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

by Nigel Collier and Koichi Takeuchi
Journal of Biomedical Informatics 37(2004) 423-435
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
- Synonomy
- 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 I^2 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 I[2] Annotation

TI - 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

FDG POS

Differential_A_ABS interactions_N_NOM_PL of_PREP [redacted]
-_- NF-kappa_N_NOM_SG [redacted] complexes_N_NOM_PL
with_PREP [redacted] 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 [redacted] -_-
NF-kappa_N_NOM_SG [redacted] complexes_N_NOM_PL with_PREP
[redacted] kappa_N_NOM_SG B_N_NOM_SG alpha_N_NOM_SG deter-
mine_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 [redacted] kappa_NN B_NNP [redacted] determine_VB
pools_NNS of_IN constitutive_JJ and_CC inducible_JJ [redacted] B_NNP
activity_NNP ._.

Brill GENIA POS

Differential_JJ interactions_NNS of_IN Rel_NN -_- NF-kappa_NN B_NN
complexes_NNS with_IN [redacted] kappa_NN B_NN [redacted] determine_VB
pools_NNS of_IN constitutive_JJ and_CC inducible_NN [redacted] B_NNP
activity_NN ._.

Part Of Speech

- 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, MAPKKK, etc.
  - Potentially very useful in biomedical domain

Orthographic Feature Values

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

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