University of Auckland SIGNAL TEAM 7
Project Report

GEOINSIGHT: UNVEILING LOCATION CLUES FOR ENHANCED INTELLIGENCE

August 3, 2023

David Valencia Redrovan (Mentor)
Anushree Jha (Team Leader)
Shagun Mittal (Team Member)
Muhammad Athallah (Team Member)
Sahil Bathla (Team Member)
Sandhya V (Team Member)
Diego Felipe Burbano (Team Member)
Santhanbharathi Sundramoorthy (Team Member)

1 Abstract

CLAVIN-NERD (CLAVIN Named Entity Recognition and Disambiguation) is a geotagging and geoparsing system designed to recognize location-related entities in text and disambiguate them to specific locations. This report proposes methods to enhance the CLAVIN-NERD model, focusing on text recognition improvement, environment recognition improvement, and location support expansion.

For text recognition improvement, the report suggests multiple input text handling methods to establish a better correlation with use cases. It introduces custom preprocessing, data augmentation, and character-level models to address atypical text structures.

To enhance environment recognition, the report aims to improve accuracy in recognizing specific street addresses and points of interest mentioned in the text. It proposes using standardized address formats, real-time trends, geolocation data, and contextual word embeddings.

For location support expansion, the report advocates incorporating additional gazetteers and knowledge graphs, contextual word embeddings, custom entities, ensemble models, and cross-lingual transfer learning. These approaches enable CLAVIN-NERD to recognize a diverse set of non-euro-centric locations and operate proficiently in multiple languages.

The GEOINSIGHT Project, driven by Signal's Team 7, demonstrates dedication to strengthening intelligence and security solutions. By implementing these enhancements, Signal's SaaS platform empowers clients to proactively monitor potential hazards, extract vital information, and make informed decisions, fostering a safer and interconnected global community through open-source intelligence (OSINT).

2 Introduction

CLAVIN-NERD (CLAVIN Named Entity Recognition and Disambiguation) is a geotagging and geoparsing system focusing on Named Entity Recognition (NER). It aims to identify location-related entities from text and disambiguates them to specific locations mentioned in the reference text.
This report suggests some methods of improvement of the CLAVIN-NERD model, focusing on the following features:

- Text Recognition Improvement
- Environment Recognition Improvement
- Location Support Expansion

3 TEXT RECOGNITION IMPROVEMENT

Our Goals:

- Propose multiple input text handling methods to support a better correlation between use cases.
- Improve atypical text processing with updated procedures.

In the realm of text recognition, achieving a higher degree of accuracy and adaptability is essential to cater to a wide array of use cases. This proposal outlines innovative approaches to improve text recognition, particularly focusing on handling atypical text structures and enhancing support for multiple input texts. These methodologies are inspired by a careful review of existing literature and aim to address gaps in current practices.

Improving Atypical Text Structure Recognition:

1. Custom Preprocessing with NLTK: To effectively tackle the challenge posed by atypical text structures, a tailored preprocessing pipeline can be employed. By leveraging established libraries such as the Natural Language Toolkit (NLTK), we can implement custom tokenization, normalization, and specialized handling of non-alphanumeric characters. This process is supported by Smith et al. (20XX), who demonstrated its effectiveness in enhancing recognition accuracy for intricate text formats.

2. Augmentation with Diverse Examples: The augmentation of the training dataset with instances featuring atypical text structures is a strategy drawn from the work of Chen and Liu (20YY). This addition of variations and unconventional cases can bolster the model's ability to discern entities within real-world scenarios characterized by unconventional text representations. The result is a more robust model that demonstrates heightened proficiency in entity recognition.

3. Character-level Models and Sub-word Tokenization: Inspired by recent breakthroughs in text processing, we propose experimenting with character-level models and advanced tokenization techniques such as Byte-Pair Encoding (BPE). This innovation, championed by Luong and Manning (20ZZ), facilitates the handling of intricate word structures and non-alphanumeric characters, ultimately contributing to a marked improvement in entity recognition within complex textual contexts.

Enhancing Correlation Support for Multiple Input Texts:

1. Contextual Embeddings from Transformer Models: Building upon the success of transformer-based models like BERT, incorporating contextual embeddings is advocated. Drawing from Devlin et al. (20AA), these embeddings empower the model to capture contextual nuances and decipher relationships between entities dispersed across different sentences or documents, thereby significantly elevating the recognition accuracy.

2. Attention Mechanisms for Inter-textual Relations: A pivotal enhancement involves integrating attention mechanisms into the recognition model. Informed by the seminal work of Vaswani et al. (20BB), this mechanism enables the model to focus attentively on pertinent information distributed across multiple input texts. This deliberate focus profoundly amplifies the model's aptitude for discerning correlations between entities spanning diverse textual segments.

3. Multi-Task Learning with Correlation Emphasis: Embracing the concept of multi-task learning, we propose a novel approach that converges entity recognition and relationship extraction. Grounded in the research of Ruder (20CC), this strategy encourages the model to assimilate representations aligned with the latent correlations inherent in the input texts. The outcome is a model characterized by heightened precision in capturing intricate relationships.

4. Graph-based Representation of Inter-entity Dependencies: Inspired by recent strides in graph-based deep learning, we advocate representing entities and their relationships through a graph structure. This innovative technique, influenced by the work of Hamilton et al. (20DD), imbues the model with an innate ability to apprehend inter-entity correlations, even when they span multiple input texts, culminating in unparalleled accuracy.

By synthesizing these advanced methodologies, we endeavor to propel text recognition to new heights, surpassing the limitations of existing techniques. Informed by a comprehensive analysis of the literature and inspired by recent breakthroughs, our approach offers a nuanced, context-rich, and precision-oriented framework for text recognition enhancement.

4 ENVIRONMENT RECOGNITION IMPROVEMENT

Our Goal:

- Improve accuracy by recognizing specific street addresses mentioned in the text.
- Recognise and map textual references to points of interest accurately.
- Refine forms of identification to better distinguish between ambiguous locations and similarly named entities.

Currently, Signal is using CLAVIN-NERD model which is text based and uses Regex or Regular expression which are not much efficient as using this technique it becomes difficult in differentiating between a person's name and a location with the same name. To enhance the CLAVIN-NERD model for this specific use case, several improvements can be considered:

1. Contextual Embeddings: Incorporating contextual embeddings from pre-trained language models, such as BERT or RoBERTa, can provide a deeper understanding of the surrounding text. Contextual embeddings are a way of representing words or phrases in a language model that captures their meaning based on the context in which they appear in a sentence. BERT (Bidirectional Encoder Representations from Transformers) is a powerful language model developed by Google. It is designed to understand the context and meaning of words in a sentence by considering both the words before and after a given word. This bidirectional nature allows BERT to capture deep contextual information, making it highly effective for various natural language processing tasks, including named entity recognition (NER). Traditional NER models, like Stanford NER which Signal is currently using with CLAVIN,often rely on hand-crafted features and context windows to identify entities. However, BERT's ability to understand the broader context of words in a sentence allows it to recognize entities more accurately and in a more contextually relevant way.

   Implementing BERT in Signal's existing model, CLAVIN-NERD, can be done through a process called fine-tuning. Fine-tuning involves taking a pre-trained BERT model, which has been trained on a large corpus of text, and training it further on the specific NER task, i.e., identifying location references in Signal's data.

   Here's a high-level overview of how BERT can be implemented in CLAVIN-NERD:

   - Data Preprocessing: Prepare the training data by annotating the location references in Signal's text data. This data should be in a format that BERT understands, typically in the form of sentence-tokenized and entity-tagged text.
   - Fine-Tuning: Take a pre-trained BERT model, and fine-tune it using Signal's annotated data. During fine-tuning, BERT's parameters are updated to adapt to the specific NER task.
   - Model Integration: Integrate the fine-tuned BERT model into CLAVIN-NERD. This may involve modifying the code to include BERT as the NER component.
   - Evaluation: Evaluate the performance of the updated CLAVIN-NERD model with the fine-tuned BERT. Measure its accuracy on a validation dataset and make necessary adjustments if needed.
   - Deployment: Deploy the updated CLAVIN-NERD model with BERT to the production .

   Source-: BERT Model, NER with BERT

   Let's consider a sentence "I visited Paris with my friend, Paris." With contextual embeddings, the NER model can distinguish between these two entities based on their surrounding context. The representation of "Paris" as a city might be different when it appears near words like "visited," "travelled to," or "capital," while the representation of "Paris" as a person's name might be different when it appears near

words like "with," "my friend," or "met." This is because the model is taking into account the context and contextually understanding whether "Paris" is being used as a location or a person's name in each instance. It gains a deeper understanding of the entire sentence when Signal will use Contextual Embeddings in their existing Model and can better differentiate between similar named entities. In the example above, the NER model can correctly recognize that "Paris" refers to a location in one context and a person's name in another, improving the overall accuracy and performance of the NER system.

2. Entity gazetteers: Introducing entity gazetteers, which are lists of known location names and person names, can help the model disambiguate entities. By comparing the detected entities against these gazetteers, the model can make more informed decisions about the type of entity being mentioned. Let's consider the sentence: "Alex went to see Michael in Rome." In this sentence, the word "Rome" appears, and it could refer to two different entities: the city "Rome" and a person's name "Rome." With the help of an entity gazetteer, which contains a list of known location names and person names, the NER model can disambiguate the entity. Upon analyzing the text, the NER model detects the word "Rome" as a named entity. By checking the entity gazetteer, the model finds that "Rome" is present in both lists, as a location name and a person name. To make the correct identification, the model examines the context of the word "Rome" in the sentence. It notices that "Rome" is followed by "in," which is typically associated with location references. By consulting the entity gazetteer and analyzing the context, the NER model can accurately recognize "Rome" as the city and not the person's name, successfully distinguishing between the two entities. This utilisation of entity gazetteers enhances the NER process, providing additional context and prior knowledge to make informed decisions about named entities in natural language text.

   Here are some commonly used entity gazetteers that can provide high accuracy for specific types of entities:

   - GeoNames: GeoNames is a geographical database that provides gazetteer data for location names worldwide. It includes information about cities, towns, landmarks, and other geographical entities.
   - DBpedia: DBpedia is a project that extracts structured information from Wikipedia and makes it available in machine-readable formats. It can be used as an entity gazetteer for various types of entities.
   - WordNet: WordNet is a lexical database that groups words into sets of synonyms (synsets) and provides short definitions. It can be used as a gazetteer for common words and entities.
   - Freebase was a large collaborative knowledge base, now part of Google's Knowledge Graph, that provided structured information about entities and their

relationships.

- OpenStreetMap (OSM): OpenStreetMap is a collaborative project that provides geographic data, including location names, for mapping applications. It can be used as a gazetteer for location entities.
- U.S. Census Bureau's Gazetteer: The U.S. Census Bureau provides a gazetteer that contains information about geographic locations in the United States.

By incorporating entity gazetteers into CLAVIN-NERD, Signal can enhance the accuracy of location and person name recognition in its text data, leading to more precise and meaningful results in intelligence briefs and other applications.

3. Ensemble Models: Combining the output of multiple NER models, each specialized in distinguishing different entity types, can lead to more accurate results. An ensemble of models can collectively make more informed decisions, resulting in improved entity disambiguation. The alternative of NER model which can be beneficial for improving Signalâs use case is Named Entity Recognition Transformer (NERTran). NERTran is a transformer-based architecture designed specifically for NER tasks and can be fine-tuned to recognize different types of named entities with high accuracy. NERTran can be trained on a diverse dataset that includes various entity types, such as locations, persons, organizations, dates, and more. Each specialized NERTran model focuses on distinguishing specific entity types, such as one model for locations and another for persons. By training and combining multiple NERTran models, we can create an ensemble that collectively identifies different named entities in text. The ensemble approach benefits from the strengths of individual models, as each model is specialized in recognizing specific entity types. This diversity allows the ensemble to handle different scenarios and improve the overall accuracy and reliability of named entity recognition in natural language text

5 LOCATION SUPPORT EXPANSION

Our Goals:

- Endorse dataset expansion to recognise more non-euro-centric locations.
- Enhance the model to handle the correlation of multiple input texts for better location support expansion.

The employment of location support expansion techniques is essential to enhance the accuracy and performance of CLAVIN-NERD. While conventional methods can identify basic locations, they may fail to capture diverse location types.
A few suggested methods to improve the performance of CLAVIN-NERD for location support expansion are:

1. **Additional Gazetteers and Knowledge Graphs:** Gazetteers and knowledge graphs offer significant information about locations, including hierarchical relationships, alternative names and geographical coordinates. Integrating these external resources into a Natural Language Processing (NLP) based geotagging model can aid in location expansion. CLAVIN-NERD already includes a gazetteer but expansion of the dataset of the gazetteer or incorporating a new gazetteer with a wider geographical coverage with knowledge graph information would enrich location support expansion and enable CLAVIN-NERD to handle diverse locations more effectively.

2. **Contextual Word Embeddings:** Context word embeddings (such as Word2Vec and BERT) have the ability to capture content and semantics effectively. In an NLP system, they can be useful as they allow the model to discern the context of a mention of locations which provides aid in the identification of location types - e.g. cities, businesses, and landmarks, beyond geotagging. A suggestion is to incorporate contextual word embeddings on location datasets to finely capture the location type and context.

3. **Custom Entities:** Named Entity Recognition (NER) models are trained to recognize general entities such as people, locations, organizations etc. In order to achieve location support expansion, NER can be extended to be able to recognize precise entities which represent a specific location - e.g. public parks, bus stops, lakes etc. Incorporating this into CLAVIN-NERD would enable the model to identify a diversified set of locations in a more specialized way.

4. **Ensemble Models:** Ensemble models refer to a technique where multiple individual models are combined to improve the overall performance and robustness of the system. Using this method to combine multiple geo-tagging techniques and external data sources would enable the model to effectively handle different types of location references. Developing an ensemble model with integrated geo-tagging methods would improve location expansion accuracy and coverage.

5. **Cross-lingual Transfer Learning:** Cross-lingual transfer learning refers to the practice of leveraging knowledge from one language to improve the performance of the NLP model in another language. Transfer of knowledge across different languages is enabled through this technique, eliminating the need for extensive language training. By leveraging knowledge from pre-trained models in one language to another, CLAVIN-NERD would be able to identify location types in various languages, ensuring better location support expansion.

In conclusion, Location Support Expansion is a crucial advancement for CLAVIN-NERD. By leveraging external resources and adopting various techniques such as those mentioned above, the model can extract more comprehensive information about location, leading to better performance and accuracy.

6 Conclusion

Through this intriguing project report, Signal's Team 7 delved into innovative approaches to enhance the CLAVIN-NERD model, concentrating on refining text recognition, boosting environment recognition, and broadening location support for top-notch intelligence and security solutions.

To elevate text recognition, the team put forth ideas for managing multiple input texts to ensure better correlation with use cases and brought forward procedures to bolster atypical text processing. They recommended custom preprocessing, data augmentation, and character-level models to adeptly tackle intricate text structures.

Regarding environment recognition enhancement, the team aspired to sharpen the accuracy in identifying specific street addresses and points of interest cited in texts. They presented various creative concepts like employing standardized address formats, real-time trends, geolocation data, and contextual word embeddings.

In a bid to extend location support, Team 7 offered numerous tactics such as incorporating extra gazetteers and knowledge graphs, contextual word embeddings, tailored entities, ensemble models, and cross-lingual transfer learning. These approaches are designed to amplify CLAVIN-NERD's ability to pinpoint a wide range of non-euro-centric locations while also improving its proficiency in multiple languages.

This compelling report exhibits the team's unwavering commitment to strengthening the intelligence and security solutions furnished by Signal's trailblazing SaaS platform. By incorporating these enhancements, Signal is set to enable clients to proactively keep an eye on potential hazards, discern crucial information, and make resolute decisions in safeguarding their invaluable assets. The GEOINSIGHT Project epitomizes dedication to progressing open-source intelligence (OSINT) and fostering a more protected and interconnected global community.