A Straightforward Approach to Zero Shot Learning for Partially Annotated Data

Ajay Anand
University of Pennsylvania
anandaj@seas.upenn.edu

Michelle Chien
University of Pennsylvania
chienm@seas.upenn.edu

Daniel Lee
University of Pennsylvania
dlee22@seas.upenn.edu

Isabel Navarro
University of Pennsylvania
ibn@seas.upenn.edu

Abstract

This paper aims to explore a straightforward approach to the task of zero shot learning using un-annotated or partially annotated data. In the cases of both the partially annotated as well an unlabelled data, there exist some classes in the test data which are never seen in the training data. This means that any classifier we use needs to be pretrained in order to make meaningful discriminations between labels that were not in the training data. The primary means chosen to achieve learning without the presence of any domain-specific training data particular to the OntoNotes dataset was to pre-train our BERT model using publicly available data from Wikipedia and WikiLinks. To minimize manual labelling, we assigned entities from WikiLinks to the given types, then pulled the associated sentences and labelled only the linked entities. This is a quick way to develop a weakly-labeled pre-training set using freely available data for Zero-shot NER.

1 Introduction

We have been tasked with assigning 18 Named Entity type labels (Table 2.) to the English Language OntoNotes 5.0 dataset, using training sets including 12, 6, or none of the labels. The most challenging part of this task is the Zero-Shot component, previously called "dataless classification" (Chang et al., 2008), which entails categorizing examples into labels that were not part of the training set. Luckily, the latter term is a misnomer, as most Zero-Shot methods depend on some form of outside data, such as Wikipedia, without which educated guesses at Zero-Shot labels could not be made (Chang et al., 2008).

2 Background Research

To carry out the given taken of making predictions from our model trained using either partially annotated or annotated data, we carried out a comparative study on several different machine learning algorithms, enumerated below:

2.1 Character Learning Models

1. "On the Strength of Character Language Models for Multilingual Named Entity Recognition" (Yu et al., 2018)
   This is an approach to Named Entity Identification, which can be considered a sub-part to the task at hand. Each word is split into characters and scored based on its likelihood that the letters in that sequence represent an entity.

2.2 Zero Shot Learning

1. "An embarrassingly simple approach to zero-shot learning(Romera-Paredes and Torr, 2015)

   Uses a general framework to carry out zero shot learning by studying the relationships between the different features, attributes, and classes to train a simple two layer linear network, in which the top layer does not get learned but is instead a product of the environment.

2. "Zero-Shot Open Entity Typing and Type-Compatible Grounding" (Zhou

et al., 2019)

This approach uses a type taxonomy of FreeBase "types" with references to typed Wikipedia entries found via WikiLinks to harvest sentence context data for the various types. Because we do not require the open typing ability of this system, we looked for a more compact method, which we found in BERT.

3. "Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach" (Yin et al., 2019)

   This work provided datasets to compare performance of our model with baseline results to determine the efficacy of our method. It also deal with the importance of Zero Shot learning and its wide applicability in real work scenarios.

4. "Importance of Semantic Representation: Dataless Classification" (Chang et al., 2008)

   In the field of machine learning text categorization had been traditionally studied as the problem of leaning from labelled data to train a classification model. Human Intelligence however is able to carry out classification tasks without the presence of any prior examples due to our implicit understanding of the meaning behind category names.The authors of this paper propose a learning protocol named "Dataless Classification" (Chang et al., 2008) that uses public world knowledge to induce classification without using any labelled data, i.e. by carrying out unsupervised training. The model is able to able to perform classification by interpreting a string of words as a group of semantic concepts, similar to human understanding. This model is able to have competitive performance to a supervised training algorithm that uses a 100 labelled training examples.

2.3 BERT based models

1. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" (Devlin et al., 2019)

   BERT, which stands for Bidirectional Encoder Representations from Transformers, was one of the most promising methods to carry out our given task. Unlike most language models, which are unable to deal well with unseen labels for classification tasks, BERT is designed to be pretrained on unlabelled text by being conditioned jointly on contextual information from both directions. Due to this pretraining BERT is able to be tuned to have better performance on specific tasks by the simple addition of an additional output layer to create accurate models for a large variety of tasks with being substantially modified for each task.

2. "Towards Lingua Franca Named Entity Recognition with BERT" (Moon et al., 2019)

   This work provided an interesting insight into multilingual classification models. The aim of the work was to help create a unified model which would be able to carry out Named Entity Recognition in multiple languages, a goal which is would help rectify the necessity of creating different models to carry out NER tasks in different languages.This solution would allow classification tasks to be carried out in many more languages by reducing the amount of training data required to allow the model to work on languages for which datasets are less widely available. The authors used a multilingual BERT, which was trained on dataset containing multiple languages simultaneous and it was found that this model post training had better accuracy than similar models which were trained on a single language. The model also performs reasonably well on unseen languages allowing NLP to be carried out in more real world applications.

3 Methods Explored

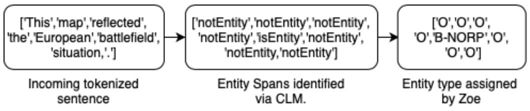
['This','map','reflected', 'the','European','battlefield', 'situation,'.']
['notEntity','notEntity','notEntity', 'notEntity','isEntity','notEntity', 'notEntity,'notEntity']
['O','O','O', 'O','B-NORP','O', 'O','O']
Incoming tokenized sentence
Entity Spans identified via CLM.
Entity type assigned by Zoe

Figure 1: Workflow for BERT-based Zero-shot NER.

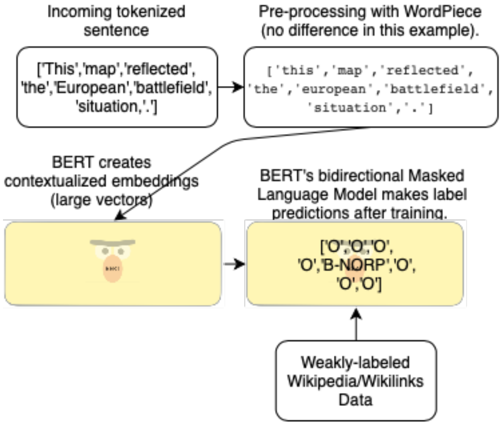
Incoming tokenized sentence
Pre-processing with WordPiece (no difference in this example).
['This','map','reflected', 'the','European','battlefield', 'situation,'.']
['this','map','reflected', 'the','european','battlefield', 'situation','.']
BERT creates contextualized embeddings (large vectors)
BERT's bidirectional Masked Language Model makes label predictions after training.
['O','O','O', 'O','B-NORP','O', 'O','O']
Weakly-labeled Wikipedia/Wikilinks Data

Figure 2: Workflow for BERT-based Zero-shot NER.

3.1 Character Language Models and ZOE

Our initial plan was to combine the results of Named Entity Identification using Character Language Models and and Zero-Shot Open Entity Typing (ZOE) (Figure 1). Due to time and resource constraints, we were only able to complete Named Entity Identification with CLM before returning to BERT which we had partially implemented for our progress report.

To do this we trained a character language model using SRILM, following one of the methods from On the Strength of Character Language Models for Multilingual Named Entity Recognition, since this had one of the better performances. Using the 18 gold training data, we compiled two lists — one of all the entities in the training data and the other of all the non-entities in the training data. For both the entity and non-entity set, we tokenized each word by characters. We processed all the words in the development set and test set in the same manner as well. We used SRILM to create a language model for the entities set and another for the non-entities set and then computed the perplexities of each word in the development and test sets using both models. To make a prediction, we compared the word's perplexity to the entity set to the its perplexity to the non-entity set and labeled the the word with the label associated with the lower complexity.

From there, we wanted to feed these results to ZOE. Since the input of CLM is text and the output of CLM is an entity or non-entity prediction (or mention predictions without any types), and ZOE takes an input of mentions without types and types them, we think we could have cleaned up the results of the CLM to identify entity spans (as opposed to single words) and provided them as input to ZOE to do the typing. However, due to time and computational constraints we weren't able to figure out how to use the existing ZOE for this project.

TYPE	Description
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FACILITY	Facility: Buildings, airports, highways, bridges, etc.
ORG	Organization: Companies, agencies, institutions, etc.
GPE	Geopolitical Entites: Countries, cities, states.
LOCATION	NonGPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. Not services.
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including "%".
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	"first", "second", etc.
CARDINAL	Numerals that do not fall under another type.

Table 2. Entity types and descriptions, from (Ont, 2012)

3.2 Implementing Zero-Shot BERT

To implement a form of 'Zero-Shot' BERT, we manually assigned a small number of entities contained within WikiLink data to various entity types. For example, we could depend on the WikiLink database containing entries containing the entity 'Mona_Lisa', which we would label as 'WORK_OF_ART'. We then pulled up to 500 sentences from the WikiLink data that contained each entity to create a database of context sentences. We simply labelled the entity span as the entity type of interest (including appropriate beginning- and inside- (B-, I-) tags) to heuristically create a weakly-labelled training dataset for BERT.

4 Experimental Evaluation

4.1 Methodology

The main criteria we used to evaluate our method are F1 score and validation accuracy, since this was what we saw in other papers and articles in the field.

4.2 Results

When we trained on the WikiLinks data only (not including any provided training data. We were able to code our strategy up, but had some confusion regarding the datasets so the statisitics we computed ended up not being accurate or meaningful. The code is available in out Google Drive of work related to the document.

Although we were not able to ultimately able to run ZOE; we've included the validation results of the task of Named Entity Identification in the following table. We were able to do this with following sets of data and calculate the accuracy and F1 score of the identification task in the following table:

Data Set	Accuracy	F1
18 Gold	0.938	0.788
12 Gold	0.930	0.700
6 Gold	0.939	0.701

Table 1: Table 1. Metrics for CLM Named Entity Identification

5 Future Directions

After implementing WikiLink-ed training and beginning to test our model, we found another interesting article (Rajasekharan (2020)) that described a simple method to pre-train BERT for NER without manually or algorithmically (weakly) labelling fresh sets of data each time an improvement is desired. This method instead takes about 5 man-hours to manually label clusters of BERT's vocabulary, then link those clusters to the entity types of interest. This ensures that future improvements can be made by adding more example sentences, which BERT can label using contextual clues. This can quickly provide an ample amount of fresh, semi-labelled data. Given more time, we would like to explore this method of labelling, as it does not depend on external availability of Wikilink-ed articles about the topic that need more data. Crucially, this would make it possible to easily create more labelled data about concepts for which the labelled Wikipedia article is unlikely to capture the 'usual' English meaning (e.g., the article "5:30" on Wikipedia does not refer to the time of day, but rather, the band). This is a drawback of our current system: certain desirable entity types simply are not very compatible with WikiLinks data.

Another, simple improvement on our current system would be further data collection. We were only able to process a limited amount of WikiLinks data for pretraining due to the time constraints. Obviously, further pretraining data is generally always useful.

Code Files and Presentation

This is the BERT-NER Colab: https://colab.research.google.com/drive/1W8VdRBHzk9rmx9jdp7zO5VL2NNNiPDAg?usp=sharing

This is the CLM Colab: https://colab.research.google.com/drive/1d01A2BBWaYY6ra6Bd4SC2cxFwKIxXlm2?usp=sharing

This is the Wikilinks data and the way we read it: https://colab.research.google.com/drive/1MedHTXUBQ8ZqhdIQbrZr8ENQT6HKb8In?usp=sharing

https://drive.google.com/file/d/1Q9MfVCKbbbiHgAJEa-3p5hGe5OH7IMIZ/view?usp=sharing

This is our video link: https://drive.google.com/file/d/19ORyR1O7UithukWINAADVTIKqW-xRbRv/view?usp=sharing

Acknowledgments

We would like to acknowledge the thoughtfulness of and the generosity in providing their time of all the TA's this semester. Our special thanks to TAs Ben Zhou, Xiaodong Yu, and Nupur Baghel who have guided us during our project.

References

2012. Ontonotes release 5.0.

Ming-Wei Chang, Lev-Arie Ratinov, D. Roth, and V. Srikumar. 2008. Importance of semantic representation: Dataless classification. In *AAAI*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Taesun Moon, Parul Awasthy, Jian Ni, and Radu Florian. 2019. Towards lingua franca named entity recognition with bert.

Ajit Rajasekharan. 2020. Unsupervised ner using bert.

Bernardino Romera-Paredes and Philip Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2152–2161, Lille, France. PMLR.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach.

Xiaodong Yu, Stephen Mayhew, Mark Sammons, and Dan Roth. 2018. On the strength of character language models for multilingual named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3073–3077, Brussels, Belgium. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Chen-Tse Tsai, and Dan Roth. 2019. Zero-shot open entity typing as type-compatible grounding.