Named Entity Recognition from Water Regulations

RAXIT GOSWAMI, University of South Carolina, USA

ACM Reference Format:
Raxit Goswami. 2022. Named Entity Recognition from Water Regulations. *Course Project - Computer Processing of Natural Language, University of South Carolina* 1, 1 (November 2022), 7 pages.

1 INTRODUCTION

Water quality monitoring is a crucial aspect of protecting water resources. Under the Clean Water Act, state, tribal, and federal agencies monitor lakes, streams, rivers, and other types of water bodies to determine water quality conditions. The data generated from these monitoring activities help water resource managers know where pollution problems exist, where to focus pollution control energies and where progress has been made. The standards and procedures prescribed are also necessary to maintain reasonable standards of purity of the drinking water of the State consistent with the public health, safety, and welfare of its citizens[1].

Named Entity Recognition (NER) helps to identify the key elements (entities) from the text. Extracting the main entities in a text helps to convert unstructured data to structured data and detect entities, which is important if you have to deal with large datasets.

2 PROBLEM

Extracting all the various chemicals and water components from the water regulations data. Develop a system to parse all the water regulations guidelines (unstructured text) and identify all the chemicals, the measurement values and the water category (e.g, drinking water, swimming water) as mentioned in the figure below.

This parser will be helpful to get the structured information from the water regulations text

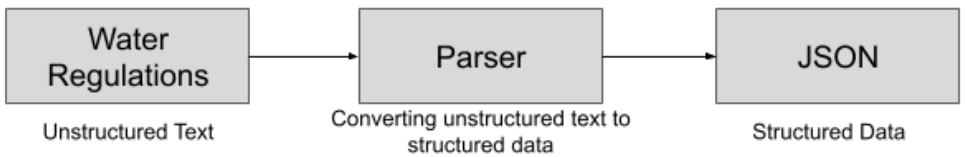

Fig. 1. Highlevel overview of the problem

and generate JSON output. Different web/mobile application (e.g, chat bot) can be developed on top of extracted data from the regulations for different use cases (e.g, Check the water quality and compare various chemical components across the states).

Author's address: Raxit Goswami, rgoswami@email.sc.edu, University of South Carolina, Columbia, SC, USA, 29201.

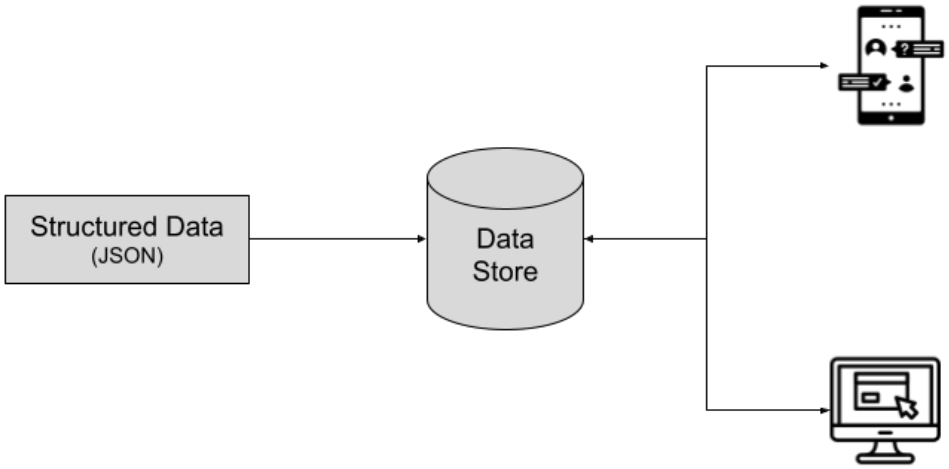

Fig. 2. Connections to the various application with NER output

3 RELATED WORK

Named Entity Recognition is a very well known problem and people have tried various approaches to get the good accuracy. A good amount of papers are published on NER using knowledge-graph/dictionary lookup based approaches. These approaches completely rely on the dictionary/knowledge graph and do not require any training dataset to train the system[7]. But it requires exhaustive lexicons in the dictionary and it fails when lexicons are not present. A large amount of research has been done using supervised machine learning methods(e.g, CRF[4]) for NER tasks. Supervised systems require more labeled dataset and the creation of the labeled training data is very costly. The Word2vec[6] and BERT[3] release have changed the entire paradigm in the area of NLP. The idea here is to get the most benefit from the raw data to improve the accuracy of the NER systems. An extensive amount of research has been published using BERT [3]specifically for the NER task in various domains. BERT with token classification[3] performs well for different domains even though BERT is trained on Book corpus and English wiki dataset[3]. There are many different variations of BERT released and people are using those for different purposes. We are introducing a domain specific NER where utilizing domain specific raw dataset to update the BERT weight and do the token classification on updated weight.

4 APPROACH

We now describe our approach to produce the NER model for water regulation data. A graphical representation of different modules which are used to create NER models are illustrated in the figure given below.

4.1 Pre-Processing

As described above water regulation data is in an unstructured format and available in PDF/Word documents. To process this data/content, the first task is to convert the data into plain text. Converting PDF/Word to text has its own challenges for e.g, content hierarchy, line break issues, special characters and tabular data. Examples of each PDF to text issues are described below. All the issues

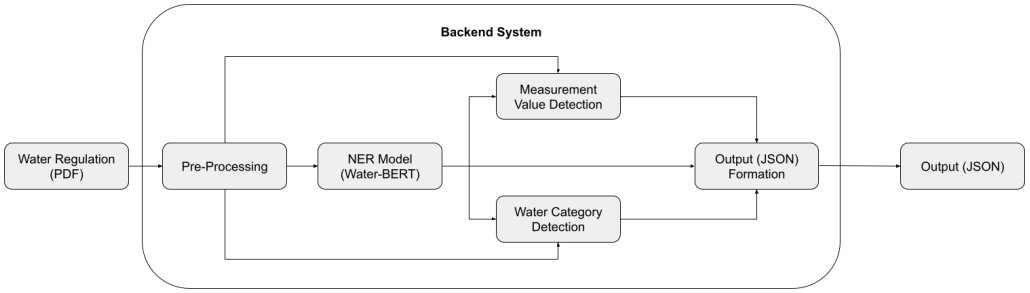

Fig. 3. Overall System Workflow

of PDF to text except content hierarchy are taken care of in the pre-processing module.

For the PDF to TXT, I have used the linux library (pdftotext) to convert PDF to TXT. In the library, there is an option to maintain the layout of the document in the text format while converting it from PDF version. But it still fails to maintain proper layout for the tabular data and the line break issues remain in the converted version. All these issues are explained in detail with examples below.

Example : Line break issues
As we can see, TTHM, HAA5 and measurement values are mentioned in the snapshot 1. When we convert this PDF to text, we can see that TTHM, HAA5 and measurement values are presented in separate lines (refer line numbers in the snapshot). It's difficult to identify line breaks are actual or occurring because of PDF to text. Rules are defined to solve this issue after analyzing the data.

Solution : Merge lines until we get two new line characters in the content. By this way, I am merging the content and converting into sentences to process further.

(b) What criteria must the Department use to determine that a profile is unnecessary? The Department may only determine that a system's profile is unnecessary if a system's TTHM and HAA5 levels are below 0.064 mg/L and 0.048 mg/L, respectively. To determine these levels, TTHM and HAA5 samples must be collected after January 1, 1998, during the month with the warmest water temperature, and at the point of maximum residence time in your distribution system. The Department may approve a more representative TTHM and HAA5 data set to determine these levels

Fig. 4. PDF format of the water regulation data

```
19 (b) What criteria must the Department use to determine that a profile is unnecessary? The Department
20 may only determine that a system's profile is unnecessary if a system's TTHM and HAA5 levels are below
21 0.064 mg/L and 0.048 mg/L, respectively. To determine these levels, TTHM and HAA5 samples must be
22 collected after January 1, 1998, during the month with the warmest water temperature, and at the point of
23 maximum residence time in your distribution system. The Department may approve a more representative
24 TTHM and HAA5 data set to determine these levels
```

Fig. 5. TXT format of the PDF data mentioned in Fig. 4

Example: Tabular Data Formatting
As per the snapshot 3, tabular data is presented in different columns and rows and the same data

converted into text and presented in snapshot 4. System will process data line by line, but as we see in snapshot 4, “Chromium (Hex), Total” and “Dissolved” are presented in different lines. Even after applying the line break solution, “Dissolved” will come after 11 e and the entire meaning of the content will be changed.

7782505	Chlorine	19	11
18540299	Chromium (Hex), Total Dissolved	16 [e]	11 [e]
16065831	Chromium (III), Total Dissolved	323 [b,e]	42 [b,e]

Fig. 6. Tabular data from the Water regulation PDF

```
4 7782505    Chlorine                                19              11
5 18540299   Chromium (Hex), Total                   16 e            11 e
6          Dissolved
7 16065831   Chromium (III), Total Dissolved         323 b,e         42 b,e
```

Fig. 7. Tabular data presented in snapshot 3, converted into text

Solution: Nowadays, OCR technology is mature enough to identify tables from the scanned documents. I treat this data as an image and used OCR to identify this as a proper tabular structure and then converted into text data where I got “Chromium (Hex), Total“ and “Dissolved” are together.

4.2 Data Annotation

Creating an accurately labeled dataset can be very time consuming. The method of auto annotation works by using various AI/ML models and dictionary lookup which in-turn speeds up the annotation time. We implemented the auto annotation approach in this project with an extra layer of "human in the loop" process to validate the quality of the machine generated output. We used various dictionary mappings to auto annotate the dataset and manually validated the outputs, made corrections and updates in the annotations wherever the auto annotation process inaccurately labeled the entities. This process helped us to expedite the labeled dataset creation without affecting the quality of the dataset.

4.3 NER Model Training

Overall training process, starting from input (PDF document) to the NER model is presented in the figure 3 above.

Bidirectional Encoder Representations from Transformers (BERT) is used to perform the NER task on the water regulation data. Fine-tuning is straightforward since the self attention mechanism in the Transformer allows BERT to model many downstream tasks, whether they involve single text or text pairs by swapping out the appropriate inputs and outputs[3]. For each task, we simply plug in the task specific inputs and outputs into BERT and finetune all the parameters end-to-end. Token Classification with BERT is represented in the Fig. 8. We have used Base-BERT and different variations for the token classification task.

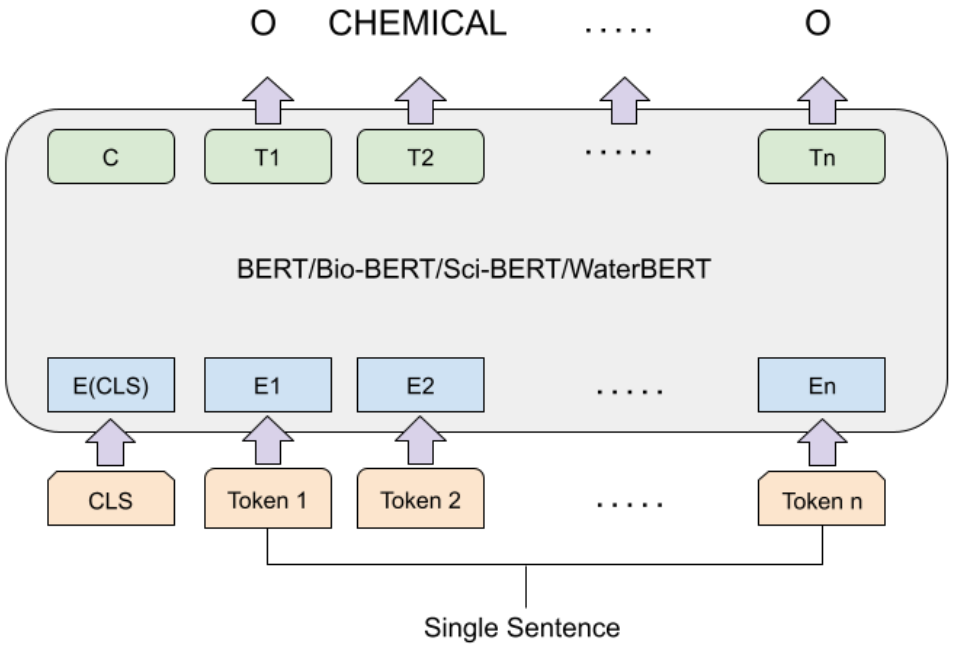

Fig. 8. Token Classification using BERT

5 EVALUATION

As described, BERT and different variations are used for token classification task on the water regulations. BERT and different variations (Bio-BER[5], Sci-BERT[2]) are pre-trained on specific corpus. Training corpus with size for different BERT variations are mentioned in the table 2.

Different study proves, BioBERT[5] and SciBERT[2] performs better than baseBERT for biomedical domain. Linguistic structure varies for different domain specific data. We also tried fine-tune BERT on water domain to verify the same hypothesis. For the fine tuning process, 130 different books as well as different water regulation data sets are considered. It covers 7,586,359 tokens in the dataset. We achieved slightly better accuracy in comparing with base-BERT, Bio-BERT[5] and Sci-BERT[2].

Precision, Recall, and F-Measure are calculated to evaluate different pre-trained language models for the NER output on water regulations and precision, recall and F1 are mentioned below in table 2.

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1 = \frac{2*Precision*Recall}{Precision+Recall}$$

Model	#Tokens	Source	Domain
BERT	2.5B	Books	General
	0.8B	Wiki	General
BioBERT	13.5B	PMC Full-text articles	Biomedical
	4.5B	PubMed Abstracts	Biomedical
	3.3B	Wiki+Books	General
SciBERT	3.17B	Scientific Papers	18% Computer Science, 82% broad biomedical

Table 1. Corpus size and domain of the pre-trained language models

Approach	Precision	Recall	F1
BERT+Token Classification	0.9359	0.9492	0.9416
BioBERT+Token Classification	0.9345	0.9707	0.9514
SciBERT+Token Classification	0.9647	0.9488	0.9565
WaterBERT+Token Classification	0.9617	0.9659	0.9631

Table 2. NER Experiment with different BERT variation on validated annotation of water regulations

6 DISCUSSION

Many different domain specific BERT variations are released day by day. Different research on domain specific tasks also proved better accuracy by using the domain specific language models. And by looking at the results for the NER on water regulations, fine tuned BERT on Water data performed better than any other BERT variation. By looking at all the facts, we can say that BERT variation in water data will be helpful to achieve very good results in different NLP tasks.

Different BERT variations are trained on very large corpus, BERT is trained on 3.3B tokens, Bio-BERT is trained on 21.3B tokens[5], Sci-BERT is trained on 3.1B tokens[3]. NER on single entity type is compared to simple tasks in NLP but fine tuned BERT with only 7.6M tokens performs slightly better than any other BERT. Using more data and creating pre-trained version of BERT will be really helpful for all the different tasks in NLP on the water domain

7 CONCLUSION

Domain specific text has the variation in terms of linguistic patterns in the content. Utilizing raw dataset to learn the linguistic pattern and perform various NLP tasks using updated language model is giving slightly better accuracy. Hence, this model is not tested with NLP tasks other than NER for single entity type. We will continue our research and validate this model for different NLP tasks (e.g, Summarization, Complex NER task, Relationship detection).

REFERENCES

[1] 2022. Water Quality Data. https://www.epa.gov/waterdata/water-quality-data

[2] Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. SciBERT: Pretrained Contextualized Embeddings for Scientific Text. *CoRR* abs/1903.10676 (2019). arXiv:1903.10676 http://arxiv.org/abs/1903.10676

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[4] John D. Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*.
[5] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *CoRR* abs/1901.08746 (2019). arXiv:1901.08746 http://arxiv.org/abs/1901.08746
[6] Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. http://arxiv.org/abs/1301.3781
[7] Vikas Yadav and Steven Bethard. 2019. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. *CoRR* abs/1910.11470 (2019). arXiv:1910.11470 http://arxiv.org/abs/1910.11470