We consider the problem of Named Entity Recognition and experiment with three data augmentation techniques based on the WikiText-2 dataset with the aim of improving NER benchmark performance. We show improved performance on the WNUT dataset using a technique inspired by back-translation.

**ABSTRACT**

**PROBLEM STATEMENT**

**Background**

Kensho is a market data analytics system that can answer hundreds of million question combinations by scanning over 90,000 customizable actions.

**Product Charter**

Understand and predict financial markets using the wealth of text data on the internet

**Project Goal**

Improved Named Entity Recognition (NER) by incorporating knowledge bases

1. Kensho is interested in improving NER benchmarks
2. NER: retrieve information from text by identifying token categories
3. Important in finance due to the need for unambiguous information

**NER Example**

Apple prices have outpaced Facebook...

Apple prices are killing the jam market.

**Conclusions**

Apple is likely a reference to the fruit.

Context matters!

**Our Datasets**

- CoNLL 2003
- WNUT 2017
- Wikidata Knowledge Base
- Text Corpus: WikiData Knowledge
  - WikiData 1.2
  - WikiData 2

**Scope of Work**

1. Task: Build an NER system which utilizes the WikiData Knowledge Graph to improve performance on a benchmark test set, when compared with existing NER approaches.
2. Training Dataset: WikiData (for the Knowledge Graph) + NER Training Dataset from one or both of (1) WNUT 2017 Emerging Entities ("WNUT") and (2) CoNLL 2003 ("CoNLL").
3. Benchmark Test Set: Test sets created out of one or both of (1) WNUT, and (2) CoNLL.
4. Evaluation: F1 score

**VOCABULARIES**

- Lists of words for each entity type used in train models
- CoNLL 2003: the canonical NER benchmark
  - 10,000 sentences
  - 20,285 (train, test) pairs
  - |V| = 21,008 words
  - LOC, ORG, PER, MISC + non-labeled words

**WikiData**

- Unambiguous:
  - We want to build a vocabulary we can rely on
  - Focus on unambiguous words (pertaining to only one entity vocabulary)
  - Don’t use heuristics to build a MISC vocabulary, stick to CoNLL definition

**RESULTS AND CHALLENGES**

- Get an average F1 score of 0.21, worse than random
- When we incorporate type priors, drops further to 0.14
- Problem: method does not scale to sparse, highly overlapping vocabularies
- Despite the name of the paper, the model is poorly suited to incorporating knowledge base data

**3. Fine-tune BERT multi-label classifier on text corpus**

Learn to classify each word in a corpus based on whether it is in each entity

**VOCABULARIES**

- Traditional (Bidirectional) Language Models
  - The company
  - Microsoft
  - Is located in Seattle
  - Data to train Model
  - Word Prediction Based on Context

**Knowledge Augmented Language Models**

- The company
  - Microsoft
  - Is located in Seattle
  - Lists of Entities
  - Word Based on Context and Type

**RESULTS AND CHALLENGES**

- Performance on CoNLL Test Set:
  - CoNLL Model
  - 0.650
  - WikiModel
  - 0.673

**4. Back-translation**

- Key issues with our other approaches: We don’t have contextual NER tags for WikiText-2!
  - Same tag for a label, regardless of context.
- New approach: inspired by “Back-translation”!
  - Used in Neural Machine Translation for language pairs where there is not enough training data.

**CONCLUSION**

**SUMMARY OF RESULTS**

1. Results from paper “Knowledge Augmented Language Models” are poor with large knowledge base vocabularies
2. Multi-label classification on top of BERT also does poorly since it encourages the model to ignore context

**WHAT WE’VE LEARNED**

1. Constructing sets of non-contextual NER tags (vocabs) from knowledge bases is extremely challenging
2. Over-reliance on non-contextual labels encourages the model to ignore context and degrades NER performance
3. Back-translation approaches don’t improve results in-sample but may improve generalization performance from CoNLL to WNUT