Learning a Part-of-Speech Tagger from Minimal Annotation

Dan Garrette

University of Texas at Austin
Low-Resource Languages

Supervised training is not an option.
Low-Resource Languages

Supervised training is not an option.

We do semi-supervised training.
Low-Resource Languages

Supervised training is not an option.

We do semi-supervised training.

Annotate some data by hand
Supervised training is not an option.

We do semi-supervised training.

Annotate some data by hand
... cheaply
Semi-Supervised Training

[Kupiec, 1992]
[Merialdo, 1994]
Semi-Supervised Training

HMM with Expectation-Maximization (EM)

[Kupiec, 1992]
[Merialdo, 1994]
Semi-Supervised Training

HMM with Expectation-Maximization (EM)

Need:

[Kupiec, 1992]
[Merialdo, 1994]
Semi-Supervised Training

HMM with Expectation-Maximization (EM)

Need:

Large raw corpus

[Kupiec, 1992]
[Merialdo, 1994]
Semi-Supervised Training

HMM with Expectation-Maximization (EM)

Need:

Large raw corpus
Tag dictionary

[Kupiec, 1992]
[Merialdo, 1994]
Semi-Supervised Training

HMM with Expectation-Maximization (EM)

Need:
- Large raw corpus ➩ know how to get this
- Tag dictionary

[Kupiec, 1992]
[Merialdo, 1994]
Semi-Supervised Training

HMM with Expectation-Maximization (EM)

Need:

- Large **raw** corpus
- Tag dictionary

[Kupiec, 1992]
[Merialdo, 1994]
A Real Tag Dictionary

# tag dict
entries
A Real Tag Dictionary

# tag dict entries

- 2,000
- 1,500
- 1,000
- 500
- 0

Labeled Corpus

2 Hours
A Real Tag Dictionary

# tag dict entries

Labeled Corpus

2 Hours
A Real Tag Dictionary

# tag dict entries

<table>
<thead>
<tr>
<th></th>
<th>Labeled Corpus</th>
<th>2 Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td># tag dict entries</td>
<td>60,000</td>
<td>0</td>
</tr>
</tbody>
</table>
A Real Tag Dictionary

Extremely low coverage means most words are unknown
A Real Tag Dictionary

Extremely low coverage means most words are unknown

⇒ Bad for learning (poorly constrained)
Our Approach

annotation → HMM
Our Approach

annotation

→ HMM
Our Approach

annotation

Tag Dict
Generalization

→ HMM
Our Approach

Tag Dict Generalization

annotation → HMM

cover the vocabulary
Our Approach

- Annotation
- Generalization
- Tag Dict
- Model Minimization

→ HMM

cover the vocabulary  remove noise
Our Approach

Tag Dict Generalization

Model Minimization

EM → HMM

cover the vocabulary  remove noise
Our Approach

Tag Dict Generalization → Model Minimization → EM → HMM

annotation

cover the vocabulary remove noise train
Our Approach

- annotation
- Tag Dict Generalization
- Model Minimization
- EM
- HMM
- cover the vocabulary
- remove noise
- train
Our Approach

Tag Dict Generalization  Model Minimization

EM → HMM

cover the vocabulary  remove noise  train

annotation
Collecting Annotations

Task #1

Up to 4 hours to create a tag dictionary
Collecting Annotations

Task #1

**Up to 4 hours** to create a tag dictionary

ordered by frequency

, the . of to a and 
only can York into after president
Collecting Annotations

Task #1
Up to 4 hours to create a tag dictionary

, DT
. .
of IN RP
to TO RP
a DT
and CC
only RB
can VB VBP MD
York NNP
into IN RP
after IN RP
president NN
Collecting Annotations

Task #2

Up to 4 hours to annotate full sentences
Collecting Annotations

Task #2

Up to 4 hours to annotate full sentences

Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.
Collecting Annotations

Task #2

Up to 4 hours to annotate full sentences

Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.
Our Approach

- Tag Dict Generalization
- Model Minimization

- EM
- HMM

- annotation

- cover the vocabulary
- remove noise
- train
Our Approach

Tag Dict
Generalization

Model
Minimization

EM
HMM

annotation

cover the vocabulary
remove noise
train
Tag Dict Generalization

These annotations are too sparse!
Tag Dict Generalization

These annotations are too sparse!

Generalize to the entire vocabulary
Tag Dict Generalization

Our strategy: Label Propagation

[Talukdar and Crammer. 2009]
Tag Dict Generalization

Our strategy: Label Propagation

• Connect annotations to raw corpus tokens

[Talukdar and Crammer. 2009]
Tag Dict Generalization

Our strategy: Label Propagation

• **Connect** annotations to raw corpus tokens

• Push tag labels to **entire corpus**

[Talukdar and Crammer. 2009]
Tag Dict Generalization
Tag Dict Generalization

Type Annotations

the  DT

dog  NN

Token Annotations

the dog walks

DT  NN  VBZ

Raw Corpus

__________

__________

__________
Tag Dict Generalization

Type Annotations
the  DT
_  
dog  NN

Token Annotations
the  TOK_thug_5
the  TOK_the_4
dog  TOK_dog_2

Raw Corpus

The dog walks
Tag Dict Generalization

Type Annotations

the **DT**
dog **NN**

Token Annotations

the dog walks **DT** **NN** **VBZ**
Tag Dict Generalization

Type Annotations

<table>
<thead>
<tr>
<th>the</th>
<th>DT</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>NN</td>
</tr>
</tbody>
</table>

Raw Corpus

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

Token Annotations

<table>
<thead>
<tr>
<th>the dog walks</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT NN VBZ</td>
</tr>
</tbody>
</table>
Tag Dict Generalization

Type Annotations

the DT

dog NN

Token Annotations

the dog walks DT NN VBZ

Raw Corpus
Tag Dict Generalization

Type Annotations

the  DT

dog NN

Token Annotations

the dog walks

DT NN VBZ

Raw Corpus

____________  
____________  
____________  

TOK_the_4  

TOK_the_1  

TOK_thug_5  

TOK_dog_2
Tag Dict Generalization

Type Annotations
- the  DT
- dog  NN

Token Annotations
- the dog walks
  DT   NN   VBZ

Raw Corpus
- 
- 
- 

TOK_the_4  TOK_the_1  TOK_thug_5  TOK_dog_2
Tag Dict Generalization

Type Annotations

_the_ DT
__dog_ NN

Raw Corpus

PREV_<b>

TOK_the_4
TOK_the_1
TOK_thug_5

PREV_the

TOK_dog_2

NEXT_walks

dog walks
DT NN VBZ
Tag Dict Generalization

Type Annotations
the DT
dog NN

Token Annotations
the dog walks
DT NN VBZ

Raw Corpus

PREV_<b>

TOK_the_4
TOK_the_1

PREV_the
TOK_thug_5
TOK_dog_2

NEXT_walks
Tag Dict Generalization

Type Annotations
the _ DT__
dog _ NN__

Token Annotations
the dog walks
DT NN VBZ

PREV_<b>

TOK_the_4
TOK_the_1

PREV_the

TOK_thug_5
TOK_dog_2

NEXT_walks

Raw Corpus

________

________

________
Tag Dict Generalization

Type Annotations

the TYPE_dog
dog NN

PREV_<b>

TOK_the_4
TOK_the_1

PREV_the

TOK_thug_5
TOK_dog_2

PREV_the

NEXT_walks

Token Annotations

the dog walks
DT NN VBZ

Raw Corpus
Tag Dict Generalization

Type Annotations

the  DT

dog  NN

Token Annotations

the dog walks

DT  NN  VBZ

PREV_the

NEXT_walks

TOK_the_1

TOK_the_4

TOK_thug_5

TOK_thug_2

TOK_dog_2
Tag Dict Generalization

Type Annotations

_the__DT__
_dog__NN__

PREV_<_b>

TOK_the_4  TOK_the_1

TYPE_the

TOK_thug_5

PREV_the

TYPE_thug

NEXT_walks

TOK_dog_2

Raw Corpus

Token Annotations

_the dog walks__DT__NN__VBZ__
Tag Dict Generalization

Type Annotations
the DT
dog NN

Token Annotations
the dog walks
DT NN VBZ

PREV_<b>

TYPE_the

TOK_the_4
TOK_the_1

Raw Corpus

PREV_the

TYPE_thug

TOK_thug_5

NEXT_walks

TOK_thug_5

TOK_dog_2
Tag Dict Generalization

Type Annotations

the  DT

dog  NN

Token Annotations

the dog walks
DT  NN  VBZ

PREV_<b>

PREV_the

NEXT_walks

TOK_the_1

TOK_thug_5

TOK_the_4

TOK_thug_5

TOK_dog_2
Tag Dict Generalization

Type Annotations

the DT

dog NN

Raw Corpus

Token Annotations

the dog walks

DT NN VBZ
Tag Dict Generalization

Any arbitrary features could be used
Tag Dict Generalization

Any arbitrary features could be used
Tag Dict Generalization

- SUF1_g
- TYPE_dog
Tag Dict Generalization

SUF1_g

TYPE_dog
Tag Dict Generalization

SUF1_g

TYPE_dog
Tag Dict Generalization

TYPE_sibatarazuka

SUF1_g

TYPE_dog
Tag Dict Generalization

Finite-State Transducer (FST)
Tag Dict Generalization

Finite-State Transducer (FST)
- Generates morphological analysis
Tag Dict Generalization

Finite-State Transducer (FST)

- Generates morphological analysis
Tag Dict Generalization

Finite-State Transducer (FST)

- Generates morphological analysis
Tag Dict Generalization

Finite-State Transducer (FST)

- Generates morphological analysis
- Hand-built by a linguist in 10 hours
Tag Dict Generalization

Finite-State Transducer (FST)

- Generates morphological analysis
- Hand-built by a linguist in 10 hours
Tag Dict Generalization

SUF1_g

TYPE_dog
Tag Dict Generalization

Type Annotations
- the | DT
- dog | NN

Token Annotations
- the dog walks | DT NN VBZ

Raw Corpus

PREV_<b>
PRE2_th
PRE1_t
SUF1_g
PREV_the
NEXT_walks
TOK_the_4
TOK_the_1
TOK_thug_5
TOK_thug_2
TOK_ptr

Tag Dict Generalization

Type Annotations

the DT
__ dog NN
____

Token Annotations

the dog walks
DT NN VBZ

PREV_<b>

PRE2_th

PRE1_t

SUF1_g

NEXT_walks

TOK_the_4

TOK_the_1

TOK_thug_5

TOK_dog_2

TYPE_the

TYPE_thug

TYPE_dog
Tag Dict Generalization

Type Annotations
the DT
__
dog NN

Token Annotations
the dog walks
DT NN VBZ

Raw Corpus

PREV_<b>
PRE1_t
PRE2_th
SUF1_g

TYPE_the
TOK_the_4
TOK_the_1

TYPE_thug
TOK_thug_5

TYPE_dog
TOK_dog_2

NEXT_walks
PREV_the
the dog walks

The diagram shows a graph with nodes labeled with tokens and their types. The tokens include `the`, `dog`, and `walks`, with their respective types `DT`, `NN`, and `VBZ`. The graph structure indicates the relationships and positions of these tokens within a sentence.
Tag Dict Generalization

Type Annotations
the DT——
dog NN——

Raw Corpus
________
________

Token Annotations
the dog walks
DT NN VBZ

TYPE_the
PRE2_th
PRE1_t
SUF1_g

TYPE_thug
PREV_the
NEXT_walks

TYPE_dog

TOK_the_4
TOK_the_1
TOK_thug_5
TOK_dog_2
Tag Dict Generalization

Type Annotations
the
dog

Token Annotations
the dog walks
DT NN VBZ

Raw Corpus

Type Annotations

PRE2_th
PRE1_t
SUF1_g

PREV_<b>
PREV_the
NEXT_walks

TOK_the_4
TOK_the_1
TOK_thug_5
TOK_dog_2

TYPE_dt
TYPE_thug
TYPE_nn

PRE1_t
SUF1_g
PREV_the
NEXT_walks
Tag Dict Generalization

Diagram:

- **DT**: Prepositions
- **NN**: Nouns
- **VBZ**: Verbs
- **TOK**: Tokens
- **PRE2, PRE1**: Previous tokens
- **SUF1**: Suffix
- **TYPE**: Token type

Diagram nodes represent the tagging and generalization of words:
- **the, dog, thug**
- **PRE低于**
- **NEXT_walks**

Diagram edges show the relationship between tokens and tags.
Tag Dict Generalization

Diagram showing connections between DT and NN nodes.
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization

TOK_the_4

TOK_the_1

TOK_thug_5

TOK_dog_2
Tag Dict Generalization
Tag Dict Generalization

TOK_the_4

TOK_the_1

TOK_thug_5

TOK_dog_2
Tag Dict Generalization

TOK_the_4

TOK_the_1

TOK_thug_5

TOK_dog_2
Tag Dict Generalization
Tag Dict Generalization

Result:

TOK_the_4
TOK_the_1
TOK_thug_5
TOK_dog_2
Tag Dict Generalization

Result:

• a tag distribution on every token (soft tagging)
Tag Dict Generalization

Result:

• a tag distribution on every token (soft tagging)
• an expanded tag dictionary (non-zero tags)
Our Approach

Tag Dict
Generalization

Model
Minimization

EM

HMM

annotation

cover the vocabulary

remove noise

train
Our Approach

Tag Dict
Generalization

Model Minimization

EM → HMM

annotation

cover the vocabulary
remove noise

train
Model Minimization

- Induce a cleaner hard tagging from a noisy soft tagging.
- Approach based on work by Sujith Ravi and Kevin Knight (ISI)

[Ravi et al., 2010; Garrette and Baldridge, 2012]
Model Minimization
Model Minimization

\[ \langle b \rangle_0 \quad \text{The}_1 \quad \text{man}_2 \quad \text{saw}_3 \quad \text{the}_4 \quad \text{saw}_5 \quad \langle b \rangle_6 \]
Model Minimization

The [b] man saw the saw

DT

NN

VBD
Model Minimization

<b>0</b>  DT  <b>1</b>  NN  VBD  NN  VBD  DT  NN  VBD  <b>6</b>

<b>The man saw the saw</b>
Model Minimization

The man saw the saw
The man saw the saw.
The man saw the saw.
The man saw the saw.
The man saw the saw.
The man saw the saw
The man saw the saw
The man saw the saw
The man saw the saw
Model Minimization

The man saw the saw
Model Minimization

The man saw the saw

Model Minimization

DT ? ? DT

<b>0 The1 man2 saw3 the4 saw5 <b>6

<b>

DT

NN

VBD
Model Minimization

DT   ?   ?   DT   ?

<b>0</b>  The<sub>1</sub>  man<sub>2</sub>  saw<sub>3</sub>  the<sub>4</sub>  saw<sub>5</sub>  <b>6</b>

DT

NN

VBD
Model Minimization

The man saw the saw
Model Minimization
Model Minimization

1.0
1.0
1.0
0.2
0.8
0.4
0.6
0.7
0.3
1.0

<b>DT</b>

<b>NN</b>

<b>VBD</b>
Model Minimization
Model Minimization

The man saw the saw!
Model Minimization

DT
NN
VBD
Model Minimization

The man saw the saw

DT 1.0

NN 1.0 0.2 0.6 0.7

VBD 0.8 0.3
Model Minimization
Model Minimization

\[ f(\text{NN} \rightarrow \text{VBD}) \]
Model Minimization

The man saw the saw.
A woman drove the car.
The dog walked.
The cat saw a bird.

\[ f(\text{NN} \rightarrow \text{VBD}) \]
Model Minimization

\[ f(\text{NN} \rightarrow \text{VBD}) \]

tag bigram occurrences

weights on their nodes
Model Minimization

\[ f( \text{DT} \rightarrow \text{NN} ) \]
Model Minimization

The man saw the saw
A woman drove the car
The dog walked
The cat saw a bird

\[ f(\text{DT} \rightarrow \text{NN}) \]
Model Minimization

\[ \text{f( DT} \rightarrow \text{ NN } \text{ )} \]
Model Minimization

f( DT \rightarrow NN ) ☑
Model Minimization

\[ f( \text{DT} \rightarrow \text{NN} ) \]
Model Minimization

The man saw the saw
A woman drove the car
The dog walked
The cat saw a bird

f( DT → NN )
Model Minimization
Model Minimization

The man saw the saw

DT
NN
VBD
Model Minimization

The man saw the saw
Model Minimization

The man saw the saw

DT  
NN  
VBD
Model Minimization

The man saw the saw
Model Minimization
The man saw the saw
The man saw the saw.
The man saw the saw
Model Minimization

\[ \text{The}_1 \text{ man}_2 \text{ saw}_3 \text{ the}_4 \text{ saw}_5 \]

DT NN VBD DT NN
Our Approach

Tag Dict Generalization

Model Minimization

EM

HMM

annotation

cover the vocabulary

remove noise

train
Our Approach

Tag Dict
Generalization
Model
Minimization

EM → HMM

annotation

cover the vocabulary
remove noise

train
EM Training

\[ \langle b \rangle_0 \quad \text{The}_1 \quad \text{man}_2 \quad \text{saw}_3 \quad \text{the}_4 \quad \text{saw}_5 \quad \langle b \rangle_6 \]

\text{DT} \quad \text{NN} \quad \text{VBD} \quad \text{DT} \quad \text{NN}
EM Training

Auto-Tagged Corpus

Expanded Tag Dictionary
EM Training

Tag Dictionary Generalization

Auto-Tagged Corpus

Expanded Tag Dictionary

---
EM Training

Tag Dictionary Generalization

Expanded Tag Dictionary

Auto-Tagged Corpus

Initial Emissions

Initial Transitions
EM Training

Tag Dictionary Generalization

Auto-Tagged Corpus

Initial Emissions

Initial Transitions

Expanded Tag Dictionary

EM
Results
Types vs. Tokens

Accuracy

90
85
80
75
70
Types vs. Tokens

<table>
<thead>
<tr>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
</tr>
<tr>
<td>85</td>
</tr>
<tr>
<td>80</td>
</tr>
<tr>
<td>75</td>
</tr>
<tr>
<td>70</td>
</tr>
</tbody>
</table>

- **Types**
- **Tokens**
Types vs. Tokens

Types

Tokens

Accuracy

90
85
80
75
70
Types vs. Tokens

Accuracy of types and tokens over time.

Types:
- Accuracy increases significantly from 70 to 90.
- Accuracy remains steady after 2:00.

Tokens:
- Accuracy increases steadily from 70 to 80.
- Accuracy remains steady after 2:00.

[English]
Morphological Analysis

accuracy

annotation time

[Kinyarwanda]
Morphological Analysis

Without FST Features

accuracy

annotation time

[Kinyarwanda]
Morphological Analysis

With FST Features

Without FST Features

accuracy

annotation time

[Kinyarwanda]
Total Accuracy

<table>
<thead>
<tr>
<th>Language</th>
<th>EM only</th>
<th>Our approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>69</td>
<td>90</td>
</tr>
<tr>
<td>Kinyarwanda</td>
<td>69</td>
<td>82</td>
</tr>
<tr>
<td>Malagasy</td>
<td>77</td>
<td>81</td>
</tr>
</tbody>
</table>

[4 hours of type annotation]
Unknown Word Accuracy

English: EM only - 38, Our approach - 61
Kinyarwanda: EM only - 32, Our approach - 70
Malagasy: EM only - 46, Our approach - 60

[2 hours of type annotation]
English Results
English Results

All of Wiktionary (Li et al., 2012)
English Results

All of Wiktionary (Li et al., 2012) 87%
English Results

All of **Wiktionary** (Li et al., 2012) 87%

**Parallel Corpus** (Täckström et al., 2013)
English Results

All of **Wiktionary** (Li et al., 2012) 87%

**Parallel Corpus** (Täckström et al., 2013) 89%
English Results

All of Wiktionary (Li et al., 2012) 87%

Parallel Corpus (Täckström et al., 2013) 89%

4-hours (Garrettte et al., 2013)
English Results

All of *Wiktionary* (Li et al., 2012) 87%

Parallel Corpus (Täckström et al., 2013) 89%

4-hours (Garrette et al., 2013) 90%
English Results

- All of **Wiktionary** (Li et al., 2012) - 87%
- **Parallel Corpus** (Täckström et al., 2013) - 89%
- **4-hours** (Garrette et al., 2013) - 90%

12 tags
English Results

12 tags

All of Wiktionary (Li et al., 2012) 87%

Parallel Corpus (Täckström et al., 2013) 89%

45 tags

4-hours (Garrette et al., 2013) 90%
Rich Morphology
Rich Morphology

Parallel Corpus (Täckström et al., 2013)
Turkish
Rich Morphology

Parallel Corpus  (Täckström et al., 2013)
Turkish  65%
Rich Morphology

Parallel Corpus (Täckström et al., 2013)
  Turkish 65%

4-hours (Garrette et al., 2013)
  Kinyarwanda
Rich Morphology

Parallel Corpus (Täckström et al., 2013)

Turkish 65%

4-hours (Garrette et al., 2013)

Kinyarwanda 82%
Current Work

- Minimally supervised CCG supertagging and parsing
- Human-provided GFL annotations
Conclusion

- Our approach is able to achieve results better that or comparable to others, but given significantly less input.
- Our annotations are available to others.
- Software available as well.