Frustratingly Easy
Domain Adaptation


Kang Ji
Language Processing for Different Domains and Genres
WS 2009/10
Overview

- Motivation
- Annotation
- Core Approach
  - Prior Works
  - Feature Annotation
  - Kernelized Version
- Some Experimental Results
A common special case

• Suppose we have a NLP system focusing on news document, and now want to migrate it into biographic domain

Would there be any difference if we

• have quite some biographic documents(target data) and lots of news documents.
• only have news documents(source data).
Rough Idea

Source Data → Combined Feature Space → New Input → ML System

Target Data
ML approaches

• Now we simplified the task to a standard machine learning problem
  • Fully supervised learning: annotated corpus
  • Semi-supervised learning: large unannotated corpus, annotated corpus from the later target data
Some Annotations

- Input space $X$
- Output space $\Psi$
- Samples: $D^s$ $D^t$

$D^s$ is a collection of $N$ examples and $D^t$ is a collection of $M$ examples (where, typically, $N \gg M$).
Some Annotations

• Distribution on the source and target domains: $\mathcal{D}^s \mathcal{D}^t$

• Learning function $h : X_i \rightarrow \Psi_i$

$X_i = \mathbb{R}^F$ and that $\Psi_i = \{-1, +1\}$
Prior works

- The SRCONLY baseline ignores the target data and trains a single model, only on the source data.
- The TGTONLY baseline trains a single model only on the target data.
- The ALL baseline simply trains a standard learning algorithm on the union of the two datasets.
Prior works

- The WEIGHTED baseline: re-weight examples from $D^S$.

in case that $N \gg M$, so if $N = a \times M$, we may weight each example from the source domain by $1/a$. 
Prior works

- The PRED baseline is based on the idea of using the output of the source classifier as a feature in the target classifier.
- The LININT baseline, we linearly interpolate the predictions of the SRCONLY and the TGTONLY models.
Prior works

• The PRIOR model is to use the SRCONLY model as a prior on the weights for a second model, trained on the target data.

• The maximum entropy classifiers model by Daumé III and Marcu (2006), learns three models and justifies on a per-example basis.
Feature Augmentation

- $\Phi^s, \Phi^t: \mathcal{X} \rightarrow \hat{\mathcal{X}}$ mapping for source and target data respectively, then define $\hat{\mathcal{X}} = \mathbb{R}^{3F}$, we get

- $\Phi^s(x) = \langle x, x, 0 \rangle$; $\Phi^t(x) = \langle x, 0, x \rangle$

- the features which are made into three: general version, source-specific version, target-specific version

- get some ideas? examples coming----> black board
a simple and pleasing result

• $\tilde{K}(x, x') = 2K(x, x')$ same domain
• $\tilde{K}(x, x') = K(x, x')$ diff. domain

• the data point from the target domain has **twice** as much influence as the data point from source domain on the prediction of the test target data.
Extension to Multi-domain adaption

- For a $K$-domain problem, we simply expand the feature space from $\mathbb{R}^{3F}$ to $\mathbb{R}^{(K+1)F}$
- “+1” stands for the “general domain”
Why better

• This model optimize the feature weights jointly, thus there’s no need to cross-validate to estimate good hyperparameters for each task as the PRIOR model does.

• Also it means that the single supervised learning algorithm that is run is allowed to regulate the trade-off between source/target and general weights.
Task Statistics

- **Table 1**: Task statistics;

- **columns** are task, domain, size of the training, development and test sets, and the number of unique features in the training set.

- Feature sets: lexical information (words, stems, capitalization, prefixes and suffixes), membership on gazetteers, etc.
<table>
<thead>
<tr>
<th>Task</th>
<th>Dom</th>
<th>SrcONLY</th>
<th>TgtONLY</th>
<th>ALL</th>
<th>WEIGHT</th>
<th>PRED</th>
<th>LININT</th>
<th>PRIOR</th>
<th>AUGMENT</th>
<th>T&lt;S</th>
<th>Win</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE-NER</td>
<td>bn</td>
<td>4.98</td>
<td>2.37</td>
<td>2.29</td>
<td>2.23</td>
<td>2.11</td>
<td>2.21</td>
<td>2.06</td>
<td></td>
<td>1.98</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>bc</td>
<td>4.54</td>
<td>4.07</td>
<td>3.55</td>
<td>3.53</td>
<td>3.89</td>
<td>4.01</td>
<td>3.47</td>
<td></td>
<td>3.47</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>nw</td>
<td>4.78</td>
<td>3.71</td>
<td>3.86</td>
<td>3.65</td>
<td>3.56</td>
<td>3.79</td>
<td>3.68</td>
<td></td>
<td>3.39</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>wl</td>
<td>2.45</td>
<td>2.45</td>
<td>2.12</td>
<td>2.12</td>
<td>2.45</td>
<td>2.33</td>
<td>2.41</td>
<td>2.12</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>un</td>
<td>2.46</td>
<td>2.46</td>
<td>2.48</td>
<td>2.40</td>
<td>2.18</td>
<td>2.10</td>
<td>2.03</td>
<td>1.91</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>cts</td>
<td>2.08</td>
<td>0.46</td>
<td>0.40</td>
<td>0.40</td>
<td>0.46</td>
<td>0.44</td>
<td>0.34</td>
<td>0.32</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>CoNLL</td>
<td>tgt</td>
<td>2.49</td>
<td>2.95</td>
<td>1.80</td>
<td>1.75</td>
<td>2.13</td>
<td>1.77</td>
<td>1.89</td>
<td>1.76</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>PubMed</td>
<td>tgt</td>
<td>12.02</td>
<td>4.15</td>
<td>5.43</td>
<td>4.15</td>
<td>4.14</td>
<td>3.95</td>
<td>3.99</td>
<td>3.61</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>CNN</td>
<td>tgt</td>
<td>10.29</td>
<td>3.82</td>
<td>3.67</td>
<td>3.45</td>
<td>3.46</td>
<td>3.44</td>
<td>3.35</td>
<td>3.37</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Treebank</td>
<td>wsj</td>
<td>6.63</td>
<td>4.35</td>
<td>4.33</td>
<td>4.30</td>
<td>4.32</td>
<td>4.32</td>
<td>4.27</td>
<td>4.11</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>swbd3</td>
<td>15.90</td>
<td>4.15</td>
<td>4.50</td>
<td>4.10</td>
<td>4.13</td>
<td>4.09</td>
<td>3.60</td>
<td>3.51</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>br-cf</td>
<td>5.16</td>
<td>6.27</td>
<td>4.85</td>
<td>4.80</td>
<td>4.78</td>
<td>4.72</td>
<td>5.22</td>
<td>5.15</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>br-cg</td>
<td>4.32</td>
<td>5.36</td>
<td>4.16</td>
<td>4.15</td>
<td>4.27</td>
<td>4.30</td>
<td>4.25</td>
<td>4.09</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>br-ck</td>
<td>5.05</td>
<td>6.32</td>
<td>5.05</td>
<td>4.98</td>
<td>5.01</td>
<td>5.05</td>
<td>5.27</td>
<td>5.41</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>br-cl</td>
<td>5.66</td>
<td>6.60</td>
<td>5.42</td>
<td>5.39</td>
<td>5.39</td>
<td>5.53</td>
<td>5.99</td>
<td>5.73</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>br-cm</td>
<td>3.57</td>
<td>6.59</td>
<td>3.14</td>
<td>3.11</td>
<td>3.15</td>
<td>3.31</td>
<td>4.08</td>
<td>4.89</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>br-cn</td>
<td>4.60</td>
<td>5.56</td>
<td>4.27</td>
<td>4.22</td>
<td>4.20</td>
<td>4.19</td>
<td>4.48</td>
<td>4.42</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>br-cp</td>
<td>4.82</td>
<td>5.62</td>
<td>4.63</td>
<td>4.57</td>
<td>4.55</td>
<td>4.55</td>
<td>4.87</td>
<td>4.78</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>br-er</td>
<td>5.78</td>
<td>9.13</td>
<td>5.71</td>
<td>5.19</td>
<td>5.20</td>
<td>5.15</td>
<td>6.71</td>
<td>6.30</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Treebank-brown</td>
<td>6.35</td>
<td>5.75</td>
<td>4.80</td>
<td>4.75</td>
<td>4.81</td>
<td>4.72</td>
<td>4.72</td>
<td>4.65</td>
<td></td>
<td></td>
<td>+</td>
</tr>
</tbody>
</table>

Table 2: Task results.
Model Introspection

- “broadcast news” contains no capitalization
- “broadcast conversation”
- “newswire”
- “Weblog”
- “usenet” may contain many email addresses and URLs
- “conversational telephone speech”

Figure 1: Hinton diagram for feature /Aa+/ at current position.
Implementation Demo

- [link](http://public.me.com/jikang/easyadapt.pl.zip) (only 10 line perl script, how elegant!)
Reference

• Hal Daumé III, 2007. Frustratingly Easy Domain Adaptation

• Hal Daume III, Daniel Marcu, 2006. Domain Adaptation for Statistical Classifiers