Named Entity Recognition in Acute Inflammatory Response Studies

Emilia Apostolova, Steven Lytinen, Daniela Raicu
School of Computer Science
DePaul University
Chicago, IL
{emilia.aposto@gmail.com, lytinen@cs.depaul.edu, dstan@cti.depaul.edu}

Abstract

This paper presents a study of the task of recognizing a specific type of Named Entities in medical texts describing trauma conditions. The study includes data gathering, named entity orthographic rule definitions, building a lexicon, and training via a machine learning method - Conditional Random Fields. This Named Entity Recognition task is viewed as the first step of information extraction of free text clinical studies describing shock, trauma, inflammation, and other related states. The end goal is to build a computer model of the acute inflammatory response using agent-based modeling. The goal of this agent-based model is to simulate the body response to shock and trauma.

1. Introduction

There is a growing number of biomedical corpora and publications, mostly in the form of free text, and with them comes the need to be able to retrieve and query easily relevant information. As a result, Natural Language Processing (NLP) of biomedical texts has received a lot of attention during the last several years.

Named Entity Recognition (NER) is a subtask of information extraction that seeks to discover and classify atomic elements within text into predefined categories such as chemical compounds, molecule names, names of cell or tissue types. NER is often viewed as a prerequisite step towards extracting structured information from unstructured documents. Named Entity Recognition as a result of the interest in NLP of biomedical texts has achieved significant attention over the past decade.

The end goal of this research is automated information extraction from clinical studies describing shock, trauma, inflammation, and other related states with the purpose of building a computerized Agent Based Model for the simulation of the effects of various trauma conditions. Figure 1 places the current paper in the overall goal of this research. Information will be extracted from publications describing trauma clinical studies by first identifying Named Entities of interest and then identifying the different types of relationships/interactions between them. The extracted information will be then saved into a structured database and used to create simulations of the human body response to inflammation and trauma.

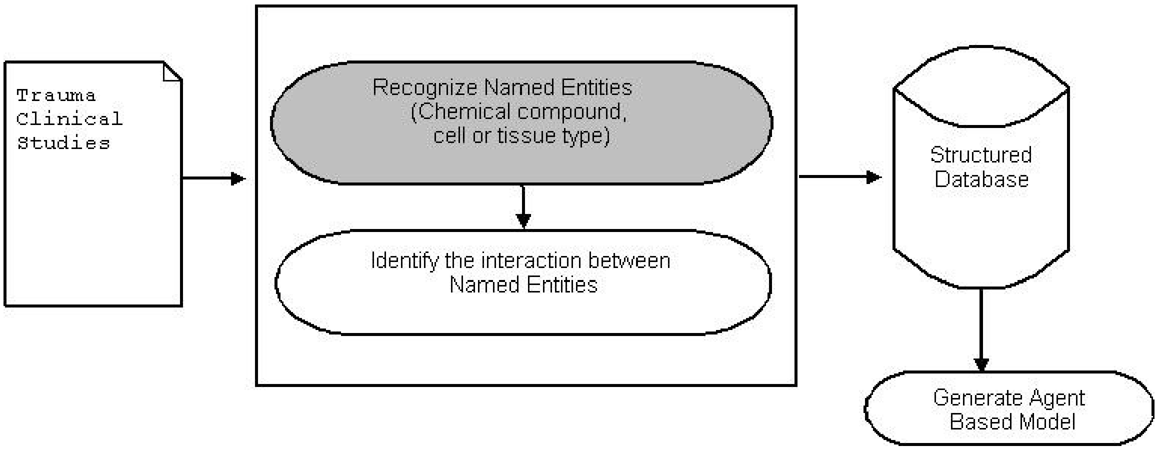

Trauma Clinical Studies
Recognize Named Entities (Chemical compound, cell or tissue type)
Identify the interaction between Named Entities
Structured Database
Generate Agent Based Model

Figure 1. Trauma clinical studies information extraction.

Experience has shown that NER in the life sciences is a rather difficult problem. Factors that contribute to these difficulties are the general lack of naming conventions in biomedical sciences, excessive use of abbreviations, frequent usage of synonyms and homonyms, and the fact that biological objects often have names consisting of many single words, such as '*a7 subunit-containing nicotinic acetylcholine receptor'.* In the case of multiword named entities it is also usually not clear where a name starts or ends, even for human readers. It is not uncommon for experts in the domain to disagree on the exact boundaries. In addition, gene names such as 'white' (symbol w), 'shaggy' (symbol ssg), or 'mind the gap' (symbol mtg) make it almost impossible to find gene-related articles using full-text search. Context dependent acronyms are yet another source of confusion. For example, the acronym ACE could stand for 'angiotensin converting enzyme', 'affinity capillary electrophoresis', 'acetylcholinesterase' and a couple of other things.

2. Related Work

Various approaches have been taken towards tackling the task of NER in biomedical texts. They can roughly be classified into three distinct groups [1] - dictionary based; rule based; and machine learning techniques. Dictionary based approaches as the name suggests match text against a fixed lexicon. Even though precision (aka specificity) is pretty high for such approaches, not surprisingly the recall (aka sensitivity) is very low as new protein and gene names, for example, are constantly introduced. Rule based approaches are usually hand crafted by experts in the field and consist of surface clues such as specific word suffixes, capital letters and digits, Greek words, etc. The disadvantage of rule-based approaches is that the process is time consuming and such approaches have difficulties handling unseen name patterns.

Various Machine Learning techniques have also been applied to solving the NER problem. Naïve Bayes is the approach taken by Nobata et al. [2] in identifying and classifying terms in biology texts. Support Vector Machines are used by the YamCha word-based classification system[3]. In Yamcha, features are defined as different types of surface clues and morpho-syntactic properties of named entities and their surrounding words, as well as matches of tokens against a dictionary. Dingare et al.[4] report on a system using a maximum entropy classifier as the basic component of maximum entropy Markov model. Using a Viterbi-style algorithm, the system predicts the most probable sequence of single classifications for the tokens of a sentence. Conditional Random Fields are another probabilistic sequence tagging framework [5]. In fact, on the BioCreAtIvE corpus [6], conditional random fields were one of the best performing methods.

Both Hidden Markov Model (HMM) and Conditional Random Fields (CRF) handle the task of assigning label sequences to a set of observation sequences. HMMs are used to identify the most likely sequence of labels for the words in any given sentence. HMM defines joint probability distribution p(X, Y) where X and Y are random variables ranging over observation sequences and their corresponding label sequences respectively. The task of defining such joint probabilities distributions involves enumerating all possible observation sequences. The problem within the domain of natural text is intractable unless observation elements are represented as isolated units, independent of other elements in the observation sequence. CRF, on the other hand, is a model that supports tractable inference and yet does not need independence assumptions. CRF defines conditional probability P (Y|x) – probability of label sequence Y given a particular observation sequence x, rather than a joint distribution over both label and observation sequences. CRF is used to label a novel observation sequence x* by selecting the label sequence y* that maximizes the conditional probability P(y*|x*). The conditional nature of such models means that no effort is wasted on modeling the observations, and one is free from having to make unwarranted independence assumptions about these sequences; arbitrary attributes of the observation data may be captured by the model, without the modeler having to worry about how these attributes are related [7][8].

Settles [9] presents a framework for simultaneously recognizing occurrences of Protein, DNA, RNA, Cell-line, and Cell-type entity classes in biomedical texts using CRF. The feature set introduced uses some basic orthographic features, for example *AlphaNumeric, HasDash, RomanNumeral*, etc, together with what he refers to as 'semantic features' – lexicons of named entities for each class. The performance in his and similar studies is measured in terms of F-score, a combined measure of precision and recall - 2*precision*recall/(precision + recall). He reports overall F-score of 69.5, observing that adding semantic features actually resulted in a slightly worse F-score, compared to F-score of 69.8 when no semantic features are used.

3. Methodology

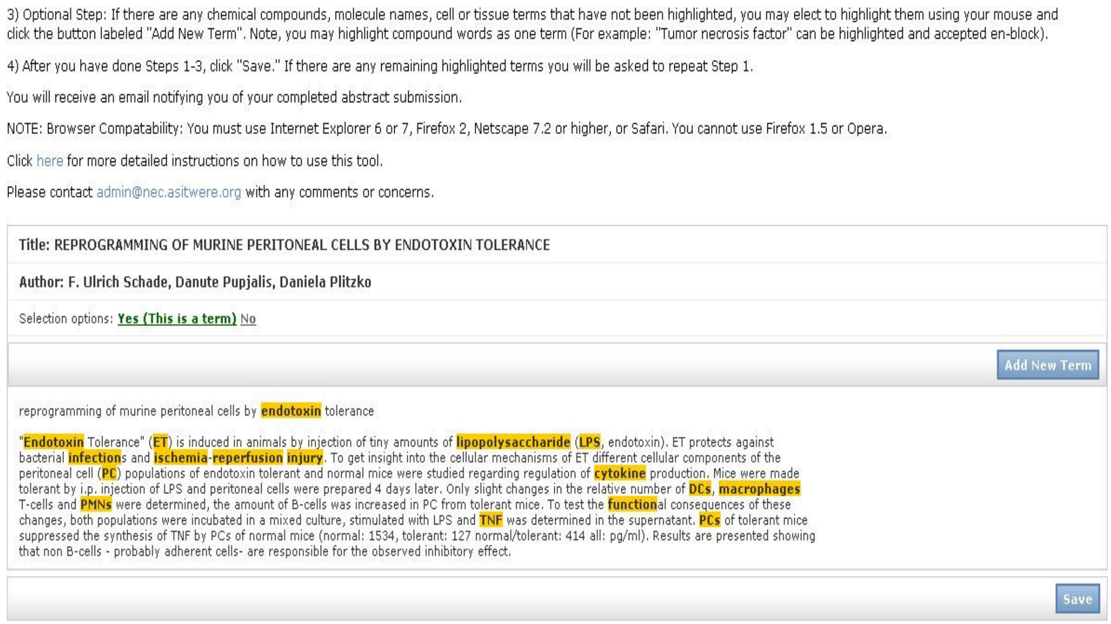
3.1 Data Gathering

One of the challenges around NER is finding appropriate data that can be used for training. The task of tagging named entity is time consuming and labor intensive. In addition, several experts in the field usually need to be involved and an agreement needs to be reached, as the boundaries of Named Entities phrases could be fuzzy and subjective.

There are in existence several tagged Named Entity biomedical text corpora. BioCreAtIvE [10] contains around 18, 000 tagged entities describing gene/protein in 15,000 sentences. GENIA [11] contains 2,500 abstracts with tagged named entities describing proteins and DNAs. Yapex [12] is yet another dataset consisting of around 200 abstracts with tagged named entities describing Protein.

The long term goal of building an agent based model of the body response to acute inflammatory illness involves finding a specific type of Named Entities - denoting chemical compound, molecule name, cell or tissue type - in a specific type of clinical studies – describing trauma conditions. None of the above mentioned corpora appears suitable for the specific type of Named Entities this study is interested in. To our knowledge there is no suitable tagged corpus that can be used to train a NER system with these specific requirements, and one of the contributions of this paper is collecting and providing this corpus.

The best-known specialized journal for studies of shock, trauma, sepsis, endotoxemia, ischemia/reperfusion, inflammation, and other related pathophysiologic states is the Shock Journal [13]. The journal has an archive of studies dating back to 2002 and publishes more than 200 studies annually. The journal abstracts are in fact the best candidate for use in building an agent-based model of trauma conditions. The Shock Journal is published by the Shock Society, which also hosts an international annual conference for similar studies. For the purpose of developing a tagged corpus in the Shock domain, the participants of the 2008 Shock Conference were asked to tag the named entities in their abstract submissions. We created a web interface to facilitate the tagging process (Figure 2 below).



3) Optional Step: If there are any chemical compounds, molecule names, cell or tissue terms that have not been highlighted, you may elect to highlight them using your mouse and click the button labeled "Add New Term". Note, you may highlight compound words as one term (For example: "Tumor necrosis factor" can be highlighted and accepted en-block).

4) After you have done Steps 1-3, click "Save." If there are any remaining highlighted terms you will be asked to repeat Step 1.

You will receive an email notifying you of your completed abstract submission.

NOTE: Browser Compatability: You must use Internet Explorer 6 or 7, Firefox 2, Netscape 7.2 or higher, or Safari. You cannot use Firefox 1.5 or Opera.

Click here for more detailed instructions on how to use this tool.

Please contact admin@nec.asitwere.org with any comments or concerns.

Title: REPROGRAMMING OF MURINE PERITONEAL CELLS BY ENDOTOXIN TOLERANCE

Author: F. Ulrich Schade, Danute Pupjalis, Daniela Plitzko

Selection options: **Yes (This is a term)** No

reprogramming of murine peritoneal cells by **endotoxin** tolerance

"**Endotoxin** Tolerance" (**ET**) is induced in animals by injection of tiny amounts of **lipopolysaccharide** (**LPS**, endotoxin). ET protects against bacterial **infection**s and **ischemia**-**reperfusion** **injury**. To get insight into the cellular mechanisms of ET different cellular components of the peritoneal cell (**PC**) populations of endotoxin tolerant and normal mice were studied regarding regulation of **cytokine** production. Mice were made tolerant by i.p. injection of LPS and peritoneal cells were prepared 4 days later. Only slight changes in the relative number of **DCs**, **macrophages** T-cells and **PMNs** were determined, the amount of B-cells was increased in PC from tolerant mice. To test the **function**al consequences of these changes, both populations were incubated in a mixed culture, stimulated with LPS and **TNF** was determined in the supernatant. **PCs** of tolerant mice suppressed the synthesis of TNF by PCs of normal mice (normal: 1534, tolerant: 127 normal/tolerant: 414 all: pg/ml). Results are presented showing that non B-cells - probably adherent cells- are responsible for the observed inhibitory effect.

Figure 2. A Web Interface provided to the 2008 Shock Conference Submitters.

Around 500 abstracts were submitted to the 2008 Shock Conference and around 70% of all submitters responded to the request to mark the named entities in their abstracts. A domain expert monitored the process and validated the data.

3.2 Named Entity Extraction

Around 350 tagged abstracts were used for this study. 70% of them were used as training data and 30% as test data. The average size of the abstracts is around 270 words.

The process can be described in terms of several subtasks. First, the data was preprocessed in the format used by the Mallet CRF open source software [14]. The text was split into sentences and each word tagged with a pipe-separated corresponding tag. The possible tags are: O - designates a non-named entity; B-NE – designates the first word of a named entity phrase; and I-NE designates subsequent words in a multiword named entity. For example, the sentence with named entities in bold:

A bolus of ***sodium selenite*** *leads to a beneficial peak of pro-oxidative* ***plasma*** *se concentration*

was transformed to the following training data sentence:

A|O bolus|O of|O sodium|B-NE selenite|I-NE leads|O to|O a|O beneficial|O peak|O of|O pro-oxidative|O plasma|B-NE se|O concentration|O

Next, orthographic features were defined and selected. The orthographic features were based on 18 rules implemented as regular expressions. Approximately half of these rules are generic orthographic expressions – *HasDigits, AlphaNumeric, HasDash, ComputePrefix, ComputeSuffix*. The other half of the features is based on rules defined by a domain expert. They include expressions that include or exclude words as Named Entities. Some sample expert defined rules are:

Rule 1: *Include words containing a hyphen with at least 4 characters on each side of the hyphen and at least one alphabetic character.*

Rule 2: *Exclude words that that contain "p<" or "p=" or "vs."*

Rule 3: *Exclude strings that end with "-day", "-week".*

Rule 4: *Include words with suffixes suffixes "cytes", "cyte", "virus" , "oid", "mRNA", "ase", "some", "sis"*

A lexicon membership feature was also used. The rule for this features checks if the word is part of a predefined lexicon of 1,300 phrases. The phrases were identified as Named

Entities by a domain expert. The CRF training algorithm outputs a weight of 1 for this feature if a word is in the lexicon, and 0 otherwise.

The Mallet CRF software was used to compute weights for the provided features. The software uses the learned feature weights (the output of the training phase) and tags the test abstracts.

Figure 3 below illustrates the process.

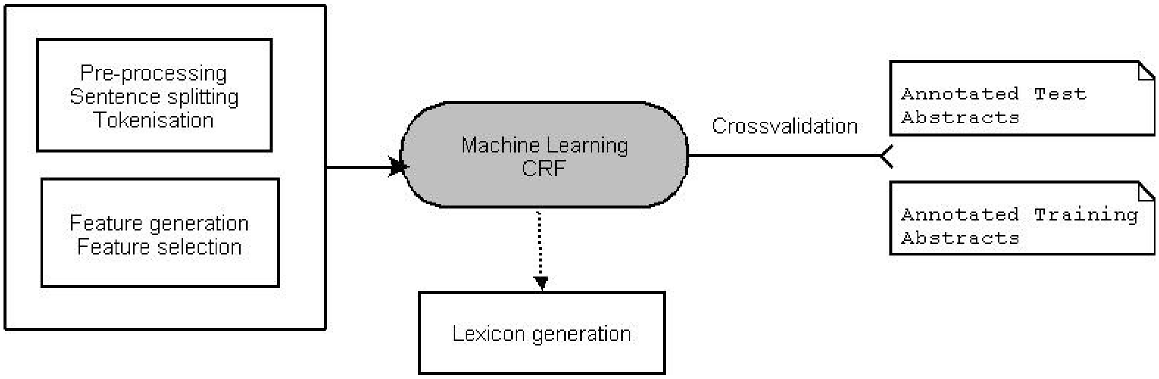


Figure 3. Named Entities extraction process.

4. Results

The algorithm achieved F-score of 71.6, a slight improvement over Settles [5] F-score of 69.5. Unlike Settles results, the lexicon-based feature improved the F-score (from 66.88 to 71.31), possibly because the lexicon was compiled by a domain expert and did not include invalid data. In addition, we predicted that the F-score could be improved by incorporating orthographic rules that are specific to this type of Named Entities. The results did not confirm this expectation, even though these rules were hand-crafted by experts in the field. That is because, although there was indeed some increase in precision, it was offset by a decrease in recall.

	Lexicon Membership Feature	*Without Lexicon Membership Feature*	**Lexicon Membership Feature**
	Domain Specific 0rthographic Features	**Domain Specific 0rthographic Features**	*Without Domain Specific 0rthographic Features*
Number of Annotated Named Entities	2,545	2,545	2,545
Number of Recognized Named Entities	1,827	2,284	1,976
True Positive	1,559	1,615	1,619
Precision	85.33	70.70	81.93
Recall	61.25	63.45	63.61
F-score	**71.31**	**66.88**	**71.62**

Table 1. Test results for three feature sets. The presence of the lexicon membership feature and the expert defined orthographic rules is varied.

5. Conclusion and Future Work

This paper presents a practical approach to finding named entities in a specific type of medical texts – trauma clinical studies, with the long-term goal of creating an agent based model for simulating body responses to trauma and shock. Conditional Random Fields, the machine learning technique demonstrating the highest success rate in existing studies, is the approach taken. Generic orthographic rules and lexicon matching features led to a reasonable for practical purposes F-score of 71.62.

Future work will incorporate general purpose shallow parsing into the CRF training. In particular, general, non-domain specific, Noun Phrase knowledge is expected to improve the CRF results. In addition, acronyms appear to be quite commonly used in these types of text and an incorporated acronym recognizer could potentially have a significant effect on performance. Integrating alternate spelling (e.g. IL-12, IL 12, IL12) which appears to also be quite common among these types of Named Entities is also expected to improve performance.

References

[1] Leser U, Hakenberg J (2005), 'What makes a gene name? Named entity recognition in the biomedical literature', 'Briefings in Bioinformatics', 6:357-369.

[2] Nobata, C., Collier, N. and Tsujii, J. (1999), 'Automatic term identification and classification in biology texts', 'Proc. Natural Language Pacific Rim Symposium' , Beijing, China.

[3] Tomohiro Mitsumori, T., Fation, S., Murata, M. et al. (2005), 'Gene/protein name recognition based on support vector machine using dictionary as features', BMC Bioinformatics, Vol. 6 (Suppl 1), pp.S8.

[4] Dingare, S., Finkel, J., Manning, C. et al. (2004), 'Exploring the boundaries: Gene and protein identification in biomedical text' in 'Proceedings of the EMBO workshop BioCreative: Critical Assessment for Information Extraction in Biology', 28th–31st March, Granada, Spain.

[5] McDonald, R. and Pereira, F. (2005), 'Identifying gene and protein mentions in text using conditional random fields', BMC Bioinformatics, Vol. 6 (Suppl 1), p. S6.

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

[7] Wallach H. (2004), 'Conditional Random Fields: An Introduction.', University of Pennsylvania Department of Computer and Information Science, Technical Report.

[8] Lafferty J., McCallum A., Pereira F. (2001), 'Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data', In Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001).

[9] Settles B. (2004), 'Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets', Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA)

[10] 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.

[11] Kim, J. D., Ohta, T., Tateisi, Y. and Tsujii, J. (2003), 'GENIA corpus – a semantically annotated corpus for bio-textmining', Bioinformatics, Vol. 19 (Suppl 1), pp. I180– I182.

[12] Franzen, K., Eriksson, G., Olsson, F. et al. (2002), 'Protein names and how to find them', Int. J. Med. Inf., Vol. 67(1–3), pp. 49–61.

[13] Shock – Injury, Inflammation, and Sepsis: Laboratory and Clinical Approaches, Copyright © 2008, Lippincott Williams & Wilkins, http://www.shockjournal.com

[14] McCallum, A. (2002), 'MALLET: A Machine Learning for Language Toolkit.' http://mallet.cs.umass.edu.