# DS-GA 1006, Fall 2017: Capstone Project and Presentation

# Roivant Sciences Drug Names Identification

Kailing Wang (kailing.wang@nyu.edu)

## 1 Introduction

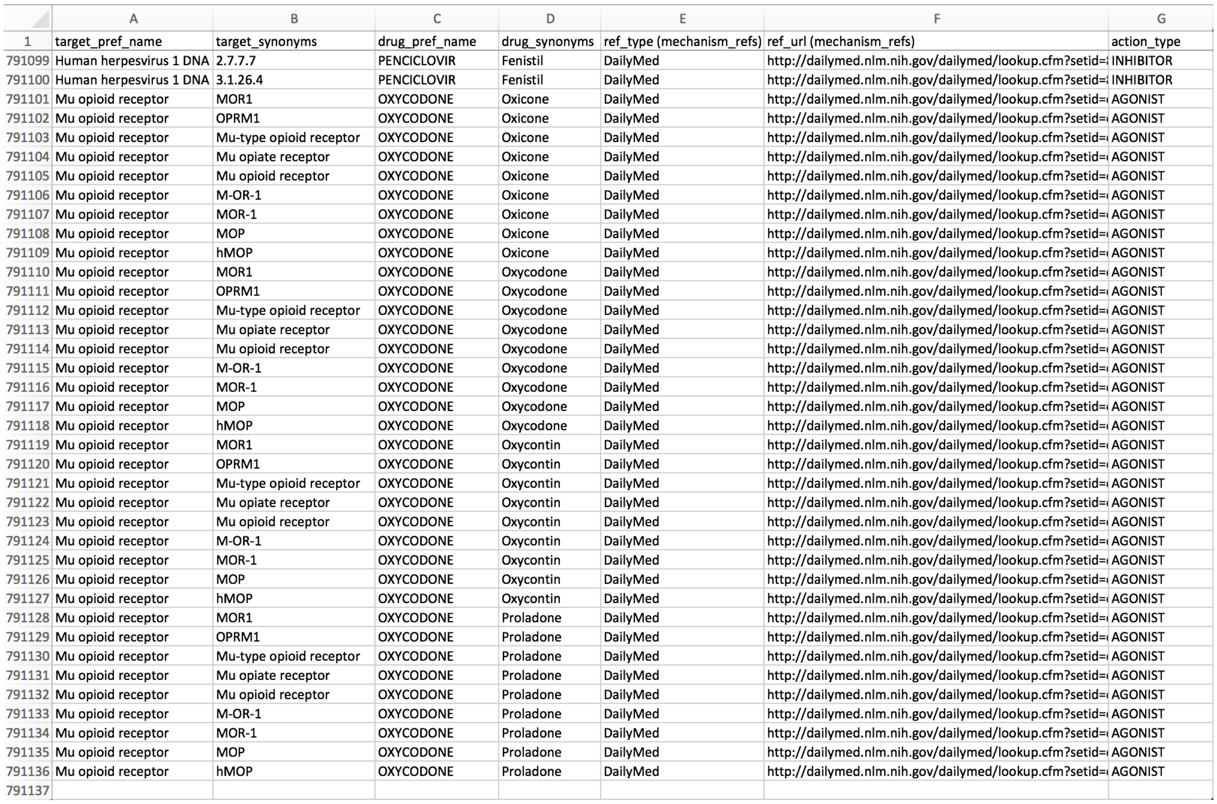
The project is to establish the document analysis pipeline where given a list of targets of interest, we improve our Named Entity Recognition (NER) to find drug names that we don't already have in our drug name dictionary from millions of medical papers. The company has made previous efforts on this problem, that is building a deep neural networks NER purely based on the syntax of documents. In this project, we will leverage ChEMBL[1] database, taking advantage of the sentence structures from those reference documents. We would expect that this additional assistance from ChEMBL could expand our drug NER capabilities. This methodology could be easily generalized into other scenarios –provided a list of entities, making recommendations of their relevant entities from unstructured data.

## 2 Data

ChEMBL is a manually curated medical database. An NDA will be required. The data is cleaned and accessible from our MySQL database.

The raw dataset contains nearly 800,000 records. However, it should be noted that a large proportion are the combinations of various drug aliases and target aliases. So if we remove those

[1] https://www.ebi.ac.uk/chembldb/



| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | target_pref_name | target_synonyms | drug_pref_name | drug_synonyms | ref_type (mechanism_refs) | ref_url (mechanism_refs) | action_type |
| 791099 | Human herpesvirus 1 DNA | 2.7.7.7 | PENCICLOVIR | Fenistil | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | INHIBITOR |
| 791100 | Human herpesvirus 1 DNA | 3.1.26.4 | PENCICLOVIR | Fenistil | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | INHIBITOR |
| 791101 | Mu opioid receptor | MOR1 | OXYCODONE | Oxicone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791102 | Mu opioid receptor | OPRM1 | OXYCODONE | Oxicone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791103 | Mu opioid receptor | Mu-type opioid receptor | OXYCODONE | Oxicone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791104 | Mu opioid receptor | Mu opiate receptor | OXYCODONE | Oxicone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791105 | Mu opioid receptor | Mu opioid receptor | OXYCODONE | Oxicone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791106 | Mu opioid receptor | M-OR-1 | OXYCODONE | Oxicone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791107 | Mu opioid receptor | MOR-1 | OXYCODONE | Oxicone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791108 | Mu opioid receptor | MOP | OXYCODONE | Oxicone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791109 | Mu opioid receptor | hMOP | OXYCODONE | Oxicone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791110 | Mu opioid receptor | MOR1 | OXYCODONE | Oxycodone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791111 | Mu opioid receptor | OPRM1 | OXYCODONE | Oxycodone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791112 | Mu opioid receptor | Mu-type opioid receptor | OXYCODONE | Oxycodone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791113 | Mu opioid receptor | Mu opiate receptor | OXYCODONE | Oxycodone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791114 | Mu opioid receptor | Mu opioid receptor | OXYCODONE | Oxycodone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791115 | Mu opioid receptor | M-OR-1 | OXYCODONE | Oxycodone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791116 | Mu opioid receptor | MOR-1 | OXYCODONE | Oxycodone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791117 | Mu opioid receptor | MOP | OXYCODONE | Oxycodone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791118 | Mu opioid receptor | hMOP | OXYCODONE | Oxycodone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791119 | Mu opioid receptor | MOR1 | OXYCODONE | Oxycontin | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791120 | Mu opioid receptor | OPRM1 | OXYCODONE | Oxycontin | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791121 | Mu opioid receptor | Mu-type opioid receptor | OXYCODONE | Oxycontin | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791122 | Mu opioid receptor | Mu opiate receptor | OXYCODONE | Oxycontin | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791123 | Mu opioid receptor | Mu opioid receptor | OXYCODONE | Oxycontin | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791124 | Mu opioid receptor | M-OR-1 | OXYCODONE | Oxycontin | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791125 | Mu opioid receptor | MOR-1 | OXYCODONE | Oxycontin | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791126 | Mu opioid receptor | MOP | OXYCODONE | Oxycontin | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791127 | Mu opioid receptor | hMOP | OXYCODONE | Oxycontin | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791128 | Mu opioid receptor | MOR1 | OXYCODONE | Proladone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791129 | Mu opioid receptor | OPRM1 | OXYCODONE | Proladone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791130 | Mu opioid receptor | Mu-type opioid receptor | OXYCODONE | Proladone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791131 | Mu opioid receptor | Mu opiate receptor | OXYCODONE | Proladone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791132 | Mu opioid receptor | Mu opioid receptor | OXYCODONE | Proladone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791133 | Mu opioid receptor | M-OR-1 | OXYCODONE | Proladone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791134 | Mu opioid receptor | MOR-1 | OXYCODONE | Proladone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791135 | Mu opioid receptor | MOP | OXYCODONE | Proladone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791136 | Mu opioid receptor | hMOP | OXYCODONE | Proladone | DailyMed | http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid= | AGONIST |
| 791137 | | | | | | | |

Figure 1: A screenshot of ChEMBL dataset

duplicates, there remains 7,320 records. It also should be noted that ChEMBL only collect a subset of actually related drug-target pairs in their listed publications.

# 3 Methodology

## 3.1 Prepocessing (Completed)

From ChemBl's list of documents, we extract two types of predicates based on the outcome of Syntaxnet's Dependency Parser and Part-of-Speech (POS) tagging: "Good Predicates": predicates that link a target to a drug; "Other Predicates": predicates that link a target to a word that is not a drug name.

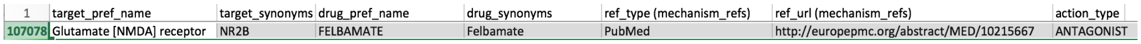
### 3.1.1 “Good Predicates” Extraction

1. From ChemBL listed publications, extract sentences containing both corresponding drugs and targets

2. From the list of sentences obtained in the previous step, apply Syntaxnet’s dependency parser to identify possible predicates, maintaining POS and dependency parameter as backup information.

   The general strategy is to trace upwards from both the drug and the target in the dependency tree until a joint node (word) is reached, and then extract all nodes (words) along the two paths. This method works for most of sentence structures. However, there exist quite a few exceptions. For example, when the drug and the target are not within a same clause, the extracted predicate would contains lots of noises. Furthermore, in some cases, the drug could be the appostive of the target, which means that there would be no verb phrase linking them together.

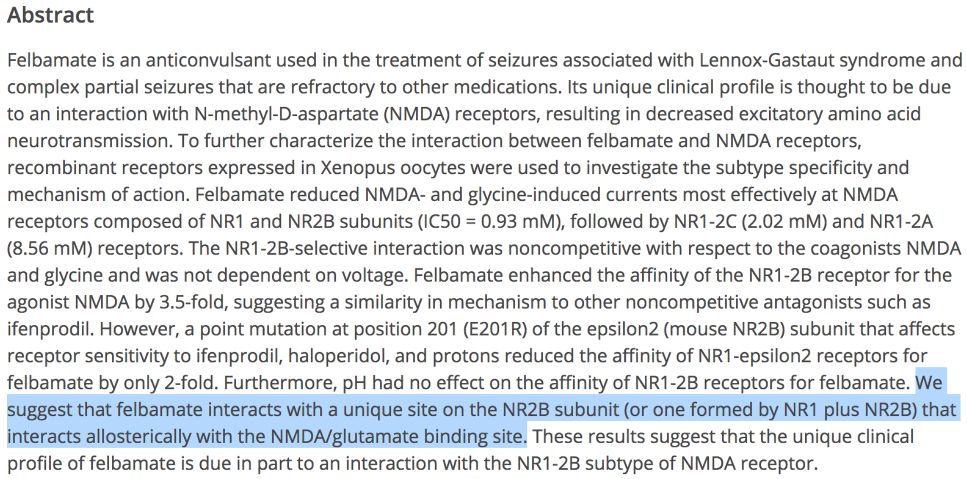
For example, the 10th line in ChemBl:



| 1 | target_pref_name | target_synonyms | drug_pref_name | drug_synonyms | ref_type (mechanism_refs) | ref_url (mechanism_refs) | action_type |
|---|---|---|---|---|---|---|---|
| 107078 | Glutamate [NMDA] receptor | NR2B | FELBAMATE | Felbamate | PubMed | http://europepmc.org/abstract/MED/10215667 | ANTAGONIST |

Figure 2: The 10th record in ChemBl

Step 1: We scrape text data from the given reference url via PubMed’s offical API, and then extract sentences of interest.



**Abstract**

Felbamate is an anticonvulsant used in the treatment of seizures associated with Lennox-Gastaut syndrome and complex partial seizures that are refractory to other medications. Its unique clinical profile is thought to be due to an interaction with N-methyl-D-aspartate (NMDA) receptors, resulting in decreased excitatory amino acid neurotransmission. To further characterize the interaction between felbamate and NMDA receptors, recombinant receptors expressed in Xenopus oocytes were used to investigate the subtype specificity and mechanism of action. Felbamate reduced NMDA- and glycine-induced currents most effectively at NMDA receptors composed of NR1 and NR2B subunits (IC50 = 0.93 mM), followed by NR1-2C (2.02 mM) and NR1-2A (8.56 mM) receptors. The NR1-2B-selective interaction was noncompetitive with respect to the coagonists NMDA and glycine and was not dependent on voltage. Felbamate enhanced the affinity of the NR1-2B receptor for the agonist NMDA by 3.5-fold, suggesting a similarity in mechanism to other noncompetitive antagonists such as ifenprodil. However, a point mutation at position 201 (E201R) of the epsilon2 (mouse NR2B) subunit that affects receptor sensitivity to ifenprodil, haloperidol, and protons reduced the affinity of NR1-epsilon2 receptors for felbamate by only 2-fold. Furthermore, pH had no effect on the affinity of NR1-2B receptors for felbamate. We suggest that felbamate interacts with a unique site on the NR2B subunit (or one formed by NR1 plus NR2B) that interacts allosterically with the NMDA/glutamate binding site. These results suggest that the unique clinical profile of felbamate is due in part to an interaction with the NR1-2B subtype of NMDA receptor.

Figure 3: Text contect on http://europepmc.org/abstract/MED/10215667

Step 2: We apply dependency parsing on the sentences with an aim to extracting predicates. Following the method mentioned above, words on the backtracking paths are extracted.

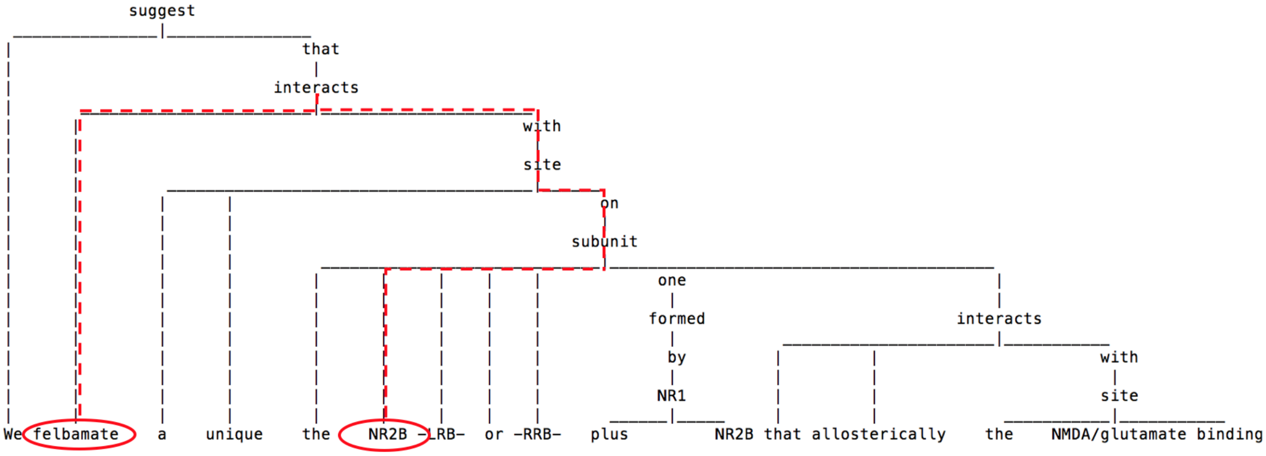


Figure 4: Dependency tree

### 3.1.2 "Other Predicates" Extraction

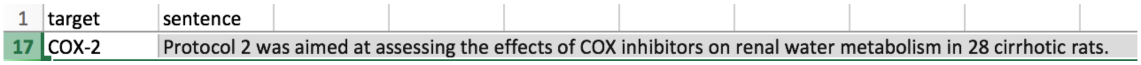
1. From ChemBL listed publications, extract sentences containing corresponding targets

2. Filter the list of sentences obtain in the previous step. We remove those sentences which contain any of the drug names in ChemBL's drug name list.

3. Apply dependency parser to the filtered list of sentences.

   The general strategy is to trace upwards from the target in the dependency tree until a node (word) whose POS tagging is "VERB" is reached. This method remains a lot to be desirable, as it only works when the target is (part of) an object in the sentence.

For example, the 17th extracted sentence



| 1 | target | sentence |
|---|---|---|
| 17 | COX-2 | Protocol 2 was aimed at assessing the effects of COX inhibitors on renal water metabolism in 28 cirrhotic rats. |

Figure 5: The 17th extracted sentence containing the target only

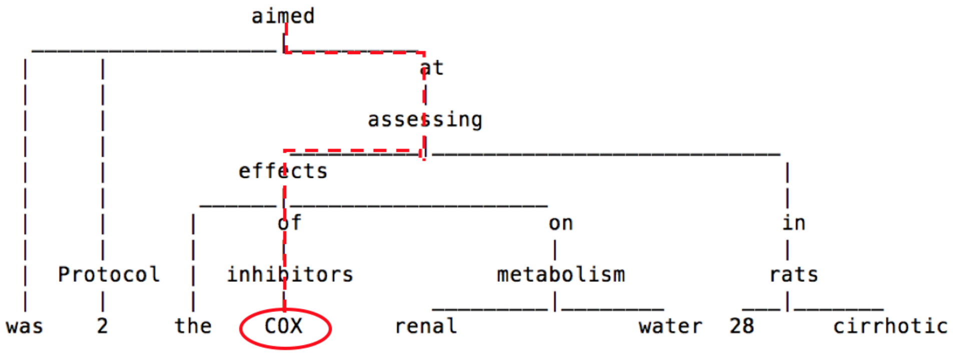


Figure 6: Dependency tree

### 3.2 Training (Completed)

We train a predicate classifier which differentiates “Good Predicates” from “Other Predicates” based on Hierarchical Attention Network (Yang et al., NAACL 2016)[2].

The attention-based RNN classifier consists of several parts: a word sequence encoder (using a pre-trained word embedding and a bidirectional GRU), a word attention layer and a softmax function. Building this classifier is basically for the purpose of reducing false positive errors when we test or generalize the model, and finding the optimum embedding representation of a predicate (as the word attention layer could identify the most informative words from a predicate).

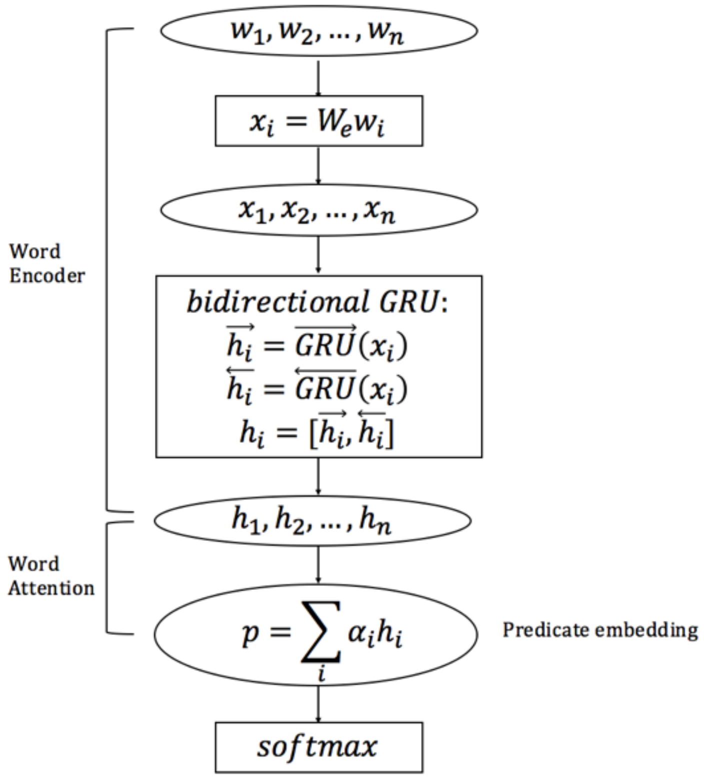

$w_1, w_2, \dots, w_n$
$x_i = W_e w_i$
$x_1, x_2, \dots, x_n$
Word Encoder
$bidirectional\ GRU$:
$\overrightarrow{h_i} = \overrightarrow{GRU}(x_i)$
$\overleftarrow{h_i} = \overleftarrow{GRU}(x_i)$
$h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]$
$h_1, h_2, \dots, h_n$
Word Attention
$p = \sum_i \alpha_i h_i$
Predicate embedding
$softmax$

Figure 7: Attention-based predicate classifier

[2] https://www.cs.cmu.edu/ diyiy/docs/naacl16.pdf

### 3.3 Testing

1. Locate the target name at a sentence level on test set

2. If the predicate in the testing sentence is determined as a “Good Predicate” by the classifier, the other noun phrase in the testing sentence will be extracted as a new drug name which acts on the given target.

## 4 Results

### 4.1 Data sets

Considering records from PubMed only, we obtain nearly 4400 predicates (around 1600 “Good Predicates” and 2800 “Other Predicates”). The ratio of training set to testing set is 7 : 3, and 20% of training set is reserved as validation set.

### 4.2 Baselines

We compare the attention-based RNN classifier with several baseline methods, including fastText model (Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T., 2016)[3], unidirectional RNN model, bidirectional RNN model without attention layer. Additionally, note that the base rate is 63%.

#### 4.2.1 Linear methods

**FastText** model represents each predicate by averaging the word-embedding vectors.

#### 4.2.2 RNN models without attention

**Unidirectional RNN** model summarizes information of predicates from only one direction. We feed the last hidden vector to a softmax function.

[3] https://arxiv.org/abs/1607.01759

| Model | training acc | validation acc | testing acc |
|---|---|---|---|
| fastText | 84.2 | 71.1 | 72 |
| uniGRU | 96.7 | 72.4 | 72 |
| uniLSTM | 96.0 | 73.4 | 73 |
| biGRU | 96.2 | 74.8 | 74 |
| biLSTM | 97.3 | 78.5 | 75 |
| Attn biGRU | 97.2 | 77.6 | 75 |
| Attn biLSTM | 96.2 | 80.2 | 78 |

Figure 8: Accuracy [%] on training, validation and testing set with the same parameters

**Bidirectional RNN without attention** model gets annotations of words by reading each predicate both forwards and backwards. The word annotations are then averaged into a predicate representation without weights.

# 5 Discussion

The testing accuracy of the classifer still remains to be raised. Basically we could improve performance of the classifer in the following three aspects: (1) set a higher bar for filtering sentences containing drug names (Step 2 in 3.1.2), as there exist quite a few false negative examples in our current predicate dataset; (2) refine the method of extracting "Other Predicates"; (3) leverage all records from ChEMBL, as now we use only PubMed data.

Once the classifier is built, we will deploy it on the testing document. The result will be compared with other NER methods. One important observation is that ChEMBL only collect a subset of all eligible target-drug pairs from their list of reference documents, which means on the testing document, precision is hardly to be calculated. So therefore, out final NER could only be evaluated by recall.