Enhancing Unlexicalized Parsing Performance using a Wide Coverage Lexicon, Fuzzy Tag-set Mapping, and EM-HMM-based Lexical Probabilities

Yoav Goldberg  Reut Tsarfaty
Meni Adler  Michael Elhadad

Ben Gurion University
University of Amsterdam

EACL 2009, Athens
Unlexicalized Hebrew Parsing
Parsing with PCFGs
Basic stuff you probably already know

Learning
- Start with a Treebank
Parsing with PCFGs
Basic stuff you probably already know

Learning

- Start with a Treebank
- Extract a Grammar

S → NP VP
NP → DT NN
VP → VB NP
... 
DT → the
NN → cat
NN → cake
NN → dog
VB → ate
VB → kicked

Yoav Goldberg, Reut Tsarfaty, Meni Adler, Michael Elhadad
Parsing with an external Lexicon
Parsing with PCFGs
Basic stuff you probably already know

Learning
- Start with a Treebank
- Extract a Grammar
- Assign probabilities to rules

Inference
Standard CKY stuff

```
S → NP VP 0.2
NP → DT NN 0.04
VP → VB NP 0.5
...
DT → the 0.1
NN → cat 0.002
NN → cake 0.005
NN → dog 0.003
VB → ate 0.08
VB → kicked 0.09
```
Parsing with PCFGs
Basic stuff you probably already know

Learning
- Start with a Treebank
- Extract a Grammar
- Assign probabilities to rules

Inference
Standard CKY stuff

S \to NP \ VP \quad 0.2
NP \to DT \ NN \quad 0.04
VP \to VB \ NP \quad 0.5
...
DT \to \text{the} \quad 0.1
NN \to \text{cat} \quad 0.002
NN \to \text{cake} \quad 0.005
NN \to \text{dog} \quad 0.003
VB \to \text{ate} \quad 0.08
VB \to \text{kicked} \quad 0.09
Parsing with PCFGs
Two kinds of rules

Syntactic Rules
- Finite (small) set of symbols
- Relative frequency estimates + some smoothing works fine

Lexical Rules
- Huge set of terminal symbols
- Problem with rare events
  - Sparsity
  - Overfitting

S → NP VP 0.2
NP → DT NN 0.04
VP → VB NP 0.5
...
DT → the 0.1
NN → cat 0.002
NN → cake 0.005
NN → dog 0.003
VB → ate 0.08
VB → kicked 0.09
Parsing with PCFGs
Two kinds of rules

**Syntactic Rules**
- Finite (small) set of symbols
- Relative frequency estimates + some smoothing works fine

**Lexical Rules**
- Huge set of terminal symbols
- Problem with rare events
  - Sparsity
  - Overfitting

**Focus of this work**

**Parsing with an external Lexicon**

- $S \rightarrow \text{NP} \text{ VP}$
- $\text{NP} \rightarrow \text{DT} \text{ NN}$
- $\text{VP} \rightarrow \text{VB} \text{ NP}$
- ...
- $\text{DT} \rightarrow \text{the}$
- $\text{NN} \rightarrow \text{cat}$
- $\text{NN} \rightarrow \text{cake}$
- $\text{NN} \rightarrow \text{dog}$
- $\text{VB} \rightarrow \text{ate}$
- $\text{VB} \rightarrow \text{kicked}$
Affixation:
- *and, from, to, the, which, as, in* are prefixes
- *possessives* are suffixed to nouns
Affixation:
- *and, from, to, the, which, as, in* are prefixes
- *possessives* are suffixed to nouns

In her net ⇒ inhernet
Affixation:
- *and, from, to, the, which, as, in* are prefixes
- *possessives* are suffixed to nouns

**In her net ⇒ inhernet**

Unvocalized writing system
- most vowels are “dropped” in writing
Affixation:
- *and, from, to, the, which, as, in* are prefixes
- *possessives* are suffixed to nouns

In her net ⇒ inhernet

Unvocalized writing system
- most vowels are “dropped” in writing

in her net ⇒ inhernet ⇒ inhrnt
Affixation:
- *and, from, to, the, which, as, in* are prefixes
- *possessives* are suffixed to nouns

In her net ⇒ inhernet

Unvocalized writing system
- most vowels are “dropped” in writing

in her net ⇒ inhernet ⇒ inhrnt

in her net? in her note? in her night? inherent?
A piece of Hebrew
In (mostly) English words

- Affixation:
  - *and, from, to, the, which, as, in* are prefixes
  - *possessives* are suffixed to nouns

In her net ⇒ inhernet

- Unvocalized writing system
  - most vowels are “dropped” in writing

in her net ⇒ inhernet ⇒ inhrnt

- Rich morphology
  - *inherent* could be inflected into different forms according to sing/pl, masc/fem properties

inhrnt, inhrnti, inhrntit, inrntiot, inhrntim
Affixation:
- *and, from, to, the, which, as, in* are prefixes
- *possessives* are suffixed to nouns

In her net \(\Rightarrow\) inhernet

Unvocalized writing system
- most vowels are “dropped” in writing

in her net \(\Rightarrow\) inhernet \(\Rightarrow\) inhrnt

Rich morphology
- *inherent* could be inflected into different forms according to sing/pl, masc/fem properties

inhrnt, inhrnti, inhrntit, inrntiot, inhrntim

Especially complex verb morphology
- Root + template morphology for verbs

ktb \(\Rightarrow\) ktb mktyb ywktd b tw ktb kw ktb yktwb ykw kwb ...
Tying it together . . .

The situation in Hebrew

- Complex, productive morphology
- Many word forms (487K distinct tokens in a 34M words corpus)
- High level of ambiguity
  - 2.7 tags/token, vs. 1.4 in English
- POS carries a lot of information
  - gender, number, tense, possesiveness, status,…

Yoav Goldberg, Reut Tsarfaty, Meni Adler, Michael Elhadad

Parsing with an external Lexicon
The situation in Hebrew

- Complex, productive morphology
- Many word forms (487K distinct tokens in a 34M words corpus)
- High level of ambiguity
  - 2.7 tags/token, vs. 1.4 in English
- POS carries a lot of information
  - gender, number, tense, possessiveness, status,

which means

Treebank derived lexicon is inadequate

- Low coverage ⇒ Many unseen events
- Hard to guess POS of unknown words
some baseline parsing performance

but first...
Our parsing setup

Data: Hebrew Treebank V2 (~ 6000 sentences)

Syntactic Rules (Goldberg and Tsarfaty 2008)

Parent annotation

Linguistically motivated state splits

$p(X \rightarrow Y)$: relative frequency estimate

Stable lexical items (seen ≥ $K$ times in treebank)

$p(tag \rightarrow word) = p_{rf}(word | tag)$

Rare/unseen lexical items (seen < $K$ times)

Yoav Goldberg, Reut Tsarfaty, Meni Adler, Michael Elhadad

Parsing with an external Lexicon
### Our parsing setup

**Data:** Hebrew Treebank V2 (~ 6000 sentences)

**Syntactic Rules** (Goldberg and Tsarfaty 2008)

- Parent annotation
- Linguistically motivated state splits
- $p(X \rightarrow Y)$: relative frequency estimate (unsmoothed)

**Stable lexical items** (seen ≥ $K$ times in treebank)

$$p(tag \rightarrow word) = p_{rf}(word | tag)$$

**Rare/unseen lexical items** (seen < $K$ times)
Our parsing setup

Data: Hebrew Treebank V2 (~ 6000 sentences)

Syntactic Rules (Goldberg and Tsarfaty 2008)
- Parent annotation
- Linguistically motivated state splits
- $p(X \rightarrow Y)$: relative frequency estimate (unsmoothed)

Stable lexical items (seen $\geq K$ times in treebank)

Rare/unseen lexical items (seen $< K$ times)
Our parsing setup

Data: Hebrew Treebank V2 (∼ 6000 sentences)

Syntactic Rules (Goldberg and Tsarfaty 2008)

- Parent annotation
- Linguistically motivated state splits
- \( p(X \rightarrow Y) \): relative frequency estimate (unsmoothed)

Stable lexical items (seen \( \geq K \) times in treebank)

\[ p(tag \rightarrow word) = p_{rf}(word|tag) \]

Rare/unseen lexical items (seen < \( K \) times)
Our parsing setup

Data: Hebrew Treebank V2 (∼ 6000 sentences)

Syntactic Rules (Goldberg and Tsarfaty 2008)
- Parent annotation
- Linguistically motivated state splits
- \( p(X \rightarrow Y) \): relative frequency estimate (unsmoothed)

Stable lexical items (seen \( \geq K \) times in treebank)

\[ p(tag \rightarrow word) = p_{rf}(word|tag) \]

Rare/unseen lexical items (seen \( < K \) times)
Our parsing setup

Data: Hebrew Treebank V2 (∼ 6000 sentences)

Syntactic Rules (Goldberg and Tsarfaty 2008)
- Parent annotation
- Linguistically motivated state splits
- \( p(X \rightarrow Y) \): relative frequency estimate (unsmoothed)

Stable lexical items (seen \( \geq \) \( K \) times in treebank)
\[ p(tag \rightarrow word) = p_{rf}(word|tag) \]

Rare/unseen lexical items (seen < \( K \) times)
???
Is the low-coverage of the TB lexicon really a problem?

**Easy baseline: assuming a segmentation Oracle**

Input Sentence: *inhrnt*
Parser sees: *in hr nt*

**Model**

- rare/unknown items replaced with RARE token
- \[ p(tag \rightarrow word) = \text{distribution over rare words:} \]
  
  \[
p(word|tag) = \begin{cases} 
p_{rf}(RARE|tag) & \text{rare} \\
p_{rf}(word|tag) & \text{otherwise} \end{cases}
\]
Is the low-coverage of the TB lexicon really a problem?

**Easy baseline: assuming a segmentation Oracle**

Input Sentence: *inhrtnt*
Parser sees: *in hr nt*

**Model**
- rare/unknown items replaced with RARE token
- \( p(tag \rightarrow word) = \) distribution over rare words:

\[
p(word|tag) = \begin{cases} 
  p_{rf}(RARE|tag) & \text{rare} \\
  p_{rf}(word|tag) & \text{otherwise}
\end{cases}
\]

\[72.24\] \(F\) (evalb score)

Yoav Goldberg, Reut Tsarfaty, Meni Adler, Michael Elhadad

Parsing with an external Lexicon
Is the low-coverage of the TB lexicon really a problem?

**Realistic** baseline: no Oracles

<table>
<thead>
<tr>
<th>Input Sentence:</th>
<th>inhrnt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parser sees:</td>
<td>inhrnt</td>
</tr>
</tbody>
</table>
Is the low-coverage of the TB lexicon really a problem?

**Realistic baseline: no Oracles**

Input Sentence: inhrt
Parser sees: inhrt

**Model**

Model of Goldberg and Tsarfaty (2008)
- lattice parser
- non-trivial treebank-based morphological analyzer
  - extended with a spellchecker wordlist
- for details, see paper
Is the low-coverage of the TB lexicon really a problem?

**Realistic baseline: no Oracles**

- Input Sentence: `inhrent`
- Parser sees: `inhrent`

**Model**

Model of Goldberg and Tsarfaty (2008)

- lattice parser
- non-trivial treebank-based morphological analyzer
  - extended with a spellchecker wordlist
- for details, see paper

72.24 \( \text{F (evalb score)} \)

67.02 \( \text{F (generalized evalb score)} \)

Yoav Goldberg, Reut Tsarfaty, Meni Adler, Michael Elhadad

Parsing with an external Lexicon
What can we do?

Look outside of the treebank

Dictionary Base Morphological Analyzer
(Developed and maintained by the Knowledge center for processing Hebrew)

Yoav Goldberg, Reut Tsarfaty, Meni Adler, Michael Elhadad

Parsing with an external Lexicon
What can we do?

Look outside of the treebank

Dictionary Base Morphological Analyzer

(Developed and maintained by the Knowledge center for processing Hebrew)
What can we do?

**Look outside of the treebank**

Dictionary Base Morphological Analyzer

(Developed and maintained by the Knowledge center for processing Hebrew)

maps word forms to their possible analyses
Treebank vs. Dictionary

Low Lexical Coverage
- 6,219 sentences
- **17,731** unique (non-affixed) word forms
- 28,349 unique tokens

High Lexical Coverage
- 25k lemmas
- **562,439** (non-prefixed) word forms
- 73 prefixes and prefixation rules
- + smart heuristic for unknown words (Adler et al 2008)
Let’s use the Dictionary for rare words!
Let’s use the Dictionary for rare words!

But the tagsets are different...
Resource Incompatibility
Treebank and Dictionary use different tagsets

NN NNT NNP PRP JJ
JJT RB RBR MOD VB
VBMD VBINF AUX AGR
IN COM REL CC QW
HAM WDT DT CD CDE
CDT AT POS

Noun NounC Proper
Pron Adj AdjC Adv Exist
Copula Conj Pref Verb
Beinoni Modal Infinitive
Prep QW Det Num
NumExp NumC At Pos
Resource Incompatibility
Treebank and Dictionary use different tagsets

NN ─── Noun
NNT ─── NounC
NNP ─── Proper
AT ─── At
...
POS ─── Pos
Resource Incompatibility
Treebank and Dictionary use different tagsets

Yoav Goldberg, Reut Tsarfaty, Meni Adler, Michael Elhadad
Parsing with an external Lexicon
Resource Incompatibility
What causes the treebank and dictionary incompatibility?

Differences in annotation perspectives

- Syntactic annotation scheme
- “If a word modifies a verb and can be replaced with an adverb, it’s an adverb”

- Lexicographic guidelines
- “If a word can have this inflection, it can be a verb”
Retag the treebank with the dictionary tagset?
Retag the treebank with the dictionary tagset?

A lesson from Arabic

- Arabic TB originally constructed with lexicon-based tags
- Switching to more syntactic tags improved results by \( \sim 2F\)-points

(Maamouri et.al 2008)

Hurt parser performance
Resource Incompatibility
Conversion?

Retag the treebank with the dictionary tagset?

And in Hebrew

- We re-tagged the treebank
  - \( \sim 90\% \) automatically, \( \sim 10\% \) manually

- Gold-morphology Oracle experiment

  Input Sentence: inhrnt
  Parser sees: IN PRP_{f,p} NN_{f,s}
Resource Incompatibility
Conversion?

Retag the treebank with the dictionary tagset?

And in Hebrew

- We re-tagged the treebank
  - ~ 90% automatically, ~ 10% manually

- Gold-morphology Oracle experiment
  
  Input Sentence:  
  
  Parser sees:  
  
  $\Rightarrow$

Hurt parser performance
Resource Incompatibility Conversion?

Retag the treebank with the dictionary tagset?

And in Hebrew

We re-tagged the treebank

∼ 90%

Gold-morphology Oracle experiment

Input Sentence: inhrnt
Parser sees: IN PRP f,p NN f,s ⇒ 83.29 F

Notice – same grammar:
Gold morphology 83.29
Gold segmentation 72.24
Full ambiguity 67.02

– morphology is informative!
– morphology is ambiguous!
– morphology is hard!

Hurt parser performance
Retag the treebank with the dictionary tagset?

And in Hebrew

- We re-tagged the treebank
  - ~ 90% automatically, ~ 10% manually

- Gold-morphology Oracle experiment

  Input Sentence:  inhrnt
  Parser sees:  IN  PRP_{f,p}  NN_{f,s}  ⇒  83.29 F
  ⇒  81.29 F

Hurt parser performance
Retag the treebank with the dictionary tagset?

Hurt parser performance

We would like to

- Keep syntactic hints of TB tagging
- Benefit from the large coverage of the Dictionary

Probabilistic Fuzzy Mapping

- Take the best of both worlds
- Define a probabilistic mapping function between the tagsets:

\[ p(T_{Dict} \mid T_{TB}) \]

- “sometimes, demonstrative pronouns function as adjective”
The fuzzy map gives rise to a simple generative process:

\[ T_{TB} \rightarrow T_{Dict} \rightarrow Word \]
<table>
<thead>
<tr>
<th>Layered Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TB</strong></td>
</tr>
<tr>
<td>:</td>
</tr>
<tr>
<td>JJ-ZY</td>
</tr>
<tr>
<td>זה</td>
</tr>
<tr>
<td>this</td>
</tr>
<tr>
<td>IN</td>
</tr>
<tr>
<td>במסגרת</td>
</tr>
<tr>
<td>“inside”</td>
</tr>
</tbody>
</table>

Yoav Goldberg, Reut Tsarfaty, Meni Adler, Michael Elhadad

 Parsing with an external Lexicon
### Layered Trees

<table>
<thead>
<tr>
<th>TB</th>
<th>Dict</th>
<th>Layered</th>
</tr>
</thead>
<tbody>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>JJ-ZY</td>
<td>:</td>
<td>JJ-ZY</td>
</tr>
<tr>
<td>ה</td>
<td>:</td>
<td>ה</td>
</tr>
<tr>
<td>this</td>
<td>:</td>
<td>this</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>IN</td>
<td>:</td>
<td>IN</td>
</tr>
<tr>
<td>במסגרת</td>
<td>Prep Noun-F-S</td>
<td>במסגרת</td>
</tr>
<tr>
<td>“inside”</td>
<td>in frame</td>
<td>in frame</td>
</tr>
</tbody>
</table>

**Mapping layer**

Yoav Goldberg, Reut Tsarfaty, Meni Adler, Michael Elhadad

Parsing with an external Lexicon
Combining fuzzy-mapping in a parser

New lexical model

Stable words (seen $\geq 2$ in training) estimated as usual:

$$p(T_{TB} \rightarrow \text{word}) = p_{rf}(\text{word}|T_{TB})$$

Rare/unseen words:

$$p(T_{TB} \rightarrow \text{word}) = p(T_{TB} \rightarrow T_{Dict}) p(T_{Dict} \rightarrow \text{word})$$

Yoav Goldberg, Reut Tsarfaty, Meni Adler, Michael Elhadad

Parsing with an external Lexicon
Combining fuzzy-mapping in a parser

New lexical model

Stable words (seen $\geq 2$ in training) estimated as usual:

$$p(T_{TB} \rightarrow \text{word}) = p_{rf}(\text{word} | T_{TB})$$

Rare/unseen words:

$$p(T_{TB} \rightarrow \text{word}) = p(T_{TB} \rightarrow T_{Dict}) p(T_{Dict} \rightarrow \text{word})$$

But ... what is $p(T_{Dict} \rightarrow \text{word})$ ?
Estimating $p(T_{Dict} \rightarrow w_{rare})$

Dictionary as Filter

Option 1: LexFilter

Use the tag-distribution over rare-words in training, but zero out analyses incompatible with the lexicon:

$$p(T_{Dict} \rightarrow w_{rare}) =$$

$$p(w_{rare} \mid T_{Dict}) = \begin{cases} 
\frac{\text{count}(RARE, T_{Dict})}{\text{count}(T_{Dict})} & T_{Dict} \in \text{Dict}(w_{rare}) \\
0 & T_{Dict} \notin \text{Dict}(w_{rare}) 
\end{cases}$$
Results

Segmentation Oracle  No Oracle

Baseline  72.24  67.02
LexFilter +  76.54

Yoav Goldberg, Reut Tsarfaty, Meni Adler, Michael Elhadad

Parsing with an external Lexicon
<table>
<thead>
<tr>
<th></th>
<th>Segmentation Oracle</th>
<th>No Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>72.24</td>
<td>67.02</td>
</tr>
<tr>
<td>LexFilter +</td>
<td>76.54</td>
<td>68.84</td>
</tr>
</tbody>
</table>

Yoav Goldberg, Reut Tsarfaty, Meni Adler, Michael Elhadad

Parsing with an external Lexicon
Yoav Goldberg, Reut Tsarfaty, Meni Adler, Michael Elhadad

Parsing with an external Lexicon

Results

Segmentation Oracle  No Oracle

Baseline  72.24  67.02

LexFilter +  76.54  68.84

Realistic performance still low... can we do better?
Hope in the face of uncertainty

YES WE CAN.
Option 2: LexProb

Consider the familiar HMM Tagging model:

\[ p(t_1, \ldots, t_n, w_1, \ldots, w_n) = \prod p(t_i|t_{i-1}, t_{i-2})p(w_i|t_i) \]
Estimating $p(T_{Dict} \rightarrow w_{rare})$

Semi-supervised estimation

**Option 2: LexProb**

Consider the familiar HMM Tagging model:

$$p(t_1, \ldots, t_n, w_1, \ldots, w_n) = \prod p(t_i | t_{i-1}, t_{i-2}) p(w_i | t_i)$$

Can be estimated from raw text using EM
Estimating $p(T_{Dict} \rightarrow w_{rare})$

Semi-supervised estimation

Option 2: LexProb

$P(t|t_{-1}, t_{-2})$

$P(w|t)$

$> 92\%$ accuracy

(Adler and Elhadad 2006, Goldberg et.al 2008)
Estimating $p(T_{Dict} \rightarrow w_{\text{rare}})$

Semi-supervised estimation

**Option 2: LexProb**

Dictionary → Raw Text

Smart Thing

$P(t|t_{-1}, t_{-2})$

$P(w|t)$

> 92% accuracy

Use as $P(T_{Dict} \rightarrow \text{word})$

(Adler and Elhadad 2006, Goldberg et.al 2008)
## Results

<table>
<thead>
<tr>
<th></th>
<th>Segmentation Oracle</th>
<th>No Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>72.24</td>
<td>67.02</td>
</tr>
<tr>
<td>LexFilter</td>
<td>76.54</td>
<td>68.84</td>
</tr>
<tr>
<td>LexProb</td>
<td>76.64</td>
<td></td>
</tr>
</tbody>
</table>

We're happy (at least until next year)

Yoav Goldberg, Reut Tsarfaty, Meni Adler, Michael Elhadad
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Segmentation Oracle</th>
<th>No Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>72.24</td>
<td>67.02</td>
</tr>
<tr>
<td>LexFilter</td>
<td>76.54</td>
<td>68.84</td>
</tr>
<tr>
<td>LexProb</td>
<td>76.64</td>
<td>73.69</td>
</tr>
</tbody>
</table>

We’re happy

(...at least until next year)
Take home message

- Treebank derived lexicons are sparse
  - Use an external dictionary / morphological analyzer

- Tagsets may differ
  - That’s OK. Tagsets may (and should) differ
  - Use a fuzzy map

- Dictionaries don’t provide probabilities
  - Semi-supervised estimation using dictionary and raw text