Creating Robust Supervised Classifiers via Web-Scale N-gram Data

Shane Bergsma
University of Alberta

Emily Pitler
University of Pennsylvania

Dekang Lin
Google, Inc.

ACL 2010
New Web-Scale N-gram Data

• Details in: [Lin et al., LREC 2010]
  – Same source as Google N-grams Version 1
  – More pre-processing: duplicate sentence removal, length+alphabetical constraints

• Includes part-of-speech tags!

flies  1643568  NNS|611646 VBZ|1031922
caught the flies ,  11  VBD|DT|NNS|,|11
plane flies really well  10  NN|VBZ|RB|RB|10
Overview

• Features from web-scale N-gram data:
  – Count(some N-gram) in web corpus

• Open questions:
  1. How well do web-scale N-gram features work when combined with conventional features?
  2. How well do classifiers with web-scale N-gram features perform on new domains?

• Conclusion: N-gram features are essential
Feature Classes

• Lex (lexical features): \( x_{\text{Lex}} \)
  – Many thousands of \textbf{binary} features indicating a property of the strings to be classified

• N-gm (N-gram count features): \( x_{\text{Ngm}} \)
  – A few dozen \textbf{real-valued} features for the \textit{logarithmic} counts of various things

• The classifier:
  \[ h(x) = w \cdot x \]
  \[ x = (x_{\text{Lex}}, x_{\text{Ngm}}) \]
Uses of New N-gram Data

• Applications:
  1. Adjective Ordering
  2. Real-Word Spelling Correction
  3. Noun Compound Bracketing
  4. Verb Part-of-Speech Tagging
     • benefits of N-grams not so clear cut (see paper)

• All experiments: linear SVM classifier, report Accuracy (%)
1. Adjective Ordering

• “green big truck” or “big green truck”?

• Used in translation, generation, etc.
• Not a syntactic issue but a semantic issue: – size precedes colour, etc.
Adjective Ordering

• As a classification problem:
  – Take adjectives in alphabetical order
  – Decision: is alphabetical order correct or not?

• Why not just most frequent order on web?
  – 87% for web order but 94% for classifier
Adjective Ordering Features

- Lex features: indicators for the adjectives
  - $\text{adj}_1$ indicated with $+1$, $\text{adj}_2$ indicated with $-1$
  - E.g. “big green”

\[ x_{\text{Lex}} = (\ldots, 0, 0, 0, 0, 0, 0, +1, 0, 0, 0, 0, \ldots) \]

Decision:

\[ h_{\text{Lex}}(x_{\text{Lex}}) = w_{\text{Lex}} \cdot x_{\text{Lex}} \]

\[ h_{\text{Lex}}(x_{\text{Lex}}) = w_{\text{big}} - w_{\text{green}} \]
Adjective Ordering Features

big green truck
Adjective Ordering Features

first big storm
Adjective Ordering Features

$W_{\text{first}}$ $W_{\text{big}}$ $W_{\text{young}}$ $W_{\text{green}}$ $W_{\text{Canadian}}$
Adjective Ordering Features

• N-gm features:

\[
\begin{align*}
&\text{Count(“big green”) } & \text{Count(“green big”)} \\
&\text{Count(“big J.*”) } & \text{Count(“green J.*”)} \\
&\text{Count(“J.* big”) } & \text{Count(“J.* green”)} \\
\end{align*}
\]

\[x_{\text{Ngm}} = (29K, 200, 571K, 2.5M, \ldots)\]
## Adjective Ordering Results

<table>
<thead>
<tr>
<th>System</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malouf (2000)</td>
<td>91.5</td>
</tr>
<tr>
<td>web $c(a_1, a_2)$ vs. $c(a_2, a_1)$</td>
<td>87.1</td>
</tr>
<tr>
<td>SVM with N-GM features</td>
<td>90.0</td>
</tr>
<tr>
<td>SVM with LEX features</td>
<td>93.0</td>
</tr>
<tr>
<td>SVM with N-GM + LEX</td>
<td>93.7</td>
</tr>
</tbody>
</table>
In-Domain Learning Curve

Accuracy (%) vs Number of training examples

- N-GM+LEX
- N-GM
- LEX

93.7%
Out-of-Domain Learning Curve

```
Accuracy (%)

N-GM+LEX
N-GM
LEX

Number of training examples
```

June 25, 2010
Slide 15
2. Real-Word Spelling Correction

• Classifier predicts correct word in context:

  “Let me know weather you like it.”
  “weather” or “whether”
Spelling Correction

• Lex features:
  – Presence of particular words (and phrases) preceding or following the confusable word
Spelling Correction

- N-gm feats: Leverage multiple relevant contexts:
  Bergsma et al., 2009

Let me know _
me know _ you
know _ you like
_ you like it

- Five 5-grams, four 4-grams, three 3-grams and two 2-grams span the confusible word
Spelling Correction

• N-gm features:
  – Count(“let me know weather you”) 5-grams
  – Count(“me know weather you like”)
    ...
  – Count(“let me know weather”) 4-grams
  – Count(“me know weather you”) 4-grams
  – Count(“know weather you like”)
    ...
  – Count(“let me know whether you”) 5-grams
    ...

Spelling Correction Results

<table>
<thead>
<tr>
<th>System</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>66.9</td>
</tr>
<tr>
<td>SVM with N-GM features</td>
<td>95.7</td>
</tr>
<tr>
<td>SVM with LEX features</td>
<td>95.2</td>
</tr>
<tr>
<td>SVM with N-GM + LEX</td>
<td>96.5</td>
</tr>
</tbody>
</table>
3. Noun Compound Bracketing

- “… bus driver”
  - female (bus driver)
  - *(female bus) driver
  - (school bus) driver

3-word case is a binary classification: **right** or **left** bracketing
Noun Compound Bracketing

• Lex features:
  – binary features for all words, pairs, and the triple, plus capitalization pattern

Vadas & Curran, 2007
Noun Compound Bracketing

- N-gm features, e.g. “female bus driver”
  - Count(“female bus”) → predicts left
  - Count(“female driver”) → predicts right
  - Count(“bus driver”) → predicts right
  - Count(“femalebus”)
  - Count(“busdriver”)
  - etc.

Nakov & Hearst, 2005
In-Domain Learning Curves

Spelling Correction

Noun Compound Bracketing

Accuracy (%)

Number of training examples

June 25, 2010	Slide 24
### Out of Domain

<table>
<thead>
<tr>
<th>Spelling Correction</th>
<th>Noun Compound Bracketing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Errors nearly double when you remove N-gram features</td>
<td>No N-gram features = BAD</td>
</tr>
</tbody>
</table>
Conclusion

• It’s good to mix standard lexical features with N-gram count features
• Domain sensitivity of NLP
Thanks

• Google, Inc.
• Johns Hopkins University
• Liam:

(getting one good picture took 45 minutes)