Assignment 3
Colorado CSCI/LING 5832
Named Entity Recognition

Group 5

Hunar Jain
huja4688@coloardo.edu

Srinivas Akhil Mallela
srma3452@colorado.edu

Aravind Bisegowda Srinivas
arbi8560@colorado.edu

November 30, 2021

Introduction

In this assignment we attempt to identify references to all genes in a set of abstracts of biomedical journal articles. The training set provided consists of sentences with genes tagged with IOB tags. We have attempted to solve the problem using 2 supervised machine learning approaches, i.e, HMM-based approach and a feature based classification approach using Conditional Random Fields (CRFs).

Methods

HMM-based approach

The training data is processed and a *Dataset* class is created. The training data is randomly split into test and training sets with a 80:20 split. We use the *Pomegranate* library to build our HMM tagger.

The transitional probabilities are based on the bigram counts of the tags and the observation probabilities are based on the unigram counts of the tags.
We use the Laplace smoothing technique to handle the cases when the observation probabilities are 0.

```
for word in data.training_set.vocab:
    for tag in data.training_set.tagset:
        if word not in count_tag_and_word[tag]:
            count_tag_and_word[tag][word] = 0
```

```
for tag, word_dict in count_tag_and_word.items():
    p_words_given_tag_state = defaultdict(float)

    for word in word_dict.keys():
        p_words_given_tag_state[word] = (count_tag_and_word[tag][word] + 1) / (tag_unigrams
 represents P(word|tag)

    emission = DiscreteDistribution(dict(p_words_given_tag_state))
    states[tag] = State(emission, name=tag)
```

We used unique word counts in the training set to establish observations probabilities for the test set.

The following results were observed for the HMM-based approach.

```
hmm score : 3242  entities in gold standard.
1531  total entities found.
913  of which were correct.
         Precision:  0.5963422599608099
         Recall:  0.28161628624305984
         F1-measure:  0.38256861512675466
```

CRF based approach

We process the data similar to the HMM based approach. We define a `word2features` function that takes a sentence as an input and extracts the features.

- Last 3 characters of a word
- Last 2 characters of a word
- Uppercase words
- Lowercase words
- CamelCase words
- Alpha-numeric words
- Stopwords
- Words ending with ASE
- Words ending with IN
- Words ending with ENE
- Word length
- Presence of digits in a word.

- BOS, if word is at the beginning of sentence.
- EOS, if word is at the end of sentence.
- True if unique in set I.
- True if unique in set O.
- True if unique in set B.

For the last 3 features stated above, we determine the unique word by keeping a track of all the unique words for each tag.

We also keep the following contexts of the previous and next word for every word in the sentence.

- Lowercase word.
- Title word.
- Uppercase word.

We make use of the `sklearn_crfsuite` library to train our CRF model. The training algorithm used is Gradient descent using the L-BFGS method with following training training parameters.

```
crf = sklearn_crfsuite.CRF(
    algorithm='lbfgs',
    c1=0.1,
    c2=0.1,
    max_iterations=1000,
    all_possible_transitions=False
)
```

The following results were observed for the CRF based approach.

```
crf score : 3272  entities in gold standard.
3001  total entities found.
2204  of which were correct.
Precision:  0.7344218593802065
Recall:  0.6735941320293398
F1-measure:  0.7026940857643871
```

`c1` and `c2` are the regularization parameters.

The top 20 features selected by the model and their associated weights and tags are listed below.

10.501352	B	BOS
7.847831	O	word.lower():release
7.548619	O	EOS
5.790113	O	word.lower():increase
5.284223	B	word[-2:]:1p
5.171570	O	word.lower():phase
5.077503	O	word.lower():contains
5.009911	B	word.lower():histone
4.874732	B	word.lower():fibrinogen
4.863599	O	word.lower():disease
4.683655	O	word.lower():min
4.597040	O	-1:word.lower():cdc28
4.577567	O	word.lower():within
4.459559	O	word[-2:]:nd
4.436698	I	word.lower():sites
4.388552	O	word.lower():kinase
4.324877	O	word.lower():decrease
4.310799	B	word.lower():osteocalcin
4.274574	O	word.lower():t1
4.248954	B	word.lower():engrailed

The normalized transition scores are as follows :-

	B	I	O
B	0	3.25	-7.69
I	0	4.81	-7.03
O	11.49	0	3.199

The token-level results for CRF based approach are listed below :-

	Precision	**Recall**	**F1-Measure**
B	0.82	0.75	0.78
I	0.81	0.75	0.78
O	0.98	0.98	0.98

Results

Running the evaluation script on the outputs of each of the approaches gave us the following results. We observed a better F1-measure with the CRF based approach on a 80:20 randomized test-train split.

	Precision	**Recall**	**F1-Measure**
HMM	0.59	0.28	0.38
CRF	0.73	0.67	0.70

Resources

https://towardsdatascience.com/named-entity-recognition-and-classification-with-scikit-learn-f05372f07ba2