ONER: Tool for Organization Named Entity Recognition from Affiliation Strings in PubMed Abstracts

Siddhartha Jonnalagadda[1,2,*], Philip Topham[1], Graciela Gonzalez[2]
[1]*Lnx Research LLC, 750 The City Drive Suite 490, Orange, CA 92868, USA*
[2]*Department of Biomedical Informatics, Arizona State University, Phoenix, AZ 85004, USA*
[*]*sjonnalagadda@lnxresearch.com*

Abstract

Automatically extracting organization names from the affiliation sentences of articles related to biomedicine is of great interest to the pharmaceutical marketing industry, health care funding agencies and public health officials. It will also be useful for other scientists in normalizing author names, automatically creating citations, indexing articles and identifying potential resources or collaborators. Today there are more than 18 million articles related to biomedical research indexed in PubMed, and information derived from them could be used effectively to save the great amount of time and resources spent by government agencies in understanding the scientific landscape, including key opinion leaders and centers of excellence. Our process for extracting organization names involves multi-layered rule matching with multiple dictionaries. The system achieves 99.6% f-measure in extracting organization names.

1. Introduction

It takes an average of 17 years for research evidence to reach practical application, and although more scientific discoveries are being achieved now than ever before, translating biomedical discoveries into practical treatments doesn't seem to occur much faster than 100 years ago [1]. One of the most important requirements to erase this gap from discovery to application is to connect "those who produce the knowledge with those who apply it". One step in that direction is to develop automated techniques that can accurately pinpoint the organizations involved in specific areas of research, such as the centers of excellence involved in the study of a particular disease, or those that use specific techniques. This can be achieved by mining the affiliation information from PubMed, and mapping it to the MeSH terms and other keywords (such as specific genes) extracted from the abstract, thus allowing public health officials and other interested researchers in pharmaceutical or biotechnology companies to rapidly identify the main players for specific advances.

On the financial side, pharmaceutical companies spend an average of 24% of their total marketing budgets in identifying the Key Opinion Leaders and centers of excellence [2, 3]. By analyzing large-scale social networks of organizations, it is possible to analyze complete scientific communities with thousands of organizations in order to discover research clusters and trends. One of the essential requirements in building such a large-scale social network is a process to automatically extract organization names.

2. Background

2.1. Organization Named Entities

Currently, there is no standard style for listing an author's affiliation – it is a free form text field with some moderate cultural preferences to list Institution, then City, State, and Country. However, there are wide variations in style. For example, a person may list Department first and then institution or visa-versa.

In a previous attempt at recognizing organization names [4], it was assumed that most of the affiliation strings in PubMed articles adhere to the format: [address component], [address component], …, [country].[email]. Their rules assumed that the affiliation sentences have the name of the country explicitly and the organization one of their keywords. However, there are many affiliation sentences which disobey these

Proceedings of the 3rd International Symposium on Languages in Biology and Medicine,
Posters, pages 129-131, Jeju Island, South Korea, 8-10 November 2009.

assumptions. For example:- In "*Centaur Science Group, 1513 28th St NW, Washington, DC 20007 USA.* (PubMed ID: 16796054)", *Centaur Science Group* is an independent company that doesn't contain any of their key words. Thus, to create a system that recognizes organization names with high accuracy, we need to apply rules at multiple levels, with each level gradually converting the unstructured input text into structured fields.

We are initially focusing our work for identifying organizations from the United States. Since the United States contributes more than 25% of the articles in PubMed and has the highest share in the pharmaceutical industry market and research funding for most of the disease areas, it is justified to start with building a highly accurate system specific to the US, rather than lose performance in attempts to generalize the process. In future, we would expand our scope, while maintaining performance, by having custom-tailored rules for each country we wish to consider.

PubMed requires the journals submitting papers to send a single field representing the organization name (http://www.ncbi.nlm.nih.gov/entrez/query/static/spec.html). Usually this is the organization of the first author, and in some cases the organization which appears first in the journal abstract. There are many organizations which are present in multiple places. For the purposes of further analysis, these organizations are different, while they could potentially have some resources in common.

2.2. Organization NER in context

There are mainly two relevant hierarchies of named entity types that have been proposed in the literature. The BBN categories [5], proposed in 2002 for a Question Answering task, consist of 29 types and 64 subtypes. Sekine's extended hierarchy [6], also proposed in 2002, is made of 200 subtypes. According to the BBN Hierarchy, our specific NER problem breaks down into identifying 4 major types of entities: a) Organization name – for classifying the names of the actual research groups, b) Geo-Political Entity name – names of country, state and city, c) Facility name – names of buildings and other man-made structures, and d) Contact Info – Address, email and URL. In MUC-7, the best system scored 93.39% of f-measure while human annotators scored 97.60% and 96.95% for this task.

Early NER systems were mostly based on hand-crafted rules which in general have a good performance but at the expense of great amount of labor done by experts familiar with the text, usually computational linguists. In recent years, many NER systems for different entities are being replaced by supervised learning systems which need lesser investment of time, and where rules are automatically created from the positive and negative examples. Unfortunately, such systems require lots of annotated data. To our knowledge, there are no publicly available annotated corpora for the different organization-related named entities discussed above. This, coupled with the demand for a high f-measure, forced us to build a rule-based system. Fortunately, since the affiliation sentences are much more structured than general English, this task was made more viable.

3. Methods

The Vedic "*Neti, Neti* – Not this, Not this" golden rule of the elimination of untruth till one lands up at the feet of truth inspired us while extracting the organization names from the affiliation sentence. This is especially true in cases when it is easier to ascertain what doesn't belong to a set we are trying to build than what belongs. The other 3 types of named entities – geopolitical entity name, facility name and contact info – conform better to the general pattern than the most important entity, organization name. This is mainly because of the idiosyncrasies of the variegated set of the peoples responsible for naming an organization. So, we first find the phrases which represent only the country, email id, URL, state, city, and address (in that order) and then consider the left-over phrases to see if they represent organization names. Fig. 1 shows the sequence.

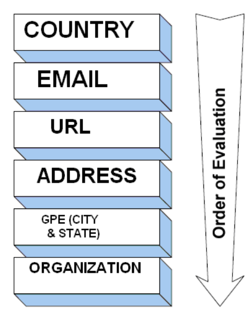

Fig 1: Order of NER rules

While this is the general framework, each subtask, say determining the name of the country, requires use of multiple rules and multiple dictionaries. In total there are 30 different manually verified dictionaries or lists generated from databases such as the Geoworldmap database [7]. For example:- Internet domain – country dictionary, stop words list, and key words for addresses.

Once the country, state, email addresses and URLs are removed, the sentences are broken into phrases. Those phrases which only represent an address or a geopolitical entity are also removed. While breaking into phrases, we take care that the common suffixes of organizations such as – LLC, Ltc, etc are not separated. Before checking whether a phrase is an organization, all the acronyms are matched by a regular expression and they are replaced by their full form using an acronym dictionary that was built using [8]. If the phrase contains an O-Key or Organization key word, then it is considered an organization. If there is no O-Key in the phrase, but the phrase has a person name or has a scientific term, then also the phrase is considered as an organization.

4. Evaluation

We obtained 6042 Medline abstracts (with affiliation) related to "Atrial Fibrillation" published between the years 2004 and 2008. Since we are evaluating a system that is expected to have near 100% performance, it is better that multiple experts carefully scrutinize each result produced by the system. So, we decided to follow the standard evaluation design suggested for such a scenario as shown in Figure 2. The metrics for evaluation such as true positive, false positive, false negative, precision, recall and f-measure follow usual definitions as in [9]. Out of the 6042 articles, 1828 were predicted to be from USA. The three data analysts commissioned with the task of validating the results reported 100.0% accuracy in determining the articles from USA and found 39 sentences with false positives and 2 sentences with false negatives. Since there are multiple organizations in a sentence, a false positive in the sentence doesn't make the whole result for the sentence a false positive if at least one organization is correctly recognized; same is the case with false negative. For the sentence which doesn't have false positives and false negatives, the precision and recall are 1.0. The average precision and average recall are calculated by finding the arithmetic mean of the precision and recall of the 1800 sentences. The average precision is 0.992, the average recall is 0.999 and so, the average f-measure is 0.996. It might be worthy to appreciate this result in comparison with the current state of art [4] of 87% accuracy for all countries.

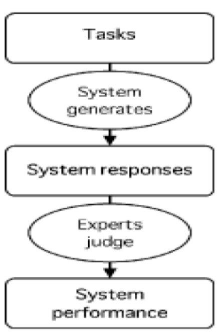

Fig 2: Evaluation Design

5. Conclusion

We have presented a highly accurate Named Entity Recognition system for organizations in the affiliation sentences in PubMed abstracts.

References

[1] E. A. Balas and S. A. Boren, "Managing clinical knowledge for health care improvement," *Yearbook Med. Info.*, 2000.

[2] Cutting Edge Information, "Pharmaceutical thought leaders: Brand strategies and product positioning," *Tech. Rep. PH64*, 2004.

[3] E. S. Knowles and J. A. Linn, *Resistance and Persuasion.* Lawrence Erlbaum Associates, 2004.

[4] W. Yu, et al., "An automatic method to generate domain-specific investigator networks using PubMed abstracts," *BMC Med. Inform. Decis. Mak.*, 2007.

[5] Brunstein Ada, "Annotation Guidelines For Answer Types," *http://www.ldc.upenn.edu/catalog/docs/ldc2005T33/BBN-Types-Subtypes.Html*, 2002.

[6] Sekine, "Sekine's extended named entity hierarchy," *http://nlp.cs.nyu.edu/ene*, 2007.

[7] G. Inc, "Geoworldmap database containing cities of the world with geographical coordinates," 2009.

[8]STANDS4 LLC, "Abbreviations," *http://abbreviations.com* 2009.

[9] R. Leaman and G. Gonzalez, "BANNER: An executable survey of advances in biomedical named entity recognition," in *Pac Symp Biocomput,* 2008.