NLP Location Identification Accuracy

Using LLMs to identify locations in unclean chat messages.

Introduction

Motivation and Objective:

- Objective: Identify locational references in unstructured texts and assign a latitude and longitude for each location.
- Motivation: Former research conducted on analyzing telegram message data from Ukrainian refugees residing in Poland.
- Goal: Utilize our proposed method for geoparsing to expand our former project and assist humanitarian data analysts in identifying commonly asked questions and concerns from message data in relation to a location.

Research Questions:

1. **ML Performance:** To what extent can an ML model accurately identify location entities in unstructured chat message data, as measured by precision, recall, and F1-score on test data?
2. **Geocoding Accuracy:** How accurately can identified toponyms be geocoded to their correct latitude and longitude coordinates, as measured by comparison to ground truth coordinates?

Quick Sidenote: Translation Accuracy

- Previously: used OpenAI's gpt4 to complete translations. Assessed accuracy with native Ukrainian speakers from the IOM mission.
- Now: using the gemma3-translator model from Ollama
  - 4B parameters
  - Can be more specific with instructions via modelfile
  - Especially accurate with repeated proper nouns (ex: Ukrainian House)

Related Work

"We propose a hybrid method, named GazPNE, which fuses rules, gazetteers, and deep learning methods without requiring any manually annotated data. It can extract place names at both coarse and fine-grained levels and place names with abbreviations."

- GazPNE: Annotation-free Deep Learning for Place Name Extraction from Microblogs Leveraging Gazetteer and Synthetic Data by Rules

"We present a sophisticated location-aware recommendation system that uses Bidirectional Encoder Representations from Transformers (BERT)"

- MDPI BERT4Loc: BERT for Location—POI Recommender System

"we presented a novel geoparsing approach based on word embeddings for toponym recognition and dynamic context identification for toponym resolution. [...]

a set of rules and facts is applied to take advantage of context to assign the most suitable geographic level to place names, and then to identify the correct locations. [...]

our proposal allows us to assign the geographical properties of specific locations not contained within the gazetteer"

- MDPI Adaptive Geoparsing Method for Toponym Recognition and Resolution in Unstructured Text

Methods: Architecture

Bert:

- ❏ bert-base-cased as foundation
- ❏ Transformers layer provides contextual embeddings for words
- ❏ Classification layer is used to predict B (beginning of location), I (inside location), or O (outside location)

GazPNE:

- ❏ Binary classification approach
- ❏ Composite word embeddings from GloVE, OSM Gazetteer, and statistics from positive/negative examples
- ❏ C-LSTM identifies embedding patterns and analysis sequence data
- ❏ Classification layer determines positive or negative

Methods: Geoparsing

Bert:

- ❏ Tags entities as IOB
- ❏ Implements dynamic context disambiguation
- ❏ Matches locations to gazetteer

GazPNE:

- ❏ Identifies embeddings as location or not a location
- ❏ Population and feature importance for disambiguation
- ❏ Matches positive locations to a gazetteer

Advantages and Disadvantages

Bert

Advantages	Disadvantages
❏ Understands Context ❏ Uses Grammatical and semantic cues ❏ Infer locations based on context	❏ Requires IOB tagged data ❏ Possible overfitting

GazPNE

Advantages	Disadvantages
❏ More available datasets	❏ Can miss context clues ❏ Ambiguity issues

Experiments

- ❏ Train each model and evaluate precision, recall, F1, at entity level on same testing data
- ❏ Implement and evaluate geoparsing for each model. Self evaluation.

Bert:

- ❏ Bert model trained and tested using an IOB tagged dataset including the following files
  - ❏ Training: 14,041 sentences
  - ❏ Testing: 3,453 Sentences
  - ❏ Validation: 3,250 Sentences

Geoparsing

- ❏ Gazetteer
  - ❏ Number of Locations: 508,720
- ❏ Testing data
  - ❏ 62 Sentences
  - ❏ 30 Locations

GazPNE:

- ❏ GazPNE trained using positive examples from an OSM gazetteer and negative examples generated in a preprocess script. Model will be tested on positive examples extracted from the same testing file used for the bert model.
  - ❏ Training Positive: 1,644,361
  - ❏ Training Negative: 179,916,430
  - ❏ Testing Positive: 505
  - ❏ Testing Negative: 505

Geoparsing Evaluation Metrics and Reasoning

- Wanted to use real data collected from Telegram channels
- Research question focuses on unstructured text and how our model performs on this type of data.
- Therefore, we do not have a formatted dataset with correct locations identified.
- Used a smaller dataset and evaluate our performance metrics by hand.
- Better understand inconsistencies in our model outputs, determine patterns in tagged entities, identify outliers, and make inferences.

Model Testing Results

Model	Precision	Recall	F1	Testing Runtime
BERT	0.9327	0.9305	0.9316	114.82 seconds
GazPNE	0.8422	0.9723	0.9026	7.03 seconds

The Bert model was more accurate at identifying locations however was slower in execution. GazPNE still achieved 90% and can be useful for cases requiring fast processing.

Geoparsing Results

Test dataset:

- 62 sentences
- 30 locations

Model	Total Locations Extracted	Total Locations Resolved	Resolved correct Locations
BERT	92	33	26
GazPNE	2391	75	20

Example text

“Wonderful day of golden autumn was received as a gift in a package with an exciting tour of the **‘Royal Lazienki’ park**. With interesting details and humor, guide Natalia Romanova acquainted us with the history of the park, at the same time we learned about Polish kings, their lovers, artists and their friends and other prominent figures in the history of **Poland** and **Warsaw** in particular. Interesting story, picturesque views, playful squirrels, wonderful company made this walk especially meaningful. In addition, we had the opportunity to practice the Polish language. Thank you Fundacja dla Wolności for the opportunity to learn more about the land where we live.

Bert Locations:

‘; Royal Lazienki ’; Poland; Warsaw

GazPNE Locations:

‘Royal Lazienki’; park. With; Natalia Romanova; Polish kings,; particular. Interesting; meaningful. In; Polish language.; Wonderful; humor,; park,; lovers,; Poland; Warsaw; story,; views,; squirrels,; addition,; Thank; Fundacja; Wolności; live.

Conclusion

Bert:

- Bert did a much better job at identifying locations within the unseen text. It made more accurate prediction.
- Bert commonly identified “home” as a location which was a location in the gazetteer causing confusion for the model.
- Excellent with gazetteer matching

GazPNE

- GazPNE commonly identified other proper nouns, such as people’s names, as locations
- Relied heavily on the Gazetteer to filter its predictions.
- Extreme amount of false positives

Conclusion

Bert:

- ❏ Less available training data
- ❏ Higher accuracy
- ❏ Slow processing
- ❏ Larger model size
- ❏ Can easily handle multi-word locations

Future:

- ❏ Bert is the more reliable model to use for location identification.
- ❏ Next step: expand gazetteer and implement more complex dynamic context disambiguation to achieve better results for unresolved locations.

GazPNE:

- ❏ Trained on more data
- ❏ Less accuracy
- ❏ Faster processing
- ❏ Smaller Model Size
- ❏ Difficulties handling locations outside of the gazetteer
- ❏ Requires candidate phrases to handle multi word locations

References

Datasets:

Kaggle: Nikhil John 3 years ago “CoNLL Locations only” (722.43 kB Training dataset). IOB format only identifies locations. Train, Test, and Validate datasets https://www.kaggle.com/datasets/nikhiljohnk/conll-locations-only?select=test.txt

GeoNames. (n.d.). Www.geonames.org. https://www.geonames.org/

MapTiler AG. (2025). Place names from OpenStreetMap. Downloadable. Ranked. With bbox and hierarchy. Ready for geocoding. OSMNames.org. https://osmnames.org/about/

Papers:

Aldana-Bobadilla, E., Molina-Villegas, A., Lopez-Arevalo, I., Reyes-Palacios, S., Muñiz-Sanchez, V., & Arreola-Trapala, J. (2020). Adaptive Geoparsing Method for Toponym Recognition and Resolution in Unstructured Text. Remote Sensing, 12(18), 3041. https://doi.org/10.3390/rs12183041

Hu, Xuke, et al. “Location Reference Recognition from Texts: A Survey and Comparison.” ACM Computing Surveys, vol. 56, no. 5, 27 Nov. 2023, pp. 1–37

Bashir, S.R.; Raza, S.; Misic, V.B. BERT4Loc: BERT for Location—POI Recommender System. Future Internet 2023, 15, 213. https://doi.org/10.3390/fi15060213

GitHub

Uhuohuy “GazPNE: Annotation-free Deep Learning for Place Name Extraction from Microblogs Leveraging Gazetteer and Synthetic Data by Rules” https://github.com/uhuohuy/GazPNE