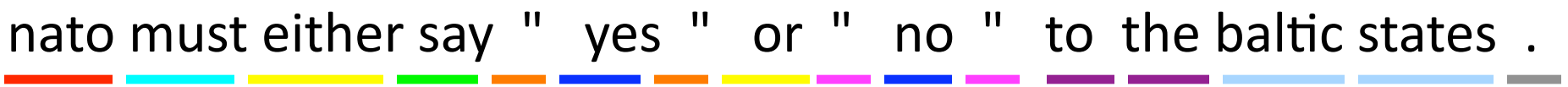
nato must either say " yes " or " no " to the baltic states .

- Generative story:
  - Generate number of word positions (n = 16)
  - Generate number of colors (m = 10)
  - Assign word positions to colors
  - Generate a lexical pattern for each color



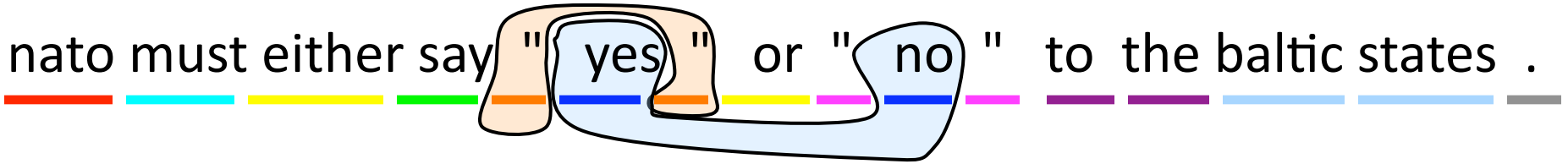
nato must either say " yes " or " no " to the baltic states .

- Generative story:
  - Generate number of word positions (n = 16)
  - Generate number of colors (m = 10)
  - Assign word positions to colors
  - Generate a lexical pattern for each color



nato must either say " yes " or " no " to the baltic states .

- Generative story:
  - Generate number of word positions (n = 16)
  - Generate number of colors (m = 10)
  - Assign word positions to colors
  - Generate a lexical pattern for each color



nato must either say " yes " or

- Generative story:
  - Generate number of wo
  - Generate number of col
  - Assign word positions to colors
  - Generate a lexical pattern for each color

A single multinomial distribution over patterns; put a Dirichlet process prior on it to encourage reuse and heavy tail.

nato must either say " yes " or " no " to the baltic states .

-- __ --	how __ ?	we __ our	his __ his
(__)	the __ (__)	over __ past	some __ others
- __ -	on __ basis	prevent __ from	may __ be
both __ and	less __ than	in __ way	as __ as
not only __ but	on __ other hand	one __ another	oil __ gas
" __ "	at __ level	political __ economic	at __ moment
more __ than	it is __ that	for __ reasons	such as __ and
either __ or	not __ , but	at __ time	question __ whether
why __ ?	play __ role	more __ more	if __ then
neither __ nor	france __ germany	the rest __ world	war __ iraq
what __ ?	he __ his	more __ less	; __ ;
rule __ law	allow __ to	in __ region	have __ been
whether __ or	for __ first time	rich __ poor	in __ cases
around __ world	china __ india	as __ whole	war __ terror
has __ been	what __ do	on __ scale	at __ cost

Punctuation

-- __ --	how __ ?	we __ our	his __ his
(__)	the __ (__)	over __ past	some __ others
- __ -	on __ basis	prevent __ from	may __ be
both __ and	less __ than	in __ way	as __ as
not only __ but	on __ other hand	one __ another	oil __ gas
" __ "	at __ level	political __ economic	at __ moment
more __ than	it is __ that	for __ reasons	such as __ and
either __ or	not __ , but	at __ time	question __ whether
why __ ?	play __ role	more __ more	if __ then
neither __ nor	france __ germany	the rest __ world	war __ iraq
what __ ?	he __ his	more __ less	; __ ;
rule __ law	allow __ to	in __ region	have __ been
whether __ or	for __ first time	rich __ poor	in __ cases
around __ world	china __ india	as __ whole	war __ terror
has __ been	what __ do	on __ scale	at __ cost

Connectives and Constructions

-- __ --	how __ ?	we __ our	his __ his
(__)	the __ (__)	over __ past	some __ others
- __ -	on __ basis	prevent __ from	may __ be
both __ and	less __ than	in __ way	as __ as
not only __ but	on __ other hand	one __ another	oil __ gas
" __ "	at __ level	political __ economic	at __ moment
more __ than	it is __ that	for __ reasons	such as __ and
either __ or	not __ , but	at __ time	question __ whether
why __ ?	play __ role	more __ more	if __ then
neither __ nor	france __ germany	the rest __ world	war __ iraq
what __ ?	he __ his	more __ less	; __ ;
rule __ law	allow __ to	in __ region	have __ been
whether __ or	for __ first time	rich __ poor	in __ cases
around __ world	china __ india	as __ whole	war __ terror
has __ been	what __ do	on __ scale	at __ cost

Agreement

-- __ --	how __ ?	we __ our	his __ his
(__)	the __ (__)	over __ past	some __ others
- __ -	on __ basis	prevent __ from	may __ be
both __ and	less __ than	in __ way	as __ as
not only __ but	on __ other hand	one __ another	oil __ gas
" __ "	at __ level	political __ economic	at __ moment
more __ than	it is __ that	for __ reasons	such as __ and
either __ or	not __ , but	at __ time	question __ whether
why __ ?	play __ role	more __ more	if __ then
neither __ nor	france __ germany	the rest __ world	war __ iraq
what __ ?	he __ his	more __ less	; __ ;
rule __ law	allow __ to	in __ region	have __ been
whether __ or	for __ first time	rich __ poor	in __ cases
around __ world	china __ india	as __ whole	war __ terror
has __ been	what __ do	on __ scale	at __ cost

Topicality

-- __ --	how __ ?	we __ our	his __ his
(__)	the __ (__)	over __ past	some __ others
- __ -	on __ basis	prevent __ from	may __ be
both __ and	less __ than	in __ way	as __ as
not only __ but	on __ other hand	one __ another	oil __ gas
" __ "	at __ level	political __ economic	at __ moment
more __ than	it is __ that	for __ reasons	such as __ and
either __ or	not __ , but	at __ time	question __ whether
why __ ?	play __ role	more __ more	if __ then
neither __ nor	france __ germany	the rest __ world	war __ iraq
what __ ?	he __ his	more __ less	; __ ;
rule __ law	allow __ to	in __ region	have __ been
whether __ or	for __ first time	rich __ poor	in __ cases
around __ world	china __ india	as __ whole	war __ terror
has __ been	what __ do	on __ scale	at __ cost

Prepositional Phrases

-- __ --	how __ ?	we __ our	his __ his
(__)	the __ (__)	over __ past	some __ others
- __ -	on __ basis	prevent __ from	may __ be
both __ and	less __ than	in __ way	as __ as
not only __ but	on __ other hand	one __ another	oil __ gas
" __ "	at __ level	political __ economic	at __ moment
more __ than	it is __ that	for __ reasons	such as __ and
either __ or	not __ , but	at __ time	question __ whether
why __ ?	play __ role	more __ more	if __ then
neither __ nor	france __ germany	the rest __ world	war __ iraq
what __ ?	he __ his	more __ less	; __ ;
rule __ law	allow __ to	in __ region	have __ been
whether __ or	for __ first time	rich __ poor	in __ cases
around __ world	china __ india	as __ whole	war __ terror
has __ been	what __ do	on __ scale	at __ cost

Gappy Phrases in Malagasy

na ___ aza

(___)

tsy ___ intsony

izany ___ izany

hoy ___ taminy @:

izao ___ izao

ireo ___ ireo

ambin' ___ folo

samy ___ avy

va ___ ?

ity ___ ity

ahoana ___ ?

fa ___ kosa

tsy ___ va ___ ?

inona ___ inona

iza ___ ?

nahoana ___ ?

lazain' ___ @:

inona ___ ?

an ___ tsaha

andriamanitry ___ isiraely

ahy ___ ahy

avokoa ___ rehetra

ny ___ i jehovah

tsy ___ akory

ry ___ o

mbola ___ ihany

ianareo ___ ianareo

tamin' ___ tamin'

io ___ io

Additional Notes

- Bilingual gappy phrases, too!
- MCMC inference.
- BLEU gains for Chinese-English MT.
- Source code available:
  `www.ark.cs.cmu.edu/MT`

	Task(s)	Statistical Model	Linguistic Expertise
1.	word alignment	latent-variable CRF with sparsity	richer lexical translation preferences
2.	POS tagging, dependency parsing	probabilistic grammars (HMMs, PCFGs)	type-level preferences in helper languages
3.	multiword expression finding	nonparametric, Bayesian	non-compositional expressions are rife

Conclusions

- Computational/learning substrates are improving in flexibility.
- Next:
  - More to do on fast *approximate* inference; expected payoffs as we integrate linguistic structure into translation models, too.
  - Anticipating exciting developments as data come available.
  - Help data team by providing half-working models?
  - Do linguists and experts think in features and priors? Regexps?