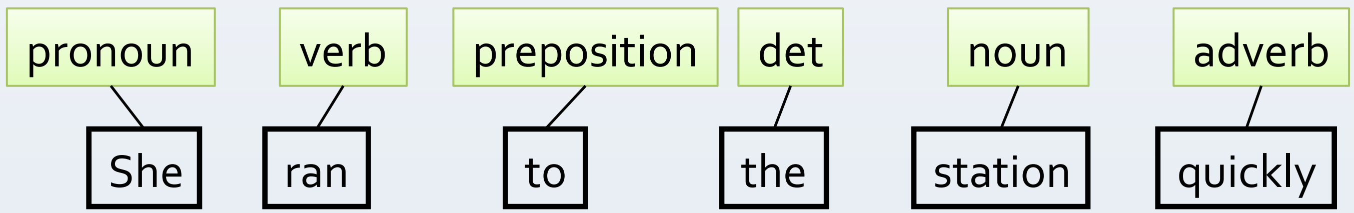
Unsupervised Approaches to Sequence Tagging, Morphology Induction, and Lexical Resource Acquisition

Reza Bosaghzadeh & Nathan Schneider

LS2 ~ 1 December 2008

Unsupervised Methods

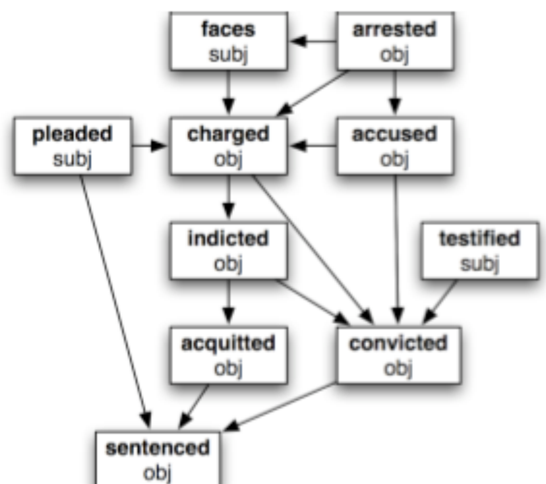
– Sequence Labeling (Part-of-Speech Tagging)



– Morphology Induction

un-supervise-d learn-ing

– Lexical Resource Acquisition

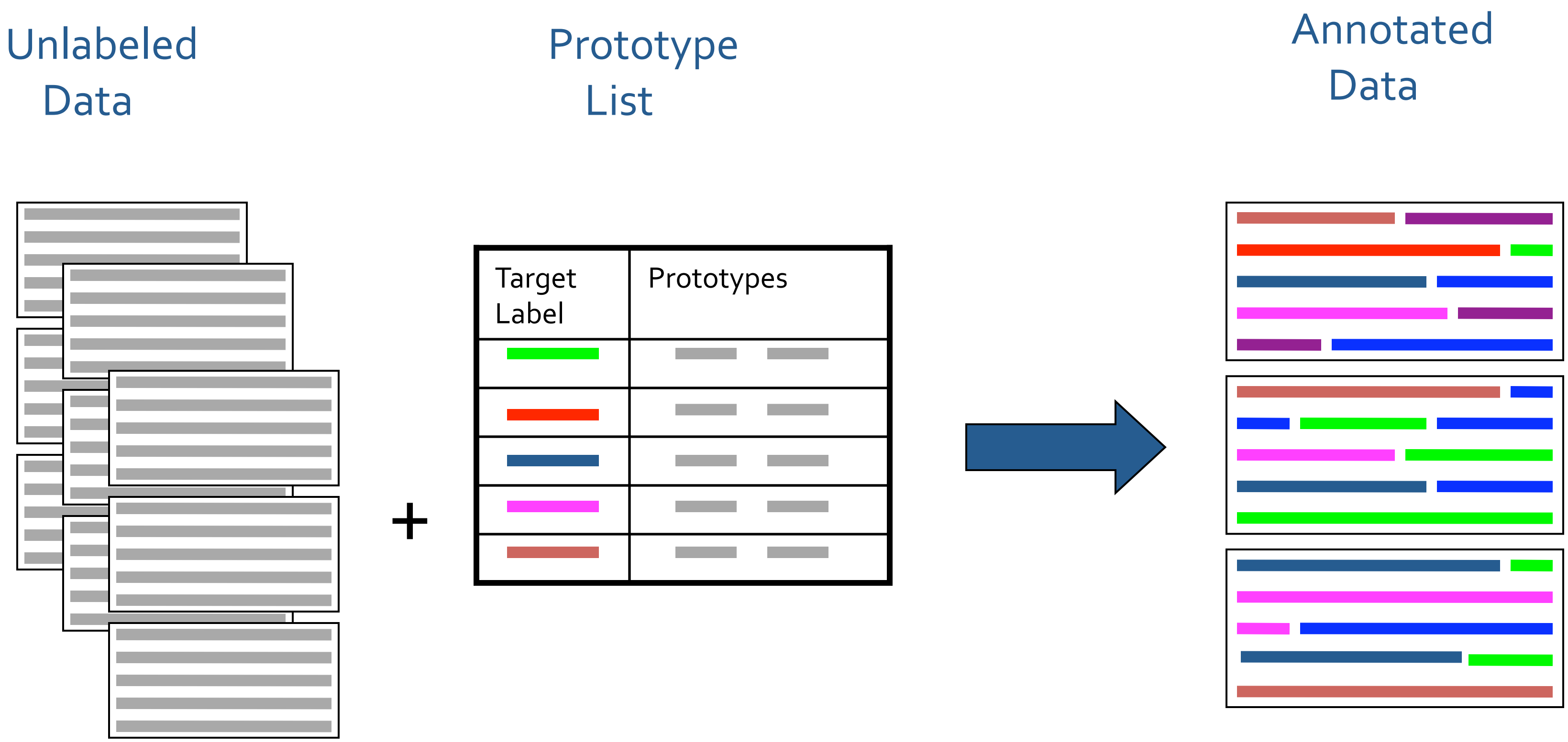


Contrastive Estimation
Smith & Eisner (2005)

- Already discussed in class
- Key idea: exploits implicit negative evidence
  - Mutating training examples often gives ungrammatical (negative) sentences
  - During training, shift probability mass from generated negative examples to given positive examples
- BUT: Requires a tagging dictionary, i.e. a list of possible tags for each word type

Prototype-driven tagging

Haghighi & Klein (2006)



slide courtesy Haghighi & Klein

Prototype-driven tagging
Haghighi & Klein (2006)

English POS

NN	VBN	CC	JJ	CD	PUNC
IN	NNS	IN	NNP	RB	DET

Newly remodeled 2 Bdrms/1 Bath, spacious upper unit, located in Hilltop Mall area. Walking distance to shopping, public transportation, schools and park. Paid water and garbage. No dogs allowed.

Prototype List

NN	president	IN	of
VBD	said	NNS	shares
CC	and	TO	to
NNP	Mr.	PUNC	.
JJ	new	CD	million
DET	the	VBP	are

slide courtesy Haghighi & Klein

Prototypes

Information Extraction: Classified Ads

Size   Restrict   Terms   Location   Features

Newly remodeled 2 Bdrms/1 Bath, spacious upper unit, located in Hilltop Mall area. Walking distance to shopping, public transportation, schools and park. Paid water and garbage. No dogs allowed.

Prototype List

FEATURE	kitchen, laundry
LOCATION	near, close
TERMS	paid, utilities
SIZE	large, feet
RESTRICT	cat, smoking

slide courtesy Haghighi & Klein

Prototype-driven tagging
Haghighi & Klein (2006)

- Trigram tagger, same features as (Smith & Eisner 2005)
  - Word type, suffixes up to length 3, contains-hyphen, contains-digit, initial capitalization
- Tie each word to its most similar prototype, using context-based similarity technique (Schütze 1993)
  - SVD dimensionality reduction
  - Cosine similarity between context vectors

slide adapted from Haghighi & Klein

Prototype-driven tagging
Haghighi & Klein (2006)

Pros

- Doesn't require tagging dictionary

Cons

- Still need a tag set
- May be hard to choose *good* prototypes

Unsupervised POS tagging
The State of the Art

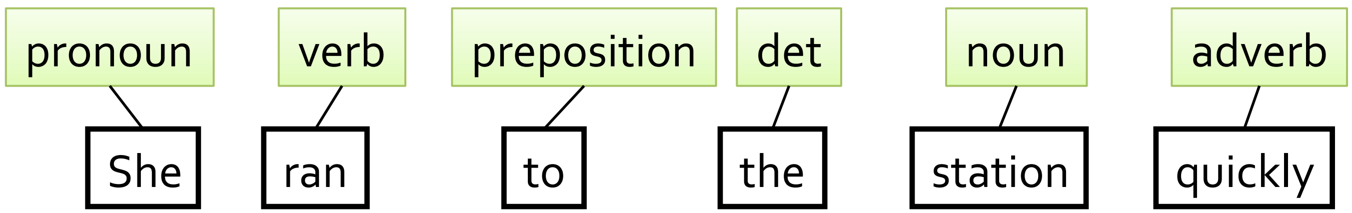
Unsupervised English POS Tagging

	24K tokens	48K tokens (2K sen.)
Haghighi and Klein (2006)		
Baseline (trigram), Treebank tagset		42.4%
Prototype-augmented, Treebank tagset		79.1%
Prototype-augmented, reduced tagset	82.2%	
Smith and Eisner (2005)		
CE, with 2125-entry tagging dictionary	79.5%	
CE, with 3362-entry tagging dictionary	88.1%	
CE, with 5406-entry tagging dictionary	90.4%	
Toutanova and Johnson (2008)		
Latent Dirichlet Allocation	93.4%	

Best supervised result (CRF): 99.5% !

Unsupervised Methods

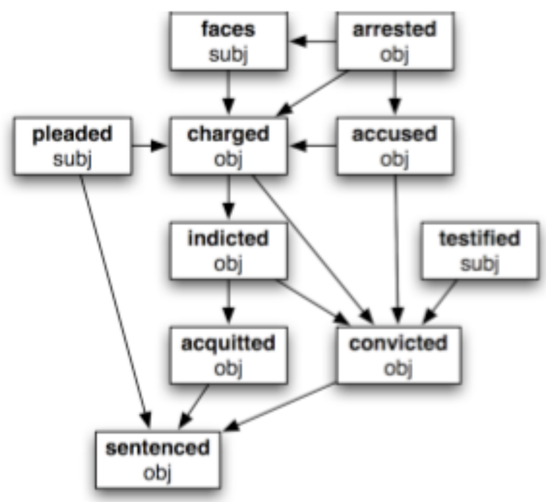
– Sequence Labeling (Part-of-Speech Tagging)



– Morphology Induction

un-supervise-d learn-ing

– Lexical Resource Acquisition



Unsupervised Approaches to Morphology

- Morphology refers to the internal structure of words
  - A **morpheme** is a minimal meaningful linguistic unit
  - **Morpheme segmentation** is the process of dividing words into their component morphemes

    un-supervise-d learn-ing

  - **Word segmentation** is the process of finding word boundaries in a stream of speech or text

    unsupervised_learning_of_natural_language

ParaMor: Morphological paradigms
Monson et al. (2007, 2008)

- Learns inflectional paradigms from raw text
  - Requires only a list of word types from a corpus
  - Looks at word counts of substrings, and proposes (stem, suffix) pairings based on type frequency
- 3-stage algorithm
  - *Stage 1*: Candidate paradigms based on frequencies
  - *Stages 2-3*: Refinement of paradigm set via merging and filtering
- Paradigms can be used for morpheme segmentation or stemming

ParaMor: Morphological paradigms
Monson et al. (2007, 2008)

speak	*dance*	*buy*
hablar	bailar	comprar
hablo	bailo	compro
hablamos	bailamos	compramos
hablan	bailan	compran
…	…	…

- A sampling of Spanish verb conjugations (inflections)

ParaMor: Morphological paradigms

Monson et al. (2007, 2008)

speak	*dance*	*buy*
hablar	bailar	comprar
hablo	bailo	compro
hablamos	bailamos	compramos
hablan	bailan	compran
...	...	...

- A proposed paradigm (correct): stems {habl, bail, compr} and suffixes {-ar, -o, -amos, -an}

ParaMor: Morphological paradigms
Monson et al. (2007, 2008)

- Two subsequent stages:
  - **Filtering** out spurious paradigms (e.g. with incorrect segmentations)
  - **Merging** partial paradigms to overcome sparsity: smoothing

ParaMor: Morphological paradigms

Monson et al. (2007, 2008)

speak	*dance*
hablar	bailar
hablo	bailo
hablamos	bailamos
hablan	bailan
…	…

- For certain sub-sets of verbs, the algorithm may propose paradigms with spurious seg-mentations, like the one at left
- The **filtering** stage of the algorithm weeds out these incorrect paradigms

ParaMor: Morphological paradigms

Monson et al. (2007, 2008)

speak	*dance*	*buy*
hablar	bailar	comprar
	bailo	compro
hablamos	bailamos	compramos
hablan		
…	…	…

- What if not all conjugations were in the corpus?

ParaMor: Morphological paradigms

Monson et al. (2007, 2008)

speak	*dance*	*buy*
hablar	bailar	comprar
	bailo	compro
hablamos	bailamos	compramos
hablan		
…	…	…

- Another stage of the algorithm **merges** these overlapping partial paradigms via clustering

ParaMor: Morphological paradigms

Monson et al. (2007, 2008)

speak	*dance*	*buy*
hablar	bailar	comprar
hablo	bailo	compro
hablamos	bailamos	compramos
hablan	bailan	compran
…	…	…

- This amounts to smoothing, or "hallucinating" out-of-vocabulary items

ParaMor: Morphological paradigms
Monson et al. (2007, 2008)

- Heuristic-based, deterministic algorithm can learn inflectional paradigms from raw text
- Currently, ParaMor assumes suffix-based morphology
- Paradigms can be used straightforwardly to predict segmentations
  - Combining the outputs of ParaMor and Morfessor (another system) won the segmentation task at MorphoChallenge 2008 for every language: English, Arabic, Turkish, German, and Finnish

Bayesian word segmentation

Goldwater et al. (2006; in submission)

- Word segmentation results – comparison

	Performance measure								
Model	P	R	F	BP	BR	BF	LP	LR	LF
NGS-u	67.7	**70.2**	68.9	80.6	**84.8**	82.6	52.9	51.3	52.0
MBDP-1	67.0	69.4	68.2	80.3	84.3	82.3	53.6	51.3	52.4
DP	61.9	47.6	53.8	**92.4**	62.2	74.3	57.0	**57.5**	57.2
NGS-b	68.1	68.6	68.3	81.7	82.5	82.1	54.5	57.0	55.7
HDP	**75.2**	69.6	**72.3**	90.3	80.8	**85.2**	**63.5**	55.2	**59.1**

Goldwater et al. Unigram DP

Goldwater et al. Bigram HDP

- See Narges & Andreas's presentation for more on this model

table from Goldwater et al. (in submission)

Multilingual morpheme segmentation Snyder & Barzilay (2008)

speak ES	*speak* FR
hablar	parler
hablo	parle
hablamos	parlons
hablan	parlent
…	…

- Considers **parallel phrases** and tries to find morpheme correspondences
- **Stray morphemes** don't correspond across languages

- **Abstract morphemes** cross languages: (ar, er), (o, e), (amos, ons), (an, ent), (habl, parl)

Morphology Papers: Inputs & Outputs

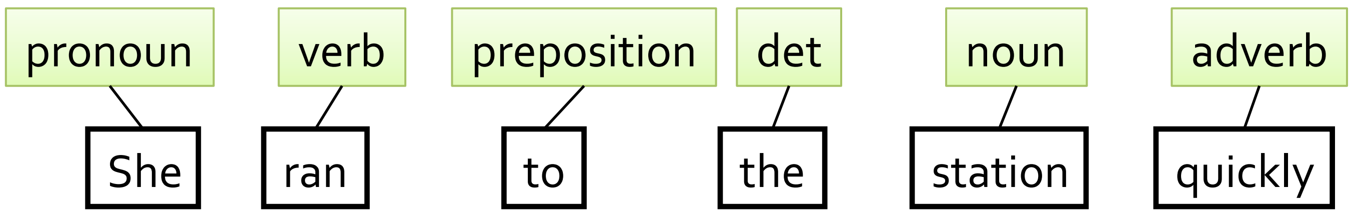
MORPHOLOGY	Monson et al.	Goldwater et al.	Snyder & Barzilay
Phrase/Document-Level			
Unsegmented text		●	
Parallel sentences			◖
Phrasal aligner			◖
Word-Level			
Vocabulary (list of word types)	●		
Sub-Word–Level			
Paradigms	◁		
Segmentations	▷	○	▷
Phonetic correspondences			(●)

Legend	training	test
input	◖	◗
output	◁	▷

- What does "unsupervised" mean for each approach?

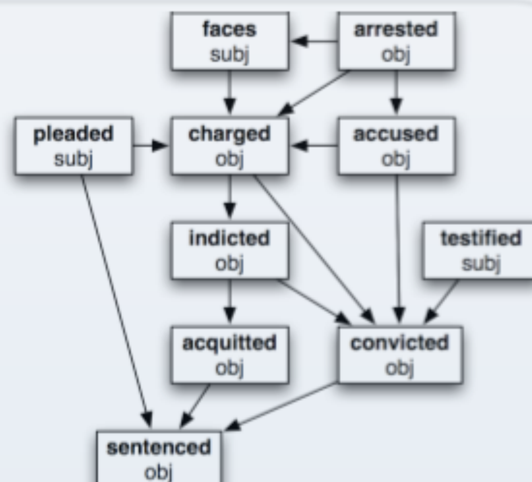
Unsupervised Methods

– Sequence Labeling (Part-of-Speech Tagging)

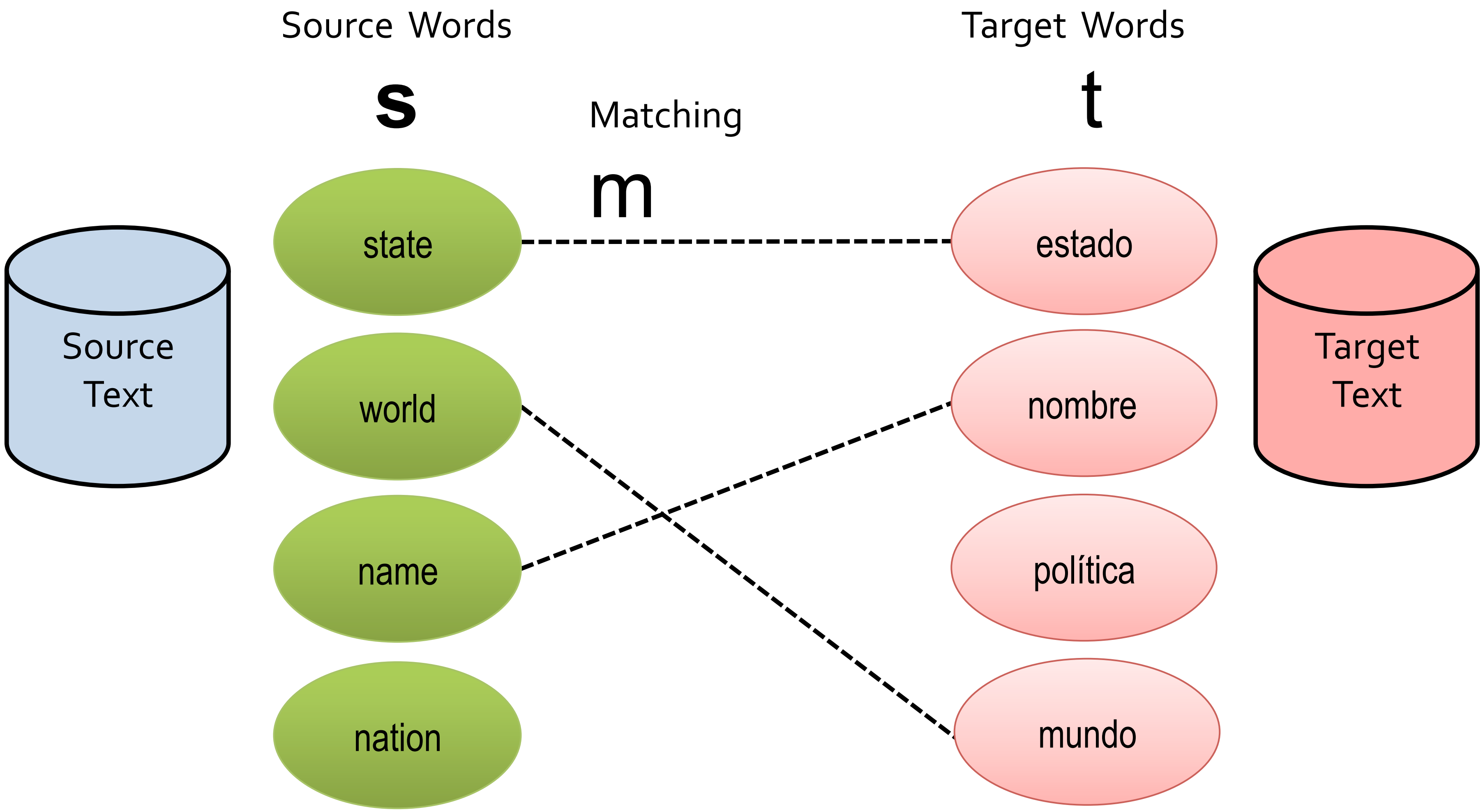


– Morphology Induction   un-supervise-d learn-ing

– Lexical Resource Acquisition



Bilingual lexicons from monolingual corpora Haghighi et al. (2008)

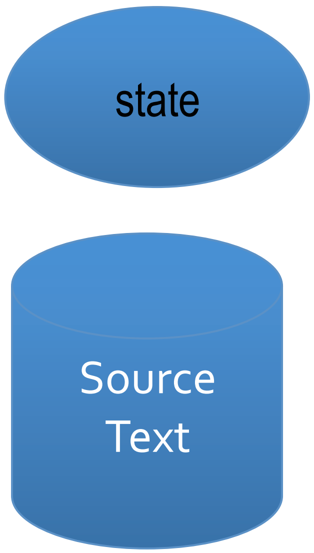


Used a variant of CCA (Canonical Correlation Analysis)

diagram courtesy Haghighi et al.

Bilingual Lexicons from Monolingual Corpora Haghighi et al. (2008)

Data Representation



Orthographic Features

#st	1.0
tat	1.0
te#	1.0

Context Features

world	20.0
politics	5.0
society	10.0

Orthographic Features

#es	1.0
sta	1.0
do#	1.0

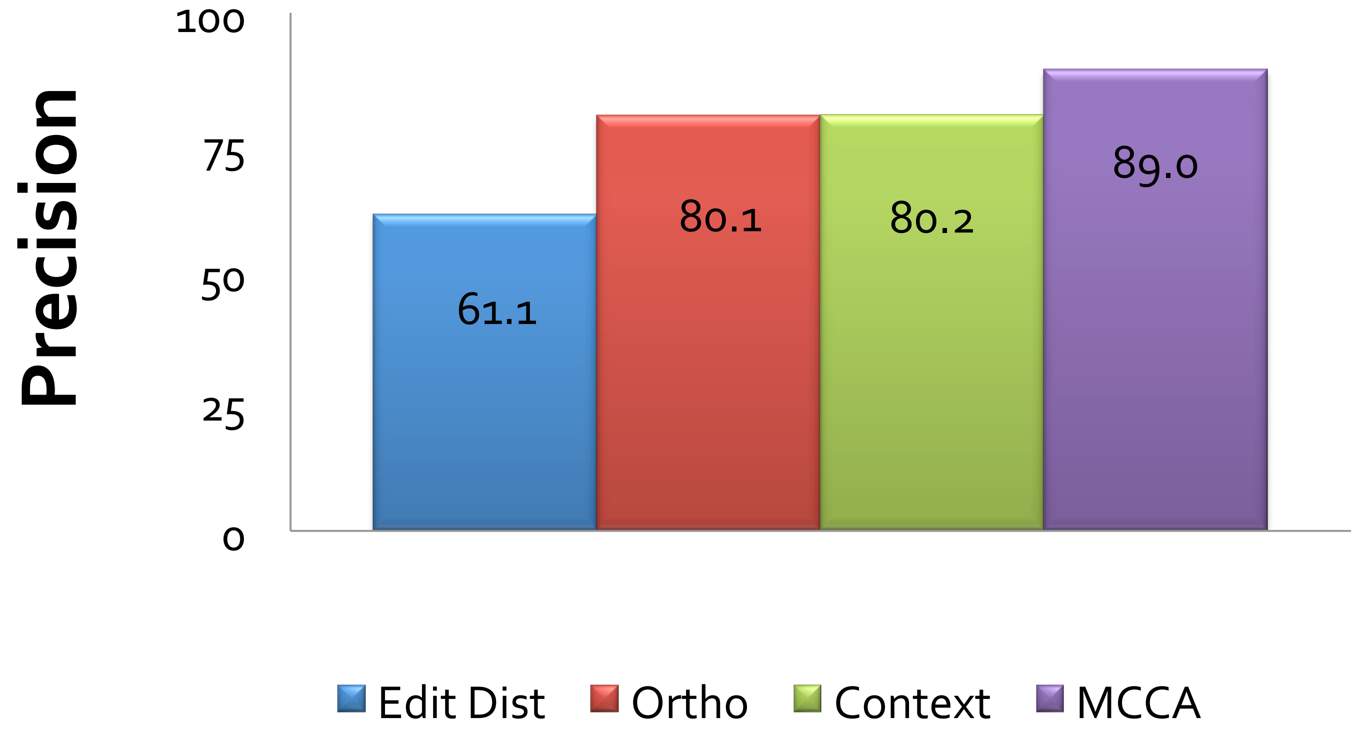
Context Features

mundo	17.0
politica	10.0
sociedad	6.0

slide courtesy Haghighi et al.

Feature Experiments

- MCCA: Orthographic and context features



4k EN-ES Wikipedia Articles

slide courtesy Haghighi et al.

Narrative events

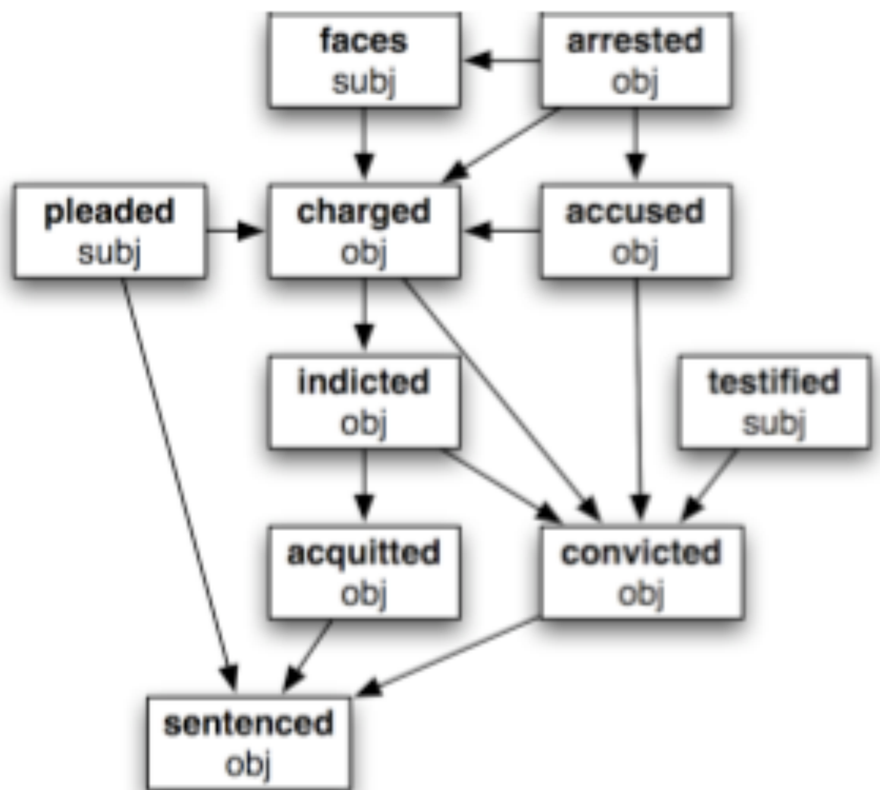
Chambers & Jurafsky (2008)

- Given a corpus, identifies related events that constitute a "narrative" and (when possible) predict their typical temporal ordering
  - E.g.: **CRIMINAL PROSECUTION** narrative, with verbs: **arrest**, **accuse**, **plead**, **testify**, **acquit**/**convict**
- Key insight: related events tend to share a participant in a document
  - The common participant may fill different syntactic/semantic roles with respect to verbs: arrest.**OBJECT**, accuse.**OBJECT**, plead.**SUBJECT**

Narrative events

Chambers & Jurafsky (2008)

- A temporal classifier can reconstruct pairwise canonical event orderings, producing a directed graph for each narrative



Statistical verb lexicon

Grenager & Manning (2006)

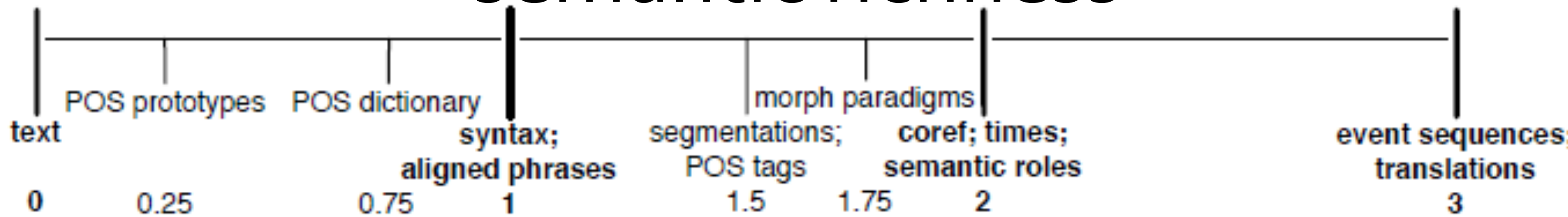
- From dependency parses, a generative model predicts for each verb:
  - PropBank-style semantic roles: ARG0, ARG1, etc. (do not necessarily correspond across verbs)
  - The roles' syntactic realizations, e.g.:

He	gave	me	a cookie
subj	verb	np#1	np#2
ARG0	give	ARG2	ARG1

He	gave	a cookie	to me
subj	verb	np#2	pp_to
ARG0	give	ARG1	ARG2

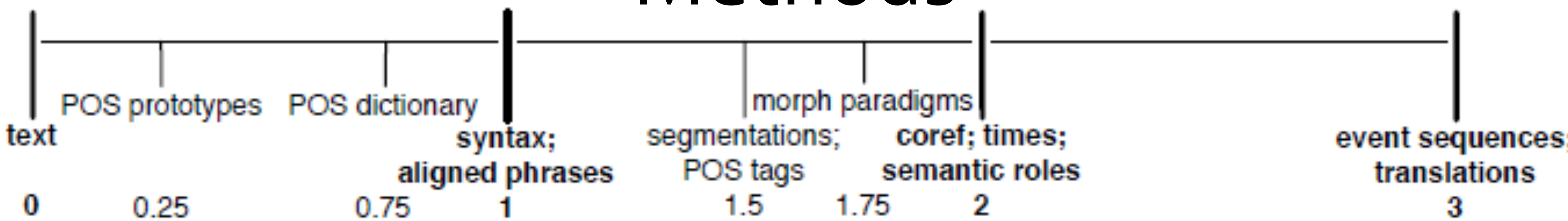
- Used for semantic role labeling

"Semanticity": Our proposed scale of semantic richness



- text < POS < syntax/morphology/alignments < coreference/semantic roles/temporal ordering < translations/narrative event sequences
- We score each model's inputs and outputs on this scale, and call the input-to-output increase **"semantic gain"**
  - Haghighi et al.'s bilingual lexicon induction wins in this respect, going from raw text to lexical translations

Semantic Gain: Comparison of Methods



	Sequences/POS		*Morphology*			*Lexical Resources*		
	S&E	**H&K**	**M+**	**G+**	**S&B**	**H+**	**G&M**	**C&J**
Input semanticity	.75	.25	0	0	1	0	1	2
Output semanticity	1.5	1.5	1.75	1.5	1.5	3	2	3
Semantic gain	.75	1.25	1.75	1.5	.5	3	1	1

Robustness to **language variation**

- About half of the papers we examined had English-only evaluations
- We considered which techniques were most adaptable to other (esp. resource-poor) languages. Two main factors:
  - Reliance on **existing tools/resources** for preprocessing (parsers, coreference resolvers, …)
  - Any **linguistic specificity** in the model (e.g. suffix-based morphology)

Summary

We examined three areas of unsupervised NLP:

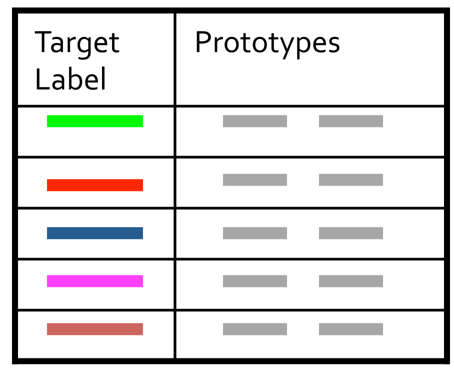
1. **Sequence tagging:** How can we predict POS (or topic) tags for words in sequence?
2. **Morphology:** How are words put together from morphemes (and how can we break them apart)?
3. **Lexical resources:** How can we identify lexical translations, semantic roles and argument frames, or narrative event sequences from text?

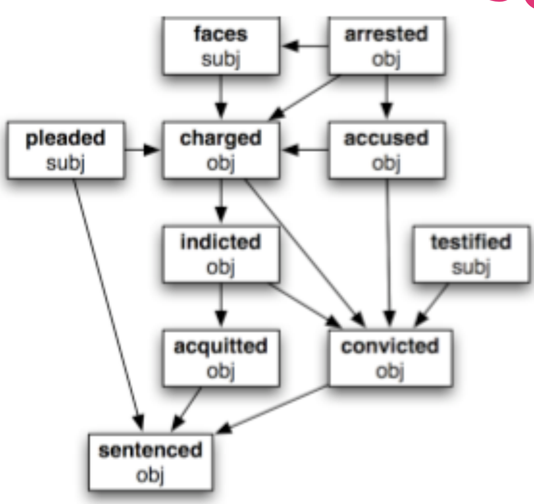
In eight recent papers we found a variety of approaches, including heuristic algorithms, Bayesian methods, and EM-style techniques.

Thanks to Noah and Kevin for their feedback on the paper; Andreas and Narges for their collaboration on the presentations; and all of you for giving us your attention!

subj=give.ARG0 verb=give np#1=give.ARG2 np#2=give.ARG1

un-supervise-d learn-ing





Questions?

hablar	bailar
hablo	bailo
hablamos	bailamos
hablan	bailan

Improvement Ideas

- POS Tagging: Learn the tag set
- Morphology: Non-agglomerative Morphology, Also parses
- Lexical Resources: Try word classes

- All: Language variability