HMM Can Find Pretty Good POS Taggers (When Given a Good Start)

Yoav Goldberg Meni Adler Michael Elhadad

Ben-Gurion University
of the Negev

ACL 2008, Columbus, Ohio

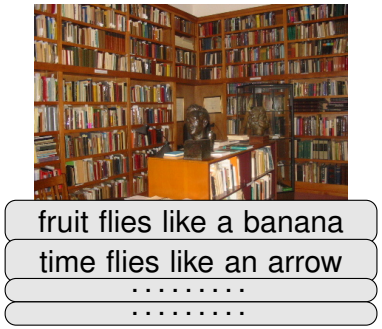
Unsupervised POS Tagging
(If you don't know what POS Tagging is, please leave the room)

Input

- Lots of (unannotated) Text
- A Lexicon
  - Maps words to their possible POS tags
  - Some words may be missing
  - Analyses for a word are not ordered

Output

- A POS Tagger



a:	DET
an:	DT
arrow:	NN
banana:	NN
flies:	NNS VB
fruit:	NN ADJ
like:	VB IN RB JJ
time:	VB NN
...	...

Previous Work – 10-15 years ago

Early Unsupervised POS Tagging

HMM

- Early works on HMM models trained with EM
- Pretty decent results (Merialdo 1994, Elworthy 1994,...)

Transformation Based Learning

- Unsupervised Transformation Based Learning (Brill, 1995)
- This also seemed to work well

Alas, it turns out they were "cheating"

- HMM – use "pruned" dictionaries:
  only probable POS tags are suggested
- Brill – assume knowledge of most-probable-tag per word

This kind of information is based on corpus Counts!

Previous Work – 10-15 years ago

Initial Conditions

- Elworthy shows that good initialization of parameters prior to EM boost results (Elworthy 1994)
- . . . but doesn't tell how it can be done automatically

Context Free Approximation from Raw Data

- Moshe Levinger proposes a way to estimate $p(tag|word)$ from raw data.
- He applies it to Hebrew. (Levinger *et al.*, CL, 1995)

Previous Work – Right About Now

EM/HMMs are Out

- "Why doesn't EM find Good HMM-POS taggers?" (Mark Johnson, EMNLP-2007)

New and **Complicated** Methods are in

- "Contrastive estimation: training log-linear models on unlabeled data" (Smith and Eisner, ACL-2005)
- "A Fully Bayesian Approach to Unsupervised Part-of-Speech Tagging" (Goldwater and Griffiths, ACL-2007)
- "A Bayesian LDA-based model for semi-supervised part-of-speech tagging" (Toutanova and Johnson, NIPS-2007)

Objective: Build a Hebrew POS-Tagger

Hebrew

- Rich Morphology
- Huge Tagset (3k tags)

Building a Hebrew Tagger

- No large annotated corpora
- A fairly comprehensive Lexicon
- An unsupervised approach is called for
- . . . but current works on English are un-realistic for us

Our Take at Unsupervised POS Tagging

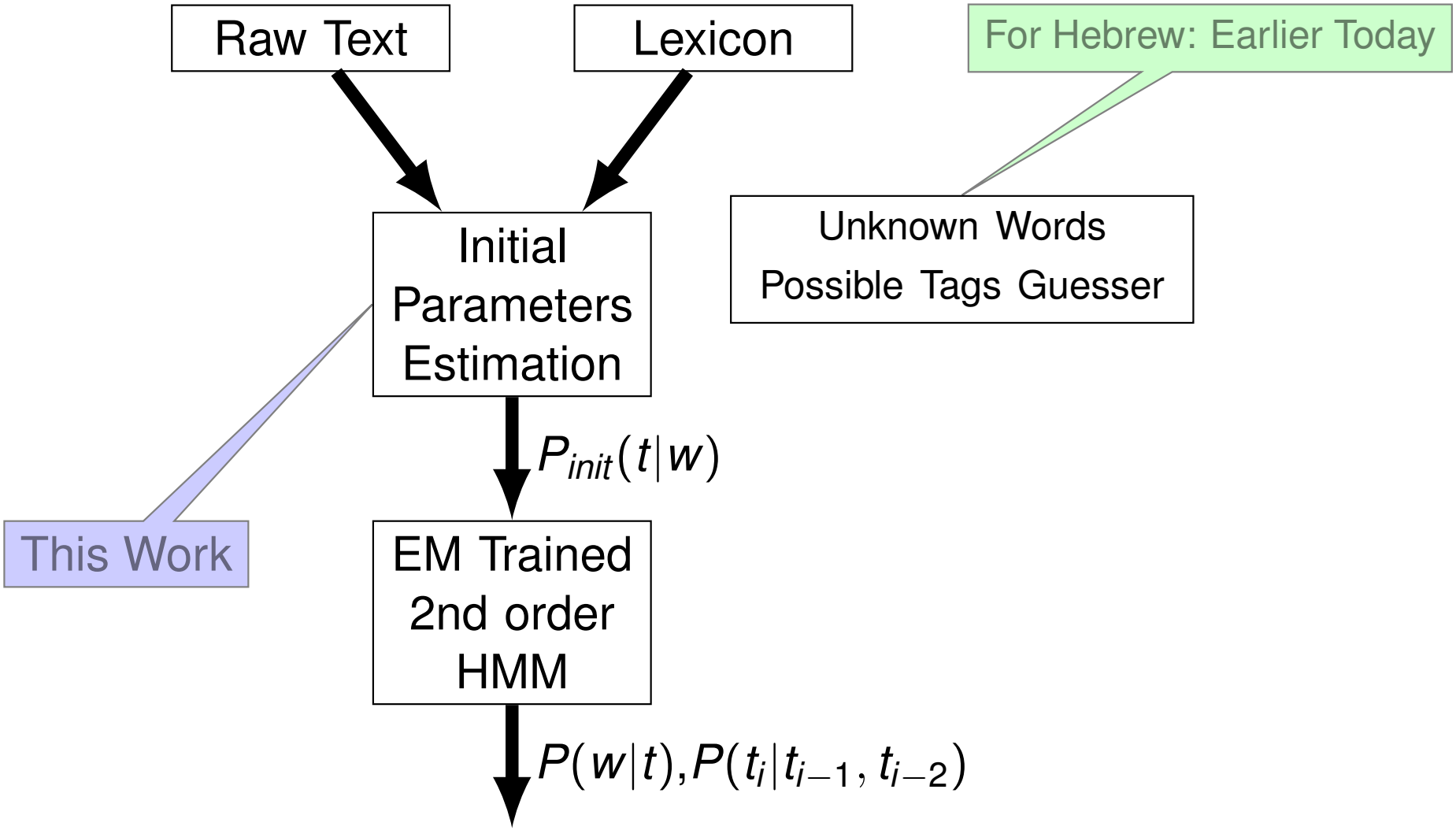
Grandma knows best!

- . . . back to EM trained HMMs
- We just need to find the right initial parameters!

Finding initial parameters

- Improved version of the Levinger algorithm
- a novel iterative context-based estimation method

Much simpler (computationally) than recent methods


Raw Text
Lexicon
For Hebrew: Earlier Today
Unknown Words
Possible Tags Guesser
Initial
Parameters
Estimation
This Work
$P_{init}(t|w)$
EM Trained
2nd order
HMM
$P(w|t), P(t_i|t_{i-1}, t_{i-2})$

Outline

- We can build a good tagger using EM-HMM if we supply good initial conditions
- It works in Hebrew and in English
- Finding initial conditions:
  - Morphology Based
  - Context Based
- Experiments
  - Hebrew
  - English

Morphology based $p(t|w)$

Levinger's "Similar Words" Algorithm

- Language specific algorithm for context-free estimation of $p(t|w)$
- Main intuitions:
  - Morphological variations of words have similar distribution
  - While a form may be ambiguous, some of its inflections aren't
  - $\Rightarrow$ Estimate based on inflected forms

Example: The Hebrew ילדה

is ambiguous between
a `Noun` (girl) and
a `Verb` (gave birth).

- Estimate $p(\texttt{Noun}|$ילדה$)$ by counting:
  הילדה (the girl)
  הילדות (the girls).
- Estimate $p(\texttt{Verb}|$ילדה$)$ by counting:
  תלד (she will give birth)
  ילדו (they gave birth)

(Would probably not work that well for English)

Context Based $p(t|w)$

The Intuition: Distributional Similarity

- Words in similar contexts have similar POS distributions

  (cf. Harris' distributional hypothesis, Schutze's POS induction, *etc.*)

- Previous work:

  what are the possible tags for a given word?

- This work:

  Possible tags are known. Let's rank them.

- In other words:

  We have a guess at $p(t|w)$. Use context to improve it.

Context Based $p(t|w)$

The Algorithm

Start with an initial $p(t|w)$
(1) Using p(t|w), estimate p(t|c)

$$\hat{p}(t|c) = \frac{\sum_{w \in W} p(t|w)\ p(w|c)}{Z}$$

(2) Using p(t|c), estimate p(t|w)

$$\hat{p}(t|w) = \frac{\sum_{c \in REL_C} p(t|c)\ p(c|w)\ \text{allow(t,w)}}{Z}$$

(3) Repeat

Context Based $p($

$p(\text{VB}|the, \_\_\_, run)p(the, \_\_\_, run|kid)+$
$p(\text{VB}|nt, \_\_\_, me)p(nt, \_\_\_, me|kid)+$
$p(\text{VB}|I, \_\_\_, you)p(I, \_\_\_, you|kid)+$
$\ldots$

p(VB | kid)

Th

Start wit an initial $p($

(1) Using p(t|w), estimate p(t|c)

$$\hat{p}(t|c) = \frac{\sum_{w \in W} p(t|w)\ p(w|c)}{Z}$$

(2) Using p(t|c), estimate p(t|w)

$$\hat{p}(t|w) = \frac{\sum_{c \in REL_C} p(t|c)\ p(c|w)\ \text{allow(t,w)}}{Z}$$

(3) Repeat

Ignore contexts with too many possible tags

Follow the Lexicon

Context Based $p(t|w)$

p(NN | the,___,run)

$$p(\text{NN}|boy)p(boy|the,\_\_\_,run)+$$
$$p(\text{NN}|fox)p(fox|the,\_\_\_,run)+$$
$$p(\text{NN}|nice)p(nice|the,\_\_\_,run)+$$
$$\ldots$$

The

Start with an initial $p(t|w)$

(1) Using p(t|w) estimate p(t|c)

$$\hat{p}(t|c) = \frac{\sum_{w \in W} p(t|w)\ p(w|c)}{Z}$$

(2) Using p(t|c), estimate p(t|w)

$$\hat{p}(t|w) = \frac{\sum_{c \in REL_C} p(t|c)\ p(c|w)\ \text{allow(t,w)}}{Z}$$

(3) Repeat

Context Based $p(t|w)$

The Algorithm

Start with an initial $p(t|w)$

(1) Using p(t|w), estimate p(t|c)

$$\hat{p}(t|c) = \frac{\sum_{w \in W} p(t|w)\, p(w|c)}{Z}$$

(2) Using p(t|c), estimate p(t|w)

$$\hat{p}(t|w) = \frac{\sum_{c \in REL_C} p(t|c)\, p(c|w)\, \text{allow(t,w)}}{Z}$$

(3) Repeat

Evaluation

Evaluating the Learned $p(t|w)$

- How does the $p(t|w)$ perform as a Context Free tagger?
  - ContextFreeTagger: $tag(w) = \arg\max_t p(t|w)$

The REAL Evaluation

- How does an EM-HMM tagger initialized with the learned $p(t|w)$ perform?

Hebrew Experiments

How good are the learned $p(t|w)$?

$P_{Unif}(t|w)$ [following the Lexicon]

Levinger's Algorithm

$p(t|c)$
$p(t|w)$

Context Free Tagger

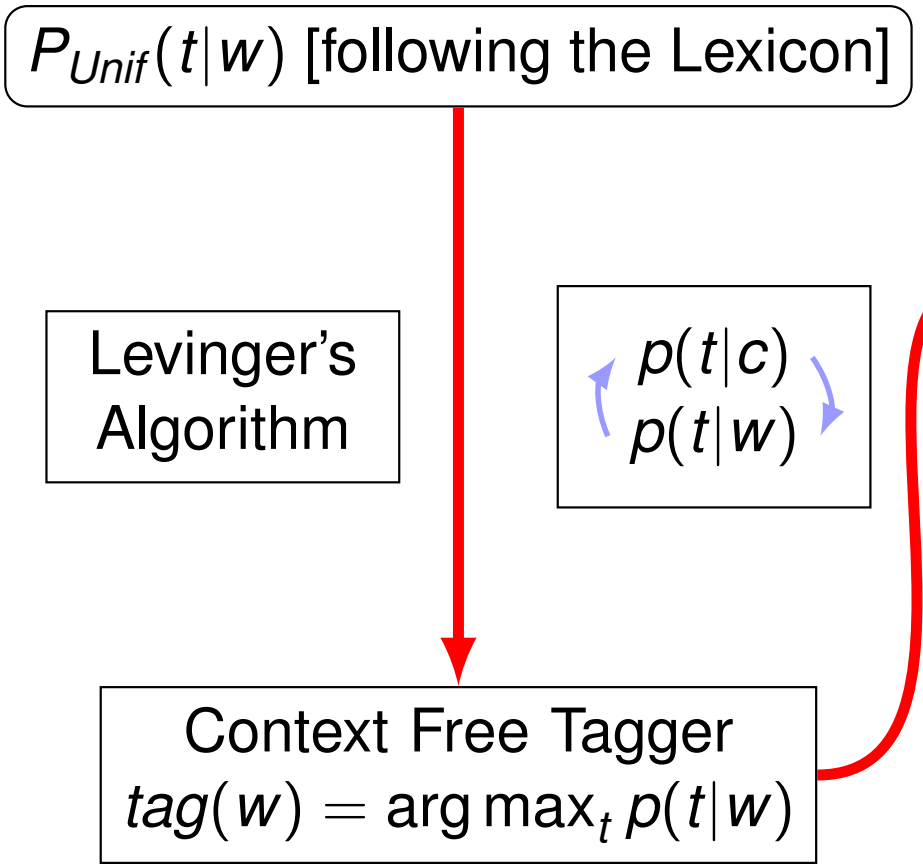
FullMorph	POS+Seg

Context Free Tagger

$tag(w) = \arg\max_t p(t|w)$

Hebrew Experiments

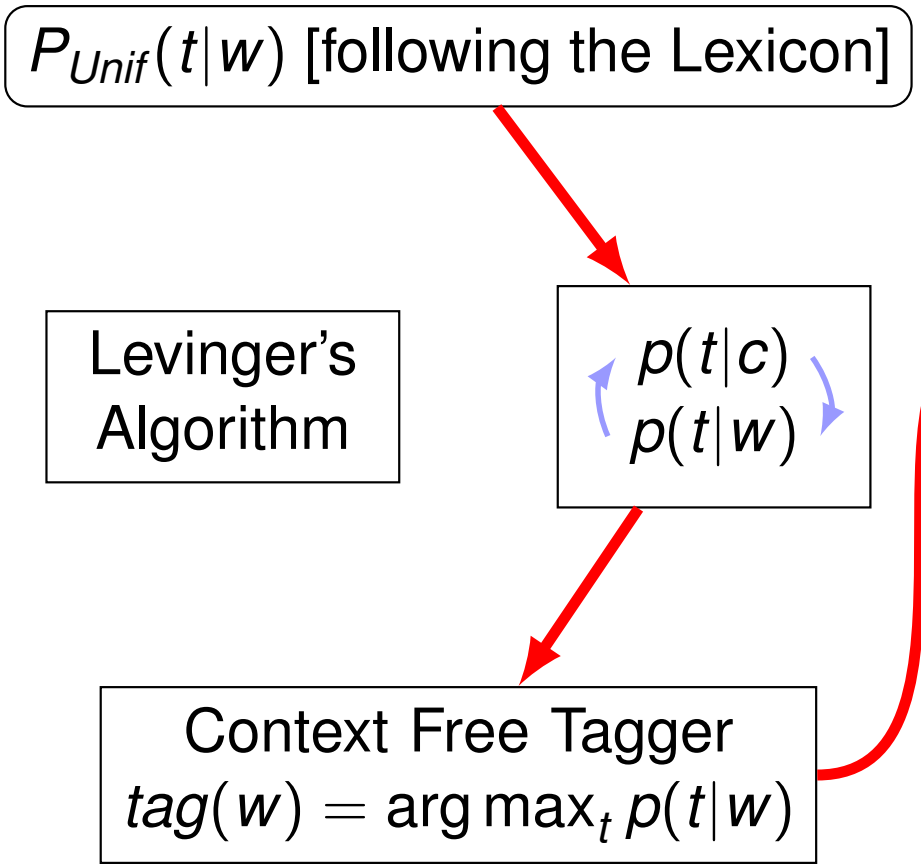
How good are the learned $p(t|w)$?



	Context Free Tagger	
	FullMorph	POS+Seg
Baseline	63.8	71.9
Context	75.4	82.6
Morphology	76.4	83.1
Morph+Cont	79.0	85.5

Hebrew Experiments

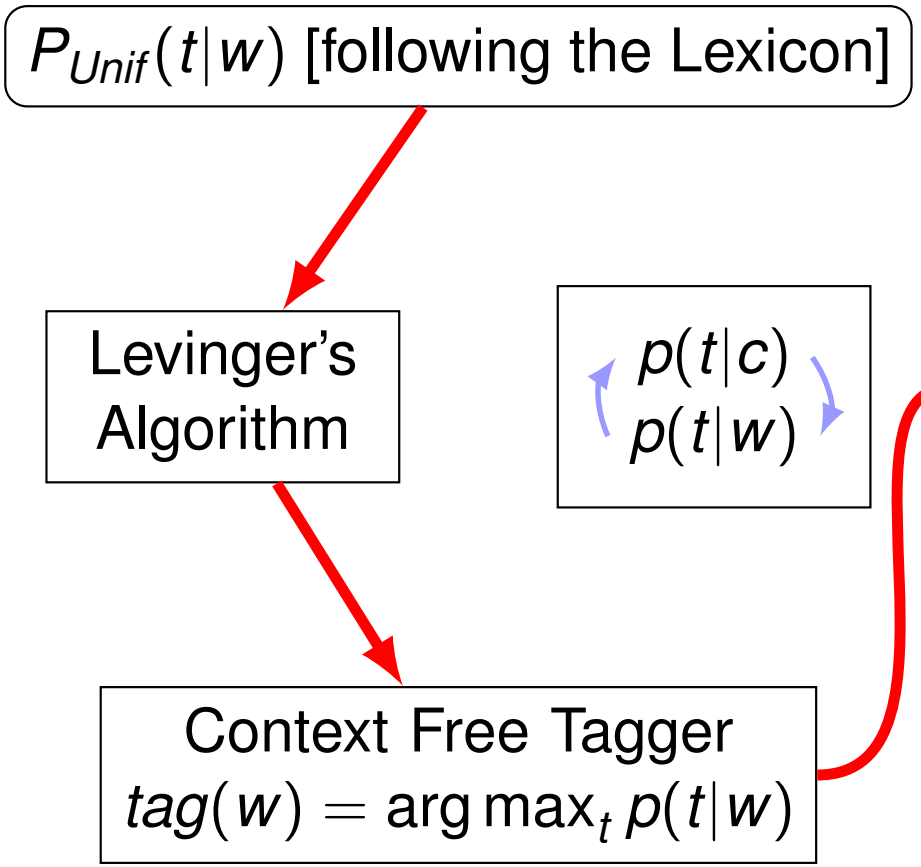
How good are the learned $p(t|w)$?



	Context Free Tagger	
	FullMorph	POS+Seg
Baseline	63.8	71.9
Context	75.4	82.6
Morphology	76.4	83.1
Morph+Cont	79.0	85.5

Hebrew Experiments

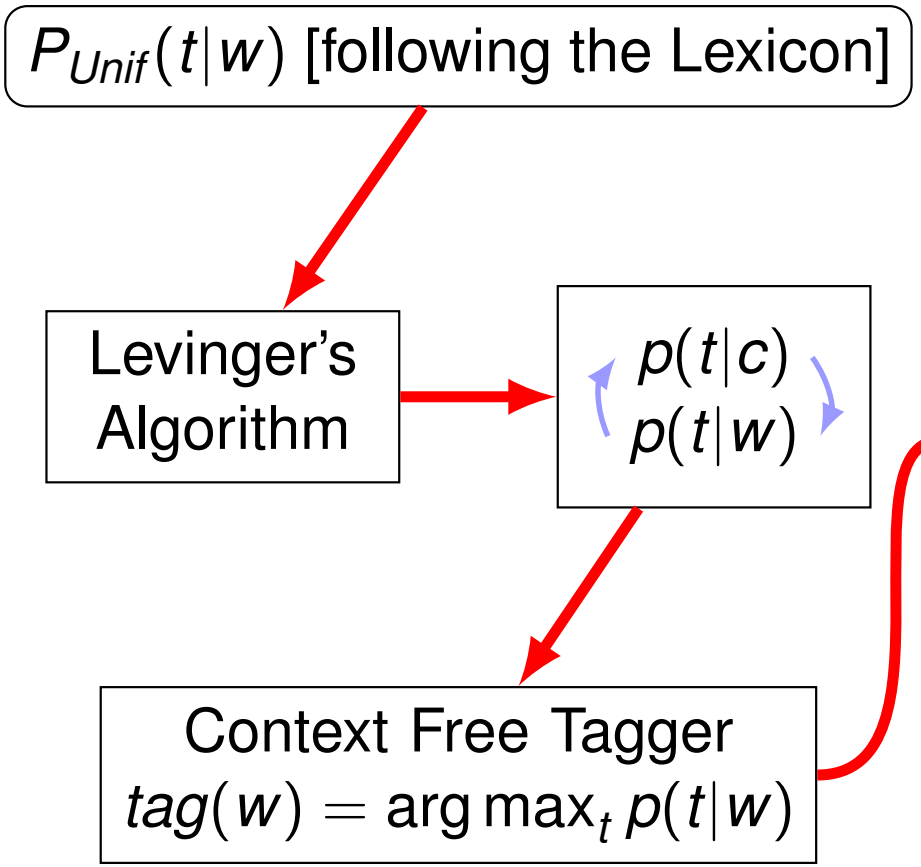
How good are the learned $p(t|w)$?



	Context Free Tagger	
	FullMorph	POS+Seg
Baseline	63.8	71.9
Context	75.4	82.6
Morphology	76.4	83.1
Morph+Cont	79.0	85.5

Hebrew Experiments

How good are the learned $p(t|w)$?



	Context Free Tagger	
	FullMorph	POS+Seg
Baseline	63.8	71.9
Context	75.4	82.6
Morphology	76.4	83.1
Morph+Cont	79.0	85.5

Hebrew Experiments
EM-HMM Tagger

$P_{Unif}(t|w)$ [following the Lexicon]

Levinger's Algorithm

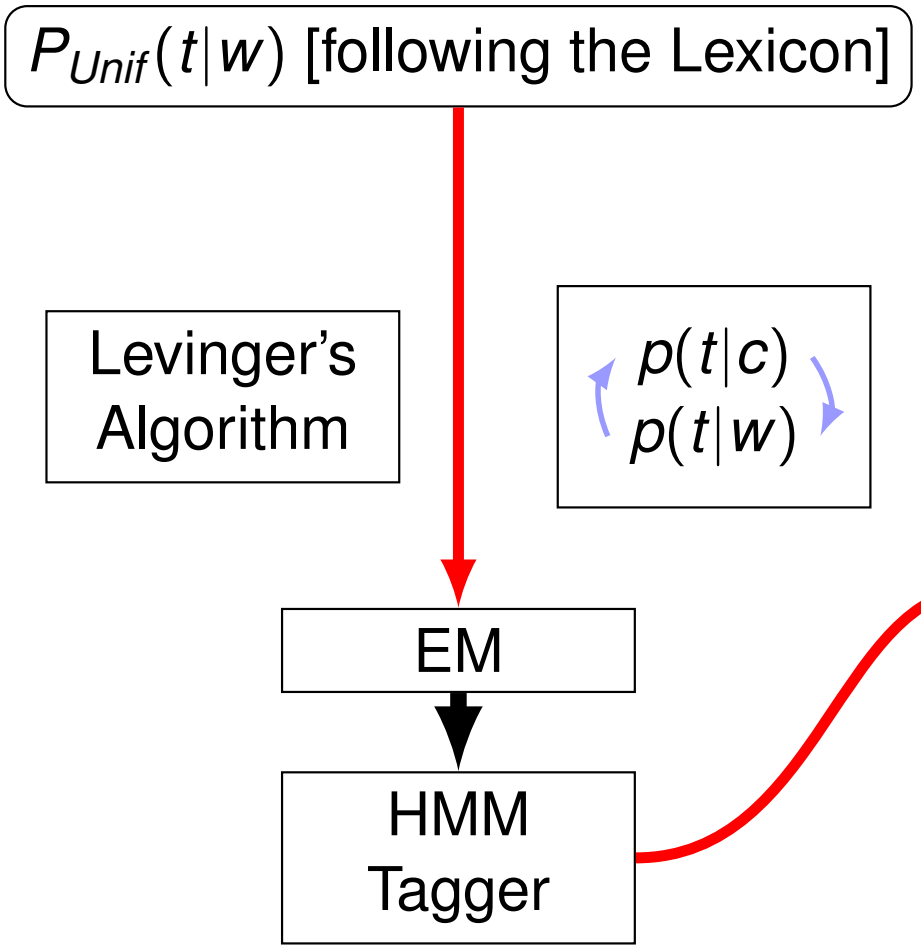
$p(t|c)$
$p(t|w)$

EM

HMM Tagger

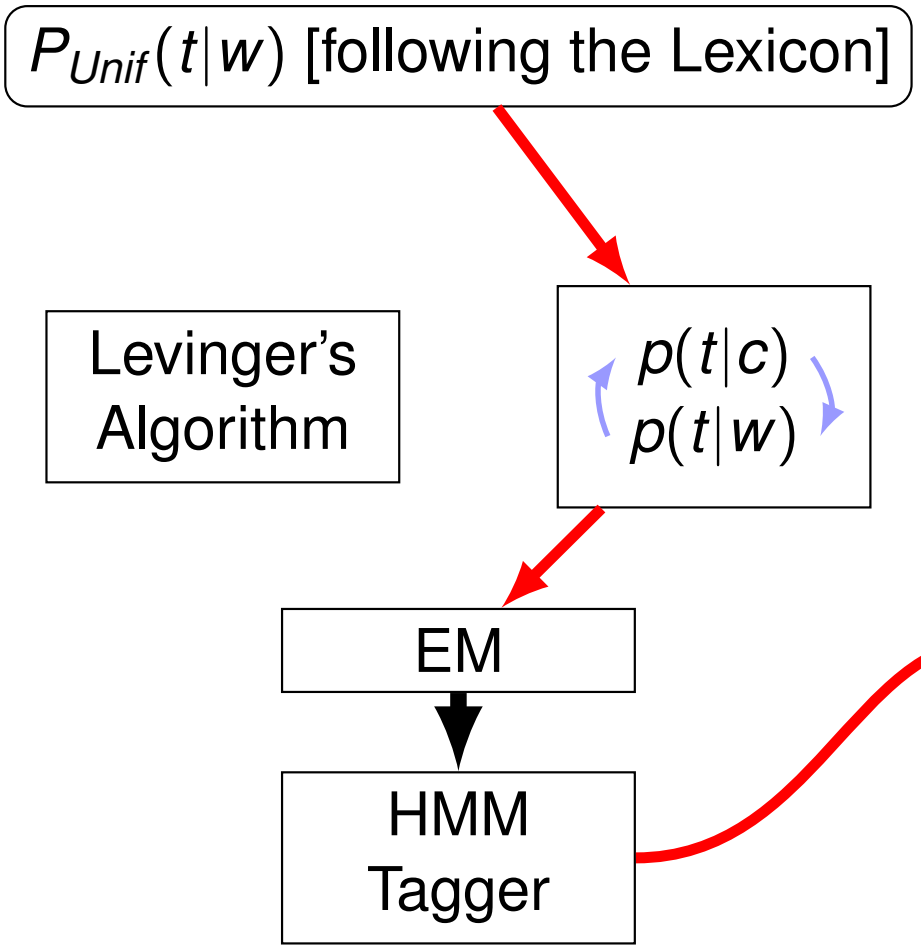
	Context Free Tagger	
	FullMorph	POS+Seg
Baseline	63.8	71.9
Context	75.4	82.6
Morphology	76.4	83.1
Morph+Cont	79.0	85.5
	EM-HMM Tagger	
Baseline	85.5	89.8
Context	85.3	89.6
Morphology	87.7	91.6
Morph+Cont	88.0	92.0

Hebrew Experiments
EM-HMM Tagger



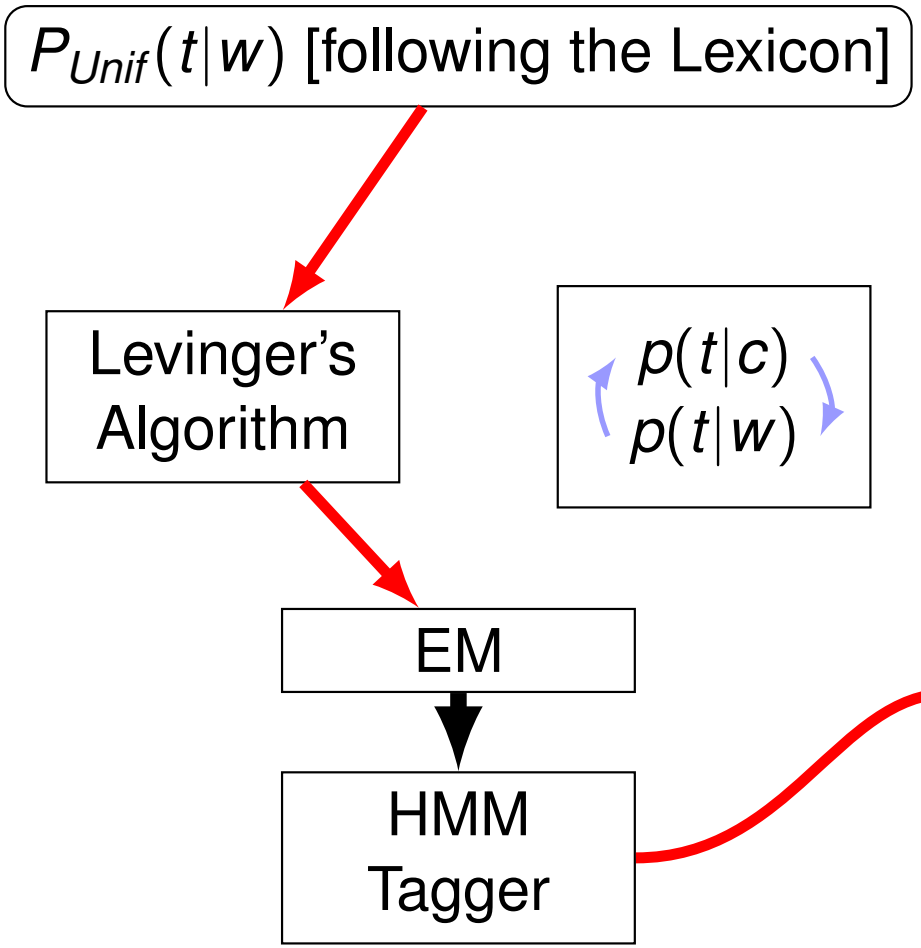
	Context Free Tagger	
	FullMorph	POS+Seg
Baseline	63.8	71.9
Context	75.4	82.6
Morphology	76.4	83.1
Morph+Cont	79.0	85.5
	EM-HMM Tagger	
Baseline	85.5	89.8
Context	85.3	89.6
Morphology	87.7	91.6
Morph+Cont	88.0	92.0

Hebrew Experiments
EM-HMM Tagger



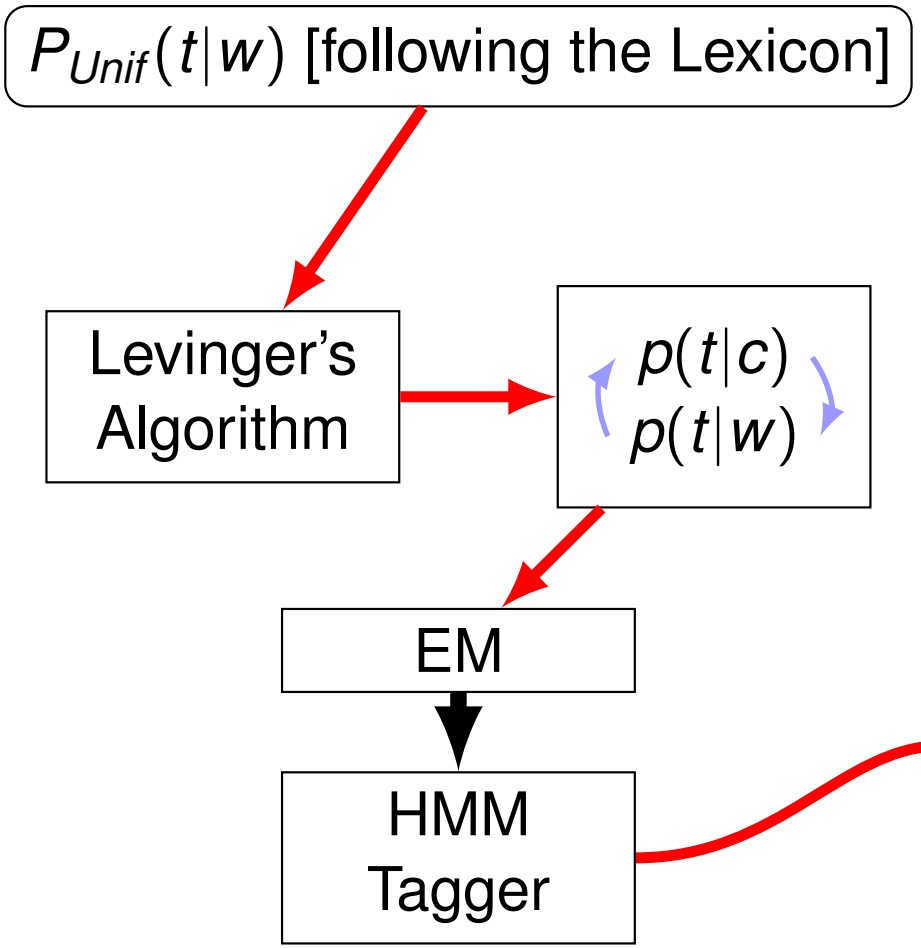
	Context Free Tagger	
	FullMorph	POS+Seg
Baseline	63.8	71.9
Context	75.4	82.6
Morphology	76.4	83.1
Morph+Cont	79.0	85.5
	EM-HMM Tagger	
Baseline	85.5	89.8
Context	85.3	89.6
Morphology	87.7	91.6
Morph+Cont	88.0	92.0

Hebrew Experiments
EM-HMM Tagger



	Context Free Tagger	
	FullMorph	POS+Seg
Baseline	63.8	71.9
Context	75.4	82.6
Morphology	76.4	83.1
Morph+Cont	79.0	85.5
	EM-HMM Tagger	
Baseline	85.5	89.8
Context	85.3	89.6
Morphology	87.7	91.6
Morph+Cont	88.0	92.0

Hebrew Experiments
EM-HMM Tagger



	Context Free Tagger	
	FullMorph	POS+Seg
Baseline	63.8	71.9
Context	75.4	82.6
Morphology	76.4	83.1
Morph+Cont	79.0	85.5
	EM-HMM Tagger	
Baseline	85.5	89.8
Context	85.3	89.6
Morphology	87.7	91.6
Morph+Cont	88.0	92.0

EM-HMM Produced a Pretty Good POS Tagger for Hebrew

How about English?

Unsupervised English POS Tagging

English is Different Than Hebrew

- Much smaller Tagset
  - Recent Supervised Work: 46 tags (WSJ)
  - Recent UN-Supervised Work: 17 tags (a subset)
- Lexicon is Derived from Corpus
- We don't have as rich morphology to rely on
  - Rely more on linear context
  - . . . but we learned from Hebrew that morphology is important to EM-HMM

Unsupervised English POS Tagging

Morphology $p(t|w)$: data driven, suffixation and function words

suff=S	word has suffix S	suff=ing
L+suff=W,S	word appear after word W, with suffix S	L+suff=have,ed
R+suff=S,W	word appear before word W, with suffix S	L+suff=ing,to
wsuff=S1,S2	word has suffix S1, same stem seen with S2	wsuff=ϵ,s
suffs=SG	word stem appear with the SG group of suffixes	suffs=ed,ing,s

Context $p(t|w)$ templates:

LL=w_{-2},w_{-1}	2 preceding words
LL=w_{+1},w_{+2}	2 following words
LL=w_{-1},w_{+1}	2 surrounding words

Morph+Cont $p(t|w)$ contexts:

The union of the two groups

All $p(t|w)$ estimates are obtained from the Context algorithm, by using different context templates.

Unsupervised English POS Tagging

Following Smith and Eisner (2005), recent works use a

17 tags tagset

```
ADJ ADV CONJ DET ENDPUNC INPUNC LPUNC RPUNC N
POS PRT PREP PRT TO V VBG VBN WH
```

In general, English does not allow V-V transitions.
But this tagset does as it include `Modals` among the `Verbs`.

Unsupervised English POS Tagging

Following Smith and Eisner (2005), recent works use a

17 tags tagset

ADJ ADV CONJ DET ENDPUNC INPUNC LPUNC RPUNC N
POS PRT PREP PRT TO V VBG VBN WH

We help the $p(t|t_{-1}, t_{-2})$ estimation by introducing a

19-tags tagset

ADJ ADV CONJ DET ENDPUNC INPUNC LPUNC RPUNC N
POS PRT PREP PRT TO V VBG VBN WH MD BE

Unsupervised English POS Tagging

Following Smith and Eisner (2005), recent works use a

17 tags tagset

ADJ ADV CONJ DET ENDPUNC INPUNC LPUNC RPUNC N
POS PRT PREP PRT TO V VBG VBN WH

We help the $p(t|t_{-1}, t_{-2})$ estimation by introducing a

19-tags tagset

ADJ ADV CONJ DET ENDPUNC INPUNC LPUNC RPUNC N
POS PRT PREP PRT TO V VBG VBN WH MD BE

We also test on the complete WSJ (+BE) tagset.

Unsupervised English POS Tagging

Results – Full Lexicon (49206 words)

(Setting as in Toutanova and Johnson 2007)

17 tags

	CF-Tag	**EM-HMM**
Baseline	81.7	88.7
Context	90.1	92.9
Morphology	82.2	88.6
Morph+Cont	89.9	93.3

- Initializations improve the baseline.
- Morphology much weaker than context.
- But their combinations is superior.

Unsupervised English POS Tagging

Results – Full Lexicon (49206 words)

(Setting as in Toutanova and Johnson 2007)

	17 tags		19 tags	
	CF-Tag	**EM-HMM**	**CF-Tag**	**EM-HMM**
Baseline	81.7	88.7	79.9	91.0
Context	90.1	92.9	88.4	93.7
Morphology	82.2	88.6	80.5	89.2
Morph+Cont	89.9	93.3	88.0	93.8

- Context Free Tagging decreases a little.
- EM-HMM tagging improves considerably.

Unsupervised English POS Tagging

Results – Full Lexicon (49206 words)

(Setting as in Toutanova and Johnson 2007)

	17 tags		19 tags		WSJ tags	
	CF-Tag	**EM-HMM**	**CF-Tag**	**EM-HMM**	**CF-Tag**	**EM-HMM**
Baseline	81.7	88.7	79.9	91.0	76.7	88.3
Context	90.1	92.9	88.4	93.7	85.5	91.2
Morphology	82.2	88.6	80.5	89.2	74.8	88.8
Morph+Cont	89.9	93.3	88.0	93.8	85.9	91.4

- Naturally, not as good as the smaller tagsets.
- But a pretty decent result.

Unsupervised English POS Tagging
Results – Small Lexicon

Following recent work, we experimented also with smaller lexicons (2141 and 1249 words).

Unknowns words Guessing

During initial $p(t|w)$ estimation:

- Allow all open-class tags for unknown words.

During EM-HMM estimation:

- Use a simple ambiguity class guesser:
  - All open class tags that appear with the word's suffix in the Lexicon.
  - suffix: the longest (up to 3 chars) suffix which also appear in the top-100 suffixes in the Lexicon.

Unsupervised English POS Tagging

Results – Small Lexicon (1249 words)

(Setting as in Toutanova and Johnson 2007)

	17 tags		19 tags		WSJ tags	
	CF-Tag	**EM-HMM**	**CF-Tag**	**EM-HMM**	**CF-Tag**	**EM-HMM**
Baseline	62.5	79.6	60.7	84.7	55.7	*
Context	78.3	85.8	76.3	86.9	70.1	82.2
Morphology	69.1	81.7	67.5	87.1	61.9	80.3
Morph+Cont	81.1	86.4	79.2	87.4	72.4	83.3

Unsupervised English POS Tagging

Results – Small Lexicon (1249 words)

(Setting as in Toutanova and Johnson 2007)

	17 tags		19 tags		WSJ tags	
	CF-Tag	**EM-HMM**	**CF-Tag**	**EM-HMM**	**CF-Tag**	**EM-HMM**
Baseline	62.5	79.6	60.7	84.7	55.7	*
Context	78.3	85.8	76.3	86.9	70.1	82.2
Morphology	69.1	81.7	67.5	87.1	61.9	80.3
Morph+Cont	81.1	86.4	79.2	87.4	72.4	83.3

- Overall consistent trends.
- As expected, results much lower.
- Morphology estimation is much more important in this setting

Unsupervised English POS Tagging

Results – Comparison

(Setting as in Toutanova and Johnson 2007)

InitEM-HMM	This work, 19tags, Morph+Cont
LDA(+AC),PLSA+AC	Toutanova and Johnson 2007. **AC**: ambiguity class model
CE+spl	Smith and Eisner 2005
BHMM	Goldwater and Griffiths 2007

Lexicon	InitEM-HMM	LDA	LDA+AC	PLSA+AC	CE+spl	BHMM
Full	**93.8**	93.4	93.4	89.7	88.7	87.3
2141	89.4	87.4	**91.2**	87.8	79.5	79.6
1249	87.4	85.0	**89.7**	85.9	78.4	71.0

Unsupervised English POS Tagging

Results – Comparison

(Setting as in Toutanova and Johnson 2007)

InitEM-HMM	This work, 19tags, Morph+Cont
LDA(+AC),PLSA+AC	Toutanova and Johnson 2007. **AC**: ambiguity class model
CE+spl	Smith and Eisner 2005
BHMM	Goldwater and Griffiths 2007

Lexicon	InitEM-HMM	LDA	LDA+AC	PLSA+AC	CE+spl	BHMM
Full	**93.8**	93.4	93.4	89.7	88.7	87.3
2141	89.4	87.4	**91.2**	87.8	79.5	79.6
1249	87.4	85.0	**89.7**	85.9	78.4	71.0

- Best results for the Full Lexicon case
- 2nd best for the small lexicons
  - The better model has a much stronger unknown words guesser

Unsupervised English POS Tagging

Results – "Realistic" Lexicon

Model	Init-HMM, Morph+Cont
Lexicon	From sections 0-18 of WSJ
Train	Complete, unannotated WSJ
Test	Sections 22-24

19 tags: 92.85%

46 tags: 91.30%

(Highest that we know of)

To Conclude

Take-home message

- EM-HMM Can Produce Pretty Good Unsupervised POS Taggers
- . . . But it needs a good starting point
- . . . which we show how to estimate

Results

- State of the art tagger for Hebrew
- State of the art unsupervised tagger for English
- Considerably raising the EM-HMM baseline

Now what?

Future

- Better unknowns guesser for English
- Different learning approaches on top of our initial parameters:
  - Bayesian
  - Prototype based learning
- Apply the Context algorithm for other problems

Questions?

(Prague, 2007)