Named Entity Recognition from Water Regulations

Student : Raxit Goswami

Project Guide : Dr. Biplav Srivastava

1. Problem

- Extract 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 (drinking water, swimming water, etc....) as mentioned in the figure below.

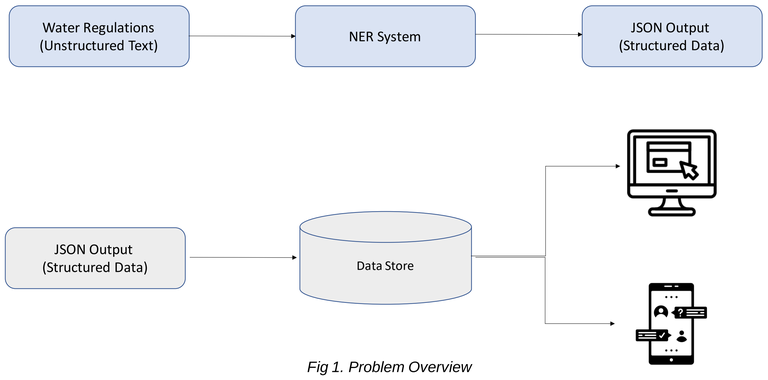


Fig 1. Problem Overview

Mississippi Data: 586 Pages

Annotation: 586 Pages, Auto Annotation with Human Validation

2. Challenges in the Data

Example 1: Line Break Issue

(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

```
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
```

Example 2: Tabular Structure

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

```
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
```

3. Approach

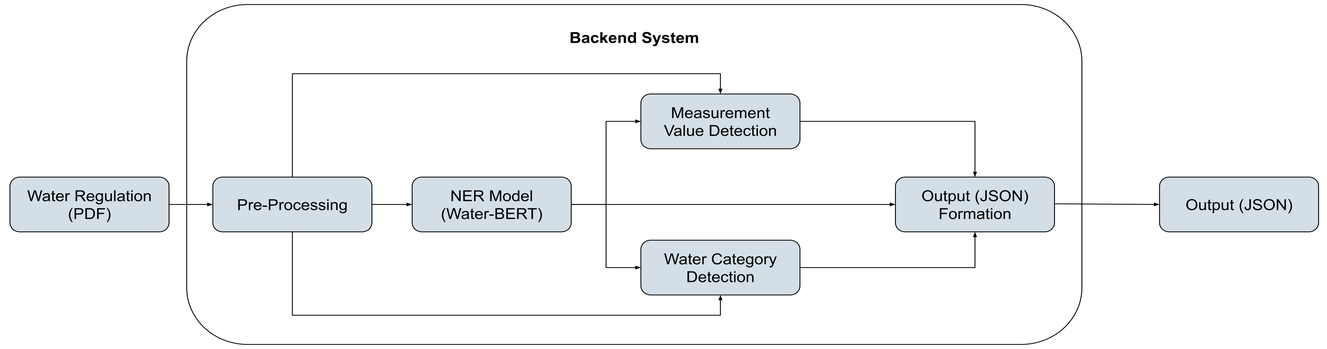


Fig 2. NLP System for Water Regulations

PDF to TXT : Linux Library is used to maintain the PDF layout

Pre-Processing : Rules to solve various issues in PDF to TXT(e.g, Line Break, Tabular Format)

NER Model : BERT and different variations of BERT are used. Finetune BERT on Books and regulations data related to water domain.

Measurement Value and Water Category : Rules to identify measurement value and water category for the entity identified by NER model

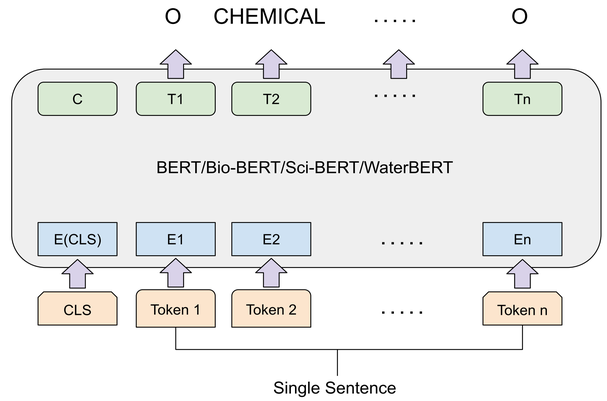
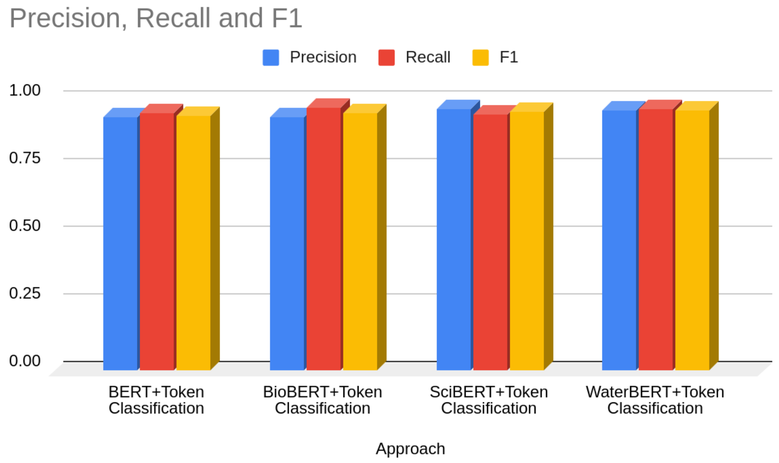


Fig 3. BERT Architecture for token Classification Task

4. Results



5. Conclusion & Future Work

- 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).
- Different BERT variations are trained on very large corpus, BERT is trained on 3.3B tokens, Bio-BERT is trained on 21.3B tokens, Sci-BERT is trained on 3.1B tokens.
- 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

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 Named Entity Recognition from Water Regulations
- [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