Adapting taggers to Twitter using not-so-distant supervision

Barbara Plank, Dirk Hovy, Ryan McDonald and Anders Søgaard
Newswire bias

The CROSS-DOMAIN GULF
Newswire bias

"domain adaptation" or "transfer learning"

The CROSS-DOMAIN GULF
Newswire bias

"domain adaptation" or "transfer learning"

The CROSS-DOMAIN GULF
#Localization #job:
Supplier / Project Manager - Localisation

Vendor - NY, NY, United States http://bit.ly/16KigBg #nlppeople
#Localization #job : Supplier / Project Manager - Localisation

Vendor - NY, NY, United States [http://bit.ly/16KigBg](http://bit.ly/16KigBg) #nlppeople
The Supplier / Project Manager performs the …
The Supplier / Project Manager performs the...

Vendor - NY, NY, United States

http://bit.ly/16KigBg

#Localization #job
Supplier / Project Manager - Localisation

#nlppeople
#Localization #job:
Supplier / Project Manager - Localisation

Vendor - NY, NY, United States http://bit.ly/16KigBg #nlppeople

The Supplier / Project Manager performs the...

URL: http://bit.ly/16KigBg
Prey Developer worked with Nintendo on project http://bit.ly/17Kbsf
Prey Developer worked with Nintendo on project http://bit.ly/17Kbsf

In a statement, Nintendo announced that …
Prey Developer worked with Nintendo on project http://bit.ly/17Kbsf

In a statement, Nintendo announced that...
Setup
Setup
Setup

tag

the New York Times
American Football Best Sport Ever!!!

tag

tag

Twitter
Setup

tag

The New York Times
American Football Best Sport Ever!!!

tag

project

Twitter
Setup

tag

add data

iterate

project
Setup

- tag
- add data
- iterate
- project
- augmented self-training
Setup

- tag
- tag
- add data
- project
- augmented self-training
- iterate
NB: URLs *not* required at testing time!
1: $X = \{ \langle x_i, y_i \rangle \}_{i=1}^{N}$ labeled tweets
2: $U = \{ \langle x_i, w_i \rangle \}_{i=1}^{M}$ unlabeled tweet-website pairs
3: $I$ iterations
4: $k = 1000$ pool size
5: $v = \text{train}(X)$ base model
6: for $i \in I$ do
7: for $\langle x, w \rangle \in \text{pool}_k(U)$ do
8: $\hat{y} = \text{predict}(\langle x, w \rangle; v)$
9: $X \leftarrow X \cup \{ \langle \hat{y}, x \rangle \}$
10: end for
11: $v = \text{train}(X)$
12: end for
13: return $v$
1: \(X = \{\langle x_i, y_i \rangle \}_{i=1}^N\) labeled tweets
2: \(U = \{\langle x_i, w_i \rangle \}_{i=1}^M\) unlabeled tweet-website pairs
3: \(I\) iterations
4: \(k = 1000\) pool size
5: \(v = \text{train}(X)\) base model
6: \(\text{for } i \in I \text{ do}\)
7: \(\text{for } \langle x, w \rangle \in \text{pool}_k(U) \text{ do}\)
8: \(\hat{y} = \text{predict}(\langle x, w \rangle, v)\)
9: \(X \leftarrow X \cup \{\langle \hat{y}, x \rangle\}\)
10: \(\text{end for}\)
11: \(v = \text{train}(X)\)
12: \(\text{end for}\)
13: \(\text{return } v\)
1: \( X = \{ \langle x_i, y_i \rangle \}_{i=1}^N \) labeled tweets
2: \( U = \{ \langle x_i, w_i \rangle \}_{i=1}^M \) unlabeled tweet-website pairs
3: \( I \) iterations
4: \( k = 1000 \) pool size
5: \( v = \text{train}(X) \) base model
6: \textbf{for} \( i \in I \) \textbf{do}
7: \textbf{for} \( \langle x, w \rangle \in \text{pool}_k(U) \) \textbf{do}
8: \( \hat{y} = \text{predict}(\langle x, w; v \rangle) \)
9: \( X \leftarrow X \cup \{ \langle \hat{y}, x \rangle \} \)
10: \textbf{end for}
11: \( v = \text{train}(X) \)
12: \textbf{end for}
13: \textbf{return} \( v \)
predict(</twitter/⟩, 🍃⟩)
predict(</twitter, 🔥/>)
predict($<$-twitter, 🍺 $>$)

plain

web
predict(⟨/twitter, 🌞⟩)

- plain
- web
- Wiktionary [ˈwɪkʃənəri] n.,

93% unambiguous
predict(< </twitter> , </dict> >)

93% unambiguous
predict(</twitter>, dict_of_pot)
Experiments
POS

Train + Test
POS Results: Which projection method?

<table>
<thead>
<tr>
<th>Method</th>
<th>Average Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>88.4</td>
</tr>
<tr>
<td>dict</td>
<td>88.5</td>
</tr>
<tr>
<td>web</td>
<td>88.8</td>
</tr>
<tr>
<td>web&lt;dict</td>
<td>89.0</td>
</tr>
<tr>
<td>dict&lt;web</td>
<td>89.1</td>
</tr>
</tbody>
</table>

Average over 3 test sets
POS Results: Which projection method?

Average over 3 test sets

- baseline: 88.4
- dict: 88.5
- web: 88.8
- web<dict: 89.0
- dict<web: 89.1

Wiki [ˈwɪkʃənəri] n.,
POS Results

Foster 91.6 92.4
Lowlands 87.5 88.4
Ritter 87.4 88.5
Test-average 88.8 89.8

WSJ+Gimpel baseline  
not-so-distant supervision
POS Results

- **Foster:**
  - WSJ+Gimpel baseline: 91.6
  - not-so-distant supervision: 92.4

- **Lowlands:**
  - WSJ+Gimpel baseline: 89.8
  - not-so-distant supervision: 88.5

- **Ritter:**
  - WSJ+Gimpel baseline: 88.4
  - not-so-distant supervision: 87.4

- **Test-average:**
  - WSJ+Gimpel baseline: 88.8
  - not-so-distant supervision: 89.8

Additional note: plain self-training
POS Results

Foster: 91.6, 92.4
Lowlands: 89.8, 88.5
Ritter: 88.4, 87.4
WSJ+Gimpel baseline
not-so-distant supervision

not-so-distant supervision
plain self-training

Test-average: 88.8, 89.8
NER Results

<table>
<thead>
<tr>
<th>Test Set</th>
<th>CoNLL+Finin baseline</th>
<th>not-so-distant supervision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ritter-test</td>
<td>77.4</td>
<td>78.5</td>
</tr>
<tr>
<td>Fromreide-test</td>
<td>82.1</td>
<td>83.9</td>
</tr>
<tr>
<td>Finin-test</td>
<td>74.0</td>
<td>75.8</td>
</tr>
<tr>
<td>Test-average</td>
<td>77.9</td>
<td>79.4</td>
</tr>
</tbody>
</table>
Analysis
## Projection Examples

<table>
<thead>
<tr>
<th></th>
<th><em>Snohomish</em></th>
<th><em>Bakery</em></th>
<th><em>Salmon-Safe</em></th>
<th><em>parks</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>initial tag</td>
<td>ADJ</td>
<td>NOUN</td>
<td>NOUN</td>
<td>NOUN</td>
</tr>
<tr>
<td>projected</td>
<td>NOUN</td>
<td>NOUN</td>
<td>ADJ</td>
<td>NOUN</td>
</tr>
</tbody>
</table>
Limitation

If I gave you one wish that will become true.

What’s your wish?...? I wish I’ll get 3 wishes from you :p URL
Limitation

If I gave you one wish that will become true.

What’s your wish ?… ? I wish i’ll get 3 wishes from you :p URL
Analysis

- improvements due to richer linguistic context

- somewhat arbitrary differences
Take-home-message
Take-home-message

✓ 20% of tweets contain URLs
Take-home-message

✓ 20% of tweets contain URLs

✓ Semi-supervised learning + not-so-distant supervision helps to bridge the gulf
Take-home-message

✓ 20% of tweets contain URLs
✓ Semi-supervised learning + not-so-distant supervision helps to bridge the gulf
✓ Models available at:

https://bitbucket.org/lowlands/ttagger-nsd
Questions?
Thanks!
Alternatives

**PROJECTION:**  NOUN-NOUN-NOUN

**HARD CONSTR.:**  VERB-NOUN-VERB

**SOFT CONSTR.:**  NOUN-VERB-NOUN

i  luv  jaxx

best path

2nd best path

most freq label in URL

2nd most freq label
NER Results

Average over 3 test sets
NER Results

Average over 3 test sets

- baseline: 77.9
- dict: 66.6
- web: 79.0
- web<dict: 67.3
- dict<web: 79.4