Exploring the Boundaries: Gene and Protein Identification in Biomedical Text

Shipra Dingare,* Jenny Finkel, Christopher Manning,****
Malvina Nissim,* Beatrice Alex*

*Institute for Communicating and Collaborative Systems
{sdingar1|mnissim|v1balex}@inf.ed.ac.uk
University of Edinburgh, United Kingdom

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

Abstract

We present a maximum-entropy based system incorporating a diverse set of features for identifying genes and proteins in biomedical abstracts. This system was entered in the BioCreative comparative evaluation and achieved the best performance in the "open" evaluation and the second-best performance in the "closed" evaluation. Central contributions are rich use of features derived from the training data at multiple levels of granularity, a focus on correctly identifying entity boundaries, and the innovative use of several external knowledge sources including full MEDLINE abstracts and web searches.

1. Introduction

The explosion of information in the biomedical domain and particularly in genetics has highlighted the need for automated information extraction techniques. MEDLINE, the primary research database serving the biomedical community, currently contains over 12 million abstracts, with 60,000 new abstracts appearing each month. There is also an impressive number of molecular biological databases covering an array of information on genes, proteins, nucleotide and amino acid sequences, both generally (GenBank, Swiss-Prot) and for particular species (FlyBase, Mouse Genome Informatics, WormBase, Saccharomyces Genome Database), and each containing entries numbering from the thousands to the millions and multiplying rapidly. All of these resources are curated by hand by expert annotators at enormous expense and the amount of information often prohibits updating previously annotated material to conform to changing annotation guidelines. This situation has naturally led to interest in automated techniques for problems such as topic classification, word sense disambiguation, and tokenization in the biomedical domain (cf. MEDLINE's Indexing Initiative). In this paper we focus on the particular problem of Named Entity Recognition (NER). As posed by the BioCreative evaluation, this task required participants to identify gene and protein mentions in medical abstracts. We present a system based on a maximum-entropy sequence tagger which achieved state-of-the-art performance in the BioCreative comparative evaluation. Below, we first describe the system (Section 2), then present its performance on the BioCreative Task 1A evaluation and development data along with an analysis of errors (Section 3), and finally close with a more general discussion of the task and our conclusions (Section 4).

2. System Description

Our entry was a machine learning system using a discriminatively trained conditional Markov model sequence tagger, based on the tagger used in (Klein et al., 2003). The system essentially uses a logistic regression model (with quadratic regularization) to classify each word, overlaid with a Viterbi-style algorithm to find the best sequence of classifications; such models are also known as maximum entropy Markov models or MEMMs. Maximum entropy models have been used with much success in NER tasks and are known for their ability to incorporate a large number of overlapping features. We devoted most of our efforts to finding useful features. The final system makes exhaustive use of clues within the sentence, as well as using various external resources. We trained our models using the tokenization supplied by the task organizers (although we had doubts of its quality). We normalized names of months and days of the week to lowercase, and mapped the British spellings of a few common medical terms to their American versions. In the following sections we describe our full feature set. We outline first the features used in the closed section (where gazetteers were not allowed) and subsequently the features used in the open section (where all external resources were allowed). We also describe a postprocessing phase aimed at reducing boundary errors. Our final system was trained on the combined training and development data using a Quasi-Newton optimization procedure and employed approximately 1.25 million features.

2.1. Features – Closed Section

The features described here were used in both the closed and open sections. The basic feature types were words, character substrings, word shapes, POS tags, abbreviations and the NE tags assigned to surrounding words. Character substrings refer to all substrings of the current word, up to a length of 6 characters. Thus the word "bio" would have features *_b*, *_bi*, *_bio*, *_bio_*, *bio_*, *io_*, *o_*, *bio*, *bi*, *io*, *b*, *i*, *o*. Word shapes refer to mappings of each word to a simplified representation that encodes attributes such as its length and whether it contains capitalization, numerals, greek letters, and so on. Thus "Varicella-zoster" would become *Xx-xxx*,

"mRNA" would become *xXXX*, and "CPA1" would become *XXXd*. POS tags were incorporated from the TnT POS tagger[1] (Brants, 2000) trained on the GENIA corpus which provides a gold standard for POS tags in biomedical text.[2] A feature encoding whether each word was an abbreviation, a long form, or neither was assigned using the method of (Schwartz and Hearst, 2003). Lastly, a parentheses-matching feature that signalled when one parenthesis was classified differently from its pair was added in an effort to eliminate errors where the tagger classified matching parentheses differently. All of these basic feature types were then used singly or combined in various ways to create new features. Word identity features were also used disjunctively on left and right contexts. The resulting feature set is summarized in Table 1 and comprises all of the features used in the closed section. Beyond standard word and POS tag features, character substring and word shape features were central players in the system of (Klein et al., 2003). We borrowed disjunctive word features from (Kazama et al., 2002), and introduced abbreviation and parentheses matching features to model key problems in this textual domain.

2.2. Features – Open Section

The features described here were used in the "open" entry and comprise various external resources including gazetteers, a web querying technique, the full abstracts corresponding to the sentences in training and test sets, the GENIA corpus, and the ABGene NE/POS tagger. The basic assumption behind and motivation for using external resources is that there are instances in the data where contextual clues do not provide sufficient evidence for confident classification. In such cases external resources may bridge the gap, either in the form of word lists known to refer to genes (gazetteers) or through examination of other contexts in which the same token appears and the exploitation of more indicative contexts (as with web-querying and use of surrounding text such as abstracts).

All external resources are vulnerable to incompleteness, noise, and ambiguity. Gazetteers are arguably subject to all three and yet have been used successfully in a number of systems. Because of its size (Google currently searches over 4,285M web pages[3]), the web is the least vulnerable to incompleteness but is highly vulnerable to noise. Nevertheless, the web has been used to good effect in various NLP tasks (see (Keller and Lapata, 2003) for an overview) from machine translation (Grefenstette, 1999) to anaphora resolution (Modjeska et al., 2003). Abstracts do not contain indicative contexts as frequently because they are so small; however their information is least vulnerable to ambiguity because a token used repeatedly within a text is likely used with the same meaning each time. Information on a word's classification elsewhere in the same text has been successfully used in other NER systems (cf. (Curran and Clark, 2003)). By incorporating all of these resources as features in a probabilistic system, we aimed to make use of their information while taking into account their reliability.

Word Features	w_i
	w_{i-1}
	w_{i+1}
	Last "real" word
	Next "real" word
	Disjunction of 4 previous words
	Disjunction of 4 next words
Bigrams	w_i + w_{i-1}
	w_i + w_{i+1}
TnT POS	POS_i
	POS_{i-1}
	POS_{i+1}
Character Substrings	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	$shape_i$
	$shape_{i-1}$
	$shape_{i+1}$
	$shape_{i-1}$ + $shape_i$
	$shape_i$ + $shape_{i+1}$
	$shape_{i-1}$ + $shape_i$ + $shape_{i+1}$
Previous NE	NE_{i-1}
	NE_{i-2}+ NE_{i-1}
Previous NE + Word	NE_{i-1}+w_i
Previous NE + POS	NE_{i-1} + POS_{i-1} + POS_i
	NE_{i-2} + NE_{i-1} + POS_{i-2} + POS_{i-1} + POS_i
Previous 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	A feature that signals when one parentheses in a pair has been assigned a different tag than the other in a window of 4 words

Table 1: Full Feature Set Used In Closed Section

Our gazetteer was compiled from lists of gene names from biomedical sites on the Web (such as Locus Link) as well as from the Gene Ontology and the data provided for Tasks 1A and 1B. The gazetteer was cleaned by removing single character entries ("A", "1"), entries containing only digits or symbols and digits ("37", "3-1"), and entries containing only words that could be found in the English dictionary CELEX ("abnormal", "brain tumour"). The final gazetteer contained 1,731,581 entries.

In using the web we built several contexts indicative of gene entities including "X gene", "X mutation" or "X antagonist". For each entity X identified as a gene by an initial run of the tagger, we submitted the instantiation of each pattern to the Web using the Google API and obtained the number of hits. If at least one of the patterns returned more than zero hits, the string was assigned a 'web' value for the Web feature. The classifier was then run again; this time incorporating the web feature. Using web-querying only on

[1]The TnT POS tagger is an HMM-based tagger; perhaps due to greater robustness, we found that it outperformed the maximum entropy POS tagger that was available to us.

[2]Testing showed that a GENIA-trained POS tagger performed much better than one trained on *Wall Street Journal* text, presumably due to the idiosyncratic nature of biomedical text.

[3]Estimate from `www.google.com`, 26.02.2004

likely candidates for genes as identified by an initial run of the tagger was more efficient than using it on all words.

To use the abstracts, we automatically located the full Medline abstract from which each BioCreative sentence was taken, and incorporated additional information by tagging the abstract and then adding to words in the corresponding sentence a feature that indicated whether the word was tagged as a gene in the abstract. We found that this feature was only helpful when combined with other information such as frequency. Presumably this was due to common words for which the abstract feature was misleading; the fact that the word "gene" was tagged as a gene in the phrase "CPA1 gene" does not indicate that it is a gene in the phrase "a gene".

The final two external resources that we incorporated were the ABGene tagger (Tanabe and Wilbur, 2002) and the GENIA corpus (Ohta et al., 2002). We found that while the ABGene tagger used alone achieved only a modest f-score of 0.62 on the BioCreative development data, use of ABGene NE output as a feature nevertheless slightly improved our recall and overall f-score. We assume that this is because its use allowed our classifier to partially exploit the various gazetteers and lists of good and bad terms incorporated into the ABGene system (see (Tanabe and Wilbur, 2002)), thereby gaining additional knowledge of gene names independent of context. We also sought to incorporate the GENIA corpus of NE-annotated MEDLINE abstracts but found this difficult because it used an entirely different tag set than the BioCreative data and the mapping between them was unclear. We gained a small improvement by training the C&C tagger (Curran and Clark, 2003) on the full NE tag set of the GENIA corpus (consisting of 37 biomedical NEs including "cell type" and "protein molecule"), then using this tagger to tag both training and test data and using its output as a feature in our final tagger.[4]

2.3. Postprocessing

We found that many of our errors stemmed from gene boundaries and addressed this issue in several ways. Boundary errors were often due to mismatched parentheses; the parentheses-matching feature described in Section 2.2. did not eliminate these errors due to instances in the training data which contained mismatched parentheses. We therefore used grep to remove genes containing mismatched parentheses from our results. We also found that we obtained different gene boundaries when we ran the classifier forwards versus backwards (reversing the order of the words) and obtained a significant improvement in recall at the expense of precision by simply combining the two sets of results. This new, larger set of genes contained instances where one gene was a substring of another gene. In those instances we kept only the shorter gene. We found that this postprocessing was highly valuable and added approximately 1% to our f-score. It was used in both the open and closed sections.

[4]The C&C tagger is another maximum entropy sequence tagger; it was used for pragmatic reasons related to memory use.

3. Results and Analysis

	Precision	Recall	F-Score
Open	0.813	0.861	0.836
Closed	0.784	0.852	0.817

Table 2: Results on Cross-Validated Training/Dev Data

	Precision	Recall	F-Score
Open	0.828	0.835	0.832
Closed	0.792	0.854	0.822

Table 3: Results on Evaluation Data

	Precision	Recall	F-Score	Δ F
Abbreviations	0.813	0.860	0.836	-0.05%
Abgene	0.810	0.861	0.834	-0.18%
Abstract	0.811	0.855	0.832	-0.39%
Gazette	0.807	0.857	0.831	-0.51%
Genia	0.806	0.857	0.831	-0.55%
Substrings	0.814	0.852	0.833	-0.37%
$POS_{i,i-1,i+1}$	0.814	0.860	0.836	-0.03%
Google Web	0.807	0.864	0.835	-0.17%
Word Shape	0.815	0.862	0.838	+0.13%
Zero Order	0.741	0.799	0.770	-6.66%
First Order	0.818	0.853	0.835	-0.15%
Second Order	0.814	0.861	0.837	+0.06%
Third Order	0.814	0.863	0.837	+0.07%

Table 4: Results on Cross-Validated Training and Development Data With One Feature Removed At a Time

The tables above show the performance of both the "open" and "closed" versions of the system on the development and evaluation data as well as lesion studies showing the individual contribution of feature classes to the overall performance. Surprisingly, the "closed" version of the system achieves performance only 1% lower than the "open" version on the evaluation data (2% on the development data). We had expected more value from extra data sources, but it may well be that they are difficult to exploit effectively because of subtly different decisions about what does and does not count as a gene. However, it is also worth noting that a 1-2% improvement is relatively significant; as the performance of the classifier gradually improved the improvements became progressively smaller so that at times features were incorporated which added only a tenth of a point. Also surprising was that removing word shape features actually increased our F-Score by 0.13%. The "zero order" and "first-order" experiments refer to how far back the classifier can see the NE tags assigned to previous words during sequence search. Thus a zero-order model can only see the classification of the current word, while a first-order model can see the classification assigned to the previous word (but not the words before). Removing second and third order features also improved our result marginally.

False Positives	**Classifier (CL)**	**Gold Standard (GS)**
General Words	homolog gene	-
Measures	kat/L	-
Possible Errors in GS	[ssDNA-] and [RNA-binding protein]	ssDNA- and [RNA-binding protein]
False Negatives	**Classifier (CL)**	**Gold Standard (GS)**
Coordination	[YAP2 uORF1] and uORF2	[YAP2 uORF1] and [uORF2]
Missing Expansion	zinc-finger protein ([THZif-1])	[zinc-finger protein] ([THZif-1])
Boundary Errors	**Classifier (CL)**	**Gold Standard (GS)**
GS NE contains CL NE(s)	AP-1 complexes	high mobility AP-1 complexes
	USH1C	USH1C disease gene
	partner of [Rac]	[partner of Rac]
CL NE contains GS NE(s)	regulator virF	virF
	Wnt pathway	Wnt
CL and GS Overlap	Serum [Fibrin Degradation Products]	[Serum Fibrin] Degradation Products

Table 5: Examples of FPs, FNs and boundary errors. In some of the examples square brackets are used to indicate the differences between the classifier's output and the annotation in the gold standard.

3.1. Sources of Error

A number of false positives (FPs) occurred when the entity tagged by the classifier was a description of a gene rather than a gene name, as with "homologue gene". FPs also occurred with several strings that were composed of characters and digits or sequences of capitalised letters, or that included symbols and punctuation. This occurred frequently with measures, such as "kat/L" (katal per litre) and acronyms for non-gene entities. Acronym ambiguity was a related source of error. The abbreviation "HAT", for instance, could stand for the gene name "histone acetyltransferase" but actually referred to "hepatic artery thrombosis" in the specific context.

False negatives (FNs) were frequently caused by gene names that had not been encountered in the training data, so that the classifier did not have information about them and contextual clues were insufficient. FNs also occurred in some coordinated NPs where the modifier was attached to only one of the phrases but modified all of the coordinated members. Abbreviations, expansions, and names in parentheses were also frequent causes of FNs.

The single largest source of error was mistaken boundaries (37% of FP and 39% of FN). In most cases, the classifier identified one correct and one incorrect boundary (i.e. either the beginning or the end). It often included left or right context as part of the entity which was not contained in the gold standard. In several instances, the classifier split a string into separate entities which in fact referred to a single entity, or tagged separate entities as a single one. Tokenisation errors sometimes triggered boundary errors, as with "PGS-2 . CAT reporter gene" where the classifier only recognized "CAT reporter" as a gene. Because many abbreviations were not genes and because the precision and recall of the gazetteer were fairly low, we believe that both abbreviation and gazetteer features helped more in identifying gene boundaries than in identifying genes.

3.2. Directions for Improvement

The learning curve in Figure 1 suggests that we can expect only very limited improvement from the availability of additional training data, given the current task and feature set. Rather we must explore other avenues, including better exploitation of existing features and resources, development of additional features, incorporation of additional external resources, or experimentation with other algorithms and strategies for approaching the task.

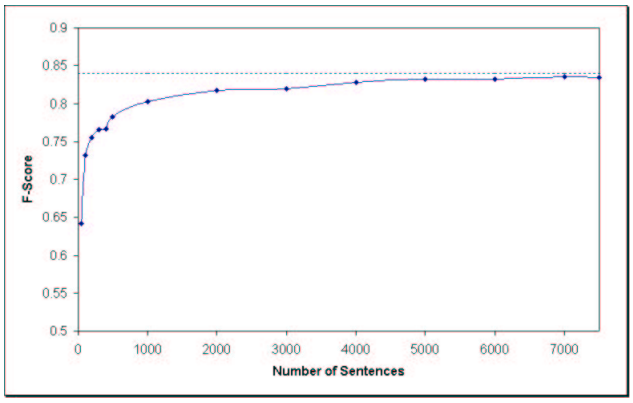

Figure 1: Performance of NER system on development data

One obvious improvement of our current system would be the incorporation of protein names into our gazetteer. Due to ambiguity in the guidelines we were unaware that protein names were to be recognized and incorporated only gene names into our gazetteers.

Secondly, boundary errors might be addressed more effectively with a more sophisticated post-processing stage. Considering only the problem of segmentation of NEs, Collins (2002) applies reranking to candidate structures generated from a maximum-entropy tagger and achieves a 17.7% relative reduction in error rate. Reranking was used to utilize features that describe the full NE identified by the tagger, such as its first and last words and attributes thereof, and whether all words between a set of quotes were given the same tag (reminiscent of the parentheses problems in our data). Such features cannot be encoded in a standard sequence tagger.

Another possible avenue would be automatic addition of conjunctions of current features (Della Pietra et al., 1997). A number of the features listed in Table 1 as well as the features used to incorporate external resources are relatively unintuitive conjunctions of other features that were

chosen by lengthy trial and error processes. Feature induction might suggest useful feature conjunctions that we have overlooked and reduce the cost of incorporating additional resources. The use of automatic feature induction would also detract from the criticism that if 25 person-weeks are necessary to develop features for a supposedly machine learning system, could one not develop a system of handcrafted rules in the same time?

4. Conclusions

We have presented in detail a machine learning system for identifying genes and proteins in text and described its feature set comprising both contextual clues and external resources. We have also presented its performance on the BioCreative development and evaluation data, analyzed its sources of error, and identified possible avenues for improvement.

Many of our features were focused on increasing the correct identification of entity boundaries. This is partly an artifact of the scoring metric: using an f-score of exact match precision and recall means that one is penalized twice, both for a FP and a FN, in cases of an incorrect boundary identification. One scores better in such cases if one suggests no entity.[5] But it equally reflects that finding correct entity boundaries in the biomedical domain is an extremely hard task, whereas in many cases it is quite trivial for people or place names in English – capitalization giving sufficient clues.

The final performance of the tagger at 0.83 f-score remains far below the best results reported for the most well-researched NER task of PERSON, LOCATION, and ORGANIZATION entities in newswire texts. Using the set of features designed for that task in CoNLL 2003 (Sang and De Meulder, 2003), our system achieves an f-score of 0.76 on the BioCreative development data, a dramatic ten points lower than its f-score of 0.86 on the CoNLL newswire data. The discrepancy in performance is a striking illustration of the greater difficulty of NER in the biomedical domain.

It is worth comparing these performance figures with levels of interannotator agreement in the biomedical domain. Interannotator agreement effectively provides a ceiling on the performance that can be expected from a system by measuring how well a human annotator performs on a task. While agreement for the MUC entities was measured at 97%, a number of results have measured agreement for biomedical NEs to be substantially lower, with f-scores in the range of 0.87 (Hirschman, 2003) to 0.89 (Demetriou and Gaizauskas, 2003). With interannotator agreement so low, it appears that we cannot currently expect to improve system performance more than a few points. This suggests that more clarity in what should be annotated (or perhaps just when a variety of answers of different extent should be counted as correct) is needed. It also may suggest that performance of 83% or improvement of just a few points is sufficient for the technology to be practically applicable.

[5] The CoNLL task used the same metric, but note that the "mid-nineties" results commonly remembered from MUC NER competitions reflect an easier metric where partial credit was given for cases of incorrect boundary identification. Note also that the BioCreative evaluation had a facility for annotators to be able to specify alternate correct answers, which ameliorated this problem by allowing as correct matches of several lengths in places where the annotators thought it appropriate.

5. Acknowledgments

We are grateful to the UK National e-Science Centre for giving us access to BlueDwarf, a p690 server donated to the University of Edinburgh by IBM. Thanks also go to Steve Clark and James Curran for the use of their tagger and particularly to James for his valuable advice at the last minute. This work was performed as part of the SEER project, which is supported by a Scottish Enterprise Edinburgh-Stanford Link Grant (R36759).

6. References

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

Michael Collins. 2002. Ranking algorithms for named entity extraction: Boosting and the voted perceptron. In *ACL*, pages 489–496.

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, Edmonton, Canada.

Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. 1997. Inducing features of random fields. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 19:380–393.

George Demetriou and Rob Gaizauskas. 2003. Corpus resources for development and evaluation of a biological text mining system, July.

Gregory Grefenstette. 1999. The WWW as a resource for example-based MT tasks. In *Proceedings of ASLIB'99 Translating and the Computer 21*, London.

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

Jun'ichi Kazama, Takaki Makino, Yoshihiro Ohta, and Jun ichi Tsujii. 2002. Biomedical name recognition: Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL 2002 Workshop on Natural Language Processing in the Biomedical Domain*, pages 1–8.

Frank Keller and Maria Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

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

N. Modjeska, K. Markert, and M. Nissim. 2003. Using the web in machine learning for other-anaphora resolution. In *Proc. of the 2003 Conference on Empirical Methods in Natural Language Processing; Sapporo, Japan, 6–7 July 2002*, pages 176–183.

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 the Human Language Technology Conference (HLT 2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2003*, pages 142–147.

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

Lorraine Tanabe and W. J. Wilbur. 2002. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18:1124–1132.