To Normalize or Not to Normalize: The Impact of Normalization on Part-Of-Speech Tagging

Rob van der Goot, Barbara Plank & Malvina Nissim
University of Groningen

07-09-2017
Gary did gd protectin SpongeBob house!
Gary did good protecting SpongeBob's house!
Problems

Experiments

- Normalization for POS tagging
- Semi-supervised adaptation of a POS tagger
- Complementary
Experimental setup

Train data

Owoputi:

<table>
<thead>
<tr>
<th>Test_O</th>
<th>Dev</th>
<th>Train (1576)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(549)</td>
<td>(249)</td>
<td></td>
</tr>
</tbody>
</table>

LexNorm:

<table>
<thead>
<tr>
<th>Test_L</th>
</tr>
</thead>
<tbody>
<tr>
<td>(549)</td>
</tr>
</tbody>
</table>

Data from: Chen Li, Yang Liu. Joint POS Tagging and Text Normalization for Informal Text. IJCAI 2015
Experimental setup

Raw data
- Wikipedia
- Tweets
- No gazetteers, hard coded rules, etc.!
Experimental setup

Bilty

https://github.com/bplank/bilstm-aux
Experimental setup

![Graph showing accuracy comparison between Bilty (vanilla) and ARK]
To Normalize

ur dat dude frm spongebob ?

orig → ur
w2v → your
aspell → Eur
lookup → your

dat → tht
that → dude
frm → from
spongebob → SpongeBob

you’re → that
dude → dude
SpongeBob → sponge bob
spongebob → sponge-bob

Random Forest

N-grams

https://bitbucket.org/robvanderg/monoise
To Normalize

Accuracy

Dev data

raw

unk

all

goldED

gold

80

82

84

86

88

90

80

82

84

86
To Normalize

<table>
<thead>
<tr>
<th>non-canonical</th>
<th>canonical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td></td>
</tr>
<tr>
<td>Train</td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td>→</td>
</tr>
</tbody>
</table>
To Normalize

<table>
<thead>
<tr>
<th>non-canonical</th>
<th>canonical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td></td>
</tr>
<tr>
<td>Train</td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td></td>
</tr>
<tr>
<td>Train</td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td></td>
</tr>
</tbody>
</table>
To Normalize

Accuracy

Train data
raw
all
union

Dev data
80
82
84
86
88
90

Accuracy

raw
unk
all
goldED
gold

Dev data
80
82
84
86
88
90

Accuracy

raw
unk
all
goldED
gold

Train data
raw
all
union
Or Not to Normalize

Word Embeddings
Or Not to Normalize

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilty</td>
<td>82</td>
</tr>
<tr>
<td>+norm</td>
<td>84</td>
</tr>
<tr>
<td>+w2v</td>
<td>90</td>
</tr>
</tbody>
</table>
Or Not to Normalize
Self training (Tweets)

- Random Tweets
- Tweets with NE
- Tweets without NE
- Tweets containing unknown words
Or Not to Normalize
Self training (Tweets)
Or Not to Normalize
Self training (EWT)

- Answers
- Reviews
- Newsgroups
- Weblog
- E-mail
Or Not to Normalize

Self training (EWT)
Combine

Accuracy

Bilty
+norm
+embeds
+comb
ARK

Owoputi
LexNorm
Conclusions

- Normalization improves the baseline tagger
- Semi-supervised learning works even better
- Combining improves performance slightly
- Performance is close to ARK tagger
Conclusions

Negative results

- Do not normalize training data
- Self-training with pre-selection is not effective
Conclusions

Future work

- Self-training with post-selection
- Domain adaptation setup (train on canonical data)
- Joint/integrated approach
Conclusions

Thx 4 ur attention