Understanding Research Articles

Rajal Nivargi
rfn5089@psu.edu
The Pennsylvania State University

Abstract

Named Entity Recognition(NER) is a technical term for a solution to a key automation problem: extraction of information from text. This project addresses the problem of scientific information extraction from research articles using NER. A new dataset called HERC is created which contains abstracts collected from 200 research articles. An implementation of a baseline model shows that this method can be useful to identify relevant keywords in scientific text by using NER.

1 Introduction

In the public health domain, a large number of studies and clinical trials address the problem of stress and anxiety while also suggesting ways to combat them. Though medication and other services are available to the public, it becomes imperative to give people agency by educating them about natural remedies for these aspects of their mental health. This can be done by analyzing scientific research and clinical evidence, gathering new evidence through data science and citizen science, and recommending to individuals optimal combinations of teas, tinctures, and powders that can be used to achieve each individual's mental health goals. This project conducts this research with the goal of allowing people to make informed decisions about their health and to increase access to healthy stress-coping alternatives. These healthy coping mechanisms can reduce reliance on harmful self-medication practices such as alcohol abuse or behavior likely to lead to opioid addiction.

This project aims to automate the process of identification and extraction of scientific information from research articles. The abstract of the paper usually contains the essential entities for defining the medicinal food and the conclusions of the study. The abstracts are open-access as compared to full-text of the papers which may be restricted. Analysing the summary provided in the abstract can be a filter to finding the research that is pertinent as a resource validating the evidence of traditional medicinal foods against stress and anxiety.

2 Related work

Information extraction is a growing area of interest in the natural language processing. Named entity recognition (NER) is the task of tagging entities in text with their corresponding type. Approaches typically use BIO notation, which differentiates the beginning (B) and the inside (I) of entities. O is used for non-entity tokens. There are multiple CoNLL datasets created for NER. The CoNLL 2002 format is used in this paper.(Tjong Kim Sang, 2002) This dataset was designed for a shared task of language-independent named entity recognition.

The SciERC dataset(Luan et al., 2018) created by annotating 500 abstracts from papers in the Semantic Scholar Corpus is a three task setup including named entities, relations and co-references. This is an extension of the previous datasets in scientific articles SemEval 2017 Task 10 (SemEval 17)(Augenstein et al., 2017) and SemEval 2018 Task 7 (SemEval 18)(Gábor et al., 2018). The annotation gudielines of (QasemiZadeh and Schumann, 2016) and BRAT interface were used for this dataset. This project follows the same guidelines along with using INCEpTION[1] as the annotation interface. The other datasets that were created using INCEpTION were Brazilian Legal Text (LeNER-BR)(Luz de Araujo et al., 2018) and Digital Athenaeus project[2] for the NER task.

[1] https://inception-project.github.io/

[2] http://www.digitalathenaeus.org/

3 HERC dataset

3.1 Data collection

The dataset contains 200 abstracts of research articles. These were collected by scraping published work from PubMed[3]. PubMed is a large repository of biomedical literature and life science journals. The papers were searched using the keywords 'stress', 'anxiety' and medicinal food name. 20 medicinal foods were considered for the purpose of this project. Their effect on stress and anxiety has been widely researched. Out of the total 375 paper abstracts scraped, a random set of 200 were selected for manual annotation and 10 other for unseen data to test the model on.

3.2 Annotation

INCEpTION[4](See figure 1) is used as the annotation tool. Its web-based platform which provides many tools to address the semantic annotation tasks. The plain text files containing abstracts of 200 research articles were annotated. 10% of the abstracts were dually annotated by 2 other annotators. They were provided with the project description and the annotation guidelines. The kappa score for annotating the entities is 79.89%. The SciERC(Luan et al., 2018) dataset annotation scheme and methods were used to define the annotation guidelines. The scope is restricted to Entities, and does not include Relations and Co-references for this project. The Named Entity layer was used for this annotation effort and a custom tag set was created to denote the values for the given span of entity. The entities are defined as follows:

1. Medicinal food: Herbs/foods the paper is based on.
2. Drug/Scientific name: Name of the drug used in the study/ The scientific name or particular active ingredient of the medicinal food.
3. Dosage: Dose for each medicinal food/drug/scientific name. The dosages for animal based studies are not considered.
4. Other_ingredients: Foods other than the medicinal foods/herbs.
5. Symptom_Workedfor: Symptom for which symptom this study implies positive results - stress/sleep/anxiety/other ailments.
6. Symptom_TestedFor: Symptom for which the study tested but does not imply positive results - stress/sleep/anxiety/other ailments.
7. Metric: What is used for measuring the symptom.
8. Duration: Duration of the study.(This not the incubation period of the medicinal food but just the amount of time the study has been conducted. The duration mentioned in the animal studies is not included.)
9. #Participants: Number of participants in the study.
10. participant_health: Health status of the participant. (What health problems the participant may have.)

Tag type	Tag counts
Medicinal foods	237
Symptoms_WorkedFor	145
Drug/Scientific Name	93
Metric	68
Symptom_TestedFor	60
Duration	41
Other_ingredients	32
participant_health	25
#participants	14
Dosage	13

4 Model implementation

For the implementation, a simple baseline of spaCy v2.0's[5] Named Entity Recognition system is used. It features a word embedding strategy using subword features and "Bloom" embeddings, a deep convolutional neural network with residual connections, and a novel transition-based approach to named entity parsing.

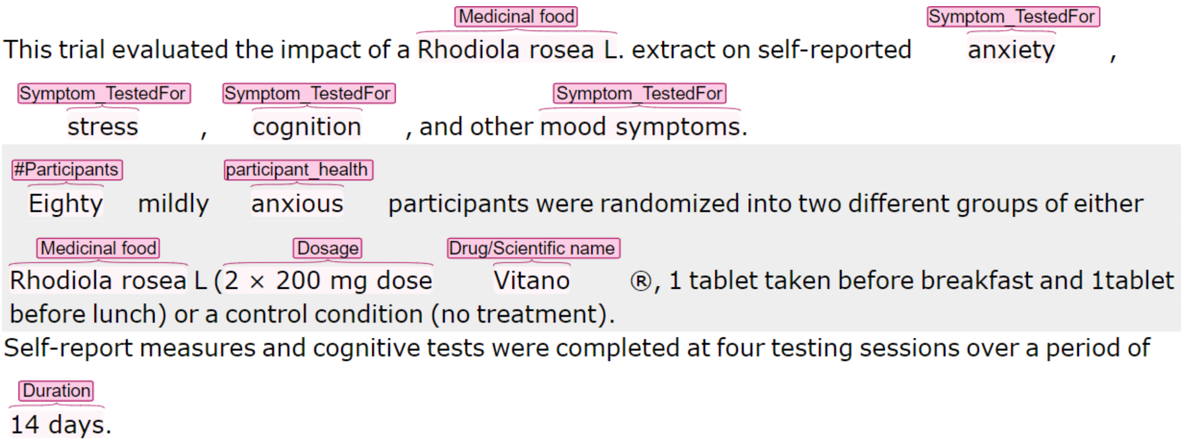
4.1 Preprocessing

The project created on INCEpTION is exported in the CoNLL 2002 format. Each annotation is saved as a .conll file.(Figure 2) As an input to the model, each line consists of two columns separated by a space. The first column contains a token and the second a tag of the entity. The sentence boundary is marked with an empty line. All the documents are combined in one CoNLL file. SpaCy takes training data in JSON format.(Figure 3) Therefore, a built-in convert command is used convert the

[3] https://pubmed.ncbi.nlm.nih.gov/

[4] https://inception-project.github.io/

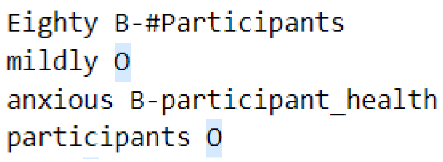
[5] https://spacy.io/



This trial evaluated the impact of a Rhodiola rosea L. extract on self-reported anxiety , stress , cognition , and other mood symptoms.
Eighty mildly anxious participants were randomized into two different groups of either Rhodiola rosea L (2 × 200 mg dose Vitano ®, 1 tablet taken before breakfast and 1tablet before lunch) or a control condition (no treatment).
Self-report measures and cognitive tests were completed at four testing sessions over a period of 14 days.

Figure 1: Annotation in INCEpTION

.conll format in the corpora to spaCy's training format. The 200 annotated abstracts were split randomly into 160 files for training data and 40 files for validation data using 80-20 split. This created 2006 training docs ad 3191 evaluation docs along with a vocabulary of 8016 unique words and 54,490 total words.



```
Eighty B-#Participants
mildly O
anxious B-participant_health
participants O
```

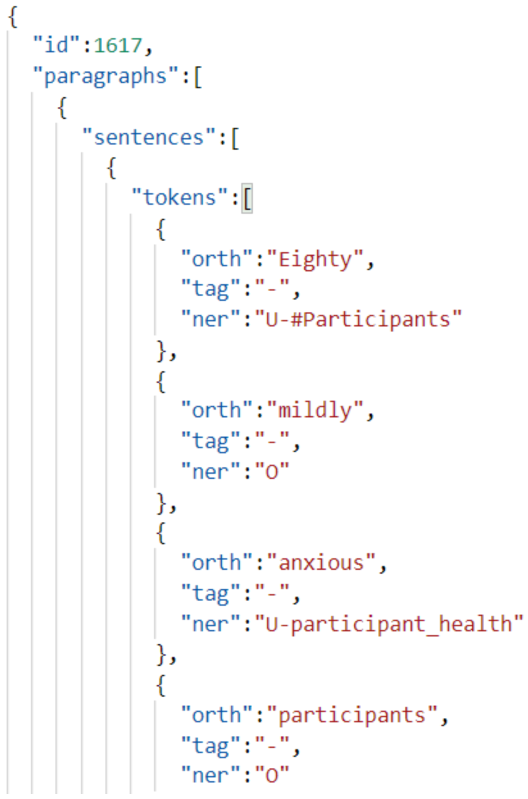
Figure 2: CoNLL format

4.2 Train NER model

The multi-task CNN trained statistical model on OntoNotes(Hovy et al., 2006) in the English language is used as base model for the Named Entity Recognition task. The base model assigns context-specific token vectors, POS tags, dependency parse and named entities. SpaCy's named entity recognizer is trained with our examples for 25 epochs. It starts off with from scratch using a blank model added to the pipeline and runs over the entire training data for a number of epochs. The model was also trained for 50 epochs to observe if the number of training steps help improve the performance. However, the model showed similar results.

4.3 Evaluation

Th trained model is evaluated on the validation dataset. The results are shown in 4.3.

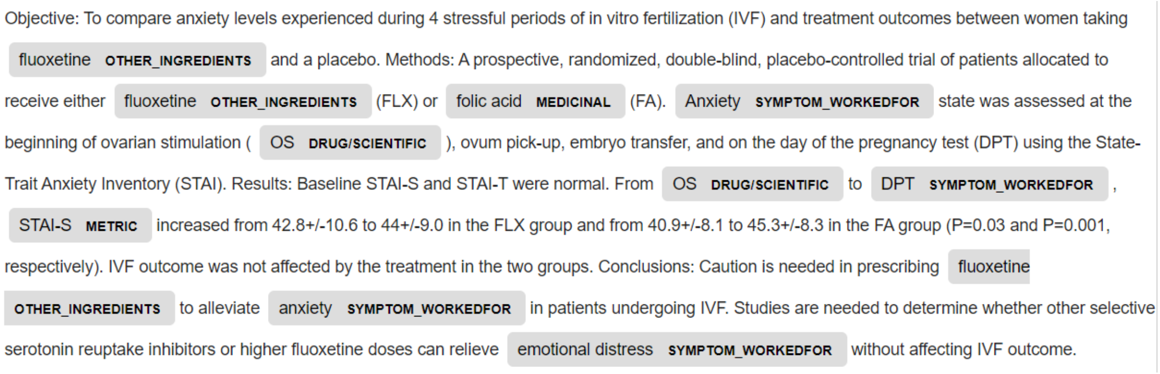


```
{
  "id":1617,
  "paragraphs":[
    {
      "sentences":[
        {
          "tokens":[
            {
              "orth":"Eighty",
              "tag":"-",
              "ner":"U-#Participants"
            },
            {
              "orth":"mildly",
              "tag":"-",
              "ner":"O"
            },
            {
              "orth":"anxious",
              "tag":"-",
              "ner":"U-participant_health"
            },
            {
              "orth":"participants",
              "tag":"-",
              "ner":"O"
```

Figure 3: spaCy json format

Time	1.44 s
Words	28894
NER P	46.28
NER R	45.15
NER F	45.71

The low performance of the model can be attributed to the low amount of data available to train on. The correlation between the F1 score and the number of annotations for the entities is shown in 5. The Scorer class in spacy is used to generate predictions on the validation dataset from the best



Objective: To compare anxiety levels experienced during 4 stressful periods of in vitro fertilization (IVF) and treatment outcomes between women taking fluoxetine OTHER_INGREDIENTS and a placebo. Methods: A prospective, randomized, double-blind, placebo-controlled trial of patients allocated to receive either fluoxetine OTHER_INGREDIENTS (FLX) or folic acid MEDICINAL (FA). Anxiety SYMPTOM_WORKEDFOR state was assessed at the beginning of ovarian stimulation (OS DRUG/SCIENTIFIC), ovum pick-up, embryo transfer, and on the day of the pregnancy test (DPT) using the State-Trait Anxiety Inventory (STAI). Results: Baseline STAI-S and STAI-T were normal. From OS DRUG/SCIENTIFIC to DPT SYMPTOM_WORKEDFOR , STAI-S METRIC increased from 42.8+/-10.6 to 44+/-9.0 in the FLX group and from 40.9+/-8.1 to 45.3+/-8.3 in the FA group (P=0.03 and P=0.001, respectively). IVF outcome was not affected by the treatment in the two groups. Conclusions: Caution is needed in prescribing fluoxetine OTHER_INGREDIENTS to alleviate anxiety SYMPTOM_WORKEDFOR in patients undergoing IVF. Studies are needed to determine whether other selective serotonin reuptake inhibitors or higher fluoxetine doses can relieve emotional distress SYMPTOM_WORKEDFOR without affecting IVF outcome.

Figure 4: The output on unseen data from the trained model(25 epochs)

performing model. These predictions are used to visualize the scatter plot in 5.

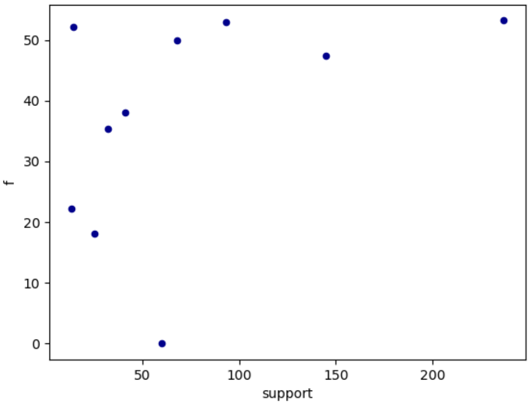


Figure 5: Plotting the F1-Score (f) versus the number of tokens for the tags shows a correlation between poor performance and shortage of training data

When tested on unseen data, the model performs reasonably well as seen in figure 4. It was able to identify the medicinal food, symptoms and other entities of importance. With a better model and extensive data, this tool can be very useful for scientific information extraction in the public health domain.

5 Challenges

- **Dataset pre-processing** The exported project can be obtained in a few formats like CoNLL(Tjong Kim Sang, 2002), XML, JSON, etc. However, to be used to train on Glove(Pennington et al., 2014) or Elmo(Peters et al., 2018) embeddings, the data has to be pre-processed in a certain way along with vocabulary files. This pre-processing is possible on this dataset and can be pursued to train better models using word embeddings.
- **Less training data** With the limited time and manual work requirement, the annotations were the most time-taking task in this project. However, with the help of few more annotators, it should be possible to annotate more data. The SciERC dataset (Luan et al., 2018) has 500 abstracts that are annotated. Extending the training set to more data can be a plausible way to improve performance.
- **Problems in training data** In scraping the abstracts, all types of research articles are extracted. A subset of these articles based on clinical trails on humans is the scope of this project. Along with these, review and analysis papers are extracted which do not provide enough entities or information. Animal studies are also a type of paper that is not important and is damaging for the training process. The animal studies mention dosages, duration, and other entities which are not relevant for human requirements. These are not annotated and thus, may distract the model. Therefore, data cleaning is another step to be added to the pipeline of this project.

6 Conclusion and future work

The project proposed an approach to the automatic information extraction of entities from published articles. This method can be used for meta-analysis of research in the public health and citizen science domain. The HERC dataset can be used for the NER task of identifying relevant data from the abstracts of papers. While the dataset has some challenges, with larger training data and data cleaning, the performance can be improved. In the future, the information extraction models like SciIE(Luan et al., 2018), SciBERT(Beltagy et al., 2019) and SpERT(Eberts and Ulges, 2019) can be used to train on the HERC dataset.

7 Acknowledgement

The list of 20 medicinal foods is obtained from the database created by CrowdDoing[6] for the project Medicinal foods for stress, sleep and anxiety.

References

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada. Association for Computational Linguistics.

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *CoRR*, abs/1903.10676.

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training. *CoRR*, abs/1909.07755.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90 In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, page 57–60, USA. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *CoRR*, abs/1808.09602.

Pedro H. Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In *International Conference on the Computational Processing of Portuguese (PROPOR)*, Lecture Notes on Computer Science (LNCS), pages 313–323, Canela, RS, Brazil. Springer.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Behrang QasemiZadeh and Anne-Kathrin Schumann. 2016. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1862–1868, Portorož, Slovenia. European Language Resources Association (ELRA).

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, page 1–4, USA. Association for Computational Linguistics.

8 Appendices

The 20 medicinal foods are :

1. Rhodiola rosea
2. Vitamin B
3. Bacopa monnieri
4. Green tea
5. Lavender
6. Holy basil
7. Turmeric
8. Chamomile
9. Bergamot
10. Ashwagandha
11. Chinese Skullcap
12. Damiana
13. Eleuthro
14. Folic acid
15. Ginkgo biloba
16. Ginseng (Siberian)
17. Magnolia Officinalis
18. Lemon balm
19. Passionflower
20. Valerian

[6] https://www.crowddoing.world/