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.

We do semi-supervised training.

Annotate some data by hand
... cheaply
HMM with Expectation-Maximization (EM)

Need:

- Large *raw* corpus
- Tag dictionary

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

# tag dict entries

- Labeled Corpus: 45,000 entries
- 2 Hours: 0 entries
A Real Tag Dictionary

Extremely low coverage means most words are unknown

⇒ Bad for learning (poorly constrained)
Our Approach

Tag Dict Generalization

Model Minimization

EM → HMM

annotation

cover the vocabulary

remove noise

train
Our Approach

- Annotation
- Cover the vocabulary
- Generalization
- Remove noise
- Model minimization
- EM
- HMM

Tag Dict

- Train
Collecting Annotations

Task #1

Up to 4 hours to create a tag dictionary

ordered by frequency

,  ,  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

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

- annotation
- cover the vocabulary
- remove noise
- train
- EM → HMM
Tag Dict Generalization

These annotations are too sparse!

Generalize to the entire vocabulary
Tag Dict Generalization

Our strategy: Label Propagation

- **Connect** annotations to raw corpus tokens
- Push tag labels to **entire corpus**

[Talukdar and Crammer. 2009]
Any arbitrary features could be used
Finite-State Transducer (FST)

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

Diagram showing the relationships between different parts of speech tags in a sentence:
- **PRE2_th**: Preposition or article
- **PRE1_t**: Preposition or article
- **SUF1_g**: Suffix
- **TOK_the_4**: Token the
- **TOK_thug_5**: Token thug
- **TOK_dog_2**: Token dog
- **PREV_the**: Previous token
- **NEXT_walks**: Next token
- **TOK_the_1**: Token the
- **TOK_dog_2**: Token dog
- **TOK_thug_5**: Token thug
- **TOK_dog_2**: Token dog

The diagram illustrates how these parts of speech tags are connected in a sentence.
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization

Result:

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

Model Minimization

Tag Dict Generalization

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)
The man saw the saw.
The man saw the saw.

Model Minimization

\[
\begin{align*}
&<b>_0 \\
&DT \\
&NN \\
&VBD \\
&NN \\
&VBD \\
&NN \\
&VBD \\
&DT \\
&NN \\
&VBD \\
&<b>_6
\end{align*}
\]
Model Minimization

The man saw the saw

<\text{DT}>_{0} \quad \text{The}_{1} \quad \text{man}_{2} \quad \text{saw}_{3} \quad \text{the}_{4} \quad \text{saw}_{5} \quad <\text{DT}>_{6}
The man saw the saw

Model Minimization
Model Minimization

The man saw the saw.
Model Minimization

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

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

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

tag bigram occurrences weights on their nodes
Model Minimization

\[ f( DT \rightarrow NN ) \]
A woman drove the car. The dog walked the cat saw a bird.
Model Minimization

The man saw the saw
Model Minimization

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

The man saw the saw

DT

NN

VBD
Model Minimization

DT
NN
VBD

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

Model Minimization
The man saw the saw
The man saw the saw.
Our Approach

Tag Dict
Generalization

Model
Minimization

EM

HMM

annotation

cover the vocabulary

remove noise

train

EM

HMM

cover the vocabulary

remove noise

train
EM Training

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

DT   NN   VBD   DT   NN
EM Training

Auto-Tagged Corpus

Initial Emissions

Initial Transitions

Tag Dictionary Generalization

Expanded Tag Dictionary

EM

Expanded Tag Dictionary

EM Training
Results
Types vs. Tokens

<table>
<thead>
<tr>
<th>Types</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>85</td>
<td>70</td>
</tr>
</tbody>
</table>

accuracy
Morphological Analysis

accuracy

With FST Features

Without FST Features

annotation time

[Kinyarwanda]
Total Accuracy

![Bar chart showing total accuracy for English, Kinyarwanda, and Malagasy]

- **English**: EM only 69, Our approach 90
- **Kinyarwanda**: EM only 69, Our approach 82
- **Malagasy**: EM only 77, Our approach 81

[4 hours of type annotation]
Unknown Word Accuracy

<table>
<thead>
<tr>
<th>Language</th>
<th>EM only</th>
<th>Our approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>38</td>
<td>61</td>
</tr>
<tr>
<td>Kinyarwanda</td>
<td>32</td>
<td>70</td>
</tr>
<tr>
<td>Malagasy</td>
<td>46</td>
<td>60</td>
</tr>
</tbody>
</table>

[2 hours of type annotation]
English Results

<table>
<thead>
<tr>
<th>Tags</th>
<th>Dataset</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>All of Wiktionary (Li et al., 2012)</td>
<td>87%</td>
</tr>
<tr>
<td>45</td>
<td>Parallel Corpus (Täckström et al., 2013)</td>
<td>89%</td>
</tr>
<tr>
<td>45</td>
<td>4-hours (Garrette et al., 2013)</td>
<td>90%</td>
</tr>
</tbody>
</table>
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.