Learning a Part-of-Speech Tagger from Two Hours of Annotation

Dan Garrette and Jason Baldridge

University of Texas at Austin
Low-Resource Languages

6,900 languages in the world

~30 have non-negligible quantities of data

No million-word corpus for any endangered language

[Maxwell and Hughes, 2006]
[Abney and Bird, 2010]
Low-Resource Languages

Kinyarwanda
Niger-Congo; morphologically-rich

Malagasy
Austronesian; spoken in Madagascar

Also, English
Low-Resource Languages

Supervised training is not an option.

We do semi-supervised training.

Annotate some data by hand

... cheaply

... like, in 2 hours
Semi-Supervised Training

HMM with Expectation-Maximization (EM)

Need:

- Large raw corpus
- Tag dictionary

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

Most previous work:

Extract from a large labeled corpus

→ too complete
→ too clean
→ too biased
A Real Tag Dictionary

Labeled Corpus

# tag dict entries

- 0
- 15,000
- 30,000
- 45,000
- 60,000

2 Hours
A Real Tag Dictionary

Extremely low coverage means most words are unknown

⇒ Bad for EM (poorly constrained)
Our Approach

Tag Dict Generalization

Model Minimization

EM → HMM

cover the vocabulary remove noise train
Our Approach

Tag Dict
Generalization

Model
Minimization

EM → HMM

Annotation

→ EM

→ HMM

cover the vocabulary
remove noise
train
Collecting Annotations

Task #1 -- 2 hours to create a tag dictionary

<table>
<thead>
<tr>
<th>,</th>
<th>DT</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td></td>
</tr>
<tr>
<td>.</td>
<td></td>
</tr>
<tr>
<td>of</td>
<td>IN RP</td>
</tr>
<tr>
<td>to</td>
<td>TO RP</td>
</tr>
<tr>
<td>a</td>
<td>DT</td>
</tr>
<tr>
<td>and</td>
<td>CC</td>
</tr>
<tr>
<td>:</td>
<td></td>
</tr>
<tr>
<td>only</td>
<td>RB</td>
</tr>
<tr>
<td>can</td>
<td>VB VBP MD</td>
</tr>
<tr>
<td>York</td>
<td>NNP</td>
</tr>
<tr>
<td>into</td>
<td>IN RP</td>
</tr>
<tr>
<td>after</td>
<td>IN RP</td>
</tr>
<tr>
<td>president</td>
<td>NN</td>
</tr>
<tr>
<td>:</td>
<td></td>
</tr>
</tbody>
</table>
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

In 2 hours:

<table>
<thead>
<tr>
<th></th>
<th># sent</th>
<th># tok</th>
<th># TD entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Sentences</td>
<td>90</td>
<td>1537</td>
<td>750</td>
</tr>
<tr>
<td>Tag Dict</td>
<td></td>
<td></td>
<td>1798</td>
</tr>
</tbody>
</table>

(for Kinyarwanda)
Our Approach

- Tag Dict Generalization

- Model Minimization

- EM

- HMM

- annotation

- cover the vocabulary

- remove noise

- train
Tag Dict Generalization

These annotations are too sparse!

Generalize to the entire vocabulary
Tag Dict Generalization

Haghighi and Klein (2006) do this with a vector space.

We don’t have enough raw data

Das and Petrov (2011) do this with a parallel corpus.

We don’t have a parallel corpus
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

Annotations

Raw Corpus

TYPE_the

TYPE_thug

TYPE_dog

...
Tag Dict Generalization

Annotations

________

________

________

Raw Corpus

________

________

________

the\textsubscript{4} thug\textsubscript{5} walks\textsubscript{6}

TOK\_the\_4 TOK\_thug\_5 TOK\_walks\_6

TYPE\_the

TYPE\_thug

TYPE\_dog

\ldots
Tag Dict Generalization

the₄  thug₅  walks₆

TOK_the_4  TOK_the_1  TOK_the_9  TOK_thug_5  TOK_dog_2

TYPE_the  TYPE_thug  TYPE_dog

Annotations

Raw Corpus
Tag Dict Generalization

**Annotations**

```
the_4
thug_5
walks_6
PREV_<b>
PREV_the
PREV_thug
```

**Raw Corpus**

```
_______
_______
_______
```

**TOK**

```
TOK_the_4
TOK_the_1
TOK_the_9
TOK_thug_5
TOK_dog_2
```

**TYPE**

```
TYPE_the
TYPE_thug
TYPE_dog
```
Tag Dict Generalization

Annotations

________

________

________

Raw Corpus

________

________

________

TYPE_thug

PREV_<b>

TOK_the_4

TOK_the_1

TOK_the_9

TYPE_the

PREV_the

TOK_thug_5

TOK_dog_2

TYPE_thug

TYPE_dog

walks_6

the_4

...
Tag Dict Generalization

Annotations

Raw Corpus

the_4 thug_5 walks_6

NEXT_thug NEXT_walks NEXT_\text{<b>}

PREV_\text{<b>}

TOK_the_4 TOK_the_1 TOK_the_9

TYPE_the

TOK_thug_5

TYPE_thug

TOK_dog_2

TYPE_dog

...
Tag Dict Generalization

Annotations

the_4 thug_5 walks_6

Raw Corpus

PREV_<b>
NEXT_thug

TOK_the_4 TOK_the_1 TOK_the_9

TYPE_the

PREV_the
NEXT_walks

TOK_thug_5 TOK_dog_2

TYPE_thug

TYPE_dog
Tag Dict Generalization

PREV_<b>

NEXT_thug

TOK_the_4

TOK_the_1

TOK_the_9

TOK_thug_5

TOK_dog_2

TYPE_the

TYPE_thug

TYPE_dog
Tag Dict Generalization
Tag Dict Generalization
Tag Dict Generalization

PREV_<b> | NEXT_thug | ... | PREV_the | NEXT_walks | ... | PRE1_t | PRE2_th | TYPE_the | TYPE_thug | TYPE_dog | PRE1_d | PRE2_do

TOK_the_4 | TOK_the_1 | TOK_the_9 | TOK_thug_5 | TOK_dog_2 | ... | PREV_the | PREV_the | PREV_the | PREV_the | PREV_the | PREV_the | PREV_the
Tag Dict Generalization

PREV_<b> <-> NEXT_thug <-> TYPE_thug <-> PRE1_t, PRE2_th

PREV_the <-> NEXT_walks <-> TYPE_thug <-> PRE1_d, PRE2_do

TOK_the_4, TOK_the_1, TOK_the_9 <-> TYPE_the

TOK_thug_5, TOK_dog_2 <-> TYPE_thug, TYPE_dog
Tag Dict Generalization

PREV_<b>
PREV_the
PREV_thug

NEXT_thug
NEXT_walks

TOK_the_1
TOK_the_4
TOK_the_9
TOK_thug_5
TOK_dog_2

TYPE_the
TYPE_thug
TYPE_dog

PRE1_t
PRE2_th
SUF1_e
SUF1_g
PRE1_d
PRE2_do
Tag Dict Generalization

PREV_<b>

NEXT_thug

PREV_the

NEXT_walks

PREV_the

PREV_<b>

PRE1_t

PRE2_th

SUF1_e

TYPE_the

TOK_the_4

TOK_the_1

TOK_the_9

TOK_thug_5

TOK_dog_2

TYPE_thug

PRE1_d

PRE2_do

TYPE_dog

SUF1_g
Tag Dict Generalization

Type Annotations

the  DT

dog  NN

PREV_ <b>  PRE2_th  PRE1_t  SUF1_g

TYPE_the  TYPE_thug  TYPE_dog

TOK_the_4  TOK_the_1  TOK_thug_5  TOK_dog_2  PREV_the  NEXT_walks
Tag Dict Generalization

Type Annotations

the

dog

- DT
  - TYPE_the
    - PREV_<b>
      - TOK_the_4
      - TOK_the_1
  - PRE2_th
  - PRE1_t
  - TYPE_thug
    - PREV_the
      - TOK_thug_5
    - SUF1_g
  - NEXT_walks
    - TOK_dog_2
Tag Dict Generalization

Type Annotations

the

dog

Token Annotations

the dog walks

DT NN VBZ
Tag Dict Generalization

Token Annotations
the dog walks

Type Annotations
the
dog
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

Tag Dict Generalization → Model Minimization

- Annotation
- Cover the vocabulary
- Remove noise
- Train

EM → HMM
Model Minimization

• Induce a cleaner hard tagging from a noisy soft tagging.

• Greedily seek the minimal set of tag bigrams that describe the raw corpus.

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

DT  NN  VBD  NN  VBD  DT  NN  VBD

<b>0  The  man  saw  the  saw  <b>6

<bspace>0  DT  VBD  VBD  DT  VBD

<bspace>DT  NN  NN  NN  NN  DT

<bspace>VBD  VBD  VBD  VBD  VBD
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

DT

NN

VBD
Model Minimization

The man saw the saw
Model Minimization

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

\[ f( \text{DT} \rightarrow \text{NN} ) \]
A woman drove
The dog walked
The cat saw a bird
The man saw the saw
..........................

Model Minimization
Model Minimization

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

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

The man saw the saw

The_1 man_2 saw_3 the_4 saw_5 <b>_6

DT NN VBD DT NN
Our Approach

Tag Dict Generalization
Model Minimization

EM
HMM

annotation

cover the vocabulary remove noise

train
Model Minimization

Tag Dictionary Generalization

Auto-Tagged Corpus

Initial Emissions

Initial Transitions

Expanded Tag Dictionary

EM
Total Accuracy

<table>
<thead>
<tr>
<th>Language</th>
<th>Tokens</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>70</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>78</td>
<td>78</td>
</tr>
<tr>
<td>Kinyarwanda</td>
<td>55</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>71</td>
<td>79</td>
</tr>
<tr>
<td>Malagasy</td>
<td>68</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>76</td>
<td>76</td>
</tr>
</tbody>
</table>

Tokens: EM only, + Our approach
Types: EM only, + Our approach
Unknown Accuracy

Remember: Very high unknown rates.
Especially for morphological-rich Kinyarwanda.
Conclusion

• Developed a semi-supervised approach to learn a tagger from realistically minimal input.

• Currently being used for further low-resource research (e.g. unsupervised dependency parsing).
ACL Preview

- Learning curves for annotation time
- Mixed types and tokens under fixed time constraints
- Morphological transducers
- 90% accuracy on full 45 tag English Penn Treebank with 4 hours of data
Software Available

Train your own low-resource taggers.

Or use our Kinyarwanda and Malagasy models.

Open source: link on my website or in the paper.