Banks Meeting: Data Selection

2009-02-24

Martha Palmer

http://www.bbn.com/NLP/OnToNotes
Motivation

- Very skewed sense distribution
- Need balanced data with the rare senses well represented
  - The fact that a sense is rare in WSJ doesn’t mean it’s also rare in some other domain
  - May need a specific (rare) sense for text mining
- Verb *add*
  - 288 instances of the predominant sense
  - 22 instances of the rare sense (7%)  
  - To get 22 instances of the rare sense, need to annotate 310 instances!
- Can we do better?
Data Selection Plan for Web Text

- 2-way translations of Arabic, Chinese and English to create parallel corpora
- Document selection
- Sentence selection
Document Selection from 2M docs

- Histograms for each doc completed and merged.
- Document profile for every doc being generated:
  - 400 nouns, 1000 verbs (440 new verbs and 530 old verbs with < 15 instances) on target list
  - 3 features generated for each doc:
    - how many verbs from list are present?
    - how many nouns from list are present?
    - is the doc weblog data?
  - Also produces wc of doc and total of hits from target list (doesn't count repetitions of the same word)
  - Pick top 2000 docs based on features

- Filtered for spam, etc., selected top 70K of docs
Sentence Selection

- Use “Document” set as test data for “sentence” data

- Select lemmas (verbs and nouns)
  - Histogram of verbs in web text
  - Histogram of verbs missing senses
  - Pick overlap that has the most instances in “Document” set

- Select sentences for top 50 verbs in overlap
  - Random sampling
  - Batch Mode Active learning
  - Language Model

- Expected results
  - 200K words of data = 10K sentences
  - Avg of 50 instances @ for 100 verbs/100 nouns
Random Sampling

- Annotate all instances of the verb
- X: Number of instances (%), Y: Rare Sense Recall
Approach 1 - Active Learning

- Run an automated system that provides confidence values
- Extract the lowest confidence instance, hand-correct it, add it to the training data
- Repeat
- Simulations using previously tagged data indicated half the additional data provides the same performance improvement as random sampling
  - Chen, Stein, Ungar, Palmer, NAACL-06
  - Zhu, J. and E.H. Hovy. IJCNLP-08, EMNLP-07
But

- Very impractical for a sense tagging project
  - Human annotators have to sit and wait while a single instance is being selected and again during retraining

- **Batch Mode Active Learning.**
  - Select the 50 lowest confidence instances at one time,
  - hand correct all of them,
  - retrain,
  - repeat if necessary
Approach 2: Language Model
Precipitating out Rare senses

- Compute a language model (wsj+brown+ebn+ectb)
- Compute probability (perplexity) for each instance of the verb
  - n-size windows around the target verb
  - logprob <instance> / total words
- Rank the instances by probability
Higher Concentration of the Rare Sense Instances at the Top

[Diagram showing a blue gradient with arrows pointing to boxes labeled 'Low probability' and 'High probability']
Verb point:
Recall vs. Number of Instances

x: Number of Instances
y: Recall
Exp #1: Top ½ of instances

<table>
<thead>
<tr>
<th>lemma</th>
<th>rare</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>account-v</td>
<td>0.12</td>
<td>0.21</td>
<td>0.93</td>
</tr>
<tr>
<td>add-v</td>
<td>0.07</td>
<td>0.10</td>
<td>0.73</td>
</tr>
<tr>
<td>admit-v</td>
<td>0.18</td>
<td>0.19</td>
<td>0.55</td>
</tr>
<tr>
<td>allow-v</td>
<td>0.06</td>
<td>0.08</td>
<td>0.69</td>
</tr>
<tr>
<td>compare-v</td>
<td>0.08</td>
<td>0.16</td>
<td>1</td>
</tr>
<tr>
<td>explain-v</td>
<td>0.10</td>
<td>0.12</td>
<td>0.6</td>
</tr>
<tr>
<td>maintain-v</td>
<td>0.11</td>
<td>0.11</td>
<td>0.53</td>
</tr>
<tr>
<td>point-v</td>
<td>0.15</td>
<td>0.29</td>
<td>1</td>
</tr>
<tr>
<td>receive-v</td>
<td>0.07</td>
<td>0.08</td>
<td>0.6</td>
</tr>
<tr>
<td>remain-v</td>
<td>0.15</td>
<td>0.20</td>
<td>0.65</td>
</tr>
<tr>
<td>worry-v</td>
<td>0.15</td>
<td>0.22</td>
<td>0.73</td>
</tr>
</tbody>
</table>

2-sense verbs
rare sense < 20%
(at least 100 instances)

average baseline: 0.11
average precision: 0.16
average recall: 0.73