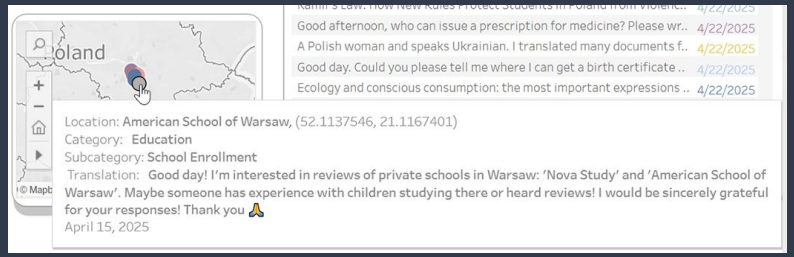
NLP Location Identification Accuracy

Using LLMs to identify locations in unclean chat messages.

Olivia Croteau, Natalie Crowell



- Our co-op and senior capstone project: LLM and data visualization software to help humanitarian professionals analyze Telegram messages between displaced Ukrainian refugees.
- Feature: Mapping messages that mention a specific location. By accurately mapping these locations, humanitarian organizations can better allocate resources and provide timely assistance to those in need.

Detecting Geospatial Location Descriptions in Natural Language Text

Main Contribution

- Determining accuracy of using NLP to identify phrases/groups of words that contain spatial references in relationship to an object. Example: The house next to the Genesee River.
- Concludes that the "meta-classifier" approach is most accurate. The first stage of this method uses three classifiers and each returns their prediction. Then, the most accurate prediction is determined out of the three classifiers and returned as the output.

Methods

- Goal of classifying text into three categories: geospatial language, other-spatial language (spatial but not geographic), and non-spatial expressions
- Used multiple different classifiers to identify parts of speech and to stack the classifiers and produce a final output
- Bag of Words classifier, Word Embedding classifier, Language Pattern classifier, Baseline classifier —> meta classifier to make the final decision.

Relevance to Research

- Provides an outline on how to use natural language processing to answer a big question in our research topic. How do we identify locational references in text when the location is not a singular word but rather a more abstract phrase?

Location Reference Recognition from Texts: A Survey and Comparison.

Main Contribution

- Evaluated the accuracy and efficiency of the 27 most used approaches for location reference recognition
- The conclusion of the paper provides useful information on what strategies performed best as well as important factors to consider when training a model for location identification.

Methods

- When evaluating the many approaches, they used the same comparison metrics, precision, recall, and F1-score
- Measured processing time for computational efficiency
- 3 formal datasets and 23 informal datasets were used on each approach
- Examined accuracy across different types of locations, cities, countries, states, streets, roads, or buildings

Relevance to Research

- Integrating a voting mechanism offers robustness, though performance will differ based on different types of texts.
- Provides information on the best geocoding methods for handling informal texts as well as ambiguous texts; references papers that tackled these problems on a large scale and smaller scale.

High Accuracy Location Information Extraction from Social Network Texts Using Natural Language Processing

Main Contribution

- Proposed NER solution accurately recognizes locations in social network message data.
- Evaluated three models on location identification
- Reevaluated models with their added tokenization method (Generalized Levenshtein Distance) and Gazetteer-based matching

Methods

- The corpus is internet and social network texts being processed by a Stanford CoreNLP model. Includes preprocessing which removes special characters and hyphenates multi-word location names, ergo only using one token.
- Tested with the proposed solution as well as three other models.

Relevance to Research

- Tests the accuracy of current NLP solutions for location identification including SpaCy and Stanford CoreNLP
- Included a custom gazetteer database specific to Burkina Faso
- Information of data of social media content to improve accuracy
- This paper provides a good model for what preprocessing we should employ in our project when working with ‘uncleaned’ message data.

Adaptive Geoparsing Method for Toponym Recognition and Resolution in Unstructured Text

Main Contribution

- Allows geoparsing design to learn from context to output more accurate locations
- Proposes a solution for understanding ambiguous words through Dynamic Context Disambiguation

Methods

- Tagged words are separated into two stacks and the model uses previous context from the correct stack to assign a location.
- Set of rules that determines that movement of the ambiguous word based on context of the previous words.
- Used three different datasets
- Measured Confusion matrix, accuracy, Distance-based ranking, Precision, Recall, F-measure and had a baseline comparison.

Relevance to Research

- Gazetteer method to match geographical locations.
- Research on geoparsing for different languages (specifically spanish)
- Answers the main question how do we now assign these tagged entities to coordinates

Our Method

Two step process:

1. Location referenced extraction
2. Applying coordinates

Will utilize BERT's built-in tokenizer and fine-tune the model to predict IOB tags for each token: B (beginning of location), I (inside location), or O (outside/not a location).
Our second step will follow an approach similar to the Dynamic Context Disambiguation mentioned in “Adaptive Geoparsing Method…”. The GeoNames gazetteer will retrieve all possible coordinates for a tagged location. Then we will implement the set of rules for dynamic context disambiguation identified in the paper. This will handle the ambiguous words and determine a location for the word based on context of the text.

Questions?