Comparison of character-level and part of speech features for name recognition in biomedical texts

by Nigel Collier and Koichi Takeuchi
DOI: 10.1016/j.jbi.2004.08.00

presented by Christopher Maier for
INLS 279: Bioinformatics Research Review
2005-03-08
The Named Entity Task

• Recognize boundaries of important terms
• Classify these terms according to an existing taxonomy
Why is Biomedical NE so Difficult?

- Large and constantly growing vocabulary
- Irregular naming conventions
  - I blame *Drosophila* researchers
- Synonymy
- Class cross-over
- Progress in the field leads to alteration of classification taxonomy
The GENIA Corpus

- http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/
- Annotated collection of MEDLINE abstracts related to transcription factors in human blood cells
- Project includes corpus, ontology, and accessory tools
- Largest and most comprehensively annotated corpus for NE in the biomedical domain
The Bio I² Corpus

- Same field as GENIA, but different articles
- Annotated to a small top-level ontology
- Smaller than GENIA (100 abstracts)
- Available online in XML format
Example: Bio I² Annotation

TL - Involvement of extracellular signal-regulated kinase module in [HIV]_{source.vi} - mediated [CD4]_{protein} signals controlling activation of [nuclear factor-kappa B]_{protein} and [AP-1]_{protein} transcription factors.

AB - Although the molecular mechanisms by which the [HIV-1]_{source.vi} triggers either [T cell]_{source.ct} activation, anergy, or apoptosis remain poorly understood, it is well established that the interaction of [HIV-1]_{source.vi} envelope glycoproteins with [cell surface]_{source.sl} [CD4]_{protein} delivers signals to the target cell, resulting in activation of transcription factors such as [NF-kappa B]_{protein} and [AP-1]_{protein}. In this study, we report the first evidence indicating that kinases [MEK-1]_{protein} ([MAP kinase/Erk kinase]_{protein}) and [ERK-1]_{protein} ([extracellular signal-regulated kinase]_{protein}) act as intermediates in the cascade of events that regulate [NF-kappa B]_{protein} and [AP-1]_{protein} activation upon [HIV-1]_{source.vi} binding to [cell surface]_{source.sl} [CD4]_{protein}. 
Support Vector Machines (SVM)

- trainable classifier for distinguishing between positive and negative examples
- a key strength is the ability to handle very large feature sets
Two Leading Approaches

- Part of Speech Tagging
- Orthographic Features
- Both are attractive because they are computationally cheap, easy to implement, and powerful
Part Of Speech

- Determine a word’s lexical class(es) based on contextual grammatical information
- Number of grammatical classes depends on annotation scheme (i.e. PTB, Brown, etc.)
- Train a POS tagger on a collection of annotated domain documents
- Important in Biomedical NE for disambiguation of word sense and boundary detection
Differential_A_ABS interactions_N_NOM_PL of_PREP NF-kappa_N_NOM_SG complexes_N_NOM_PL with_PREP kappa_N_NOM_SG B_ABBR_NOM_SG alpha_N_NOM_SG determine_V_PRES pools_N_NOM_PL of_PREP constitutive_A_ABS and_CC inducible_A_ABS NF-kappa_N_NOM_SG B_ABBR_NOM_SG activity_N_NOM_SG ...

FDG GENIA
Differential_A_ABS interactions_N_NOM_PL of_PREP NF-kappa_N_NOM_SG complexes_N_NOM_PL with_PREP kappa_N_NOM_SG B_N_NOM_SG alpha_N_NOM_SG determine_V_PRES pools_N_NOM_PL of_PREP constitutive_A_ABS and_CC inducible_A_ABS NF-kappa_N_NOM_SG B_N_NOM_SG activity_N_NOM_SG ...

Brill WSJ POS
Differential_JJ interactions_NNS of_IN Rel_NNP NF-kappa_NNP B_NNP complexes_NNS with_IN kappa_NN B_NNP determine_VB pools_NNS of_IN constitutive_JJ and_CC inducible_JJ B_NNP activity_NNP ...

Brill GENIA POS
Differential_JJ interactions_NNS of_IN Rel_NN NF-kappa_NN B_NN complexes_NNS with_IN kappa_NN B_NN determine_VB pools_NNS of_IN constitutive_JJ and_CC inducible_NN B_NNP activity_NN ...
Some NE tasks have found that POS does not improve system performance (mostly non-bio, though).

Genia-derived POS in biomedical domain can lead to big performance gains, however.
Orthographic Features

- What does the word “look like”?
- Very effective in news domain (e.g. initial capitals)
- wnt, NF-κB, IRF-7, p53, MAPK
- Potentially very useful in biomedical domain
# Orthographic Feature Values

<table>
<thead>
<tr>
<th>Feature</th>
<th>Example</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>GreekLetter</td>
<td>kappa</td>
<td>3</td>
<td>145</td>
<td>6</td>
<td>0.96</td>
</tr>
<tr>
<td>CapsDigitHyphen</td>
<td>Oct-1</td>
<td>6</td>
<td>560</td>
<td>24</td>
<td>0.96</td>
</tr>
<tr>
<td>CapsAndDigits</td>
<td>STAT1</td>
<td>5</td>
<td>514</td>
<td>63</td>
<td>0.91</td>
</tr>
<tr>
<td>SingleCap</td>
<td>B</td>
<td>5</td>
<td>442</td>
<td>49</td>
<td>0.90</td>
</tr>
<tr>
<td>LettersAndDigits</td>
<td>p105</td>
<td>2</td>
<td>186</td>
<td>21</td>
<td>0.90</td>
</tr>
<tr>
<td>LowCaps</td>
<td>pre-BI</td>
<td>5</td>
<td>149</td>
<td>30</td>
<td>0.83</td>
</tr>
<tr>
<td>OneDigit</td>
<td>2</td>
<td>4</td>
<td>62</td>
<td>24</td>
<td>0.72</td>
</tr>
<tr>
<td>TwoCaps</td>
<td>EBV</td>
<td>8</td>
<td>975</td>
<td>505</td>
<td>0.66</td>
</tr>
<tr>
<td>InitCap</td>
<td>Sox</td>
<td>7</td>
<td>302</td>
<td>843</td>
<td>0.26</td>
</tr>
<tr>
<td>HyphenDigit</td>
<td>95-</td>
<td>2</td>
<td>6</td>
<td>36</td>
<td>0.14</td>
</tr>
<tr>
<td>LowerCase</td>
<td>kinases</td>
<td>10</td>
<td>2049</td>
<td>15,942</td>
<td>0.11</td>
</tr>
<tr>
<td>HyphenBackslash</td>
<td>-</td>
<td>5</td>
<td>65</td>
<td>530</td>
<td>0.11</td>
</tr>
<tr>
<td>Punctuation</td>
<td>(</td>
<td>4</td>
<td>118</td>
<td>2404</td>
<td>0.05</td>
</tr>
<tr>
<td>DigitSequence</td>
<td>98401159</td>
<td>1</td>
<td>1</td>
<td>135</td>
<td>0.01</td>
</tr>
<tr>
<td>TwoDigit</td>
<td>37</td>
<td>0</td>
<td>0</td>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td>FourDigit</td>
<td>1997</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>NucleotideSequence</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- **a**: classes in which value was used
- **b**: number of tokens tagged with this value
- **c**: number of non-NE tokens tagged with value
- **d**: predictive power of value
A little something extra: Soundex

- Phonetically similar, but orthographically different, words should indicate similar objects
- Algorithm is computationally simple, based on a simple LUT of phonetic codes
- e.g. JAK1, JAK2, JAK3 all map to ‘J200’
- But what about “Interleukin-7” and “interaction”?
The Big Question

- How do variations of and interactions between these representations affect performance in the NE task?
Experiment

- SVM with variable window size
- Combinations of orthographic and POS techniques
- 10-fold cross-validation
- Compare precision, recall, and F-score
Results: Orthography

• BaseNDO with a \(-1+1\) window performed best

• Soundex performs above base, but does not contribute as much as orthographic features, due to noise

• Windows larger than \(-1+1\) have degraded performance
Results: POS

• Again, a -1+1 window has best performance
• Brill GENIA is best, followed closely by FDG and FDG GENIA
Results: Combination

- “POS and Orthographic features do not mix well”
Discussion

- Noun phrases have difficult-to-detect boundaries
- Noun phrases with embedded words of different classes are hard
- Sometimes orthography can bias against rare occurrences
- Long phrases are hard
- Embedded abbreviations
Conclusions

• POS not as useful as orthography because of complex interplay between boundaries, syntax, and semantics

• POS tagging algorithm might affect this, though