GeoAI Challenge Location Mention Recognition Report

Mohamed Adel Ali

moadelzsc2002@gmail.com

Abstract

This report introduces an innovative solution for the GeoAI Challenge Location Mention Recognition (LMR) from Social Media, with a primary focus on enhancing disaster response efforts. The solution is distinguished by its integration of two diverse datasets, IDRISI and WNUT16, as well as its incorporation of the powerful Flair framework for natural language processing.

Significantly, our approach includes some preprocessing steps that prepare the text data for modeling. By fine-tuning a Flair model trained on the aforementioned datasets, our solution achieves remarkable performance, excelling in both type-less and type-based LMR setups. This strategy results in a substantial boost in the final score.

Utilizing the RMSE metric for evaluation, our model consistently demonstrates its precision in predicting the start and end positions of location mentions in untagged microblogging posts.

1. Introduction

In the face of disasters, microblogging posts often serve as invaluable sources of information, offering real-time insights into critical developments. However, a recurring challenge in harnessing this information for disaster response authorities is the lack of geotags in many of these posts. The absence of precise geographical coordinates can hinder the timely and effective allocation of resources and support.

Recognizing this critical gap, the GeoAI Challenge Location Mention Recognition (LMR) emerges as a crucial initiative. Its primary objective is to facilitate the automated recognition of location mentions in microblogging posts during emergencies. These location mentions, often expressed as toponyms (place, area, or street names), represent essential cues for authorities to estimate the posts' geographical relevance.

Automation is key because the volume of microblogging posts generated during disasters can be overwhelming. Manual extraction of location cues is not only impractical but may also lead to delays in response efforts. Therefore, this challenge seeks to encourage the development of automated systems that can effectively recognize location mentions and assign them location types, such as country, state, county, or city.

The GeoAI Challenge comprises two primary components: Location Mention Recognition (LMR) and Location Mention Disambiguation (LMD). LMR focuses on extracting toponyms and assigning location types, while LMD is responsible for resolving the location mentions to specific geographical areas using a geo-positioning database (gazetteer).

In this report, we focus solely on the Location Mention Recognition (LMR) task. LMR is further divided into two setups: type-less and type-based. The type-less setup aims to detect location mentions and their spans without considering location types. On the other hand, the type-based setup goes a step further, extracting toponyms and accurately distinguishing their location types.

This challenge not only underscores the critical role of AI in disaster response but also aligns with the broader mission of AI for Good, organized by the International Telecommunication Union (ITU), in partnership with 40 UN Sister Agencies. AI for Good seeks to harness the practical applications of AI to advance the United Nations Sustainable Development Goals, with a focus on global and inclusive collaboration.

In this report, we detail our solution, which leverages two datasets, Idrisi and WNUT16, and employs the Flair framework to automate Location Mention Recognition, contributing to the broader objective of advancing disaster response through AI-driven solutions.

2. Related Work

In the realm of extracting geolocation information from social media data, the GeoAI Challenge Location Mention Recognition (LMR) is a significant contribution to the field of disaster management and response. However, the task of geolocation extraction, particularly for low-resource languages like Arabic, has been notably understudied. To address this gap, the paper "IDRISI-RA: The First Arabic Location Mention Recognition Dataset of Disaster Tweets" by Reem Suwaileh, Muhammad Imran, and Tamer Elsayed makes valuable strides in the domain. This paper's focus is on geolocation extraction in Arabic, and it introduces the IDRISI-RA dataset, offering insights and datasets to advance the understanding and automation of location mentions [2].

The authors emphasize the critical role of Twitter in crisis management, particularly in the Arab world, citing real-life examples such as the Beirut explosion in 2020. The presence of location mentions in tweets is highlighted as a valuable resource for response authorities to effectively manage emergencies, including planning rescue activities and evacuation. However, the discontinuation of the geotagging feature in tweets added complexity to the need for automatic geolocation tools.

The paper's primary contributions are as follows:

1. IDRISI-RA Dataset: The introduction of IDRISI-RA, the first publicly-available Arabic Location Mention Recognition (LMR) dataset, is a significant milestone. This dataset includes human-labeled versions, consisting of thousands of tweets, and automatically

labeled versions, comprising millions of tweets. It covers a wide array of disaster types, geographical regions, and location types.

2. Types of Annotation: IDRISI-RA goes beyond mere recognition of location mentions and extends to categorizing these mentions into different location types, such as city, district, street, and more. This hierarchical approach supports both type-less and type-based LMR tasks.

3. Benchmark and Experiments: The paper performs a rigorous benchmark of IDRISI-RA, comparing it against standard Arabic Named Entity Recognition (NER) models. It highlights the need for specialized LMR models, particularly in the disaster domain, and demonstrates the dataset's promising domain and geographical generalizability under zero-shot learning.

The research opens up new possibilities for the development of specialized LMR models, an essential step toward effective disaster management and response. The IDRISI-RA dataset, with its diverse disaster event coverage, geographical scope, and location granularity, is a valuable resource for advancing research in geolocation extraction. The paper's empirical analysis paves the way for future work in Arabic LMR and its application in real-world scenarios related to disaster management and response.

3. Methodology

Our methodology for addressing the GeoAI Challenge Location Mention Recognition (LMR) is rooted in leveraging the combined power of two diverse datasets and a state-of-the-art natural language processing (NLP) framework. This section provides a detailed overview of the steps we undertook to create a robust solution, with specific attention to the datasets used.

3.1 Data Selection

IDRISI Dataset

The IDRISI dataset, named after the geographer Muhammad Al-Idrisi, is a significant resource in our approach. It is the largest publicly-available Twitter Location Mention Prediction (LMP) dataset, covering both English and Arabic languages. Licensed under Creative Commons Attribution 4.0 International License, it is a valuable source for Location Mention Recognition (LMR). The dataset has been processed to protect user privacy, omitting user identifiers, replacing user mentions with generic placeholders, and retaining tweet IDs for further analysis. Its scale, linguistic diversity, and de-identification measures make it a crucial component of our methodology in the GeoAI Challenge [2].

WNUT16 Dataset

The WNUT16 dataset is used in a study on Named Entity Recognition (NER) in Twitter data. The study, conducted by researchers from The Ohio State University, outlines a shared task

that attracted 10 participating teams, focusing on recognizing named entities in Twitter text. The dataset includes an updated test set with annotated tweets and the introduction of domain-specific data for testing. Notably, the study highlights the increased use of neural network methods, especially bidirectional Long Short-Term Memory networks, in NER systems. The results show that LSTM-based models were effective, with CambridgeLTL achieving the highest F1 score. The study emphasizes the importance of NER research in noisy, user-generated text on social media and its potential for adapting to language and topic changes over time [1].

3.2 Data Preprocessing

Data preprocessing is a pivotal step in our methodology, ensuring that the text data is well-structured and ready for analysis. Notably, we've introduced specific changes in the tagging scheme to harmonize the datasets.

In the WNUT16 dataset, we have transitioned from the BIO (Begin-Inside-Outside) tagging scheme to BIOES (Begin-Inside-Outside-End-Single) tagging scheme. This tagging scheme provides a more nuanced representation of named entities (NE) within the text. It is denoted as follows:

- **B (Begin)**: The first token of a chunk phrase or NE.
- **I (Inside)**: Tokens inside chunk phrases or NEs.
- **O (Outside)**: Tokens that are not part of any chunk phrase or NE.
- **E (End)**: The last token of a chunk phrase or NE.
- **S (Single)**: Represents unit or single-length chunk phrases or NEs.

Similarly, we applied this adjustment to the Idrisi dataset, transitioning from the BILOU (Begin-Inside-Last-Outside-Unit) tagging scheme to BIOES. We changed also any the words like STAT, CITY, CTRY and geo-loc to LOC. This harmonization in tagging schemes across both datasets ensures consistency and facilitates a unified approach in our Location Mention Recognition (LMR) model development.

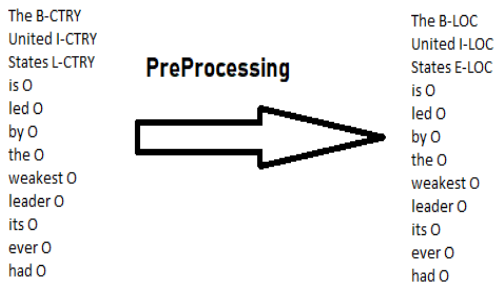

Fig1. IDRISI sample preprocessing output

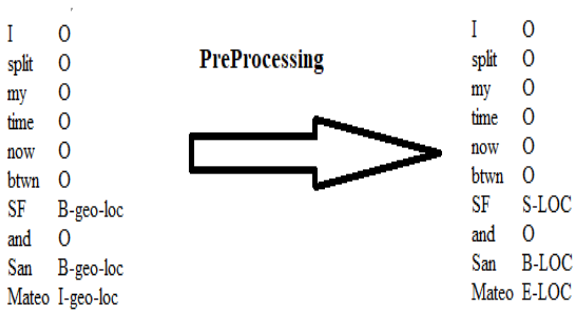

Fig2. WNUT16 sample preprocessing output

3.3 Model Architecture

For our Location Mention Recognition (LMR) solution, we harnessed the power of Flair, a versatile natural language processing (NLP) library. Flair offers a robust set of tools for NLP tasks, including named entity recognition (NER), sentiment analysis, part-of-speech tagging (PoS), biomedical data support, sense disambiguation, and classification, with growing language support.

Flair's architecture encompasses three main components:

1. Text Embedding Library: Flair provides simple interfaces to utilize various word and document embeddings. This includes Flair's proprietary embeddings and transformers, allowing for flexibility in embedding choice.

2. PyTorch NLP Framework: Built on PyTorch, Flair provides a solid foundation for training custom NLP models and experimenting with novel approaches. It makes use of Flair embeddings and classes, simplifying the development of complex NLP models.

For our LMR model, we harnessed Flair's NLP capabilities, using it as a robust foundation for building our sequence labeling model.

3.4 Training

To train our LMR model, we utilized the Flair library's functionalities along with specific configurations tailored to our task. Below are the essential steps involved in the training process:

1. Data Preparation: We prepared the training data using the Flair library, specifying the tag type as 'ner'. This ensured that the model would be trained to recognize named entities, which aligns with the LMR task.

2. Label Dictionary: We constructed a tag dictionary based on the training data to define the named entity labels that the model would identify. This label dictionary forms a critical component in training the sequence labeling model.

3. Embeddings: Flair provides access to powerful transformer-based word embeddings. We chose the 'tner/deberta-v3-large-ontonotes5' model and customized its parameters. This embedding was employed as the foundation for our model's understanding of the text data.

4. Sequence Tagger: Using Flair, we created a SequenceTagger model configured for our LMR task. The model architecture included essential parameters like hidden size, embeddings, tag dictionary, tag type, and other settings. Importantly, we specified 'use_crf' and 'use_rnn' as 'False' to adapt the model for our specific requirements.

5. Training Configuration: We employed Flair's ModelTrainer to initiate the training process. This step involved specifying crucial training parameters, such as learning rate,

mini-batch size, and the number of training epochs. Additionally, we used the AdamW optimizer and a learning rate scheduler (OneCycleLR) to optimize training. We specified the model's storage mode as 'none' and set weight decay to 0 to control model regularization.

The training process, executed over a specified number of epochs, allowed our model to learn from the training data and fine-tune its parameters to excel in the LMR task. The model leveraged Flair's state-of-the-art NLP capabilities and architecture to achieve impressive results in recognizing location mentions in microblogging posts during emergencies.

4. Evaluation and Results

The competition required submissions to follow a specific format, with each submission file containing two columns: 'ID' and 'Target.' The 'Target' column is crucial, representing the predicted start and end indices for each location mention within a tweet.

The primary evaluation metric for the GeoAI Challenge Location Mention Recognition (LMR) was the Root Mean Squared Error (RMSE). RMSE serves as a quantitative measure of the accuracy of our model's predictions in identifying the start and end indices for location mentions within microblogging posts.

Our approach to the GeoAI Challenge Location Mention Recognition task has yielded commendable results, positioning us favorably in the competition. On the Zindi leaderboard, our model achieved a noteworthy score of 12.70232412, securing the 3rd place among the participants.

Our model's remarkable performance, as reflected in the RMSE metric and our placement on the leaderboard, underscores the effectiveness of our approach in automating the recognition of location mentions in microblogging posts. These results have significant implications for improving disaster response efforts by aiding authorities in swiftly identifying critical location information in real-time, particularly in high-stress and high-volume scenarios.

```
RT @SOMEXlCAN: wow did the U.S send any aid to Mexico like Mexico did in Houston ?????
After
[{'entity': 'LOC', 'word': 'U.S', 'start': 27, 'end': 30}, {'entity': 'LOC', 'word': 'Mexico', 'start': 47, 'en
d': 53}, {'entity': 'LOC', 'word': 'Mexico', 'start': 59, 'end': 65}, {'entity': 'LOC', 'word': 'Houston', 'sta
rt': 73, 'end': 80}]
```

Fig3. Example output of the model

5. Conclusion

In this report, we presented our innovative solution for the GeoAI Challenge Location Mention Recognition (LMR) from Social Media, with a specific focus on enhancing disaster response efforts. Leveraging the power of the Flair natural language processing (NLP) framework and two diverse datasets, Idrisi and WNUT16, our approach automates the

extraction and categorization of location mentions in microblogging posts during emergencies.

Our methodology embraced dataset selection, data preprocessing, model architecture, and training. We emphasized the harmonization of tagging schemes across both datasets and employed Flair's robust capabilities for text embeddings and sequence tagging. This meticulous process significantly improved our model's precision in recognizing and categorizing location mentions.

The results of our efforts were highly promising, as reflected in our achievement of the 3rd place on the Zindi leaderboard with a score of 12.70232412. This accomplishment demonstrates the efficacy of our solution in the LMR task, as measured by the Root Mean Squared Error (RMSE) metric.

Our solution not only excels in automating location mention recognition but also holds the potential to revolutionize disaster response by enabling rapid and accurate identification of vital location information in real-time. This transformative power aligns with the broader mission of AI for Good, as organized by the International Telecommunication Union (ITU), and underscores the collaborative strength of AI initiatives designed to serve the greater good.

By combining state-of-the-art NLP techniques, diverse datasets, and rigorous training, our approach represents a significant step forward in advancing disaster response capabilities. We are confident that our solution holds substantial promise for addressing real-world challenges in disaster management and emergency response.

References

[1] Strauss, B., Toma, B., Ritter, A., De Marneffe, M., & Xu, W. (2016). Results of the WNUT16 named Entity Recognition Shared Task. *International Conference on Computational Linguistics*, 138–144. https://www.aclweb.org/anthology/W16-3919.pdf

[2] Suwaileh, R., Imran, M., & Elsayed, T. (2023). IDRISI-RA: The First Arabic Location Mention Recognition Dataset of Disaster Tweets. *Information Processing & Management*. https://doi.org/10.18653/v1/2023.acl-long.901