Gene Entity Recognition of Full Text Articles

Manuel Noll*
Université de Liège
4, Chemin de la Vallée
B-4000 Liège, Belgium
+32(0)4366 - 5012
mnoll@student.ulg.ac.be

Jonathan Lete*
Université de Liège
4, Chemin de la Vallée
B-4000 Liège, Belgium
+32(0)4366 - 5012
jonathan.lete@student.ulg.ac.be

Patrick E. Meyer
Université de Liège
4, Chemin de la Vallée
B-4000 Liège, Belgium
+32(0)4366 - 3030
patrick.meyer@ulg.ac.be

ABSTRACT

Biomedical scientific literature is an unexploited treasure. Due to the staggering number of publications it is literally intractable to gather manually all information. Automatized information extraction (IE) is therefore key. An important subtask is the recognition of names in the text as specific entities (named entity recognition, NER). NER for genes in biomedical literature is a challenging task. This paper reports preliminary results for the identification of gene names in full text with the naive Bayes, support vector machine and random forest algorithms, showing that there is no loss on performance compared to the gene NER restricted to abstracts.

CCS Concepts

• Information systems~Chemical and biochemical retrieval • Applied computing~Health care information systems • Applied computing~Bioinformatics

Keywords

Bibliomics; Automated gene-name identification; NER; Machine learning; R.

1. INTRODUCTION

At the time of writing, MEDLINE, the most prominent online database for life sciences and biomedical information, cited to 23,343,329 scientific abstracts as of 2015 compared to 13,476,222 in 2005, with a growth rate of about 800, 000 articles per year [17].

An abundance of biological information is stored in there. However, for a biomedical scientist it is literally impossible to keep up with this speed.

Automatized information extraction (IE) is a sine qua non instrument for a modern researchers. It is able to gather potential information on all kind of biological relevant processes, like gene functions, gene-gene interactions, gene-protein interaction, chemical compound relations, cell compound classification and many more.

Submitted March 2017
Revised August 2017
For ICCBS'17

The former painstaking attention to detail being necessary for model organisms as *Drosophila melanogaster* or *Escherichia coli* can now be done in an automatized manner for all reported organisms with a high precision within no time. An abyss of data can become a nutshell. For instance, gene – gene networks or gene-protein networks can be extracted from literature. All it requires is the identification of entities and concepts, such as proteins and gene named entities. The identification of entities is one of the central tasks to successfully integrate biomedical literature into the infrastructure of structured biomedical data resources. Up to this point, corresponding experiments were only conducted on abstracts. This paper focuses therefore on the identification of gene names in full text articles.

However, retrieving and processing this information is challenging due to the natural-language narrative in biomedical literature. Although, the increasing interest in bioinformatics and annotated corpora, like GENIA[12], BioCreative[8], PennBio, FsuPrge and CRAFT [1] have accelerated this research. Nowadays' methods combine sophisticated methods from machine learning to natural language processing and outperform already human precision at some tasks. In this paper two contributions to this noble goal are made,

- the random forest is successfully introduced for gene entity recognition
- full text articles are consulted for gene entity recognition

2. NAMED ENTITY RECOGNITION

In the Sixth Message Understanding Conference (MUC-6), researchers were focusing on IE tasks where structured information is extracted from unstructured text. In defining this task, it became apparent that is an essential sub-task to recognize information units like names and was called „Named Entity Recognition and Classification (NERC)“. Previously unknown entities have to be discovered what hinges upon the classification rules triggered by features associated to positive and negative examples. That is either hand-crafed rules, supervised machine learning as a way to automatically induce rule-based systems or a hybrid of both.

NER Challenges for Biomedical Literature

Applying NERC systems to biomedical literature started around 15 years ago[6].It turned out that the major problem is the lack of convention for naming unambiguously biological concepts, such as genes or proteins. Many genes and proteins have more than one name. Furthermore, especially in the beginning of the genomic era, gene names were not distinguished from normal language. For example, for one of the first genomes studied, *Drosophila*, many genes are named after a specific phenotype of a mutant and have names such as 'white' (abbreviated by 'w'), 'shaggy' ('s') or 'mind the gap' ('mtg')[14]. Even if standards would be introduced and widely used, there is still a large amount of existing publications containing 'old' names. Furthermore, it is very hard to distinguish between gene and protein names. Another challenge, the detection of word boundaries, has turned out to be one of the most difficult NER tasks (e.g. “HZF-7” vs. “HZF-7 protein”).

BILOU Model

One approach to tackle the boundary detection challenge is to label such multi-token chunks of gene names according to the scheme beginning, inside and last token, labeled by 'B', 'I', and 'L', respectively. Gene names that consist of only one token are considered as unit-length chunks and labeled by 'U'. All non gene name tokens are tagged by an 'O' for outside [18].

NER Features

Dictionary-based

Dictionaries are large lists of words that represent examples for a specific entity class. They can be easily build from databases and can either be matched exactly against names in the text or with any inexact pattern matching algorithm in order to compensate the very low recall of the former.

Rule-based

Rule-based approaches are related to the character makeup of words. They build on the definition of rules to separate different classes, describing word case, punctuation, alphanumerical appearance etc. So play digit pattern an important role in identifying gene names, though they appear also dates or intervals. A common approach is checking for morphological features of the word, like common word endings, like 'ist' for human professions (*journalist, cyclist*). However, such endings are apparently not useful for identifying gene names, but rather looking for patterns and so-called summarized patterns. Those rules allow to map a range of words onto a small subset of patterns over character types. It is therefore a possibility to map all uppercase letters to 'A', all lowercase letters to 'a', all digits to '0' and all punctuations to '-'. In a further step consecutive character types can be summarized, i.e. '000' becomes simply '0'.

This hand-crafting of rules is time-consuming and has a couple of drawbacks. Different rules might not be mutually exclusive and lead therefore to interferences. Furthermore, as they are not comprehensive, they are not robust towards unseen patterns.

3. MACHINE LEARNING

Machine leaning is a highly interdisciplinary field building upon ideas from many different kinds of fields such as artificial intelligence, optimization theory, information theory, statistics, cognitive science, optimal control, and many other disciplines of science, engineering, and mathematics [8].

It can be roughly divided into two sub-domains [10] : supervised learning and unsupervised learning. Briefly, supervised learning requires training with labeled data which has inputs and desired outputs. In contrast with the supervised learning, unsupervised learning does not require labeled training data and the environment only provides inputs without desired targets [8]. Let us focus in this paper only on the former one, as it is the way of learning that our methods is using. From externally supplied labeled examples, a supervised machine learning reasons to produce a general hypotheses, called model, of the distribution of labels in terms of predictor features. Those labels can be assigned to classes to create a classifier that is able to predict to future instances or lables of unseen examples. Therefore, machine learning algorithms overcome following strictly static program instructions, i.e. the usage of hand-crafted rules. The NER task is basically a classification problem where each word in the text has to be assigned to a given class. In our case it is gene or not-gene. Unfortunately, the main shortcoming of supervised learning biomedical literature is the need for a large annotated corpus. There have been used a couple of classification algorithms for biomedical NER tasks. We discuss the Random Forest algorithm as it has not been used, to the best to our knowledge, for classifying names as genes in biomedical literature.

Naive Bayes

A naive Bayes is a simple classification algorithm that serves typically as baseline for evaluation. Let a classification task with k possible classes$C_{1,}C_{2,}\ldots,C_k$be given. An instance, represented by a vector$x = (x_{1,}x_2,\ldots,x_n)$, is then classified by the maximum a posteriori rule

$$argmax_{K=1,2,\ldots,k}\, p(C_k) \prod p\,(x_i \vee C_k)$$

This simplification is due to the “naive” conditional independence assumption, assuming that each feature is independent given the class.

Support Vector Machine

A support vector machine [11] is a supervised learning algorithm typically applied to classification tasks of two classes. The given labeled instances are represented as points in a possibly high-dimensional vector space. If the two classes are linearly separable, then the algorithm will find the best separating (hyper-)plane between the two sets of points. That is, the gap between them is maximized. New instances are then labeled according to the side of the (hyper-)plane they are falling to.

Random Forest

A random forest is a state-of-the-art machine learning algorithm typically used for making new predictions (in both classification and regression tasks). Random Forests can perform non-linear predictions and, thus, those often outperform linear models. Since its introduction it has been widely used in many fields from gene regulatory network inference to generic image classification. Random forest relies on growing a multitude of decision trees, a prediction algorithm that has shown good performances by itself but, when combined with other decision trees (hence the name forest), returns predictions that are much more robust to outliers and noisy data.

4. NATURAL LANGUAGE PROCESSING (NLP)

Natural language processing (NLP) is a field of computer science, artificial intelligence, and linguistics concerned with the computational processing of natural (human) languages. One of the task involved in NLP is natural language understanding. It became apparent that this involves the classifications of tokens in a sentence into nouns, verbs, adjectives, prepositions, etc. This is called part-of-speech (POS) tagging, which is considered to be the most basic form of linguistic corpus annotation.

This meta-informaion generated by grammatically analysing sentences can improve the ability to extract compound names, for which reason that many NER approaches include a POS tagger.

For the purpose of gene/protein name extraction, POS information can also be used in rule-based systems for rule conditions and/or error recovery, or as features in machine learning algorithms.

5. GOLD STANDARD CORPORA

In order to validate a gene-name identified in text a manually annotated corpus of biomedical literature is consulted. Schuhman et al[Evaluating gold standard corpora against gene/protein tagging solutions and lexical resources] specify a list of gold standard corpora for this purpose : (1) Jnlpba which stems from the Genia corpus [12], (2) BioCreative-II (2007) for human PGNs [8;9], (3) PennBio corpus (2006–2007) about oncology, (4) FsuPrge corpus (2009) on gene-regulatory events, and (5) the CRAFT corpus (2016)[1]. The latter one is the only corpus containing full text articles, which stems from the PubMed Central Open Acces Subset.

Leser and Hakenberg[15] report that the best performing systems in the 2004 BioCreative competition yielded F1-score around 83%, and the best performing system in the 2004 BioNLP/NLPBA competition reached about 70% F1-score. And furthermore, they gave cause for serious concern that when the best current NER systems reach an F-measure around 85 per cent, there is a real danger that all systems reporting better results will only represent an overfitting of the method to the particular gold standard, ie annotator. While Schuhman et al. evaluated gold standard corpora against gene tagging, they could not consider the CRAFT corpus in their studies as it and was not yet available during the experimental phase of their work in 2013. To the best of our knowledge, there has neither been any study performed on gene tagging consulting the CRAFT corpus nor using full text articles at all. The corpus contains 67 full text articles, > 560,000 Tokens, > 21,000 sentences and approximately 100,000 annotations to 7 different biomedical taxonomies/ontologies such as GO, NCBI and Entrez Gene (see Table 1).

Table 1. Basic statistics about the CRAFT corpus

Number of Articles	67
Type of Article	Full Article
Tokens	> 560,000
Sentences	> 21,000
Annotations	100,000
Ontologies / Taxonomies	• Chemical Entities of Biological Interest • Cell Ontology • Entrez Gene • Gene Ontology (biological process, cellular component, and molecular function) • NCBI Taxonomy • Protein Ontology • Sequence Ontology

6. METHOD

The methods' base is composed of statistical rules about the character makeup of the token and its textual neighborhood. The gene name boundary challenge was taken into account by using the BILOU model. In order to compensate the (highly) imbalanced classes of genes and not genes, stratification of the data set was applied [2;16].

Statistical Features in Detail

Frequency based rules

The articles in the CRAFT corpus are not related to each other, apart from being biomedical literature. Therefore, the articles can be considered independent; in the sense that an inferred textual neighborhood structure appearing in several articles is not due to the same author's use of language, but rather a real significant relationship that goes beyond the CRAFT corpus. Therefore, the CRAFT corpus has been used to infer statistical propositions about the character makeup of the token.

The statistical rules cover several textual neighborhood structures and makeup patterns of a gene name. Those rules catch the obvious gene names deviating appearance from 'common' names in that they consider token makeups were all letters are capitalized like for the gene HBB, or having digit(s) and special character(s) as in the SERPINA1 or the β-Klotho gene. HBB exhibits also the characteristic of some gene names as 'unpronounceable" words, wherefore the number of vowels, consonants and consecutive consonants seems to serve as a significant identifier.

See Table 3 for the statistical token makeup rules.

Table 2. Statistical Rules For a Token ('#' number of)

upper case letters
lower case letters
letters
digits
punctuation
vowel
consonants
Greek letters
consecutive consonants

The pure number of occurrence reflects an overall importance of this property for this token. In order to reflect an impact of this property on this token, a normalized version for each feature is generated i.e. dividing the total occurrence by the length of the word.

Pattern based rules

Pattern features were introduced by Collins [4] as a mean to map tokens onto a small set of patterns over character types. The pattern feature map alls uppercase letters to 'A', all lowercase letters to 'a', all digits to '0' and all punctuation to '-'.

The abridged pattern feature is a condensed form of it in which consecutive character types are not repeated in the mapped string.

Table 3 shows some examples.

Table 3. Pattern mapping and abridged patterns

Token	Pattern	Abridged
Hoi	Aaa	Aa
pipeloi	aaaaaaa	a
Gal41-B	Aaa00-A	Aa0-A

A common approach to map tokens to a pattern is called stemming, consisting of stripping off both inflectional and derivational suffixes before it is matched [3]. For this purpose the *R* package SnowballC is used (CRAN SnowballC, 2014).
Proper identifications of prefixes and suffixes improve the precision of classifier by excluding false positives. Apparently, prefixes like 'an' and 'anti', and likewise suffixes might include each other and are not easily identifiable by a dictionary.-less machine learning algorithm. Therefore, each 1-gram up to 5-grams of each token's beginning and ending letter sequence is considered as a feature. Additionally a binary feature states whether the last character of the token is a roman number.

POS tagging

The Stanford CoreNLP tagger [13] is consulted to incorporate grammatical structures. Each token in the CRAFT corpus is labeled with its corresponding part-of-speech tag. This information is fed into a machine learning algorithm to build the classifier.

7. EXPERIMENTS AND RESULTS

The goal is to test whether a high degree of precision can be achieved for the automatized identification of gene names in full text.

For this reason the CRAFT corpus has been chosen in this experiment. It was mutually exclusively divided into a training of 47 full articles and a text set of 20 full articles.

The automating is implemented by three supervised machine learning algorithm. A naive Bayes (NB), which will serve as a baseline for measuring the precision, a support vector machine (SVM) and a random forest (RF). The *R* packages e1071 (CRAN e1071, 2016), for NB and SVM, and randomForest (CRAN randomForest, 2015) for RF have been used for this purpose.

The test set contained 139.690 tokens, where 4687 refer to a gene. For the evaluation, the true positives (tp), false positives (fp), and false negatives (fn) are counted and finally the precision, recall and F1-measure are computed:

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$F1score = 2\frac{Precision \cdot Recall}{Precision + Recall}$$

In order to deal with the identification of the boundary of a gene name the BILOU token labeling is applied. The labeling 'B', 'I', 'L' and 'U' designate a gene token. A soft-matching is used to evaluate the true positives, false positives and false negatives. If parts of a gene are predicted to be 'B', 'I', 'L' or 'U', then it is considered as a true positive, while the part wrongly classified accounts for a false negative. False positive are non-gene tokens classified as either 'B', 'I', 'L' or 'U'. Clearly, correctly identified non-gene tokens, labeled by 'O' , account also for the true positives.

The contextual neighborhood structure is taken into account by considering the above described 33 features for a token window size of *1*, i.e. for the token before the current and one afterwards.

POS vs. Non-POS

In a first conducted experiment series the effect of a POS tagger was evaluated. First, the combined statistical rules were invoked to generate the features for each token in the training text, without incorporating any POS information. This was fed into the NB, SVM and RF each. The RF has been trained with 200 trees using the Gini index.

The running time of the tests showed significant differences. While the SVM needed 1h 19m, the NB took 2m 31s and the RF 10m 21s on a 1.2GHz Intel Core Duo Processor.. Table 4 shows the corresponding outcome for the precision, recall and the F1-score.

Table 4. Precision, Recall and F1 score, without POS

Method	Precision	Recall	F1-score	Time
Naive Bayes	5.76%	93.08%	10.86%	2m 31s
SVM	66.68%	39.48%	49.59%	1h 19m
Random Forest	69.15%	60.16 %	64.34%	10m 21s

In the second experiment, the Stanford CoreNLP POS tagger labeled additionally the training set, before it was fed into the algorithms. The annotation of the data set took additionally 1h 45m. The running time of the algorithms altered not in magnitude, NB 3m 30s, RF 8m 5s, and the SVM in 1h 23m. The POS-tagged model is performing better than the not tagged one, which is in accordance with our expectations. Although, its just minimal increase might be due to the fact that the feature set is already informative enough to explain the POS tagging. However, it dropped the running time of the RF as it helped to earlier classify the tokens in a decision tree (see Table 5). Although, this presumably has happened due to overfitting, a deeper analysis and corresponding experiment has to be conducted.

Table 5. Precision, Recall and F1 score, with POS

Method	Precision	Recall	F1-score	Time
Naive Bayes	7.29%	92.69%	13.52%	3m 30s
SVM	67.00%	39.33%	49.57%	1h 23m
Random Forest	69.82%	61.28 %	65.27%	8m 5s

The best F1 score of 65.27% and a precision of 69.82% was achieved by a RF of *200* trees using the Gini index, POS tagging and data stratification.

Tuning the Number of Trees

In a second conducted experiment series the sensitivity of the performance on the number of trees was evaluated by a step-wise increase of the number of trees. In order to guarantee to become not as pricey as a SVM in terms of running time, a maximum number of *1000* trees has been tested, taking already ~1h compared to 2m10s for *50* trees. Table 6 shows that the trend of the F1-score and precision are stable, for the number of trees *n =50, 100, 200, 500, 1000,* being around 65%, 69% respectively.

Table 6. Precision, Recall and F1 score, without POS

Trees	Precision	Recall	F1-score	Time
50	68.72	61.53	64.93	2m 10s
100	69.86	61.45	65.39	4m 12s
200	69.82	61.28	65.27	8m 32s
500	69.98	61.49	65.46	2m 20s
1000	70.79	61.78	65.98	59m 22s

8. RELATED WORK

Comparing gene name identification methods is difficult, because some distinguish between genes, proteins and enzymes and others do not. In between is a wide range of possible combinations. The PROPER system of Fukuda et. el [5] achieved a precision of 95 % and a recall of 99 % using 30 abstracts about the SH3 protein domain. Kazama et. el [11] achieved on the whole GENIA corpus of abstracts an F1-score of 45.99% for the identification of all biomedical entities available in this corpus. Leser and Hakenberg[15] reported that the best performing systems in the 2004 BioCreative competition [9] yielded F-measures around 83%, and the best performing system in the 2004 BioNLP/NLPBA competition reached about 70% F1-score.

9. CONCLUSIONS

The CRAFT corpus contains 67 full articles from the PubMed Central Open Acces Subset. The articles are not directly related to each other, and can therefore be considered independent. Hence, learning on a subset and testing on the remaining articles is not overfitting the model to a certain article type. I.e. the propositions A and B found as being significant for classifying a token as a gene is not restricted to the CRAFT corpus.

The perfomance of the Random Forest can therefore be to some extend extrapolated to other corpora. It is expected that due to more examples the performance can be improved. A more comprehensive test is planed. Unfortunately is the amount of fully annotated articles sparse and future experiments will consider the corpora Jnlpba, BioCreative, PennBio and FsuPrge, which comprise only abstracts of biomedical articles.

While the statistical rules for the makeup pattern of a gene name cover a wide range of possible alphanumeric combinations for gene names, the textual neighborhood of a gene name might not be well captured by the very same features on a corpus of 67 articles. It is expected that a larger corpus will improve those rules, as the later has to cover a wide range of semantically meaningful word/verb-tokens surrounding a gene name.

Furthermore, a more technical issue has to be addressed concerning the loss of information about the character makeup by downloading articles in plain text. For instance, Gene names are nowadays mostly written in italic. In combination with the publication date of an article this might further improve the performance.

The nature of the results in this paper is twofold. First, it is shown that a random forest performs well on the automated gene-name extraction task. Purely abstract based gene name extraction reach a F1-score of about 65%. Therefore, second, the automated gene name extraction can successfully be extended to full texts without loss of performance.

10. ACKNOWLEDGMENTS

Our thanks to ACM SIGCHI for allowing us to modify templates they had developed. Manuel Noll is grateful to the F.R.S.- FNRS for a doctoral grant (1.A.320.16F).

11. REFERENCES

[1] Bada M, Eckert M, Evans D, Garcia K, Shipley K, Sitnikov D, Baumgartner WA, Cohen KB, Verspoor K, Blake JA, Hunter LE. Concept annotation in the CRAFT corpus. BMC Bioinformatics. 2012;4:161.

[2] Chen, C., Liaw, A., Breiman, L.: Using Random Forest to Learn Imbalanced Data. (2004)

[3] Coates-Stephens, Sam. *The Analysis and Acquisition of Proper Names for the Understanding of Free Text.* Computers and the Humanities. 1992: 26.441-456, San Francisco: Morgan Kaufmann Publishers.

[4] Collins, Michael. *Ranking Algorithms for Named–Entity Extraction: Boosting and the Voted Perceptron.* 2002. In Proc. Association for Computational Linguistics.

[5] Fukuda et al. *Toward information extraction: identifying protein names from biological papers.*1998. Proceedings of the Pacific Symposium on Biocomputing (PSB98).

[6] Grishman, Ralph; Sundheim, B. *A Brief History. In Proc. International Conference on Computational Linguistics.1996.* Message Understanding Conference – 6

[7] Hahn U, Beisswanger E, Buyko E, Poprat M, Tomanek K, Wermter J. Semantic annotations for biology–a corpus development initiative at the Jena University Language & Information Engineering (JULIE) Lab. LREC 2008–Proceedings of the 6th International Conference on Language Resources and Evaluation. 2008. pp. 28–30.

[8] Hunter L, Carpenter B, Tsai R, Dai HJ, Liu F, Chen Y, Sun C, Katrenko S, Adriaans P, Blaschke C, Torres R, Neves M, Nakov P. et al. Overview of BioCreative II gene mention recognition. Genome Biol. 2008;4(Suppl 2):S2.

[9] Hirschman, L., Yeh, A., Blaschke, C. and Valencia, A. (2005), 'Overview of BioCreAtIvE: critical assessment of information extraction for biology', BMC Bioinformatics, Vol. 6 (Suppl 1), p. S1.

[10] Junfrei Qiu et al. 2016. *A survey of machine learning for big data processing.* EURASIP Journal on Advances in Signal Processing 2016:67

[11] Kazama et al. 2002. *Tuning support vector machines for biomedical named entity recognition* BioMed '02 Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain - Volume 3

[12] Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus–semantically annotated corpus for bio-textmining. Bioinformatics. 2003;4(Suppl 1):i180–i182.

[13] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, pp. 252-259.

[14] Kulick S, Bies A, Liberman M, Mandel M, McDonald R, Palmer M, Schein A, Ungar L. Integrated annotation for biomedical information extraction. HLT-NAACL 2004 Workshop: Biolink 2004, "Linking Biological Literature, Ontologies and Databases". 2004. pp. 61–68

[15] Leser U., Hakenberg J.,. *What makes a gene name? Named entity recognition in the biomedical literature.* Brief Bioinform. 2005. 6(4) 357-69.

[16] Liu, Y., Loh, H. T., Sun, A.: Imbalanced text classification: A term weight approach. Expert Systems with Applications, 36, pp. 690–701, (2009)

[17] MEDLINE. Citation statistics. 2015. https://www.nlm.nih.gov/bsd/licensee/2015_stats/2015_LO.html

[18] Ratinov, L., & Roth, D. *Design challenges and misconceptions in named entity recognition.* In Proceedings of the Thirteenth Conference on Computational Natural Language Learning. 2009. pp. 147-155. Association for Computational Linguistics.