Machine learning for sequential data:
A comparative study with applications to natural language processing

Sander Canisius
S.V.M.Canisius@uvt.nl
ILK / Language and Information Science
Tilburg University

(Joint work with Antal van den Bosch and Walter Daelemans)
Predicting label sequences

input  a  b  c  d  e

output A  B  C  D  E
Olympic champion Agassi meets Karim Alami of Morocco in the first round.
Sequences in NLP

preexisting

prigistin

c00cc0000i00

[c][c][i]
Machine learning methods

Conditional Random Fields (Lafferty et al., 2001)
Cycling Dependency Networks (Toutanova et al., 2003)
Max-margin Markov Networks (Taskar et al., 2003)
Conditional Markov Models (Ratnaparkhi, 1996)
Maximum-entropy Markov Models (McCallum et al., 2000)
Discriminatively trained Hidden Markov Models (Collins, 2002)
Stacked Sequential Learning (Cohen, 2004)
Constraint Satisfaction Inference (Canisius et al., 2006)

...
Benchmark data sets

Natural language processing

• Word-level
  – CELEX Morphological segmentation / parsing
  – CELEX Grapheme-phoneme conversion

• Sentence-level
  – CoNLL-2000 Syntactic chunking
  – CoNLL-2002/3 Named-entity recognition
  – GENIA Named-entity recognition

• Document level
  – FAQ segmentation
Benchmark data sets

Bioinformatics

- Protein secondary structure prediction
- Gene prediction

Suggestions?

- ...
Case study: (bio)medical named-entity recognition

Named-entity recognition in Medline abstracts

[DNA_part li kappa B-1 ] is a [DNA positive regulatory element ] in [cell_line B-cell lines ] and in the [cell_line li-expressing T-cell line ] , [cell_line H9 ] , but acts as a [DNA negative regulatory element ] in [cell_line myelomonocytic ] and [cell_line glia cell lines ] .
Case study: (bio)medical named-entity recognition

Named-entity recognition in Dutch medical encyclopedias

Learning method

- Maximum-entropy models
  - a.k.a log-linear models

\[ P(c | d, \lambda) = \frac{\exp\left(\sum_i \lambda_i f_i(c, d)\right)}{\sum_{c'} \exp\left(\sum_i \lambda_i f_i(c', d)\right)} \]
Sequence prediction methods

- Sliding window
- Recurrent sliding window
- Stacking
- Constraint satisfaction inference
- Conditional Markov models
- Maximum-entropy Markov models
- Conditional random fields
Features

- Simple features only
  - 3-1-3 sliding window of words and POS tags
# Results

**GENIA**

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>$F_{\beta=1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sliding window</td>
<td>54.9 ±1.16</td>
<td>54.1 ±1.23</td>
<td>54.5 ±1.02</td>
</tr>
<tr>
<td>Rec. sliding window</td>
<td>67.3 ±1.04</td>
<td>57.6 ±1.25</td>
<td>62.1 ±1.11</td>
</tr>
<tr>
<td>Stacking</td>
<td>57.8 ±1.21</td>
<td>55.3 ±1.07</td>
<td>56.5 ±1.11</td>
</tr>
<tr>
<td>CSI</td>
<td>64.1 ±1.06</td>
<td>56.6 ±1.10</td>
<td>60.1 ±1.03</td>
</tr>
<tr>
<td>CMM</td>
<td>67.7 ±0.96</td>
<td>57.9 ±1.07</td>
<td>62.4 ±1.01</td>
</tr>
<tr>
<td>MEMM</td>
<td>67.1 ±1.14</td>
<td>57.7 ±1.13</td>
<td>62.1 ±1.15</td>
</tr>
<tr>
<td>CRF</td>
<td>66.8 ±1.10</td>
<td>59.2 ±1.14</td>
<td>62.8 ±1.08</td>
</tr>
</tbody>
</table>
# Results

**Dutch medical encyclopedia**

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>$F_{\beta=1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sliding window</td>
<td>62.3 ±1.12</td>
<td>60.8 ±1.06</td>
<td>61.5 ±0.98</td>
</tr>
<tr>
<td>Rec. sliding window</td>
<td><strong>68.5 ±1.16</strong></td>
<td><strong>60.0 ±1.13</strong></td>
<td><strong>63.9 ±0.89</strong></td>
</tr>
<tr>
<td>Stacking</td>
<td>63.2 ±1.23</td>
<td>60.8 ±1.13</td>
<td>62.0 ±1.10</td>
</tr>
<tr>
<td>CSI</td>
<td><strong>68.6 ±1.15</strong></td>
<td>59.9 ±1.11</td>
<td><strong>63.9 ±1.02</strong></td>
</tr>
<tr>
<td>CMM</td>
<td><strong>68.8 ±1.26</strong></td>
<td>59.6 ±1.09</td>
<td><strong>63.9 ±0.99</strong></td>
</tr>
<tr>
<td>MEMM</td>
<td><strong>68.8 ±1.09</strong></td>
<td><strong>59.3 ±1.26</strong></td>
<td><strong>63.7 ±1.09</strong></td>
</tr>
<tr>
<td>CRF</td>
<td>66.8 ±1.14</td>
<td>60.2 ±1.14</td>
<td>63.4 ±0.99</td>
</tr>
</tbody>
</table>
Observations

• Sequence methods tend to favour precision over recall
  – In named-entity recognition tasks, entities are predicted more conservatively

• Very similar performance with many sequence methods

• Recurrent sliding window and its probabilistic version CMM have almost exactly the same performance
  – Doesn't the extra inference step add anything?
Recurrent sliding window

input sequence

output sequence

window 1

window 2

...
Ratnaparkhi's conditional Markov models

- Label sequence conditional probability
  \[ P(y_1, y_2, \ldots, y_n | x_1, x_2, \ldots, x_n) = \sum_i p(y_i | h_i, x_i) \]

- \( h_i \) corresponds to the history features in the recurrent sliding window method

- Beam search is used to select the most likely label sequence
FAQ segmentation
McCallum et al., 2000

<prolog> This section of the FAQ is about the electronic support network that exists for 386bsd and its off-spring.
<prolog>
<prolog>
<prolog>
<prolog>
<question>1.0 I just downloaded all of 386bsd version 0.1 and I can't get [some feature] to work? Do you have any suggestions?
<question>
<answer> Yes. Get FreeBSD, OpenBSD, or NetBSD.
<answer>
<answer>
<answer>
<question>1.1 Minimum hardware configuration recommended
<answer>
<answer>
<answer>
<answer>
<answer>
There has been considerable debate about what the REAL minimum configuration for *BSD is. Some would claim that it is the
Features for FAQ segmentation
McCallum et al., 2000

- begins-with-number
- begins-with-ordinal
- begins-with-punctuation
- begins-with-question-word
- begins-with-subject
- blank
- contains-alphanum
- contains-bracketed-number
- contains-http
- contains-non-space
- contains-number
- contains-pipe
- contains-question-mark
- contains-question-word
- ends-with-question-mark
- first-alpha-is-capitalized
- indented
- indented-1-to-4
- indented-5-to-10
- more-than-one-third-space
- only-punctuation
- prev-is-blank
- prev-begins-with-ordinal
- shorter-than-30
## Results

### FAQ segmentation

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>$F_{\beta=1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default maxent</td>
<td>21.3</td>
<td>46.6</td>
<td>27.7</td>
</tr>
<tr>
<td>Rec. sliding window</td>
<td>70.8</td>
<td>73.6</td>
<td>70.7</td>
</tr>
<tr>
<td>CMM</td>
<td>74.2</td>
<td>78.0</td>
<td>74.9</td>
</tr>
</tbody>
</table>
Discussion

• Recurrent sliding window (CMM, beam size: 1) vs. CMM
  – Hardly any difference on two domain-specific entity recognition tasks
  – CMM outperforms recurrent sliding window on FAQ segmentation
  – What causes these differences?
    • Do the properties that favour CMMs actually occur in real-world NLP tasks?
    • So far, various potential explanations have been explored, none proved to be true
Summary

• Presented plans and preliminary results for a large-scale empirical evaluation of sequence prediction methods in the context of natural language processing.
• Suggestions for relevant/informative data sets are welcome.
• Small case study on domain-specific entity recognition:
  – Sequence prediction methods tend to improve F-score mainly by improving precision, not recall.
  – Inference methods on top of probabilistic (maxent) classifiers did not prove to have a large advantage over simpler methods.
• However, there may be data sets where this advantage does exist (e.g. FAQ segmentation).