GeoAI Challenge Location Mention Recognition from Social Media by ITU

Olufemi Victor Tolulope
femiosinkolu@gmail.com

Abstract

Location Mention Recognition (LMR) from social media data during crisis events is a critical task for enabling effective disaster response. This report presents an approach to LMR for the GeoAI Challenge. The task was to predict the starting and ending positions of location words in tweets using the IDRISI-RE dataset. Leveraging some pre-processing and post-processing techniques we significantly improved the quality of predictions from open source LMR models. Our approach includes the implementation of three key functions: the Mentions Destroyer, Fix Entity, and Post Process Predictions functions. The Mentions Destroyer eliminates Twitter username mentions to avoid misclassification as locations, ensuring more accurate predictions. The Fix Entity function rectifies predictions that may be broken into tokens, thereby providing complete words. The Post Process Predictions function enhances the proximity analysis between word boundaries to correct instances where words are wrongly separated. In this report, we detail the methodology and approach. By integrating these key functions, our solution achieved exceptional results in the competition. We also discuss the importance of Dr Reem Ali Suwaileh's work, which laid the foundation for our approach.

1. Introduction

Social media has become an invaluable resource for disaster response and management, providing real-time insights and situational awareness during crises. As the digital age continues to evolve, the need to extract geolocation information from the vast troves of social media data has become increasingly critical for first responders, disaster relief organizations, and researchers. Recognizing location mentions (LMR) within crisis-related tweets is a fundamental step in this process.

The GeoAI Challenge presented a unique opportunity to address this challenge, offering the IDRISI-RE dataset and tasking participants with predicting the starting and ending positions of location words in tweets. The competition aimed to advance the state-of-the-art in LMR from social media during disaster events, with an emphasis on the English language.

2. Related Work

Location Mention Recognition (LMR) in the context of crisis management is a multifaceted research area that has garnered considerable attention in recent years. Recognizing location mentions from social media data is pivotal for gaining situational awareness during disasters and delivering timely aid. The GeoAI Challenge, which focused on this problem, builds upon an evolving body of work that has set the stage for advancing LMR in crisis response.

2.1 Dr. Reem Ali Suwaileh's Influential Work:

Dr Reem Ali Suwaileh, a researcher at Qatar University, has made notable contributions to the field of LMR from social media data during crisis events. Her extensive research has explored the challenges and opportunities in this domain which aimed to further advance the state-of-the-art in LMR.

2.2 Baseline Models:

As part of her work, Dr. Suwaileh created baseline models, which are publicly available on the Hugging Face model hub. These models are tailored to address the LMR problem in crisis-related tweets. The models are optimized for various aspects, including temporal information and location types.

2.3 Broader Research Landscape:

Beyond the GeoAI Challenge, the research landscape in LMR from social media data is vibrant. Researchers have explored various aspects of this problem, including pre-processing techniques, model selection, evaluation metrics, and domain adaptation.

Notable models in the broader landscape include those based on BERT, a popular language model, and other advanced machine-learning techniques. Researchers continue to investigate ways to enhance model performance, particularly under conditions where labelled training data is limited

3. Methodology

The methodology of our approach to Location Mention Recognition (LMR) in crisis-related tweets involves a multi-faceted strategy. Primarily, our methodology encompasses pre-processing, post-processing, and an ensemble strategy to improve the accuracy of location-mention predictions.

3.1 Baseline Models:

For our solution, we adopted the use of open-source baseline models made available by Dr Reem Ali Suwaileh on the Hugging Face model hub. Specifically, we utilized two models:

1. IDRISI-LMR-EN-timebased-typeless
2. IDRISI-LMR-EN-random-typebased

These models are tailored to the LMR problem in English tweets, with variations in the consideration of temporal information and location types. We leveraged these models to generate initial predictions.

3.2 Pre-processing Techniques:

1. **Mentions Destroyer Function**
   We implemented pre-processing techniques to enhance the quality of the input data for our models. The most notable pre-processing function is the **Mentions Destroyer**. This function addresses a unique challenge in the GeoAI Challenge dataset. Upon inspection of the dataset annotation instructions, we found that Twitter username mentions were not to be considered as location mentions. To prevent the model from erroneously classifying Twitter usernames as locations, the Mentions Destroyer function replaces mentions with question marks of the same length. This technique ensures that Twitter username mentions are appropriately masked.

Mentions Destroyer:

Tweet Sample: @Canadawatch there seems to be a hailstorm in Lagos Nigeria today, the entire Lagos market has closed.

Mentions Destroyer Function

Output: ???????????? there seems to be a hailstorm in Lagos Nigeria today, the entire Lagos market has closed.

Advantage: Helps to avoid model confusion and preserve text length

3.3 Post-processing Techniques:

Post-processing is a crucial component of our solution. We implemented two key post-processing functions to refine the predictions made by the baseline models:

1. **Fix Entity Function**: The Fix Entity function addresses instances where model predictions may be incomplete words, as the models are primarily designed for token classification. It checks the returned prediction to ensure it forms a complete word in the tweet. If not, the function completes the word and adjusts the end index to match.

Sample tweet

HURRICANE FLORENCE HEROES RESCUE 6 TRAPPED DOGS FROM RISING WATERS OR THEY WOULD HAVE DROWNED Rescued six dogs in **Leland**, NC, after the owner LEFT THEM locked in an outdoor cage that filled with flood

High Confidence Score

Model Output: [{'entity': 'U-CITY', 'score': 0.91648316, 'index': 54,'word': 'Le','start': 114,'end': 116}, ...

Incomplete word

Entity Fixer Function: Find the word in the tweet and complete it

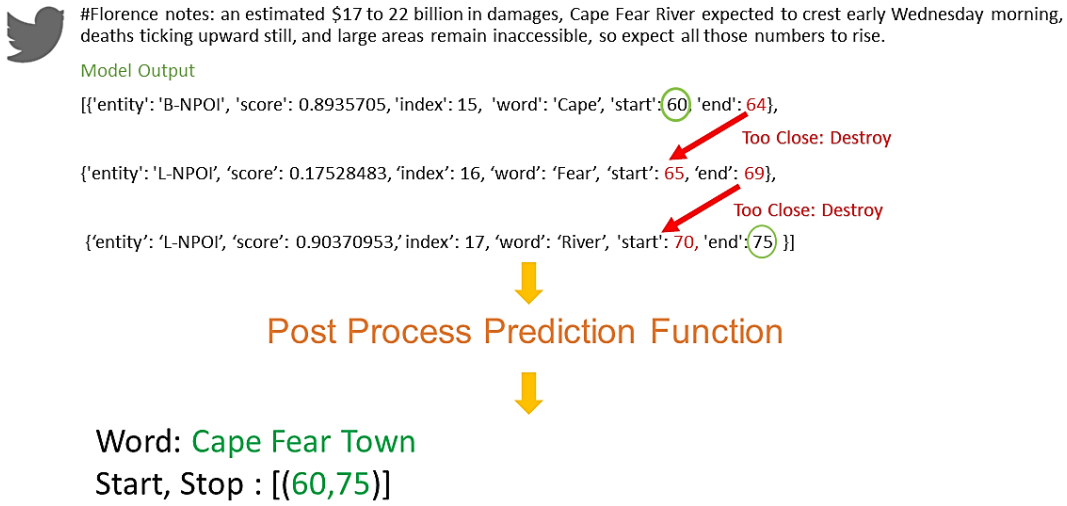
Function: [{'entity': 'U-CITY', 'score': 0.91648316, 'index': 54, 'word': ' **Leland** ', 'start': 114,'end': Start+len(word)},.....

Which is [{'entity': 'U-CITY', 'score': 0.91648316, 'index': 54,'word': ' **Leland** ','start': 114,'end': 114+6},.....

Output: [{'entity': 'U-CITY', 'score': 0.91648316, 'index': 54,'word': ' **Leland** ','start': 114,'end': 120},.....

Advantage: Fixes word, Fixes end position

2. **Post Process Predictions Function**: This function focuses on the proximity of the end index of one word and the start value of the next word. When these values are extremely close, indicating that they likely represent the same word, this function helps correct instances where words are wrongly separated by the model.



3.4 Ensemble Strategy:

1. An important component of our methodology involves the ensemble of results from the two baseline models mentioned above. We found that combining the predictions from multiple models improved the overall performance of our solution.

This methodology allowed us to address the challenges of the GeoAI Challenge effectively, enhancing the accuracy of LMR in crisis-related tweets.

4. Experimentation and Results

In our pursuit of improving Location Mention Recognition (LMR) from crisis-related tweets for the GeoAI Challenge, we conducted a series of experiments with the primary objective of enhancing model predictions.

4.1 RMSE Scores:

Our approach culminated in a final ensemble score with a Root Mean Square Error (RMSE) of 11.9151 on the GeoAI Challenge leaderboard. This score reflected the accuracy of our model predictions, with lower RMSE values indicating more precise predictions.

Individual model performance on the leaderboard was noteworthy:

- **IDRISI-LMR-EN-timebased-typeless**: Achieved an RMSE of 11.9282.
- **IDRISI-LMR-EN-random-typebased**: Attained an RMSE of 12.7298.

The results of our approach in the GeoAI Challenge demonstrated the effectiveness of our methodology. By leveraging baseline models, implementing pre-processing and post-processing

techniques, and applying ensemble strategies, we achieved remarkable RMSE in LMR from crisis-related tweets.

Notably, without the Fix Entity function and the Mentions Destroyer, we observed an RMSE of 13.99 when utilizing the single model IDRISI-LMR-EN-timebased-typeless. The introduction of these functions led to a substantial improvement in prediction accuracy, underscoring their significance in our solution.

5. Conclusion

5.1 Potential for Further Research:

The GeoAI Challenge highlighted the complexities and importance of LMR from crisis-related tweets. While our approach achieved competitive results, there is still ample room for further research in this field. Future investigations may focus on fine-tuning models for even better performance, exploring additional pre-processing and post-processing techniques, and expanding the dataset diversity for improved generalizability. We look forward to future research endeavours in this domain, with a commitment to refining and expanding the capabilities of LMR models in the critical field of crisis response.

5.2 Acknowledgements

First and foremost, we extend our heartfelt gratitude to Dr Reem Ali Suwaileh, whose pioneering research in the domain of Location Mention Recognition from crisis-related tweets served as the inspiration for this challenge. Dr. Suwaileh's extensive work, including the creation of baseline models and datasets, formed the cornerstone of our approach. Her commitment to open research and collaboration significantly enriched our methodology.

We would like to express our appreciation to the organizers of the GeoAI Challenge and the International Telecommunication Union (ITU) for providing this platform and opportunity to apply our skills and knowledge to a real-world problem with significant humanitarian implications.

References

[1] S. Reem, E. Tamer and I. Muhammad, "IDRISI-RE: A generalizable dataset with benchmarks for location mention recognition on disaster tweets," *Information Processing & Management,* 1 May 2023.

[2] S. Reem, E. Tamer, I. Muhammad and S. Hassan, "When a disaster happens, we are ready: Location mention recognition from crisis tweets," *International Journal of Disaster Risk Reduction,* vol. 78, pp. 103-107, 1 August 2022.

[3] S. Reem, I. Muhammad, E. Tamer and S. Hassan, "Are we ready for this disaster? Towards location mention recognition from crisis tweets," in *Proceedings of the 28th international conference on computational linguistics*, 2022.