Kaushik Acharya at WNUT 2020 Shared Task-1: Conditional Random Field(CRF) based Named Entity Recognition(NER) for Wet Lab Protocols

Kaushik Acharya
Philips India Ltd. / Bangalore, India
acharya.kaushik@gmail.com

Abstract
Detecting named entities in user generated text is a challenging task. Lab protocols specify steps in performing a lab procedure. The majority of wet lab protocols are written in noisy, dense, and domain-specific natural language. There is a growing need of automatic or semi-automatic conversion of protocols into machine-readable format to benefit biological research.

The paper describes how a classifier model built using Conditional Random Field[1] detects named entities in wet lab protocols. The model trained on the training data showed precision, recall and F1-score of 0.762, 0.743 and 0.752 respectively on the development set. When applied to unseen test data, the model showed 0.737, 0.640 and 0.685 respectively.

1 Introduction
Wet laboratories are laboratories for conducting biology and chemistry experiments. These require handling of various types of chemicals and potential "wet" hazards. These experiments are guided by a sequence of instructions collectively referred as wet lab protocols.

The instructions are mostly composed of imperative statements which are meant to describe an action. Figure 1 shows a representative wet lab protocol. Figure 2 shows BRAT annotations (entities and relations) on two sentences from the representative protocol. For each protocol, annotators had identified and marked every span of text corresponding to action or one of the 17 types of entities. Table 1 shows a few typical examples for each of these classes. For detailed description of entities please refer Kulkarni et al’s [2] Annotation Guidelines.

1 https://github.com/kaushikacharya/wet_lab_protocols
2 https://sklearn-crfsuite.readthedocs.io/en/latest/

Figure 1: An example wet lab protocol

Figure 2: Example sentences from the wet lab protocol example shown in figure 1 as shown in the BRAT annotation interface.

2 Named Entity Recognition Methodology
A Conditional Random Fields (CRF) classifier was trained to recognize named entities. The CRF NER model was implemented using sklearn-crfsuite

---

286
Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 286–289
Online, Nov 19, 2020. ©2020 Association for Computational Linguistics
which is a Python wrapper over C++ based CRF-suite \(^3\). It utilized L-BFGS [3], a limited memory quasi-Newton algorithm for large scale numerical optimization. The classifier was trained with both L1 and L2 regularization.

2.1 Features

Three types of features have been extracted using Python library spaCy [4].

- **Lexical features**
  - Unigrams
  - Lemmas

- **Parts of speech (POS) features**
  - Current word’s POS
  - Prev and Next word’s POS
  - Governor word’s POS

- **Dependency parse features**
  - Governor words
  - Dependency type
  - Dependency type of Governor word

As an example, *microfuge* in the sentence shown in Figure 3 produces the following features:

- **current word POS**: NOUN
- **dependency tag**: pobj
- **parent dependency tag**: prep

![Figure 3: POS and dependency parse for the sentence shown in figure 2. Dependency parse tree visualized using spaCy’s displaCy.](image)

Table 1: Top frequent examples of Action and Entities

| Tag               | Examples                                    |
|-------------------|---------------------------------------------|
| Action            | add, incubate, mix                          |
| Amount            | 50 l, equal volume                          |
| Concentration     | 1x                                          |
| Device            | filter, vacuum, microfuge                   |
| Location          | 30-kd, several times, 100v tube, plate, well|
| Measure-Type      | volume, concentration                       |
| Mention           | It, them, this                              |
| Method            | up and down, extraction                     |
| Modifier          | each, gently, at least one, 3, several, several times |
| Numerical         | ph 8.0, ph8.0                               |
| Reagent           | cells, supernatant                          |
| Seal              | Lid, cap, aluminum foil                    |
| Size              | 0.02 m, 12 x 75 mm                          |
| Speed             | 14,000xg, 10,000 rpm                        |
| Temperature       | room temperature, overnight                 |
| Time              | 5 minutes                                   |

Table 2: Wet Lab Protocols dataset statistics

| Dataset  | Protocols | Sentences |
|----------|-----------|-----------|
| Training | 370       | 8444      |
| Development | 122    | 2839      |
| Test     | 123       | 2813      |

3 Experiments

The experiments were based on the datasets provided by the organizers of W-NUT 2020 shared task on Entity and Relation Recognition over Wet Lab Protocols [5]. The dataset (Table 2) was annotated in both StandOff and CoNLL formats. Entities and relations of 615 protocols were annotated in brat with 3 annotators with 0.75 inter-annotator agreement, measured by span-level Cohen’s Kappa.

3.1 Results

For the experiments, the classifier was trained on the training data and evaluated on development and test data.

The reported averages are defined as follows:

- **Macro average**: averaging the unweighted mean per label.
- **Micro average**: averaging the total true positives, false negatives and false positives.
- **Weighted average**: averaging the support-weighted mean per label

Table 3 and Table 4 show results at token and entity level respectively.

Table 5 compares my results *(KaushikAcharya)* to the other systems participating in the shared task on the unseen test data.
Table 3: Token level metrics on development and test sets (P: Precision, R: Recall, F1: F1 score). This includes non-entity tokens also as one of the classes.

| Dataset | Average | P     | R     | F1    |
|---------|---------|-------|-------|-------|
| Dev     | Macro   | 0.677 | 0.633 | 0.651 |
|         | Weighted| 0.809 | 0.817 | 0.812 |
| Test    | Macro   | 0.691 | 0.646 | 0.666 |
|         | Weighted| 0.833 | 0.839 | 0.835 |

Table 4: Entity level metrics on development and test sets (P: Precision, R: Recall, F1: F1 score).

| Dataset | Average | P     | R     | F1    |
|---------|---------|-------|-------|-------|
| Dev     | Micro   | 0.762 | 0.743 | 0.752 |
|         | Macro   | 0.755 | 0.743 | 0.748 |
| Test    | Micro   | 0.782 | 0.766 | 0.774 |
|         | Macro   | 0.777 | 0.766 | 0.771 |

Table 5: Comparison of system results on both exact and partial match (F1 score)

| Team                | Exact | Partial |
|---------------------|-------|---------|
| BITEM               | 77.99 | 81.67   |
| PublishInCovid19    | 77.57 | 81.75   |
| mgsorhab            | 76.6  | 80.5    |
| Kabir               | 75.35 | 80.08   |
| Winners             | 74.91 | 79.54   |
| BIO-BIO             | 74.59 | 79.03   |
| Fancy Man Launches Zippo | 73.92 | 78.71   |
| SudeshnaTCS         | 73.16 | 77.8    |
| B-NLP               | 70.25 | 76.46   |
| **KaushikAcharya**  | 68.48 | 73.73   |
| IBS                 | 67.9  | 72.89   |
| DSC-IITISM          | 60.42 | 64.49   |
| mahab               | 51.54 | 56.57   |

Table 6: Entity level classification metrics per entity type

| Entity               | Precision | Recall | F1 score |
|----------------------|-----------|--------|----------|
| Action               | 0.871     | 0.885  | 0.878    |
| Amount               | 0.865     | 0.824  | 0.844    |
| Concentration        | 0.688     | 0.730  | 0.708    |
| Device               | 0.613     | 0.584  | 0.598    |
| Generic-Measure      | 0.255     | 0.205  | 0.227    |
| Location             | 0.726     | 0.678  | 0.701    |
| Measure-Type         | 0.562     | 0.502  | 0.530    |
| Mention              | 0.662     | 0.589  | 0.623    |
| Method               | 0.473     | 0.395  | 0.430    |
| Modifier             | 0.588     | 0.502  | 0.541    |
| Numerical            | 0.544     | 0.480  | 0.510    |
| pH                   | 0.853     | 0.784  | 0.817    |
| Reagent              | 0.742     | 0.781  | 0.761    |
| Seal                 | 0.769     | 0.778  | 0.773    |
| Size                 | 0.625     | 0.385  | 0.476    |
| Speed                | 0.881     | 0.784  | 0.830    |
| Temperature          | 0.919     | 0.919  | 0.919    |
| Time                 | 0.919     | 0.907  | 0.913    |

Table 7: Frequent confuser entity classes

- Predicted entity text span matches truth but entity class is incorrect.
- Entity text span mismatches.
  - Partial match: Example shown in Table 8.
  - Complete mis-match: Example shown in Table 9.

Table 8 and Table 9 show examples of misclassification for the highlighted text portion of the corresponding sentences.

Table 8: Expected: Two entities for the highlighted phrase: a) Modifier: *lab grade* b) Reagent: *water*. Whereas the system predicted a single entity(Reagent) over the entire text span.

Table 9: Expected: Entity(Method) for the highlighted phrase. Whereas the system predicted two

3.2 Error Analysis

Entity type wise performance metrics on development dataset are available in Table 6. This is based on strict evaluation mode of matching as defined in SemEval’13 [6]. As per strict evaluation, a predicted entity is correct only if it matches with gold-standard in both exact boundary and type. Used seqeval for the evaluation.

Table 7 shows the poorly performing entity classes along with their frequent confusers.

Errors are of primarily two types:

• Predicted entity text span matches truth but entity class is incorrect.
• Entity text span mismatches.
  - Partial match: Example shown in Table 8.
  - Complete mis-match: Example shown in Table 9.
entities: a) Action: without lysing  b) Reagent: erythrocytes.

| Text              | Entity     | Truth/Predicted |
|-------------------|------------|-----------------|
| lab grade         | Modifier   | Truth           |
| water             | Reagent    | Truth           |
| lab grade water   | Reagent    | Predicted       |

Table 8: Mis-classification #1:
Sentence: Rinse slides with lab grade water.

| Text              | Entity     | Truth/Predicted |
|-------------------|------------|-----------------|
| without lysing    | Method     | Truth           |
| erythrocytes      | Action     | Predicted       |
| without lysing    | Action     | Predicted       |
| erythrocytes      | Reagent    | Predicted       |

Table 9: Mis-classification #2:
Sentence: Prepare cells from your tissue of interest without lysing erythrocytes.

Figure 4: Global structured information in the dependency tree. The figure shows the dependency sub-tree for the Reagent entity: 10g/L Fe Stock Solution. The entity’s text span is covered by the subtree having Solution as its root and of as its head.

4 Conclusion
This paper has proposed a CRF-based named entity extraction system to extract Action and 17 Entities of wet lab protocols. Future plan:

- Analyse the errors in more detail and extract richer features.
- Extract global structured information features of the dependency trees[7] as shown in Figure 4. Currently as the system only uses local dependency features, it predicts Fe Stock Solution as Reagent entity and misses 10g/L.
- Develop Long Short-Term Memory (LSTM) recurrent neural network model [8].

References
[1] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. pages 282–289, 01 2001.
[2] Alan Ritter Raghu Machiraju Chaitanya Kulkarni, Wei Xu. An Annotated Corpus for Machine Reading of Instructions in Wet Lab Protocols. In Proceedings of NAACL-HLT, 2018.
[3] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. Mathematical Programming, 45:503–528, 1989.
[4] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.
[5] Jeniya Tabassum, Sydney Lee, Wei Xu, and Alan Ritter. WNUT-2020 Task 1 Overview: Extracting Entities and Relations from Wet Lab Protocols. In Proceedings of EMNLP 2020 Workshop on Noisy User-generated Text (WNUT), 2020.
[6] Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIEXtraction 2013). In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 341–350, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.
[7] Zhanming Jie, Aldrian Obaja Muis, and Wei Lu. Efficient dependency-guided named entity recognition, 2018.
[8] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 260–270, San Diego, California, June 2016. Association for Computational Linguistics.