Transformer Model Performance on Diversified Named Entity Recognition Datasets

Pierce Coggins, Bhuvnesh Sharma
W266: Natural Language Processing
UC Berkeley School of Information
pierce.coggins@ischool.berkeley.edu, bhuvneshsharma@berkeley.edu

ABSTRACT

This project will employ the HuggingFace Transformers framework and Simple Transformers as the basis for developing a Transformer-based model for NER. Most transformer models have only been tested against the primary CoNLL-2003 NER dataset and not other more complex NER datasets. In this project we look to reconfigure and test the BERT pre-trained base model against more recent NER benchmarks that better represent the diversity of modern day NER tasks.

1. INTRODUCTION

Named Entity Recognition (NER) [3] aims to recognize mentions of rigid designators from text belonging to predefined semantic types such as person, location, organization etc. NER not only acts as a standalone tool for information extraction (IE), but also plays an essential role in a variety of natural language processing (NLP) applications such as information retrieval , automatic text summarization , question answering , machine translation , and knowledge base construction.

Named Entity Recognition [3] acts as an important preprocessing step for a variety of downstream applications such as semantic search. Semantic search refers to a collection of techniques, which enable search engines to understand the concepts, meaning, and intent behind user queries[21]. According to Guo et al., about 71% of search queries contain at least one named entity. Recognizing named entities in search queries would help us to better understand user intents, hence to provide better search results.[22] To incorporate named entities in search, entity-based language models [21], which consider individual terms as well as term sequences that have been annotated as entities (both in documents and in queries), have been proposed by Raviv et al. [23]. There are also studies utilizing named entities to enhance user experience, such as query recommendation [24], query autocompletion [25], [26] and entity cards [27],[28].

1.1. Evolution of NER

The term "Named Entity" (NE) was first used at the sixth Message Understanding Conference (MUC-6), as the task of identifying names of organizations, people and geographic locations in text, as well as currency, time and percentage expressions. Since MUC6 there has been increasing interest in NER, and various scientific events (e.g., CoNLL03, ACE , IREX, and TREC Entity Track) devote much effort to this topic.

A basic definition of a named entities is "a proper noun, serving as a name for something or someone".[14] This varies significantly across disciplines, but the fundamental multi-classification problem remain the same. Over the past decade, various models have set the state-of-the-art performance in NER, including LSTM and CNN models. However, the emergence of the transformer model triggered a paradigm shift across various NLP tasks, including Named Entity Recognition (NER). However, majority of NER research is conducted against the CoNLL-2003 dataset: the question is whether or not these models are able to generalize to more domain specific corpora.

Our goal in this paper is to quickly outline and ultimately build upon the existing research around the generalization of the transformer-based NER model performance against existing domain specific benchmarks such as the GENIA medical NER dataset[29] and the W-NUT 2017 Emerging Entities dataset.

2. PROJECT OVERVIEW

2.1. Datasets Used

High quality annotations are critical for both model learning and evaluation**.** For this project, we are planning to use the following is datasets:

- CoNLL-2003 (English) [7][8]
- WNUT 2017 Emerging Entities task [1]
- GENIA [29]

2.2. CoNLL-2003 (English) [7][8]

The CoNLL-2003 dataset set a standard amongst the NER research community by generating the first large publicly available NER dataset. The dataset contains 4 entities (persons, organizations, locations and miscellaneous), and the data was taken from the Reuters Corpus2 . This corpus consists of Reuters news stories between August 1996 and August 1997. The following tables summarize the size and distribution of the corpus:

English data	Articles	Sentences	Tokens
Training set	946	14,987	203,621
Development set	216	3,466	51,362
Test set	231	3,684	46,435

Table 1: Number of articles, sentences and tokens in each data.

English data	LOC	MISC	ORG	PER
Training set	7140	3438	6321	6600
Development set	1837	922	1341	1842
Test set	1668	702	1661	1617

Table2: Number of named entities per data file

2.2.1. CoNLL03 - Data Format [7][8]

The CoNLL-2003 shared task data files contain four columns separated by a single space. Each word has been put on a separate line and there is an empty line after each sentence. The first item on each line is a word, the second a part-of-speech (POS) tag, the third a syntactic chunk tag and the fourth the named entity tag. Here is an example:

U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

2.3. W-NUT 2017 Emerging Entities task[1]

The W-NUT 2017 Emerging Entities task [1] operates over a wide range of user-generated text and focuses on generalisation to complex user slang and text types (e.g. emojis). Scores are given both over entity chunk instances, and unique entity surface forms, to normalise the biasing impact of entities that occur frequently.

2.3.1. W-NUT 2017 - Data Format [1]

Based on this dataset, annotations include the following entity types:

person – Names of people (e.g. Virginia Wade).
location – Names that are locations (e.g. France). Fictional locations can be included
corporation – Names of corporations (e.g.Google).
product – Name of products (e.g. iPhone).
creative-work – Names of creative works (e.g. Bohemian Rhapsody).
group – Names of groups (e.g. Nirvana, San Diego Padres).

Metric	Dev	Test
Documents	1,008	1,287
Tokens	15,734	23,394
Entities	835	1,040
person	470	414
location	74	139
corporation	34	70
product	114	127
creative-work	104	140
group	39	150

Table3: The emerging entity dataset statistics

2.4. GENIA Medical Dataset[29]

The GENIA [29] Corpus covers various medical journals and clinical texts, and has annotated a subset of the substances and the biological locations involved in reactions of proteins, based on the GENIA ontology of the biological domain.The base abstracts and sentences are selected from the search results of keywords “Human”, “Blood”, "Cells” and “Transcription Factors”.

3. Background & Methodology

3.1. NER Evaluation Metrics [3]

NER systems are usually evaluated by comparing their outputs against human annotations. The comparison can be quantified by either exact-match or relaxed match. In this paper we will be using exact-match evaluation.

3.1.1. Exact-match Evaluation [3]

NER involves identifying both entity boundaries and entity types. With “exact-match evaluation”, a named entity is considered correctly recognized only if both boundaries and type match ground truth. Precision, Recall, and F-score are computed on the number of true positives (TP), false positives (FP), and false negatives (FN). Precision measures the ability of a NER system to present correct entities, Recall measures the ability of a NER system to recognize all entities in a corpus.

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

F-score is the harmonic mean of precision and recall, and the balanced F-score is most commonly used:

$$\text{F-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Huggingface.co transformer [30]

Huggingface Transformers [30] (formerly pytorch-transformers) provide state-of-the-art general-purpose architectures (*BERT, GPT-2, RoBERTa, XLM, DistilBert, XLNet, CTRL...*) for Natural Language Understanding (NLU) and Natural Language Generation (NLG) with over 32+ pretrained models in 100+ languages and deep interoperability between TensorFlow 2.0 and PyTorch.

3.2. Simple Transformer

The Simple Transformers library was conceived to make Transformer models easy to use. Simple Transformers enabled the application of Transformer models to Sequence Classification tasks. In this paper we utilized Simple Transformers NER model to simplify the instantiation of our BERT-Base model.

4. MODEL ARCHITECTURE

The objective of this paper is to assess how generalizable the BERT-Base model is across various NER datasets. Although we haven’t made any modifications to the basic BERT-Base model architecture depicted in Fig. 2, it’s worth mentioning that it was derived from the multi-head attention transformer model (see: Fig. 1) popularized by Vaswani et al. [17]

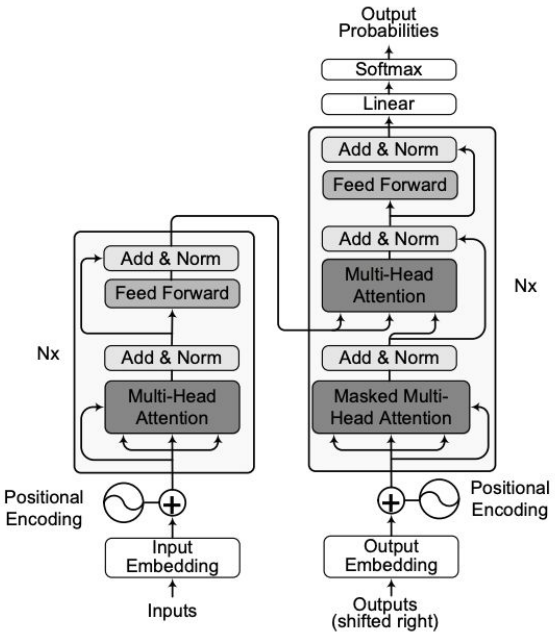

Fig 1: Multi-Head Attention Transformer model

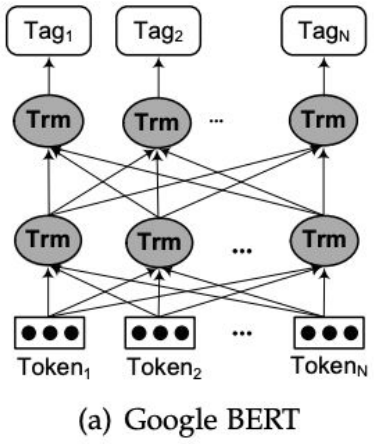

Fig 2: Google BERT Architecture

5. RESULTS

5.1. CoNLL-2003 Baseline

For the CoNLL-2003 dataset we were able to reproduce results similar to those generated by Devlin et al. using the BERT-base model. Without extensive training on NER data, we were able to achieve the following results:

Eval Metric	Precision	Recall	F1 Score
Test Results	90.113	91.407	90.755

This F-1 score is just 1.6 points below the performance of Devlin et al. using the BERT-Base model, and just 2.75 points below the current state-of-the-art set by Edonuv et al.[31] As a reminder, the goal of our project is not to outperform the state-of-the-art, but instead to test the generalizability of the pre-trained BERT-Base model. This performance gives us confidence that our model performs similarly to the model proposed in Google's original paper. To reiterate, the CoNLL-2003 dataset only contains four named entity types, and is derived from news data exclusively. Although the model performs well against this benchmark, there are more complex NER datasets that better represent modern day NER tasks.

With this baseline established, we look to test this model against more diverse datasets to assess whether or not the BERT base model will generalize to more complicated NER datasets.

5.2. W-NUT Performance

After a literary review it was evident that transformer models had not been used often on the W-NUT dataset. Thus, in this section we will look to compare the performance of our BERT-Base model against the state-of-the-art for this specific dataset. Our results are as follows:

Eval Metric	Precision	Recall	F1 Score
Test Results	35.52	38.12	37.01

Relative to the CoNLL dataset, our results are underwhelming. However, the current state-of-the-art performance on the W-NUT dataset is a F-1 Score of 49.59 set by Akbik et al. Thus, our performance is in-line with expectations despite being lower than the best-in-class model by F1 of 12.58 or ~ 25%. In this case, it seems the BERT-Base model did not perform well against the noisy user-generated W-NUT dataset. This indicates that the BERT-Base model may need more slang specific pre-training to perform anywhere close to the standard set on this dataset.

5.3. GENIA Performance

Much like the W-NUT dataset, the GENIA medical NER dataset has not often been tested using the BERT-Base model. We will assess performance against the standard set by the medical research community. Our performance is as follows:

Eval Metric	Precision	Recall	F1 Score
Test Results	48.59	64.18	55.30

The previous best performance against the GENIA dataset is reported to be (recall: 76.0 | precision: 69.4 | F1: 72.6).[32] Relative to this standard, our BERT-Base model performed well against precision but much worse against F1 and recall. This indicates with additional training the BERT-Base model may be generalizable enough to be trained on medical related NER datasets.

5.4. Overall Performance

Results against training and test sets are collated below:

	Training Data			
Milestone	**Loss**	**Precision**	**Recall**	**F1**
CoNLL03	**0.0226**	97.10	97.04	**97.07**
W-NUT	**0.0221**	40.32	41.18	**40.82**
GENIA	**0.1983**	64.96	74.90	**69.57**

	Test Data			
Milestone	**Loss**	**Precision**	**Recall**	**F1**
CoNLL03	0.1085	90.11	91.40	**90.75**
W-NUT	0.0759	35.52	38.12	**37.01**
GENIA	0.2804	48.58	64.18	**55.30**

Overall, it seems the BERT base model did not generalize well to the user-generated text of the W-NUT dataset. This dataset includes complex text such as emoji's, hyperlinks and slang terms that were likely not commonly found in the BERT pre-training dataset . However, our model performed much better against the GENIA medical dataset as these terms likely occur in common datasets such as medical journals and wikipedia.

6. CONCLUSION

As an overview, we developed a transformer-based model using HuggingFace Transformers and Simple Transformer packages that performed markedly similar to the original BERT-Base model on the CoNLL-2003 dataset.

After establishing that our model performed to a similar standard, the model was tested against the W-NUT Emerging Entities dataset, which contains noisy user-generated text for various online sources, and the GENIA medical NER dataset containing domain specific medical text sourced from biology and clinical texts. From this assessment, this project demonstrated that the BERT-Base model retains strong generalizable properties within datasets that contain proper english language. However, when tested against the noisier W-NUT dataset, our model was not able to perform anywhere near the sort of level set by more specialized models.

Future research on the generalizability of the BERT-Base model should include an assessment of performance against a set of domain specific dataset instead of basing this on the performance against one fundamental dataset. Additionally, we would recommend for those interested in researching the generalizability of BERT NER models to test their model performance against a wider range of publicly available dataset. Our hope is that our findings would be validated by further research.

We also found in completing this research that the NER public datasets are all based upon three fundamental generic corpora: news, wikipedia and medical (PubMed etc.). Given the state of publicly available datasets, we believe there is a lot of work to be done in order to assess NER performance across all the domains that realistically employ it. Given that NER is foundational to such a diverse set of broadly utilized NLP applications (i.e. information retrieval , automatic text summarization , question answering , etc.), generating a truly generalizable NER model could have a dramatic impact on all down-stream NLP applications

REFERENCES

[1] Leon Derczynski, Eric Nichols, Marieke van Erp, Nut Limsopatham, *Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition*

[2] Thilina Rajapakse , *Simple Transformers — Named Entity Recognition with Transformer Models*

[3] Jing Li, Aixin Sun, Jianglei Han, Chenliang Li,*A Survey on Deep Learning for Named Entity Recognition*

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova,*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*

[5] Jana Straková, Milan Straka, Jan Hajic,*Neural Architectures for Nested NER through Linearization*

[6] D. Nadeau and S. Sekine*, "A survey of named entity recognition and classification," Lingvisticae Investigationes, vol. 30, no. 1, pp. 3–26, 2007.*

[7] Erik F. Tjong Kim Sang and Fien De Meulder *, Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition*

[8] https://www.clips.uantwerpen.be/conll2003/ner/

[9] http://cemantix.org/data/ontonotes.html

[10] Sameer Pradhan , Alessandro Moschitti, Nianwen Xue , Hwee Tou Ng ,Anders Bjorkelund ¨ , *Olga Uryupina , Yuchen Zhang4 and Zhi Zhong5 , Towards Robust Linguistic Analysis Using OntoNotes*

[11] R. Grishman and B. Sundheim, *"Message understanding conference-6: A brief history," in Proc. COLING, vol. 1, 1996.*

[12] S. A. Kripke, "Naming and necessity," *in Semantics of natural language. Springer, 1972, pp. 253–355.*

[13] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang, *"Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes," in Proc. EMNLP, 2012, pp. 1–40.*

[14] V. Yadav and S. Bethard, *"A survey on recent advances in named entity recognition from deep*

learning models," in Proc. COLING, 2018, pp. 2145–2158.

[15] M. Collins and Y. Singer, *"Unsupervised models for named entity classification," in Proc. EMNLP, 1999.*

[16] K. Balog, "Entity-oriented search," 2018.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *"Attention is all you need," in Proc. NIPS, 2017, pp. 5998–6008.*

[18] N. Kitaev and D. Klein, *"Constituency parsing with a selfattentive encoder," in Proc. ACL, 2018, pp. 2675–2685.*

[19] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, *"Generating wikipedia by summarizing long sequences," arXiv preprint arXiv:1801.10198, 2018.*

[19] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, *"Improving language understanding by generative pre-training," Technical report, OpenAI, 2018.*

[20] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, *"Deep contextualized word representations," in Proc. NAACL-HLT, 2018, pp. 2227–2237.*

[21] K. Balog, *"Entity-oriented search," 2018.*

[22] J. Guo, G. Xu, X. Cheng, and H. Li, *"Named entity recognition in query," in Proc. SIGIR, 2009, pp. 267–274.*

[23] H. Raviv, O. Kurland, and D. Carmel, *"Document retrieval using entity-based language models," in Proc. SIGIR, 2016, pp. 65–74.*

[24] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna, *"The query-flow graph: model and applications," in Proc. CIKM, 2008, pp. 609–618.*

[25] F. Cai, M. De Rijke et al., *"A survey of query auto completion in information retrieval," Foundations and TrendsR in Information Retrieval, vol. 10, no. 4, pp. 273–363, 2016.*

[26] Z. Bar-Yossef and N. Kraus, *"Context-sensitive query autocompletion," in Proc. WWW, 2011, pp. 107–116.*

[27] G. Saldanha, O. Biran, K. McKeown, and A. Gliozzo, *"An entity focused approach to generating company descriptions," in Proc. ACL, vol. 2, 2016, pp. 243–248.*

[28] F. Hasibi, K. Balog, and S. E. Bratsberg, *"Dynamic factual summaries for entity cards," in Proc. SIGIR, 2017, pp. 773–782.*

[29] GENIAProject *http://www.geniaproject.org/shared-tasks/bionlp-jnlpba-shared-task-2004*

[30] https://medium.com/huggingface

[31] A. Baevski, S. Edunov, Y. Liu, L. Zettlemoyer, M. Auli, *Cloze-drive Pretraining of Self-attention Networks.*

[32] J. Kim, T. Ohta, Y. Tsuruoka, Y. Tatisi: *Introduction to the Bio-Entity Recognition Task at JNLPBA*