INTERNSHIP REPORT

BACHELOR OF TECHNOLOGY

in

ENGINEERING PHYSICS

by

Karanveer Singh Sirohi

2K20/EP/061

Under Supervision of

Dr. Ashika, Scientist F

CAIR, DRDO

Bangalore.

(Duration: 21/04/2023 to 06/10/2023)

Department of Applied Physics,

DELHI TECHNOLOGICAL UNIVERSITY
formerly Delhi College of Engineering
(under Delhi Act 6 of 2009, Govt. of NCT of Delhi)
Accredited with 'A' Grade (CGPA 3.22 out of 4.0) by NAAC (1st Cycle)
ISO 9001: 2015 Certified

REPORT

Machine Learning Paradigms for Information Extraction

Karanveer Singh Sirohi

INDEX

1. Introduction

1.1 What is Named Entity Recognition?

Named Entity Recognition is the process of locating a word or a phrase that references a particular entity within a text. The NER task first appeared in MUC-6 and involved the recognition of Entity names (people and organizations), place names, temporal expressions, and numerical expressions.

In MUC-6, Named Entities were categorized into three types of labels, each of which uses specific attributes for a particular entity type. Entities and their labels were defined as follows:

- **ENAMEX**: person, organization, location

- **TIMEX**: date, time

- **NUMEX**: money, percentage, quantity

Named Entity Recognition (NER) plays a crucial role in natural language processing by identifying and classifying named entities within textual data. Named entities, such as persons, organizations, locations, dates, and more, carry valuable information that is vital for various language-based applications, including information extraction, question-answering, and text summarization. The ability to accurately detect and categorize these entities has become essential in enabling machines to understand and process human language effectively.

This report aims to explore the concept of Named Entity Recognition and investigate the implementation of various methods for this task. We will delve into traditional rule-based approaches, statistical and machine learning methods, as well as the advancements achieved through neural network architectures. By understanding and comparing these approaches, we aim to gain insights into their strengths, limitations, and potential applications.

In the following sections, we will first provide a comprehensive background on NER, discussing its significance, the different types of named entities, and the challenges associated with the task. Subsequently, we will explore traditional rule-based approaches, where handcrafted patterns and linguistic rules are employed. We will then delve into statistical and machine learning techniques, focusing on algorithms such as Hidden Markov Models, Conditional Random Fields, and Support Vector Machines. Lastly, we will explore neural network approaches, including recurrent neural networks, long short-term memory networks, and transformer models, which have revolutionized the field with their ability to capture complex contextual information.

To evaluate the performance of these methods, we will discuss commonly used evaluation metrics and benchmark datasets for NER tasks. By conducting experiments and analyzing the results, we aim to gain insights into the strengths and weaknesses of each method and provide a comparative analysis.

This report seeks to contribute to the existing body of knowledge on NER and provide a comprehensive understanding of its implementation through different methods. By highlighting the advancements and challenges in this field, we hope to shed light on current trends and future directions for NER research.

2. Literature Survey

Introduction to Named Entity Recognition (NER) in Biomedical Texts

NER is pivotal in biomedical studies, aiding in areas like drug discovery, analyzing genetic links in diseases, and managing vast patient data. It streamlines knowledge synthesis by extracting key entities from biomedical texts, thus enhancing hypothesis creation and decision-making in this field.

Challenges in Biomedical NER

- **Diverse Biomedical Terminology**: Biomedical texts are laden with specialized terminology, acronyms, and jargon, distinct from everyday language, posing a significant challenge due to its complexity and continual evolution.

- **Ambiguity and Multiple Meanings**: Biomedical language often contains words with multiple meanings, adding complexity to entity recognition. For example, 'cold' might indicate a viral infection or a sensation related to temperature.

- **Context-Dependent Meanings**: In biomedical texts, the significance of terms frequently relies on their specific context. For instance, a reference to a drug could pertain to its therapeutic use, adverse effects, or its pharmacological properties, necessitating context awareness for precise entity identification.

Role of NER in Biomedical Informatics

- **Enhancing Data Mining and Knowledge Extraction**: NER is crucial for deriving valuable insights from biomedical literature, health records, and clinical data, thus playing a substantial role in biomedicine's knowledge expansion.
- **Supporting In-depth Research**: NER assists in processing and organizing extensive datasets, enabling researchers to uncover new findings, like potential drug targets or genetic indicators for diseases.

Evolution of NER Techniques

Initial Approaches to NER

- **Rule-Based Systems:** Early NER systems were built on custom-made rules, utilizing lexicons, pattern matching, and linguistic guidelines for entity identification. These systems worked well within certain areas but lacked versatility and required significant manual input for adaptation to new areas or changes in language.

- **Machine Learning Approaches**: Following rule-based systems, there was a shift towards machine learning methods such as Support Vector Machines (SVM) and Hidden Markov Models (HMM). These approaches trained on annotated datasets to recognize patterns, although they necessitated intricate feature engineering and faced challenges in handling the nuanced aspects of language processing.

Advancements with Statistical Models

Advancements in NER Techniques

- Conditional Random Fields (CRF): CRFs marked a significant advancement in NER by accounting for context and label dependencies, enhancing their performance in sequence labeling, particularly in handling biomedical language nuances.

- Limitations of CRFs: Despite their success, CRFs were resource-intensive and depended on manually created features, which hindered their flexibility in adapting to the evolving biomedical lexicon.

Deep Learning Evolution in NER

- Neural Networks and Embeddings: The emergence of deep learning, especially through neural networks using word embeddings like Word2Vec and GloVe, revolutionized feature learning from large text datasets.

- RNNs and LSTMs: Recurrent Neural Networks, notably Long Short-Term Memory networks, gained prominence for their ability to process sequential data and context, vital in biomedical NER.

State-of-the-Art Developments

- Transformers and BERT: Recent breakthroughs in NER are led by transformer models like BERT, excelling in capturing deep, bidirectional contexts. Pre-training on extensive data followed by domain-specific fine-tuning has significantly advanced biomedical NER.

Impact on Biomedical NER

- Specialization Trend: The progression from rule-based to deep learning models reflects a shift towards specialized biomedical NER systems, adept at handling specific text challenges.

- Improved Accuracy and Scalability: Each stage of NER technology evolution has enhanced the precision, efficiency, and scalability of information extraction from complex biomedical literature.

Feature Engineering in Biomedical NER

- **Lexical Features**: These encompass the text's words, their capitalization, and structural patterns like prefixes and suffixes, which are often indicative of specific biomedical entities (like diseases or proteins).

- **Syntactic Features**: Derived from sentence structure, including parts of speech and sentence segmentation, these features help determine each word's role, crucial for entity identification.

- **Semantic Features**: Critical in biomedical texts, these include the word's biological function, interactions, and context-specific meanings, often informed by specialized ontologies and lexicons.

- **Contextual Features**: Focusing on the words surrounding a term, these features are essential in biomedical texts where meanings can vary greatly based on context.

Challenges in Biomedical Feature Engineering

- **Complex Biomedical Language**: The extensive specialized terminology and jargon in biomedicine pose significant challenges in feature creation.
- **Evolving Biomedical Knowledge**: Rapid changes in biomedical terms require ongoing updates to feature sets.
- **Ambiguity and Multiple Meanings**: The presence of ambiguous terms or those with multiple meanings complicates the assignment of accurate entity types based on linguistic features alone.

Feature Engineering in CRF-based Biomedical NER

- **CRFs and Feature Interdependency:** In CRFs, the quality of features is crucial for modeling the label's conditional probability given the data. Their ability to handle complex dependencies between features makes them ideal for biomedical NER.
- **Feature Granularity Balance:** Effective biomedical NER requires a careful balance in feature detail. While overly broad features may overlook domain-specific nuances, excessively detailed features risk overfitting, reducing the model's general applicability.

Evolution with Deep Learning

Shift in Focus with Neural Models in NER

With the rise of deep learning, particularly RNNs and transformer models, the emphasis on manual feature engineering in NER has significantly diminished. These advanced neural models are adept at autonomously learning intricate data representations, reducing the necessity for detailed, domain-specific feature engineering.

References:

1. Sun, L., Patel, R., Liu, J., Chen, K., Wu, T., Li, J. and Ye, J.
   (2009) Mining brain region connectivity for Alzheimer’s disease study via sparse inverse covariance estimation.
2. Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li (2018)A Survey on Deep Learning for Named Entity Recognition.

3. Dataset

The dataset consists of annotated biomedical texts from the WhiteText corpus, which includes diverse biomedical terms and complex sentence structures, presenting a challenging scenario for entity recognition.

1. The dataset used to train the models was WhiteTextNegFixFull.xml
2. The dataset used to test the models was WhiteTextUnseenEval.xml

Both of these can be downloaded from this link:

1. http://dx.doi.org/10.5683/SP2/AARXSN
2. http://dx.doi.org/10.5683/SP2/4J5NHT

4. Experiment Setup

The experiment was run in Google Colab.
Here are a few examples of how to Implement each of the three methods:

For MEM:

```
import nltk
from nltk.classify import MaxentClassifier
from nltk.metrics import accuracy


training_data = [({'feature1': 'value1', 'feature2': 'value2'}, 'label1'),
                 ({'feature1': 'value3', 'feature2': 'value4'}, 'label2')]
testing_data = [({'feature1': 'value1', 'feature2': 'value2'}, 'label1')]


classifier = MaxentClassifier.train(training_data, 'IIS', trace=0,
max_iter=1000)

accuracy = accuracy(classifier, testing_data)
print(f"Model Accuracy: {accuracy}")
```

For CRF:

```
import sklearn_crfsuite

X_train = [[{'feature1': 'value1'}, {'feature1': 'value2'}]]
y_train = [['label1', 'label2']]

crf = sklearn_crfsuite.CRF(algorithm='lbfgs')
crf.fit(X_train, y_train)


X_test = [[{'feature1': 'test_value'}]]
predicted = crf.predict(X_test)
print("Predicted Labels:", predicted)
```

For SVM:

```
from sklearn import svm
from sklearn.feature_extraction.text import CountVectorizer

X_train = ['text data example 1', 'text data example 2']
y_train = ['label1', 'label2']
X_test = ['test data example']

vectorizer = CountVectorizer()
X_train_counts = vectorizer.fit_transform(X_train)

clf = svm.SVC()
clf.fit(X_train_counts, y_train)

X_test_counts = vectorizer.transform(X_test)
predicted = clf.predict(X_test_counts)
print("Predicted Labels:", predicted)
```

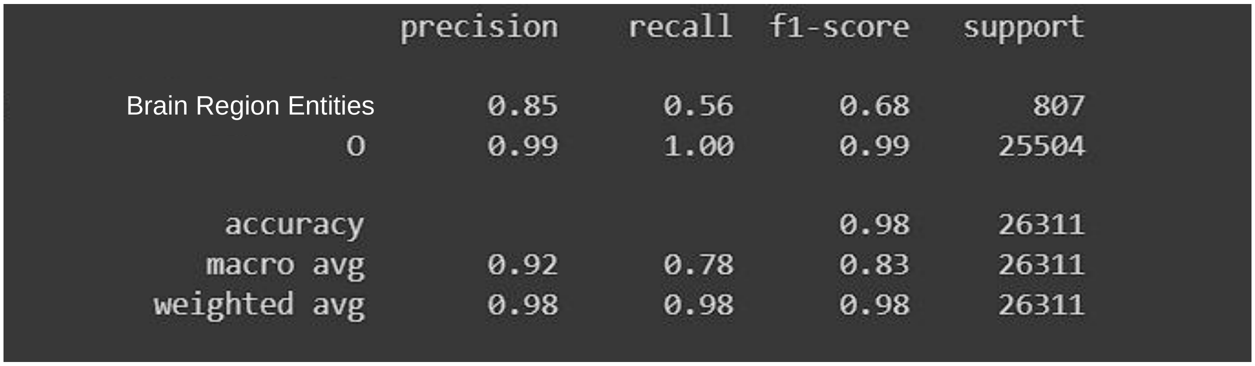
IMPLEMENTATION IN GOOGLE COLAB:

This is the link of the code and results obtained in google colab.

bconn_ext.ipynb

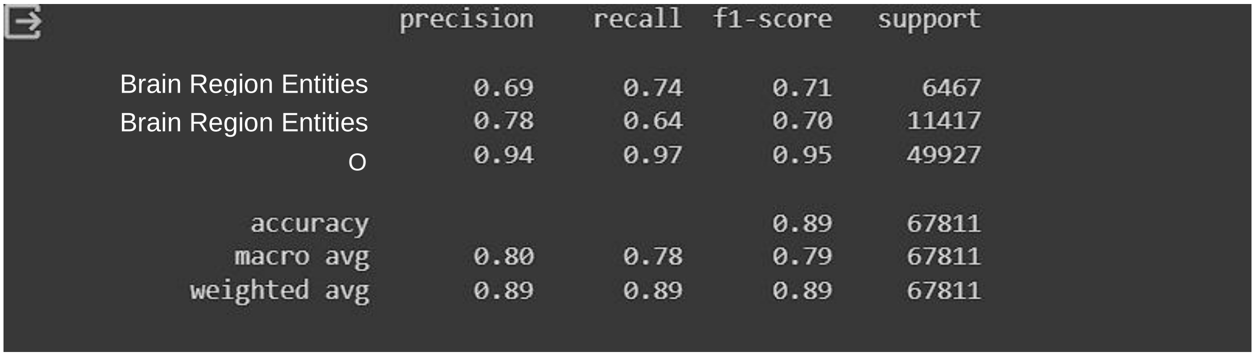
5. Results

Results for Conditional Random Field:



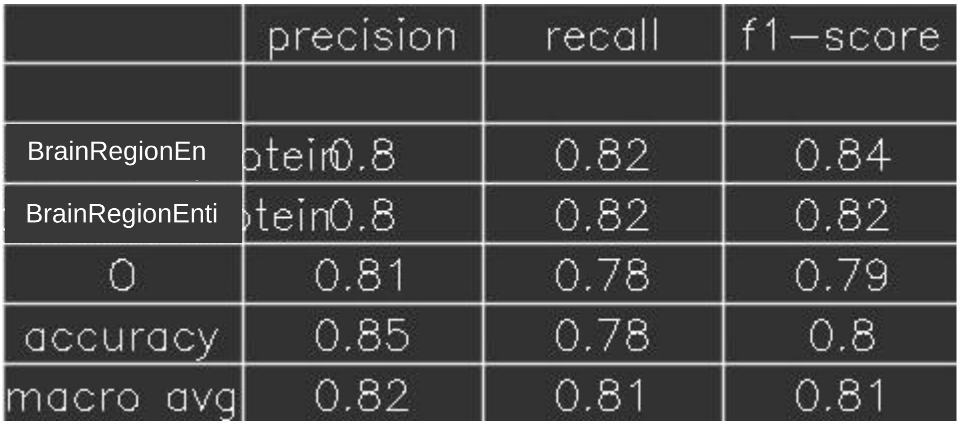
```
                       precision    recall  f1-score   support

  Brain Region Entities       0.85      0.56      0.68       807
                      O       0.99      1.00      0.99     25504

               accuracy                           0.98     26311
              macro avg       0.92      0.78      0.83     26311
           weighted avg       0.98      0.98      0.98     26311
```

Results for Maximum Entropy Model:



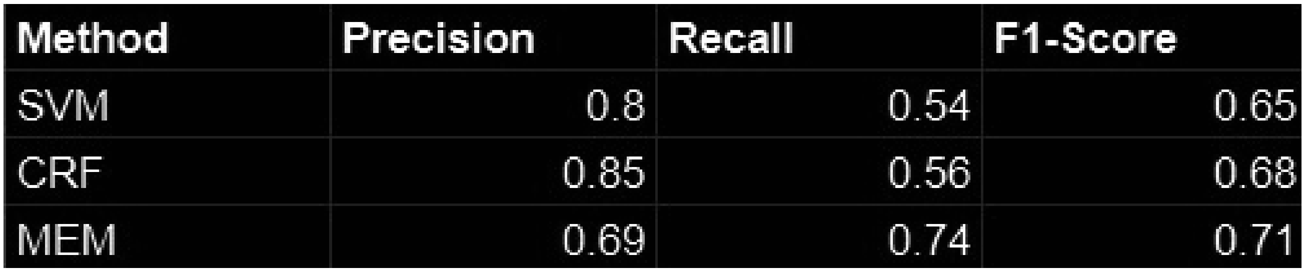
```
                       precision    recall  f1-score   support

  Brain Region Entities       0.69      0.74      0.71      6467
  Brain Region Entities       0.78      0.64      0.70     11417
                      O       0.94      0.97      0.95     49927

               accuracy                           0.89     67811
              macro avg       0.80      0.78      0.79     67811
           weighted avg       0.89      0.89      0.89     67811
```

Results for Support Vector Machines:



	precision	recall	f1−score
BrainRegionEn	pteir0.8	0.82	0.84
BrainRegionEnti	tein0.8	0.82	0.82
O	0.81	0.78	0.79
accuracy	0.85	0.78	0.8
macro avg	0.82	0.81	0.81

Final comparison of all the results and values:



Method	Precision	Recall	F1-Score
SVM	0.8	0.54	0.65
CRF	0.85	0.56	0.68
MEM	0.69	0.74	0.71

6. Discussion on Results

The results of this experiment show that CRF with hand-picked features as a supervised machine learning model yielded the best results.

Precision:

For **"B-Individual_protein"**, the precision is 0.85, meaning that 85% of the instances predicted by the model as a "B-Individual_protein" were correct.

For non-entities **"O"**, the precision is very high at 0.99, indicating that the model is very effective at correctly identifying tokens that are not entities.

Recall:

The recall for "**B-Individual_protein**" is 0.56, suggesting that the model identified 56% of all actual "B-Individual_protein" entities in the dataset.

The recall for "**O**" is perfect at 1.00, which means every instance that was a non-entity was identified by the model.

F1-Score:

The F1-score for **"B-Individual_protein"** is 0.68, which is the harmonic mean of precision and recall for the entity. It reflects a moderate balance between precision and recall but indicates room for improvement, particularly in recall.
The F1-score for **"O"** is nearly perfect at 0.99.

Support:

Support refers to the number of actual occurrences of the class in the dataset. There are 807 instances of "B-Individual_protein" and 25504 instances of non-entities "O".

The model has a high accuracy of **0.98**, which indicates the overall ability of the model to correctly label tokens. However, this high accuracy is likely influenced by the high number of non-entities, which are much easier to identify correctly.

Analysis of Results:

The CRF model performs very well on non-entities, almost perfectly classifying them. This is expected since the majority of tokens in a typical text are non-entities and are easier to classify.

The performance on "B-Individual_protein" is good in terms of precision but needs improvement in recall. This could suggest that while the model is conservative in its prediction of entities (few false positives), it misses a significant number of actual entities (false negatives).

The disparity between precision and recall for "B-Individual_protein" may suggest the model could be improved with additional training data, better feature engineering, or by adjusting the class weight to pay more attention to the minority class.

The model’s overall accuracy is high, but this metric can be misleading due to the imbalance in the dataset. The F1-score is a more reliable indicator of the model's performance on the entity recognition task, especially for the "B-Individual_protein" class.

7. Conclusion

The performance metrics suggest that CRF is adept at capturing the nuances of biomedical entity recognition, but there's room for improvement in recognizing actual entities. The high precision with lower recall for "B-Individual_protein" indicates a potential need for further refinement of the model or additional labeled data to enhance the model's ability to identify true entity instances without increasing false positives.