Disease Mention Recognition with Specific Features

Md. Faisal Mahbub Chowdhury [†‡] **and Alberto Lavelli** [‡]
[‡] Human Language Technology Research Unit, Fondazione Bruno Kessler, Trento, Italy
[†] ICT Doctoral School, University of Trento, Italy
{chowdhury,lavelli}@fbk.eu

Abstract

Despite an increasing amount of research on biomedical named entity recognition, there has been not enough work done on disease mention recognition. Difficulty of obtaining adequate corpora is one of the key reasons which hindered this particular research. Previous studies argue that correct identification of disease mentions is the key issue for further improvement of the disease-centric knowledge extraction tasks. In this paper, we present a machine learning based approach that uses a feature set tailored for disease mention recognition and outperforms the state-of-the-art results. The paper also discusses why a feature set for the well studied gene/protein mention recognition task is not necessarily equally effective for other biomedical semantic types such as diseases.

1 Introduction

The massive growth of biomedical literature volume has made the development of biomedical text mining solutions indispensable. One of the essential requirements for a text mining application is the ability to identify relevant entities, i.e. named entity recognition. Previous work on biomedical named entity recognition (BNER) has been mostly focused on gene/protein mention recognition. Machine learning (ML) based approaches for gene/protein mention recognition have already achieved a sufficient level of maturity (Torii et al., 2009). However, the lack of availability of adequately annotated corpora has hindered the progress of BNER research for other semantic types such as diseases (Jimeno et al., 2008; Leaman et al., 2009).

Correct identification of diseases is crucial for various disease-centric knowledge extraction tasks (e.g. drug discovery (Agarwal and Searls, 2008)). Previous studies argue that the most promising candidate for the improvement of disease related relation extraction (e.g. disease-gene) is the correct identification of concept mentions including diseases (Bundschus et al., 2008).

In this paper, we present a BNER system which uses a feature set specifically tailored for disease mention recognition. The system[1] outperforms other approaches evaluated on the Arizona Disease Corpus (AZDC) (more details in Section 5.1). One of the key differences between our approach and previous approaches is that we put more emphasis on the contextual features. We exploit syntactic dependency relations as well. Apart from the experimental results, we also discuss why the choice of effective features for recognition of disease mentions is different from that for the well studied gene/protein mentions.

The remaining of the paper is organized as follows. Section 2 presents a brief description of previous work on BNER for disease mention recognition. Then, Section 3 describes our system and Section 4 the feature set of the system. After that, Section 5 explains the experimental data, results and analyses. Section 6 describes the differences for the choice of feature set between diseases and genes/proteins. Finally, Section 7 concludes the paper with an outline of our future research.

2 Related Work

Named entity recognition (NER) is the task of locating boundaries of the entity mentions in a text and tagging them with their corresponding semantic types (e.g. person, location, gene and so on). Although several disease annotated corpora have been released in the last few years, they have been annotated primarily to serve the purpose of relation extraction and, for different reasons, they

[1] The source code of our system is available for download at http://hlt.fbk.eu/people/chowdhury/research

Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, pages 83–90,
Uppsala, Sweden, 15 July 2010.

are not suitable for the development of ML based disease mention recognition systems (Leaman et al., 2009). For example, the BioText (Rosario and Hearst, 2004) corpus has no specific annotation guideline and contains several inconsistencies, while PennBioIE (Kulick et al., 2004) is very specific to a particular sub-domain of diseases. Among other disease annotated corpora, EBI disease corpus (Jimeno et al., 2008) is not annotated with disease mention boundaries which makes it unsuitable for BNER evaluation for diseases. Recently, an annotated corpus, named as Arizona Disease Corpus (AZDC) (Leaman et al., 2009), has been released which has adequate and suitable annotation of disease mentions following specific annotation guidelines.

There has been some work on identifying diseases in clinical texts, especially in the context of CMC Medical NLP Challenge[2] and i2b2 Challenge[3]. However, as noted by Meystre et al. (2008), there are a number of reasons that make clinical texts different from texts of biomedical literature, e.g. composition of short, telegraphic phrases, use of implicit templates and pseudo-tables and so on. Hence, the strategies adopted for NER on clinical texts are not the same as the ones practiced for NER on biomedical literature.

As mentioned before, most of the work to date on BNER is focused on gene/protein mention recognition. State-of-the-art BNER systems are based on ML techniques such as conditional random fields (CRFs), support vector machines (SVMs) etc (Dai et al., 2009). These systems use either gene/protein specific features (e.g. Greek alphabet matching) or post-processing rules (e.g. extension of the identified mention boundaries to the left when a single letter with a hyphen precedes them (Torii et al., 2009)) which might not be as effective for other semantic type identification as they are for genes/proteins. There is a substantial agreement in the feature set that these systems use (most of which are actually various orthographical and morphological features).

Bundschus et al. (2008) have used a CRF based approach that uses typical features for gene/protein mention recognition (i.e. no feature tailoring for disease recognition) for disease, gene and treatement recognition. The work has been evaluated on two corpora which have been annotated with those entities that participate in disease-gene and disease-treatment relations. The reported results show F-measure for recognition of all the entities that participate in the relations and do not indicate which F-measure has been achieved specifically for disease recognition. Hence, the reported results are not applicable for comparison.

To the best of our knowledge, the only systematic experimental results reported for disease mention recognition in biomedical literature using ML based approaches are published by Leaman and Gonzalez (2008) and Leaman et al. (2009).[4] They have used a CRF based BNER system named BANNER which basically uses a set of orthographic, morphological and shallow syntactic features (Leaman and Gonzalez, 2008). The system achieves an F-score of 86.43 on the BioCreative II GM corpus[5] which is one of the best results for gene mention recognition task on that corpus.

BANNER achieves an F-score of 54.84 for disease mention recognition on the BioText corpus (Leaman and Gonzalez, 2008). However, as said above, the BioText corpus contains annotation inconsistencies[6]. So, the corpus is not ideal for comparing system performances. The AZDC corpus is much more suitable as it is annotated specifically for benchmarking of disease mention recognition systems. An improved version of BANNER achieves an F-score of 77.9 on AZDC corpus, which is the state of the art on ML based disease mention recognition in biomedical literature (Leaman et al., 2009).

3 Description of Our System

There are basically three stages in our approach – pre-processing, feature extraction and model training, and post-processing.

3.1 Pre-processing

At first, the system uses GeniaTagger[7] to tokenize texts and provide PoS tagging. After that, it corrects some common inconsistencies introduced by GeniaTagger inside the tokenized data (e.g. GeniaTagger replaces double inverted commas with

[2]http://www.computationalmedicine.org/challenge/index.php
[3]https://www.i2b2.org/NLP/Relations/Main.php

[4]However, there are some work on disease recognition in biomedical literature using other techniques such as morpho-syntactic heuristic based approach (e.g. MetaMap (Aronson, 2001)), dictionary look-up method and statistical approach (Névéol et al., 2009; Jimeno et al., 2008; Leaman et al., 2009).
[5]As mentioned in http://banner.sourceforge.net/
[6]http://biotext.berkeley.edu/data/dis_treat_data.html
[7]http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

two single inverted commas). These PoS tagged tokized data are parsed using Stanford parser[8]. The dependency relations provided as output by the parser are used later as features. The tokens are further processed using the following generalization and normalization steps:

- each number (both integer and real) inside a token is replaced with *'9'*
- each token is further tokenized if it contains either punctuation characters or both digits and alphabetic characters
- all letters are changed to lower case
- all Greek letters (e.g. alpha) are replaced with *G* and Roman numbers (e.g. iv) with *R*
- each token is normalized using SPECIALIST lexicon tool[9] to avoid spelling variations

3.2 Feature extraction and model training

The features used by our system can be categorized into the following groups:

- general linguistic features (Table 1)
- orthographic features (Table 2)
- contextual features (Table 3)
- syntactic dependency features (Table 4)
- dictionary lookup features (see Section 4)

During dictionary lookup feature extraction, we ignored punctuation characters while matching dictionary entries inside sentences. If a sequence of tokens in a sentence matches an entry in the dictionary, the leftmost token of that sequence is labeled with B-DB and the remaining tokens of the sequence are labeled with I-DB. The label B-DB indicates the beginning of a dictionary match. If a token belongs to several dictionary matches, then all the other dictionary matches except the longest one are discarded.

The syntactic dependency features are extracted from the output of the parser while the general linguistic features are extracted directly from the preprocessed tokens. To collect the orthographic features, the original tokens inside the corresponding sentences are considered. The contextual features are derived using other extracted features and the original tokens.

Tokens are labeled with the corresponding disease annotations according to the IOB2 format. Our system uses Mallet (McCallum, 2002) to train a first-order CRF model. CRF is a state-of-the-art ML technique applied to a variety of text processing tasks including named entity recognition (Klinger and Tomanek, 2007) and has been successfully used by many other BNER systems (Smith et al., 2008).

3.3 Post-processing

Once the disease mentions are identified using the learned model, the following post-processing techniques are applied to reduce the number of wrong identifications:

- *Bracket mismatch correction:* If there is a mismatch of brackets in the identified mention, then the immediate following (or preceding) character of the corresponding mention is checked and included inside the mention if that character is the missing bracket. Otherwise, all the characters from the index where the mismatched bracket exists inside the identified mention are discarded from the corresponding mention.
- *One sense per discourse:* If any instance of a character sequence is identified as a disease mention, then all the other instances of that character sequence inside the same sentence are also annotated as disease mentions.
- *Short/long form annotation:* Using the algorithm of Schwartz and Hearst (2003), *"long form (short form)"* instances are detected inside sentences. If the short form is annotated as disease mention, then the long form is also annotated and vice versa.
- *Ungrammatical conjunction structure correction:* If an annotated mention contains comma (,) but there is no "and" in the following character sequence (from the character index of that comma) of that mention, then the annotation is splitted into two parts (at the index of the comma). Annotation of the original mention is removed and the splitted parts are annotated as two separate disease mentions.

[8] http://nlp.stanford.edu/software/lex-parser.shtml
[9] http://lexsrv3.nlm.nih.gov/SPECIALIST/index.html

- *Short and long form separation:* If both short and long forms are annotated in the same mention, then the original mention is discarded and the corresponding short and long forms are annotated separately.

4 Features for Disease Recognition

There are compelling reasons to believe that various issues regarding the well studied gene/protein mention recognition would not apply to the other semantic types. For example, Jimeno et al. (2008) argue that the use of disease terms in biomedical literature is well standardized, which is quite opposite for the gene terms (Smith et al., 2008).

After a thorough study and extensive experiments on various features and their possible combinations, we have selected a feature set specific to the disease mention identification which comprises features shown in Tables 1, 2, 4 and 3, and dictionary lookup features.

Feature name	Description
PoS	Part-of-speech tag
NormWord	Normalized token (see Section 3.1)
Lemma	Lemmatized form
charNgram	3 and 4 character n-grams
Suffix	2-4 character suffixes
Prefix	2-4 character prefixes

Table 1: General linguistic features for token$_i$

Feature name	Description
InitCap	Is initial letter capital
AllCap	Are all letters capital
MixCase	Does contain mixed case letters
SingLow	Is a single lower case letter
SingUp	Is a single upper case letter
Num	Is a number
PuncChar	Punctuation character (if token$_i$ is a punctuation character)
PrevCharAN	Is previous character alphanumeric

Table 2: Orthographic features for token$_i$

Like Leaman et al. (2009), we have created a dictionary with the instances of the following nine of the twelve UMLS semantic types from the semantic group "DISORDER"[10] from UMLS Metathesaurus (Bodenreider, 2004): (i) *disease or syndrome*, (ii) *neoplastic process*, (iii) *congenital abnormality*, (iv) *acquired abnormality*, (v) *experimental model of disease*, (vi) *injury or poisoning*, (vii) *mental or behavioral dysfunction*, (viii) *pathological function* and (ix) *sign or symptom*. We have not considered the other three semantic types (*findings*, *anatomical abnormality* and *cell or molecular Dysfunction*) since these three types have not been used during the annotation of Arizona Disease Corpus (AZDC) which we have used in our experiments.

Previous studies have shown that dictionary lookup features, i.e. name matching against a

Feature name	Description
Bi-gram$_{k,k+1}$ for $i-2 \leq k < i+2$	Bi-grams of normalized tokens
Tri-gram$_{k,k+1,k+2}$ for $i-2 \leq k < i+2$	Tri-grams of normalized tokens
CtxPoS$_{k,k+1}$ for $i \leq k < i+2$	Bi-grams of token PoS
CtxLemma$_{k,k+1}$ for $i \leq k < i+2$	Bi-grams of lemmatized tokens
CtxWord$_{k,k+1}$ for $i-2 \leq k < i+2$	Bi-grams of original tokens
Offset conjunctions	Extracted by Mallet from features in the range from token$_{i-1}$ to token$_{i+1}$

Table 3: Contextual features for token$_i$

Feature name	Description
dobj	Target token(s) to which token$_i$ is a direct object
iobj	Target token(s) to which token$_i$ is an indirect object
nsubj	Target token(s) to which token$_i$ is an active nominal subject
nsubjpass	Target token(s) to which token$_i$ is a passive nominal subject
nn	Target token(s) to which token$_i$ is a noun compound modifier

Table 4: Syntactic dependency features for token$_i$. For example, in the sentence "Clinton defeated Dole", "Clinton" is the *nsubj* of the *target token* "defeated".

[10] http://semanticnetwork.nlm.nih.gov/SemGroups/

dictionary of terms, often increase recall (Torii et al., 2009; Leaman et al., 2009). However, an unprocessed dictionary usually does not boost overall performance (Zweigenbaum et al., 2007). So, to reduce uninformative lexical differences or spelling variations, we generalize and normalize the dictionary entries using exactly the same steps followed for the pre-processing of sentences (see Section 3.1).

To reduce chances of false and unlikely matches, any entry inside the dictionary having less than 3 characters or more than 10 tokens is discarded.

5 Experiments

5.1 Data

We have done experiments on the recently released Arizona Disease Corpus (AZDC)[11] (Leaman et al., 2009). The corpus has detailed annotations of diseases including UMLS codes, UMLS concept names, possible alternative codes, and start and end points of disease mentions inside the corresponding sentences. These detailed annotations make this corpus a valuable resource for evaluating and benchmarking text mining solutions for disease recognition. Table 5 shows various characteristics of the corpus.

Item name	Total count
Abstracts	793
Sentences	2,783
Total disease mentions	3,455
Disease mentions without overlaps	3,093
Disease mentions with overlaps	362

Table 5: Various characteristics of AZDC.

For the overlapping annotations, (e.g. "endometrial and ovarian cancers" and "ovarian cancers") we have considered only the larger annotations in our experiments. There remain 3,224 disease mentions after resolving overlaps according to the aforementioned criterion. We have observed minor differences in some statistics of the AZDC reported by Leaman et al. (2009) with the statistics of the downloadable version[12] (Table 5). However, these differences can be considered negligible.

5.2 Results

We follow an experimental setting similar to the one in Leaman et al. (2009) so that we can compare our results with that of the BANNER system. We performed 10-fold cross validation on AZDC in such a way that all sentences of the same abstract are included in the same fold. The results of all folds are averaged to obtain the final outcome. Table 6 shows the results of the experiments with different features using the exact matching criterion.

As we can see, our approach achieves significantly higher result than that of BANNER. Initially, with only the general linguistic and orthographic features the performance is not high. However, once the contextual features are used, there is a substantial improvement in the result. Note that BANNER does not use contextual features. In fact, the use of contextual features is also quite limited in other BNER systems that achieve high performance for gene/protein identification (Smith et al., 2008).

Dictionary lookup features provide a very good contribution in the outcome. This supports the argument of Jimeno et al. (2008) that the use of disease terms in biomedical literature is well standardized. Post-processing and syntactic dependency features also increase some performance.

We have done statistical significance tests for the last four experimental results shown in Table 6. For each of such four experiments, the immediate previous experiment is considered as the baseline. The tests have been performed using the approximate randomization procedure (Noreen, 1989). We have set the number of iterations to 1,000 and the confidence level to 0.01. According to the tests, the contributions of contextual features and dictionary lookup features are statistically significant. However, we have found that the contributions of post-processing rules and syntactic dependency features are statistically significant only when the confidence level is 0.2 or more. Since AZDC consists of only 2,783 sentences, we can assume that the impact of post-processing rules

[11] Downloaded from http://diego.asu.edu/downloads/AZDC/ at 5-Feb-2009

[12] Note that *"Disease mentions (total)"* in the paper of Leaman et al. (2009) actually refers to the *total disease mentions after overlap resolving* (Robert Leaman, personal communication). One other thing is, Leaman et al. (2009) mention 794 abstracts, 2,784 sentences and 3,228 (overlap resolved) disease mentions in the AZDC. But in our downloaded version of AZDC, there is 1 abstract missing (i.e. total 793 abstracts instead of 794). As a result, there is 1 less sentence and 4 less (overlap resolved) disease mentions than the originally reported numbers.

and syntactic dependency features has been not so significant despite of some performance improvement.

5.3 Error analysis

One of the sources of errors is the annotations having conjunction structures. There are 94 disease mentions in the data which contain the word "and". The boundaries of 11 of them have been wrongly identified during experiments, while 39 of them have been totally missed out by our system. Our system also has not performed well for disease annotations that have some specific types of prepositional phrase structures. For example, there are 80 disease annotations having the word "of" (e.g. "deficient activity of acid beta-glucosidase GBA"). Only 28 of them are correctly annotated by our system. The major source of errors, however, concerns abbreviated disease names (e.g. "PNH"). We believe one way to reduce this specific error type is to generate a list of possible abbreviated disease names from the long forms of disease names available in databases such as UMLS Metathesaurus.

6 Why Features for Diseases and Genes/Proteins are not the Same

Many of the existing BNER systems, which are mainly tuned for gene/protein identification, use features such as token shape (also known as word class and brief word class (Settles, 2004)), Greek alphabet matching, Roman number matching and so forth. As mentioned earlier, we have done extensive experiments with various feature combinations for the selection of disease specific features. We have observed that many of the features used for gene/protein identification are not equally effective for disease identification. Table 7 shows some of the results of those experiments.

This observation is reasonable because gene/protein names are much more complex than entities such as diseases. For example, they often contain punctuation characters (such as parentheses or hyphen), Greek alphabets and digits which are unlikely in disease names. Ideally, the ML algorithm itself should be able to utilize information from only the useful features and ignore the others in the feature set. But practically, having non-informative features often mislead the model learning. In fact, several surveys have argued that the choice of features matter at least as much as the choice of the algorithm if not more (Nadeau and Sekine, 2007; Zweigenbaum et al., 2007).

One of the interesting trends in gene/protein mention identification is to not utilize syntactic dependency relations (with the exception of Vlachos (2007)). Gene/protein names in biomedical literature are often combined (i.e. without being separated by space characters) with other characters which do not belong to the corresponding mentions (e.g. *p53*-mediated). Moreover, as mentioned before, gene/protein mentions commonly have very complex structures (e.g. *PKR(1-551)K64E/K296R* or *RXRalphaF318A*). So, it is a common practice to tokenize gene/proten names adopting an approach that split tokens as much as possible to extract effective features (Torii et al., 2009; Smith et al., 2008). But while the extensive tokenization boosts performance, it is often difficult to correctly detect dependency relations for the tokens of the gene/protein names in the sentences where they appear. As a result, use of the syntactic dependency relations is not beneficial in such approaches.[13] In comparison, disease mentions are less complex. So, the identified dependencies for disease mentions are more reliable and hence may be usable as potential features (refer to our experimental results in Table 6).

The above mentioned issues are some of the reasons why a feature set for the well studied gene/protein focused BNER approaches is not necessarily suitable for other biomedical semantic types such as diseases.

7 Conclusion

In this paper, we have presented a single CRF classifier based BNER approach for disease mention identification. The feature set is constructed using disease-specific contextual, orthographic, general linguistic, syntactic dependency and dictionary lookup features. We have evaluated our approach on AZDC corpus. Our approach achieves significantly higher result than BANNER which is the current state-of-the-art ML based approach for disease mention recognition. We have also explained why the choice of features for the well studied gene/protein does not apply for other semantic types such as diseases.

[13] We have done some experiments on Biocreative II GM corpus with syntactic dependency relations of the tokens, which are not reported in this paper, and the results support our argument.

System	Note	Precision	Recall	F-score
BANNER	(Leaman et al., 2009)	80.9	75.1	**77.9**
Our system	Using general linguistic and orthographic features	74.90	71.01	72.90
Our system	After adding contextual features	82.15	75.81	78.85
Our system	After adding post-processing	81.57	76.61	79.01
Our system	After adding syntactic dependency features	82.07	76.66	79.27
Our system	After adding dictionary lookup features	83.21	79.06	**81.08**

Table 6: 10-fold cross validation results using exact matching criteria on AZDC.

Experiment	Note	Precision	Recall	F-score
(i)	Using general linguistic, orthographic and contextual features	82.15	75.81	78.85
(ii)	After adding *WC* and *BWC* features in *(i)*	82.08	75.57	78.69
(iii)	After adding *IsGreekAlphabet*, *HasGreekAlphabet* and *IsRomanNumber* features in *(i)*	82.10	75.69	78.76

Table 7: Experimental results of our system after using some of the gene/protein specific features for disease mention recognition on AZDC. Here, *WC* and *BWC* refer to the "word class" and "brief word class" respectively.

Future work includes implementation of disease mention normalization (i.e. associating a unique identifier for each disease mention). We also plan to improve our current approach by including more contextual features and post-processing rules.

Acknowledgments

This work was carried out in the context of the project "eOnco - Pervasive knowledge and data management in cancer care". The authors would like to thank Robert Leaman for sharing the settings of his experiments on AZDC.

References

Agarwal, P., Searls, D. 2008. Literature mining in support of drug discovery. *Brief Bioinform*, 9(6):479–492.

Aronson, A. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings AMIA Symposium*, pages 17–21.

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–270, January.

Bundschus, M., Dejori, M., Stetter, M., Tresp, V., Kriegel, H. 2008. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics*, 9:207.

Dai, H., Chang, Y., Tsai, R., Hsu, W. 2009. New challenges for biological text-mining in the next decade. *Journal of Computer Science and Technology*, 25(1):169–179.

Jimeno, A., Jimnez-Ruiz, E., Lee, V., Gaudan, S., Berlanga, R., Rebholz-Schuhmann, D. 2008. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(S-3).

Klinger, R., Tomanek, K. 2007. Classical Probabilistic Models and Conditional Random Fields. Technical Report TR07-2-013, Department of Computer Science, Dortmund University of Technology, December.

Kulick, S., Bies, A., Liberman, M., Mandel, M., McDonald, R., Palmer, M., Schein, A., Ungar, L. 2004. Integrated annotation for biomedical information extraction. In *Proceedings of HLT/NAACL 2004 BioLink Workshop*, pages 61–68.

Leaman, R., Gonzalez, G. 2008. Banner: An executable survey of advances in biomedical named entity recognition. In *Proceedings of Pacific Symposium on Biocomputing*, volume 13, pages 652–663.

Leaman, R., Miller, C., Gonzalez, G. 2009. Enabling recognition of diseases in biomedical text with machine learning: Corpus and benchmark. In *Proceedings of the 3rd International Symposium on Languages in Biology and Medicine*, pages 82–89.

McCallum, A. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu,.

Meystre, S., Savova, G., Kipper-Schuler, K., Hurdle, J. 2008. Extracting information from textual documents in the electronic health record: a review of

recent research. *IMIA Yearbook of Medical Informatics*, pages 128–44.

Névéol, A., Kim, W., Wilbur, W., Lu, Z. 2009. Exploring two biomedical text genres for disease recognition. In *Proceedings of the BioNLP 2009 Workshop*, pages 144–152, June.

Nadeau, D., Sekine, S. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26.

Noreen, E.W. 1989. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley-Interscience.

Rosario, B., Hearst, M. 2004. Classifying semantic relations in bioscience texts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*.

Schwartz, A., Hearst, M. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proceedings of Pacific Symposium on Biocomputing*, pages 451–62.

Settles, B. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 104–107.

Smith, L., Tanabe, L., Ando, R., Kuo, C., et al. 2008. Overview of BioCreative II gene mention recognition. *Genome Biology*, 9(Suppl 2).

Torii, M., Hu, Z., Wu, C., Liu, H. 2009. BioTagger-GM: a gene/protein name recognition system. *Journal of the American Medical Informatics Association : JAMIA*, 16:247–255.

Vlachos, A. 2007. Tackling the BioCreative2 gene mention task with conditional random fields and syntactic parsing. In *Proceedings of the 2nd BioCreative Challenge Evaluation Workshop*, pages 85–87.

Zweigenbaum, P., Demner-Fushman, D., Yu, H., Cohen, K. 2007. Frontiers of biomedical text mining: current progress. *Brief Bioinform*, 8(5):358–375.