Exploiting Context for Biomedical Entity Recognition: From Syntax to the Web

Jenny Finkel,* Shipra Dingare,† Huy Nguyen,*
Malvina Nissim,† Christopher Manning,* and Gail Sinclair†

*Department of Computer Science
Stanford University
Stanford, CA 93405-9040
United States
{jrfinkel|htnguyen|manning}
@cs.stanford.edu

†Institute for Communicating and
Collaborative Systems
University of Edinburgh
Edinburgh EH8 9LW
United Kingdom
{sdingar1|mnissim|csincla1}
@inf.ed.ac.uk

Abstract

We describe a machine learning system for the recognition of names in biomedical texts. The system makes extensive use of local and syntactic features within the text, as well as external resources including the web and gazetteers. It achieves an F-score of 70% on the Coling 2004 NLPBA/BioNLP shared task of identifying five biomedical named entities in the GENIA corpus.

1 Introduction

The explosion of information in the fields of molecular biology and genetics has provided a unique opportunity for natural language processing techniques to aid researchers and curators of databases in the biomedical field by providing text mining services. Yet typical natural language processing tasks such as named entity recognition, information extraction, and word sense disambiguation are particularly challenging in the biomedical domain with its highly complex and idiosyncratic language. With the increasing use of shared tasks and shared evaluation procedures (e.g., the recent BioCreative, TREC, and KDD Cup), it is rapidly becoming clear that performance in this domain is markedly lower than the field has come to expect from the standard domain of newswire. The Coling 2004 shared task focuses on the problem of Named Entity Recognition, requiring participating systems to identify the five named entities of protein, RNA, DNA, cell line, and cell type in the GENIA corpus of MEDLINE abstracts (Ohta et al., 2002). In this paper we describe a machine learning system incorporating a diverse set of features and various external resources to accomplish this task. We describe our system in detail and also discuss some sources of error.

2 System Description

Our system is a Maximum Entropy Markov Model, which further develops a system earlier used for the CoNLL 2003 shared task (Klein et al., 2003) and the 2004 BioCreative critical assessment of information extraction systems, a task that involved identifying gene and protein name mentions but not distinguishing between them (Dingare et al., 2004). Unlike the above two tasks, many of the entities in the current task do not have good internal cues for distinguishing the class of entity: various systematic polysemies and the widespread use of acronyms mean that internal cues are lacking. The challenge was thus to make better use of contextual features, including local and syntactic features, and external resources in order to succeed at this task.

2.1 Local Features

We used a variety of features describing the immediate content and context of each word, including the word itself, the previous and next words, word prefixes and suffix of up to a length of 6 characters, word shapes, and features describing the named entity tags assigned to the previous words. Word shapes refer to a mapping of each word onto equivalence classes that encodes attributes such as length, capitalization, numerals, greek letters, and so on. For instance, "Varicella-zoster" would become *Xx-xxx*, "mRNA" would become *xXXX*, and "CPA1" would become *XXXd*. We also incorporated part-of-speech tagging, using the TnT tagger(Brants, 2000) retrained on the GENIA corpus gold standard part-of-speech tagging. We also used various interaction terms (conjunctions) of these base-level features in various ways. The full set of local features is outlined in Table 1.

2.2 External Resources

We made use of a number of external resources, including gazetteers, web-querying, use of the surrounding abstract, and frequency counts from the British National Corpus.

Word Features	w_i, w_{i-1}, w_{i+1}
	Disjunction of 5 prev words
	Disjunction of 5 next words
TnT POS	POS_i, POS_{i-1}, POS_{i+1}
Prefix/suffix	Up to a length of 6
Abbreviations	$abbr_i$
	$abbr_{i-1}$ + $abbr_i$
	$abbr_i$ + $abbr_{i+1}$
	$abbr_{i-1}$ + $abbr_i$ + $abbr_{i+1}$
Word Shape	$shape_i$, $shape_{i-1}$, $shape_{i+1}$
	$shape_{i-1}$ + $shape_i$
	$shape_i$ + $shape_{i+1}$
	$shape_{i-1}$ + $shape_i$ + $shape_{i+1}$
Prev NE	NE_{i-1}, NE_{i-2} + NE_{i-1}
	NE_{i-3} + NE_{i-2} + NE_{i-1}
Prev NE + Word	NE_{i-1} + w_i
Prev NE + POS	NE_{i-1} + POS_{i-1} + POS_i
	NE_{i-2} + NE_{i-1} + POS_{i-2} + POS_{i-1} + POS_i
Prev NE + Shape	NE_{i-1} + $shape_i$
	NE_{i-1} + $shape_{i+1}$
	NE_{i-1} + $shape_{i-1}$ + $shape_i$
	NE_{i-2} + NE_{i-1} + $shape_{i-2}$ + $shape_{i-1}$ + $shape_i$
Paren-Matching	Signals when one parenthesis in a pair has been assigned a different tag than the other in a window of 4 words

Table 1: Local Features (+ indicates conjunction)

2.2.1 Frequency

Many entries in gazetteers are ambiguous words, occasionally used in the sense that the gazetteer seeks to represent, but at least as frequently not. So while the information that a token was seen in a gazetteer is an unreliable indicator of whether it is an entity, less frequent words are less likely to be ambiguous than more frequent ones. Additionally, more frequent words are likely to have been seen often in the training data and the system should be better at classifying them, while less frequent words are a common source of error and their classification is more likely to benefit from the use of external resources. We assigned each word in the training and testing data a frequency category corresponding to its frequency in the British National Corpus, a 100 million word balanced corpus, and used conjunctions of this category and certain other features.

2.2.2 Gazetteers

Our gazetteer contained only gene names and was compiled from lists from biomedical websites (such as LocusLink) as well as from the Gene Ontology and the data provided for the BioCreative 2004 tasks. The final gazetteer contained 1,731,496 entries. Because it contained only gene names, and for the reasons discussed earlier, we suspect that it was not terribly useful for identifying the presences of entities, but rather that it mainly helped to establish the exact beginning and ending point of multi-word entities recognized mainly through other features.

2.2.3 Web

For each of the named entity classes, we built indicative contexts, such as "X mRNA" for RNA, or "X ligation" for protein. For each entity X which had a frequency lower than 10 in the British National Corpus, we submitted instantiations of each pattern to the web, using the Google API, and obtained the number of hits. The pattern that returned the highest number of hits determined the feature value (e.g., "web-protein", or "web-RNA"). If no hits were returned by any pattern, a value "O-web" was assigned. This value was also assigned to all words whose frequency was higher than 10 (using yet another value for words with higher frequency did not improve the tagger's performance).

2.2.4 Abstracts

A number of NER systems have made effective use of how the same token was tagged in different parts of the same document (see (Curran and Clark, 2003) and (Mikheev et al., 1999)). A token which appears in an unindicative context in one sentence may appear in a very obvious context in another sentence in the same abstract. To leverage this we tagged each abstract twice, providing for each token a feature indicating whether it was tagged as an entity elsewhere in the abstract. This information was only useful when combined with information on frequency.

2.3 Deeper Syntactic Features

While the local features discussed earlier are all fairly surface level, our system also makes use of deeper syntactic features. We fully parsed the training and testing data using the Stanford Parser of (Klein and Manning, 2003) operating on the TnT part-of-speech tagging – we believe that the unlexicalized nature of this parser makes it a particularly suitable statistical parser to use when there is a large domain mismatch between the training material (*Wall Street Journal* text) and the target domain, but have not yet carefully evaluated this. Then, for each word in the sentence which is inside a noun phrase, the head and governor of the noun phrase are extracted. These features are not very useful when identifying only two classes (such as `GENE` and `OTHER` in the BioCreative task), but they were quite useful for this task because of the large number of classes which the system needed to distinguish between. Because the classifier is now choosing between classes where members can look very similar, longer distance information can provide a better representation of the context in which the word appears. For instance, the word *phosphorylation* occurs in the training corpus 492 times, 482 of which it is was classified as `other`. However, it is the governor of 738 words, of which 443 are `protein`, 292 are `other` and only 3 are `cell line`.

We also made use of abbreviation matching to help ensure consistency of labels. Abbreviations and long forms were extracted from the data using the method of (Schwartz and Hearst, 2003). This data was combined with a list of other abbreviations and long forms extracted from the BioCreative 2004 task. Then all occurrences of either the long or short forms in the data was labeled. These labels were included in the system as features and helped to improve boundary detection.

2.4 Adjacent Entities

When training our classifier, we merged the `B-` and `I-` labels for each class, so it did not learn how to differentiate between the first word of a class and internal word. There were several motivations for doing this. Foremost was memory concerns; our final system trained on just the six classes had 1.5 million features – we just did not have the resources to train it over more classes without giving up many of our features. Our second motivation was that by merging the beginning and internal labels for a particular class, the classifier would see more examples of that class and learn better how to identify it. The drawback of this move is that when two entities belonging to the same class are adjacent, our classifier will automatically merge them into one entity. We did attempt to split them back up using NP chunks, but this severely reduced performance.

3 Results and Discussion

Our results on the evaluation data and a confusion matrix are shown in Tables 2 and 4. Table 4 suggests areas for further work. Collapsing the B- and I- tags does cost us quite a bit. Otherwise confusions between some named entity and being nothing are most of the errors, although protein/DNA and cell-line/cell-type confusions are also noticeable.

Entity	Precision	Recall	F-Score
Fully Correct			
protein	77.40%	68.48%	72.67%
DNA	66.19%	69.62%	67.86%
RNA	72.03%	65.89%	68.83%
cell line	59.00%	47.12%	52.40%
cell type	62.62%	76.97%	69.06%
Overall	**71.62%**	**68.56%**	**70.06%**
Left Boundary Correct			
protein	82.89%	73.34%	77.82%
DNA	68.47%	72.01%	70.19%
RNA	75.42%	68.99%	72.06%
cell line	63.80%	50.96%	56.66%
cell type	63.93%	78.57%	70.49%
Overall	**75.72%**	**72.48%**	**74.07%**
Right Boundary Correct			
protein	84.70%	74.96%	79.53%
DNA	74.43%	78.29%	76.31%
RNA	78.81%	72.09%	75.30%
cell line	70.2%	56.07%	62.34%
cell type	71.68%	88.10%	79.05%
Overall	**79.65%**	**76.24%**	**77.91%**

Table 2: Results on the evaluation data

Analysis of performance in biomedical Named Entity Recognition tends to be dominated by the perceived poorness of the results, stemming from the twin beliefs that performance of roughly ninety percent is the state-of-the-art and that performance of 100% (or close to that) is possible and the goal to be aimed for. Both of these beliefs are questionable, as the top MUC 7 performance of 93.39% (Mikheev et al., 1998) in the domain of newswire text used an easier performance metric where incorrect boundaries were given partial credit, while both the biomedical NER shared tasks to date have used an exact match criterion where one is doubly penalized (both as a FP and as a FN) for incorrect boundaries. However, the difference in metric clearly cannot account entirely for the performance discrepancy between newswire NER and biomedical NER. Biomedical NER appears to be a harder task due to the widespread ambiguity of terms out of context, the complexity of medical language, and the apparent need for expert domain knowledge. These are problems that more sophisticated machine learning systems using resources such as ontologies and deep processing might be able to overcome. However, one should also consider the inherent "fuzziness" of the classification task. The few existing studies of inter-annotator agreement for biomedical named entities have measured agreement between 87% (Hirschman, 2003) and 89% (Demetriou and Gaizauskas, 2003). As far as we know there are no inter-annotator agreement results for the GENIA corpus, and it is necessary to have such results before properly evaluating the performance of systems. In particular, the fact that BioNLP sought to distinguish between gene and protein names, when these are known to be systematically ambiguous, and when in fact in the GENIA corpus many entities were doubly classified as "protein molecule or

gold\ans	DNA B-	DNA I-	RNA B-	RNA I-	cell line B-	cell line I-	cell type B-	cell type I-	protein B-	protein I-	O
B-DNA	723	39	0	0	1	0	0	0	154	1	138
I-DNA	52	1390	0	0	0	0	0	0	19	71	257
B-RNA	1	0	89	3	0	0	0	0	14	0	11
I-RNA	0	1	5	164	0	0	0	0	2	0	15
B-cell_line	3	0	0	0	319	41	37	5	12	1	82
I-cell_line	0	6	0	0	24	713	5	104	0	14	123
B-cell_type	1	0	0	0	164	22	1228	90	31	5	380
I-cell_type	0	0	0	0	13	383	88	2101	8	27	371
B-protein	48	5	10	3	20	1	19	3	4200	192	566
I-protein	6	66	0	11	0	10	2	25	245	3630	779
O	170	240	25	26	85	142	184	132	1042	656	78945

Table 3: Our confusion matrix over the evaluation data

human	B-cell_type	
monocytes	I-cell_type	
human	O	
monocytes	B-cell_type	
macrophages	B-cell_type	
primary	B-cell_type	JJ
T	I-cell_type	NN
lymphocytes	I-cell_type	NNS
primary	O	JJ
peripheral	B-cell_type	JJ
blood	I-cell_type	NN
lymphocytes	I-cell_type	NNS

Table 4: Examples of annotation inconsistencies

region" and "DNA molecule or region", suggests that inter-annotator agreement could be low, and that many entities in fact have more than one classification.

One area where GENIA appears inconsistent is in the labeling of preceding adjectives. The data was selected by querying for the term *human*, yet the term is labeled inconsistently, as is shown in Table 4. Of the 1790 times the term *human* occurred before or at the beginning of an entity in the training data, it was not classified as part of the entity 110 times. In the test data, there is only on instance (out of 130) where the term is excluded. Adjectives are excluded approximately 25% of the time in both the training and evaluation data. There are also inconsistencies when two entities are separated by the word *and*.

4 Acknowledgements

This paper is based on work supported in part by a Scottish Enterprise Edinburgh-Stanford Link Grant (R36759), as part of the SEER project, and in part by the National Science Foundation under the Knowledge Discovery and Dissemination program.

References

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *ANLP 6*, pages 224–231.

James R. Curran and Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-03)*, pages 164–167.

George Demetrious and Rob Gaizauskas. 2003. Corpus resources for development and evaluation of a biological text mining system. In *Proceedings of the Third Meeting of the Special Interest Group on Text Mining*, Brisbane, Australia, July.

Shipra Dingare, Jenny Rose Finkel, Christopher Manning, Malvina Nissim, and Beatrice Alex. 2004. Exploring the boundaries: Gene and protein identification in biomedical text. In *Proceedings of the BioCreative Workshop*.

Lynette Hirschman. 2003. Using biological resources to bootstrap text mining.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *ACL 41*, pages 423–430.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In *CoNLL 7*, pages 180–183.

Andrei Mikheev, Claire Grover, and Mark Moens. 1998. Description of the LTG system used for MUC-7. In *Proceedings of MUC-7*.

Andrei Mikheev, Marc Moens, and Claire Grover. 1999. Named entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 1–8. Association for Computational Linguistics.

Tomoko Ohta, Yuka Tateisi, Hideki Mima, and Jun'ichi Tsujii. 2002. GENIA corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of he Human Language Technology Conference*, pages 73–77.

Ariel Schwartz and Marti Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pacific Symposium on Biocomputing*, Kauai, Jan.