Digital Assignment – 1

NATURAL LANGUAGE PROCESSING

SURVEY

TEAM MEMBER - 1

NAME: OMPRAKASH

REG NO: 21BCE1950

TEAM MEMBER - 2

NAME : MOHAMED RIYAAS

REG NO: 21BCE5828

TITLE:

NAMED ENTITY RECOGNITION (NER) MODEL:

1) Using Bidirectional Encoder Representations from Transformers(BERT) to classify traffic crash severity types

Amir Hossein Oliaeea, Subasish Dasb, Jinli Liuc, M. Ashifur Rahmand

The application of Bidirectional Encoder Representations from Transformers (BERT) for classifying traffic crash severity types has garnered attention in recent research. The study by demonstrates how BERT's contextual embeddings enhance the classification accuracy of crash severity by effectively capturing the nuances in textual data related to traffic incidents. The authors emphasize the model's ability to leverage large datasets and fine-tuning techniques, resulting in improved performance over traditional machine learning methods. This approach showcases BERT's potential in real-world applications, particularly in traffic safety analysis.

2)Semantics aware intelligent framework for content-based e-learning recommendation

Hadi Ezaldeen, Sukant Kishoro Bisoy, Rachita Misra, Rawaa Alatrash

It integrates semantic analysis with machine learning techniques to enhance the personalization of learning experiences. By leveraging user preferences and content characteristics, the framework aims to improve the accuracy of recommendations. The authors highlight the importance of understanding the semantics of educational content to provide tailored suggestions, ultimately aiming to increase learner engagement and effectiveness in e-learning environments. This approach represents a significant advancement in adaptive learning technologies.

3)Rare words in text summarization

Danila Morozovskii, Sheela Ramanna

Their study proposes novel techniques for enhancing summarization effectiveness by leveraging contextual embeddings and semantic analysis. The authors emphasize the importance of addressing rare word occurrences to improve the quality of generated summaries, ultimately aiming to create more coherent and informative outputs that retain the essence of the original text.

4)Predicting Facebook sentiments towards research

Murtuza Shahzad, Cole Freeman, Mona Rahimi, Hamed Alhoori

By employing machine learning techniques, the authors analyze user-generated content to discern patterns in sentiment expression. The study highlights the potential of social media as a valuable resource for understanding public perceptions of research, providing insights that can inform researchers and policymakers about community engagement and the impact of academic work on societal views.

5) Improving aspect-based sentiment analysis with contrastive learning

Lingling Xu, Weiming Wang*

Their method focuses on distinguishing between similar sentiments associated with different aspects, thereby improving the model's ability to identify nuanced opinions. The authors demonstrate that contrastive learning significantly boosts performance on benchmark datasets, showcasing its effectiveness in refining sentiment classification and providing more accurate insights into consumer opinions on specific product features.

6) DzNER: A large Algerian Named Entity Recognition dataset

Abdelhalim Hafedh Dahoua, Mohamed Amine Cheraguib

The dataset includes a diverse range of entities relevant to the region, facilitating the development of localized NER models. The authors emphasize the importance of such datasets in improving NER performance for underrepresented languages and domains, ultimately contributing to the advancement of NLP applications tailored to specific cultural and linguistic contexts.

7) Do cues in a video help in handling rare words in a machine translation system underalow-resource setting?

Loitongbam Sanayai Meeteia, Alok Singhb, Thoudam Doren Singha, Sivaji Bandyopadhyayc

The authors explore how visual information can complement textual data to better translate uncommon words and phrases, addressing challenges posed by limited parallel corpora. The study highlights the benefits of multimodal approaches in enhancing machine translation performance for under-resourced language pairs and domains.

8) Development and verification ofauser-friendly software for Germantext simplification focused on patients with cerebral palsy

Simon Strübbea, Irina Sidorenkoa, Susmita Roya, Alexander T.D. Grünwalda, Renée Lampea,b

Emphasize the importance of accessible content and the challenges faced by this target audience in comprehending complex language. The software incorporates various simplification techniques and is evaluated for usability and effectiveness, contributing to improved reading experiences for patients with cerebral palsy.

9)Detecting abusive comments at a fine-grained level in a low-resource language

Bharathi Raja Chakravarthia, Ruba Priyadharshinib, Shubanker Banerjeec, Manoj Balaji Jagadeeshand, Prasanna Kumar Kumaresane, Rahul Ponnusamyf, Sean Benhurg, John Philip McCraee

Develop robust models capable of identifying different types of abusive comments, such as hate speech, profanity, and cyberbullying, even in the absence of large annotated datasets. The study highlights the importance of addressing online harassment in underrepresented languages and proposes techniques to overcome data scarcity challenges.

10) Named Entity Recognition— datasets,tools,and methodologies

Basra Jehangir, Saravanan Radhakrishnan, Rahul Agarwal

Emphasize the significance of NER in diverse applications, ranging from social media to biomedical domains, and address the challenges faced by NER systems. The survey serves as a valuable resource for researchers and practitioners interested in the field of NER.

11) ASPER:Attention-based approach to extract syntactic patterns denoting semantic relations in sentential context

Md. Ahsanul Kabira,∗, Tyler Phillipsa, Xiao Luob, Mohammad Al Hasana

Demonstrate the effectiveness of their method in identifying and classifying semantic relationships, such as part-whole and causal relations, using attention mechanisms. ASPER's ability to capture contextual information and handle complex linguistic structures contributes to improved performance in semantic relation extraction tasks.

12) A comprehensive review of State-of-The-Art methods for Javacode generation from Natural Language Text

Jessica López Espejel∗, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, Walid Dahhane, El Hassane Ettifouri

Discuss various approaches, including rule-based systems, statistical models, and deep learning techniques, highlighting their strengths and limitations. The survey provides insights into the challenges and recent advancements in the field of code generation from textual descriptions, aiming to facilitate the development of more efficient and accurate code generation systems.

13)Recent advances in named entity recognition

Imed Keraghel, Stanislas Morbieu, Mohamed Nadif

Various methodologies, including traditional rule-based approaches and modern deep learning techniques, highlighting the evolution of NER systems. They emphasize the importance of datasets and tools used in NER, as well as the challenges faced in accurately identifying entities across different domains. The survey aims to synthesize current research trends and propose future directions for enhancing NER effectiveness.

14)Named entity recognition and classification

David Nadeau, Satoshi Sekine

They categorize NERC systems based on the type of machine learning employed, such as supervised, semi-supervised, and unsupervised learning. The survey also examines the impact of language and domain on NER performance, as well as the challenges posed by ambiguous and nested entities. Nadeau and Sekine emphasize the importance of NER in various applications, including information extraction, question answering, and text summarization.

15)Named Entity Recognition: A Survey for Indian Languages

Krishnanjan Bhattacharjee, Shiva Karthik S

Discuss the unique linguistic challenges posed by the diverse scripts and dialects in India, emphasizing the need for language-specific models and datasets. They review various methodologies, including rule-based and machine learning approaches, and highlight the importance of annotated corpora for training NER systems.

16)A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts

Priyankar Bose

The challenges of handling domain-specific terminology and the importance of context in medical data. The author examines various methodologies, including deep learning approaches, and their effectiveness in improving the accuracy of entity recognition and relationship extraction in clinical settings.

17)Active learning with feature matching for clinical named entity recognition

panelLinh Le , Gianluca Demartini , Guido Zuccon , Genghong Zhao , Xia Zhang

Present a method that iteratively selects the most informative samples for labeling, enhancing model performance with minimal labeled data. They demonstrate the effectiveness of their approach through experiments, showing significant improvements in NER accuracy for clinical texts.

18)An evaluation of Google Translate for Sanskrit to English translation via sentiment and semantic analysis

panelAkshat Shukla , Chaarvi Bansal , Sushrut Badhe , Mukul Ranjan, Rohitash Chandr
Analyze translation accuracy by comparing translated outputs with original sentiments and meanings, highlighting challenges in handling the nuances of Sanskrit. The study reveals that while Google Translate performs well in general, it struggles with context-specific translations and rare vocabulary, suggesting areas for improvement in machine translation systems for low-resource languages.

19) Bipol: A novel multi-axes bias evaluation metric with explainability for NLP

lLama Alkhaled , Tosin Adewumi , Sana Sabah Sabry

The importance of explainability in understanding model behavior and mitigating bias effects. Bipol provides a structured framework for evaluating bias in various NLP tasks, enabling researchers to identify and address potential biases in model outputs.

20) Named Entity Recognition and Relation Detection for Biomedical Information Extraction

Nadeesha Perera ,Matthias Dehmer,Frank Emmert-Streib[1]

Review existing methodologies and propose a framework that integrates NER with relation extraction to enhance the identification of complex biomedical relationships. The study highlights the challenges posed by domain-specific terminology and the need for robust models to improve the accuracy of biomedical information retrieval, ultimately contributing to better healthcare outcomes.

A SURVEY:

Named Entity Recognition (NER) is a natural language processing (NLP) task that involves identifying and classifying named entities in text into predefined categories such as person names, organizations, locations, dates, and more. It's a crucial step in various NLP applications like information extraction, question answering, and sentiment analysis. BERT (Bidirectional Encoder Representations from Transformers) is a powerful pre-trained language model developed by Google that has revolutionized NLP tasks. It utilizes a transformer architecture, allowing it to capture bidirectional contextual information from input text, which is particularly beneficial for tasks like NER. In NER with BERT, the model is fine-tuned on labeled NER datasets, where it learns to predict the entity type for each token in the input text. By leveraging BERT's contextual embeddings and fine-tuning on task-specific data, NER models achieve state-of-the-art performance, surpassing traditional methods.

TOOLS/TECHNIQUES USED:

SpaCy: SpaCy is an open-source natural language processing (NLP) library designed for efficient and fast processing of natural language text. Developed by Explosion AI, SpaCy is written in Python and is widely used for various NLP tasks, including tokenization, part-of-speech tagging, named entity recognition, syntactic parsing, and more. It is known for its simplicity, speed, and state-of-the-art capabilities.

Methodologies

NER methodologies can be broadly categorized into:

- Rule-Based Approaches: These rely on handcrafted rules and patterns to identify entities.

- Supervised Learning: This approach uses labeled datasets to train models, allowing them to learn patterns and make predictions on unseen data.

- Unsupervised Learning: In this method, models identify entities without labeled data, often using clustering techniques.

- Deep Learning: Recent advancements have seen the adoption of deep learning techniques, including Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) networks, which have significantly improved NER performance.

Datasets for NER

Several datasets are crucial for training and evaluating NER systems:

- CoNLL Datasets: These include CoNLL-2002 and CoNLL-2003, which provide multilingual data focused on various entity types.
- WNUT-2017: This dataset is tailored for social media, containing diverse entity categories from user-generated content.
- OntoNotes: A large corpus annotated across multiple languages and genres, useful for training robust NER models.
- Biomedical Datasets: Specific datasets like NCBI Disease and BioCreative focus on identifying entities relevant to the biomedical domain.

Tools and Frameworks

Several tools facilitate NER implementation:

- SpaCy: A popular Python library that offers pre-trained models and easy integration for NER tasks.
- Stanford NER: A Java-based tool that provides a robust framework for NER with customizable models.
- Transformers: Libraries like Hugging Face's Transformers offer state-of-the-art pre-trained models that can be fine-tuned for specific NER tasks.

Challenges and Future Directions

Despite advancements, NER systems face challenges such as:

- Ambiguity: Entities can have multiple meanings depending on context, complicating identification.
- Domain Adaptation: Models trained on general datasets may not perform well in specialized domains without additional training.
- Data Scarcity: High-quality annotated datasets are often limited, hindering model development.

Future research may focus on improving transfer learning techniques, enhancing model interpretability, and developing more robust systems capable of handling diverse languages and domains.In conclusion, NER is a foundational component of NLP with numerous applications across various fields. Ongoing research and advancements in methodologies and tools continue to enhance its effectiveness and applicability.