# Towards Robust Named Entity Recognition via Temporal Domain Adaptation and Entity Context Understanding

**Oshin Agarwal**

University of Pennsylvania
oagarwal@seas.upenn.edu

## Abstract

Named Entity Recognition models perform well on benchmark datasets but fail to generalize well even in the same domain. The goal of my thesis is to quantify the degree of in-domain generalization in NER, probe models for entity name vs. context learning and finally improve their robustness, focusing on the recognition of ethnically diverse entities and new entities over time when the models are deployed.

## Introduction

Named Entity Recognition (NER) is the task of recognizing phrases of given semantic types such as organizations, locations, products and events in text. State-of-the-art neural models for NER achieve impressive performance on benchmark datasets. It is natural to ask whether these models can generalize to diverse names and contexts, even within the same genre of text. Consider the following sentences–

0. *Jane* is a computer scientist.
1. *Samarpita* is a computer scientist.
2. Her name is *Jane*.
3. *Jane* was released in theatres in 2017.

These sentences represent three categories of in-domain generalization. The first two sentences consist of different names of the same semantic type (person) in the same context—one is a common American name and the second is an uncommon Indian name. The first and the third sentence have the same name of the same type (person) but in different contexts. Lastly, the first and the fourth sentence have the same name but with different semantic types (person vs film) that can be inferred from their context. To claim in-domain generalization, a model should be able to correctly identify the entities and their types in all these scenarios.

The goal of my thesis is to first quantify the generalization capabilities of NER models by establishing appropriate benchmarks, metrics and best practices for task setup. Several recent works including my work on *entity-switched* datasets (Agarwal et al. 2020) have already made strides towards such a quantification and demonstrated that NER models are brittle. A closely related question aims to understand the reason for the lack of robustness by probing whether models memorize entity names or they can recognize predictive contexts. If a model depends extensively on names, it will not be able to recognize new names and contexts. In Agarwal et al. (2021b), I explored the degree of reliance of the models on names vs contexts as well as the feasibility of relying on context for NER.

Finally, with the developed resources and the insights gained from the analysis, I seek to improve the robustness of NER models, focusing on the recognition of ethnically diverse entities and new entities over time when these models are deployed in practice. My work will answer research questions pertaining to the robustness of these models. It will also have practical implications, particularly the techniques for improving robustness to evolving entities and language over time, can be adopted by NER practitioners.

## Benchmarks for Estimating Robustness

An analysis of two popular datasets, English CoNLL 2003 and Ontonotes, shows that several entities are repeated in the test set and many of them also appear in the training set with the same type. This points to the need to develop more challenging datasets that can better capture the state of generalization. Derczynski et al. (2017) developed such a dataset through manual human annotations of emerging, rare and temporally diverse entities. I developed a simple inexpensive method with minimal manual intervention to create datasets with varied entities that capture Type-1 and Type-2 generalization as shown in the examples above. I replaced entities in existing datasets with entities from various countries of origin while retaining the rest of the text and taking care of consistency to maintain textual coherence, thereby creating *entity-switched* datasets (Agarwal et al. 2020). American and Indian entities were recognized well with the state-of-the-art models, but not Vietnamese and Indonesian entities.

Another direction for developing benchmarks to capture generalization is to collect temporally stratified examples which mimic model deployment since language and entities change over time and may thus result in a drift in model performance. TTC (Rijhwani and Preotiuc-Pietro 2020) is a temporally stratified dataset of tweets annotated with named entities. Similarly, I am collecting a dataset with temporally stratified sentences from the New York Times and annotating them with named entities. On the collected pilot data, I observed that while the overall drift in the news is much

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

slower, the rate of drift varies across the type of news. News comprises of several sub-domains, some of which do not change much over time (eg national, sports) whereas others evolve rapidly (art, lifestyle). Sometimes new sub-domains get added (eg software technology). I found that the rate of drift is much higher for new and evolving sub-domains. The successful collection of this dataset will provide another type of benchmark for the evaluation of temporal robustness.

## Temporal Domain Adaptation

A reasonable way to counteract temporal drift in models is to retrain them on new and recent labeled data. Collection of manually labeled gold standard data can be time-consuming and expensive. I am working on methods such as intermediate pre-training and self-labelling (Agarwal and Nenkova 2021) for temporal domain adaptation using old labeled data in conjunction with new unlabeled data. I have found the latter method to be successful on tweets, in some cases even outperforming models trained on recent labeled data. I will test these methods on the news as well once I have finished collecting the NYT temporally stratified data, to ensure that the methods work across datasets and genres. I will also explore an alternative to retraining on new annotated sentences, by developing methods that can use gazetteers that possibly could be updated more quickly and cheaply.

The increasing amount of new data, labeled or unlabeled, poses a challenge with model training. If the amount of training data keeps growing, model training will need more resources, both in terms of time and hardware. An alternative to randomly sampling data continuously is to optimize the data selection such that only examples which provide new information are selected. I will work on optimizing the data selection process by utilizing the syntactic and semantic properties of examples to find sentences that are significantly different than the training data. The goal of this process would be to improve the coverage of new entities and contexts. While seemingly similar to active learning, we may not use the trained model to select new data. Active learning may not be a viable option since a preliminary analysis showed us that the NER models are often confident even when they make incorrect predictions.

## Entity Context Understanding

Temporal domain adaptation via retraining offers a simple practical solution. However, to build robust models that do not require frequent retraining, we need to understand what they are learning – i) whether they are learning artifacts in the training data, or ii) correct but shallow reasoning, or iii) the proper reasoning required to perform the intended task.

Consider, the earlier example “Jane is a computer scientist”. The correct identification of “Jane” as a person may be due to knowing that *Jane* is a fairly common name, or a competent user of language would know that only a person may be a computer scientist. Such probing of the reasons behind a prediction is needed to develop robust models. In Agarwal et al. (2021b), I focused on the interplay between learning names and recognizing constraining contexts. We define a constraining context as the sentence level context that determines the entity type regardless of the exact entity string. I found that while systems obtain high performance using just the word identity, the same is not true when just the context is used. I further performed human evaluation to determine if the failure cases for the context-only system had constraining contexts and there was scope for better reasoning, or if there were only ambiguous contexts and the task was indeed hard. While the majority of the cases were ambiguous contexts, about a quarter were constraining. Ideally, a system should be able to recognize such contexts and determine the entity type correctly, irrespective of the exact entity string. While learning the entity strings is a correct way to identify names, it is not sufficient for generalization, in cases like “Jane was released in theatres in 2017” where the most common type for Jane isn’t correct for the context.

As next steps, I plan to develop models that identify constraining contexts explicitly. I will use the entity-switched datasets to determine the degree of predictiveness of contexts as proportion of entities correctly identified in the same context. I will also use my work on converting knowledge graphs into synthetic natural language sentences (Agarwal et al. 2021a). The sentences generated in this work are succinct and correspond to a specific set of knowledge graph triples with entity of known types, thus providing distantly supervised data for named entity recognition with likely constraining contexts. Lastly, I may collect human annotations of constraining contexts to compare it to the first two methods of automatically finding constraining contexts.

## Future Work

The completion of my dissertation involves three avenues of work—complete the collection of the temporally stratified news dataset, optimize data selection for temporal domain adaptation, and incorporate constraining contexts in models.

## References

Agarwal, O.; Ge, H.; Shakeri, S.; and Al-Rfou, R. 2021a. Knowledge Graph Based Synthetic Corpus Generation for Knowledge-Enhanced Language Model Pre-training. In *Proceedings of NAACL-HLT*, 3554–3565. Online: ACL.

Agarwal, O.; and Nenkova, A. 2021. Temporal Effects on Pre-trained Models for Language Processing Tasks. *arXiv preprint arXiv:2111.12790.*

Agarwal, O.; Yang, Y.; Wallace, B. C.; and Nenkova, A. 2020. Entity-Switched Datasets: An Approach to Auditing the In-Domain Robustness of Named Entity Recognition Models. *arXiv preprint arXiv:2004.04123*.

Agarwal, O.; Yang, Y.; Wallace, B. C.; and Nenkova, A. 2021b. Interpretability Analysis for Named Entity Recognition to Understand System Predictions and How They Can Improve. *Computational Linguistics*, 47(1): 117–140.

Derczynski, L.; Nichols, E.; van Erp, M.; and Limsopatham, N. 2017. Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition. In *Proceedings of the 3rd WNUT*. Copenhagen, Denmark: ACL.

Rijhwani, S.; and Preotiuc-Pietro, D. 2020. Temporally-Informed Analysis of Named Entity Recognition. In *Proceedings of ACL*, 7605–7617. Online: ACL.