Supplementary material of Fine-grained Entity Recognition with Reduced False Negatives and Large Type Coverage

Abhishek ABHISHEK.ABHISHEK@IITG.AC.IN
Indian Institute of Technology Guwahati,
Guwahati, Assam, India

Sanya Bathla Taneja* SANYABT11@GMAIL.COM
Garima Malik* ANNU.2353@GMAIL.COM
Indira Gandhi Delhi Technical University for Women,
Kashmere Gate, Delhi, India

Ashish Anand ANAND.ASHISH@IITG.AC.IN
Amit Awekar AWEKAR@IITG.AC.IN
Indian Institute of Technology Guwahati,
Guwahati, Assam, India

1. Model Evaluation, Training and Hyper-parameters

1.1 Model Evaluation and Training

Performance of the trained models were evaluated on the FIGER corpus and 1k-WFB-g corpus. Since FIGER corpus is quite small we use the complete corpus for testing. We split 1k-WFB-g into two parts with similar type distribution. We use one as the development set and the other as testing. We train each learning model five times and report the best result. To ensure fairness, the absolute number of sentence (for FgED) and entity mentions (for FgEC) processed by any model remains similar during training.

1.2 Model Hyper-parameters:

All the models used the following hyper-parameters:[1]
FgED model: Bi-directional LSTM hidden-layer: 100, learning rate: 0.001, character embedding: 50.
FgEC model: Bi-directional LSTM hidden-layer: 200, learning rate: 0.002, character embedding: 200, character LSTM hidden-layer: 50.
Common parameters: Dropout: 0.5, batch size: 500. We use 300 dimensional, 42B pre-trained word embeddings shared by [Pennington et al., 2014]. These were not updated during model training.

∗. The authors contributed to the work during their internship at the Indian Institute of Technology Guwahati.
1. We did not conduct a hyper-parameter search for any dataset, same parameter were used irrespective of dataset. The selected values are quite standard.

2. Analysis of discarded and retained sentences

In this section, we analysed the discarded and retained sentences from the HAnDS framework on the following parameters:

1. Lengths of the discarded sentences: whether the discarded sentences were longer on average?

2. Lengths of entity mentions: whether the entity mentions in the discarded sentences longer on average?

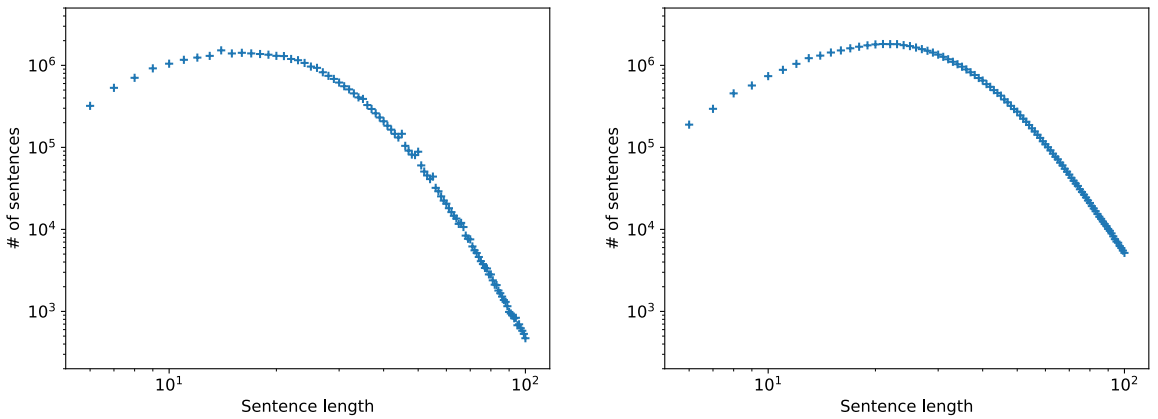
3. Distribution of token and entity mention: whether there is a fundamental change in the token and entity mention distribution of discarded and retained sentences?

This analysis is done while generating the WikiFbF dataset. The number of sentences in retained corpus is 31.92 million whereas the number of sentences in the discarded corpus is 50.33 millions.

2.1 Sentence length distribution analysis

Figure 1 illustrates the sentence length distribution among the discarded and retained sentences. We observe that sentence shorter than 6 tokens and greater than 100 tokens were mostly caused by errors in sentence segmentation. Thus we have plotted the distribution for the sentences in between 6 and 100 tokens.

In Figure 1 we can observe that the discarded sentences are longer in length. The mean length for discarded sentences is 27.29 whereas the mean length for retained sentences is 21.63.



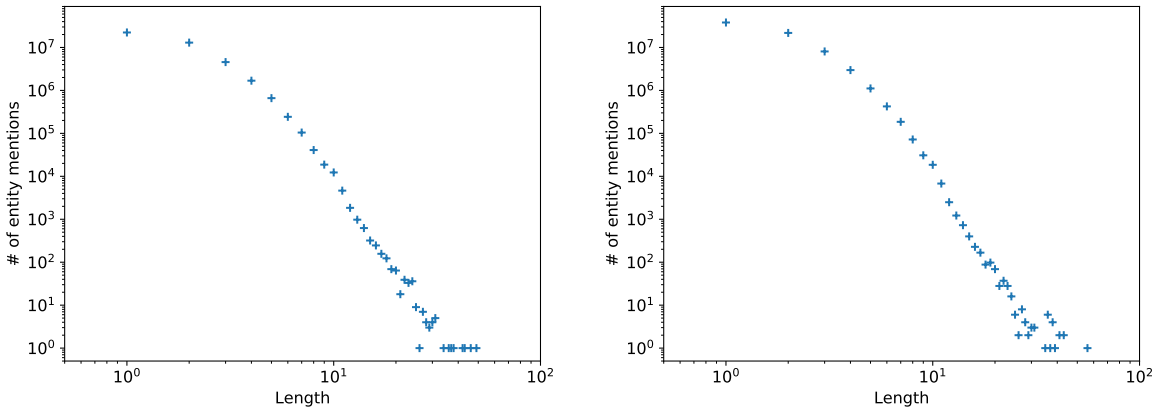
(a) Sentence length distribution in the retained sentences.

(b) Sentence length distribution in the discarded sentences.

Figure 1: Distribution of sentences of length between 6 and 100 on a log-log plot.

2.2 Entity length distribution analysis

Figure 2 illustrates the entity length distribution among the discarded and retained sentences. We can observe that there is no notable difference between these two plots. In both these corpus there are about 10k entity mentions with length 10 tokens.



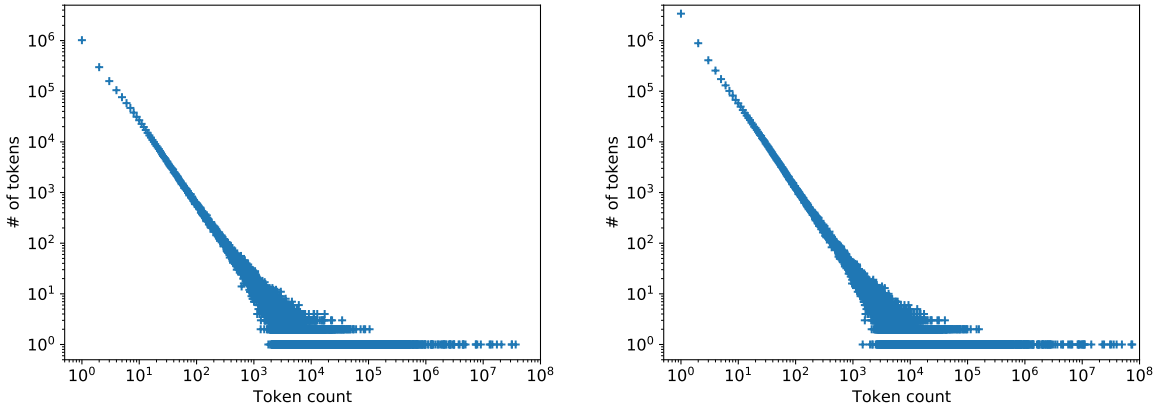
(a) Entity length distribution in the retained sentences.

(b) Entity length distribution in the discarded sentences.

Figure 2: Entity length analysis on log-log scale.

2.3 Token distribution analysis

Figure 3 illustrates the token distribution among the discarded and retained sentences. We can observe that other than a slight change in slope and absolute magnitude, there is no notable difference between these two plots. This can be attributed to the fact that the retained corpus has 31.92 million sentences and the discarded corpus has 50.33 million sentences.



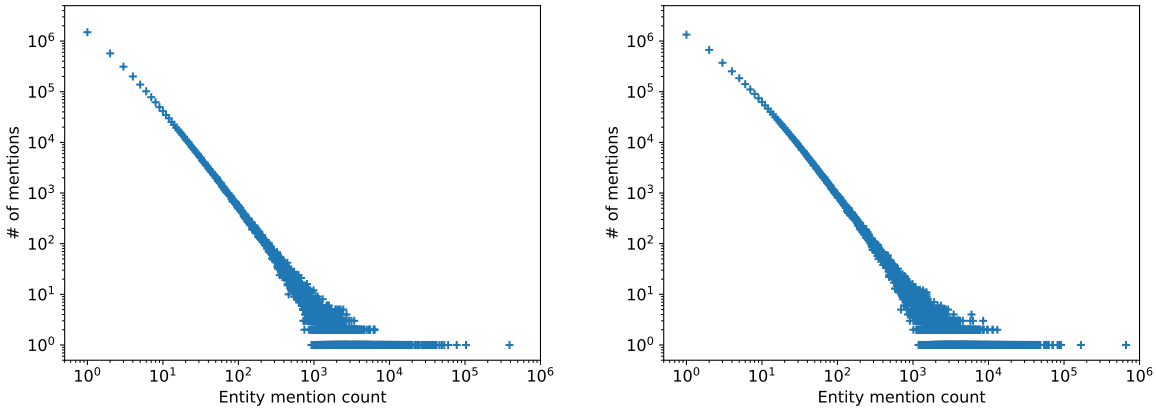
(a) Token distribution in the retained sentences.

(b) Token distribution in the discarded sentences.

Figure 3: Token distribution analysis on log-log scale.

2.4 Entity mention distribution analysis

Figure 4 illustrates the entity mention distribution among the discarded and retained sentences. We can observe that there is no notable difference between these two plots.



(a) Entity mention distribution in the retained sentences.

(b) Entity mention distribution in the discarded sentences.

Figure 4: Entity mention distribution analysis on log-log scale.

2.5 Sentence length distribution comparison across multiple datasets

In Figure 5 and Figure 6 we compare the sentence length distribution of the retained and discarded sentences with five NER datasets namely CoNLL [Tjong Kim Sang and De Meulder, 2003], OntoNotes [Weischedel et al., 2013], BBN [Weischedel and Brunstein, 2005], CADEC [Karimi et al., 2015] and BC5CDR [Li et al., 2016]. These datasets have different writing styles as mentioned below:

1. CoNLL dataset: Sentences sampled from news articles in Reuters corpus.
2. OntoNotes dataset: Sentences sampled from news, conversation text, broadcast conversation and weblogs.
3. BBN dataset: Sentences sampled from news article in Wall Street Journal.
4. CADEC dataset: Sentences sampled from a medical forum discussion reated to adverse drug reaction.
5. BC5CDR dataset: Sentences sampled from clinical abstracts.

In Figure 5 there is no restriction on sentence length. We can observe that the most notable distribution change occurs at either shorter sentences or longer sentences.

In Figure 6 we only consider sentence whose length are in between 6 and 100. Here we can observe that the distribution of sentence length in five NER datasets are close to the retained sentence length distribution.

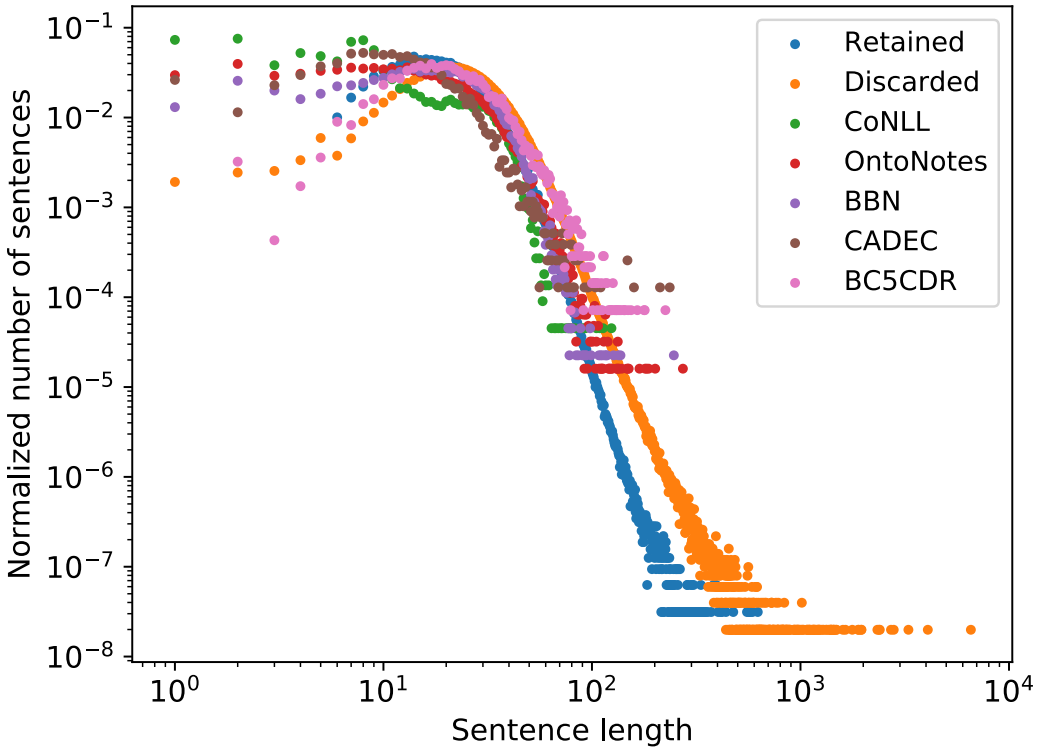


Figure 5: Distribution of sentence length compared across five NER datasets with the retained and discarded sentences.

References

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81, 2015.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016, 2016.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Erik F Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics, 2003.

Ralph Weischedel and Ada Brunstein. Bbn pronoun coreference and entity type corpus. *Linguistic Data Consortium, Philadelphia*, 112, 2005.

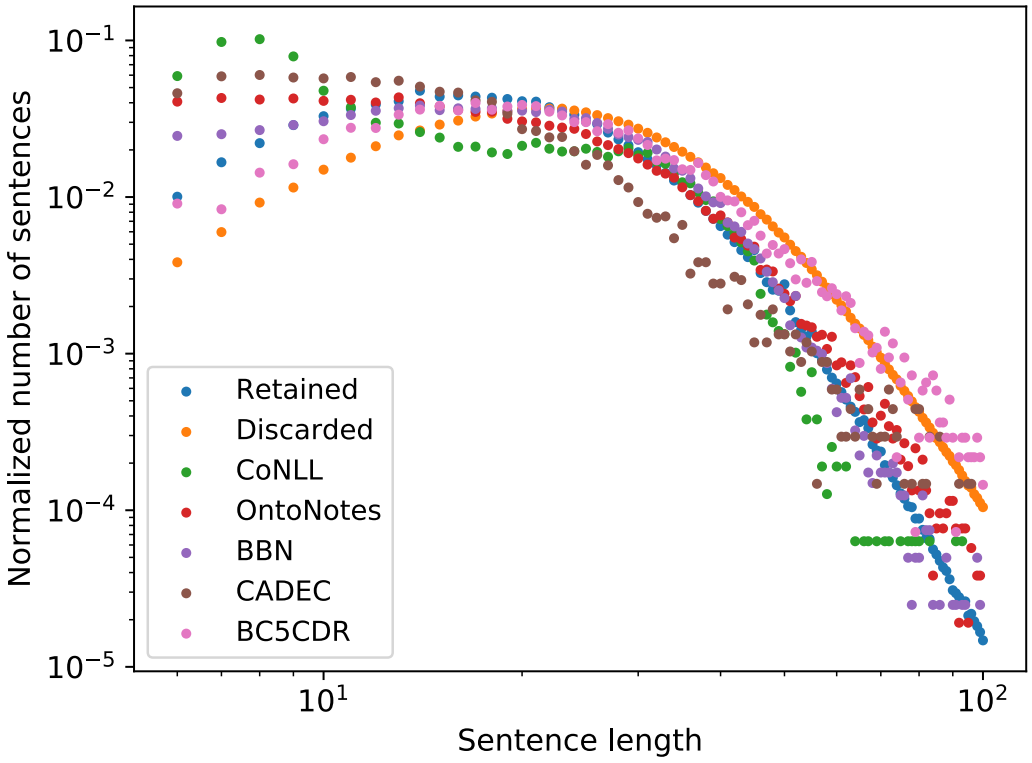


Figure 6: Distribution of sentence length between 6 to 100 compared across five NER datasets with the retained and discarded sentences.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 2013.