

NAMED ENTITY RECOGNISATION

Sairam Thadakamalla, Dr. Narsappa Reddy

¹B.tech Student, Department Of Electronics and Computer Engineering, J.B Institute of Engineering and Technology

²Associate Professor and HOD, Department Of Electronics and Computer Engineering, J.B Institute of Engineering and Technology

Abstract: With the increase in availability of data, extraction of useful information from this data has become most important activity across all domains. When the data is available as documents written in natural language, information extraction becomes more challenging. Named Entity recognition (NER) is a technique used extensively for automatic extraction of useful information from unstructured natural language document collections. Used both, for web applications as well as stand-alone systems, NER is considered as one of the major step in Natural Language Processing (NLP) for analysis of text. This paper discusses basics of NER, various algorithms used for NER and major applications and challenges in the field of NER.

Keywords: Named Entity Recognition, Information Extraction, Natural Language Processing.

I. INTRODUCTION

Named Entity Recognition (NER) is a natural language processing (NLP) task that involves identifying and classifying entities (objects, people, locations, organizations, dates, monetary values, percentages, etc.) in text data. The goal of NER is to extract structured information from unstructured text and assign predefined labels to specific entities.

The problem can be formally defined as follows:

Given a sequence of words (tokens) in a text document, the task of Named Entity Recognition is to identify and classify spans of words that correspond to specific named entities. Each identified entity is assigned a label from a predefined set of categories, such as person names, organizations, locations, dates, etc.

II. LITERATURE SURVEY

Richa Sharma, Sudha Morwal, Basant Agarwal published the paper “Named Entity Recognition” (NER) plays an important role in various Natural Language Processing (NLP) applications to extract the key information from a huge amount of unstructured text data. NER is a task of identifying and classifying the named entities into predefined categories for a given text. Recently, language models are highly appreciable in several NLP tasks as these state-of-the-art models result better even in resource scarcity. In this paper, we perform NER task on the Hindi language by incorporating the recently released multilingual language model MuRIL which stands

for Multilingual Representation for Indian Languages. MuRIL is specially trained for 16 Indian languages. We develop a Hindi NER system using MuRIL with a conditional random field (CRF) layer and fine-tune the model on the ICON 2013 Hindi NER dataset.

Rachna Jain, Abhishek Sharma, Gouri Sankar Mishra, Parma Nand and Sudeshna Chakraborty published the paper “Named entity Recognition” is a word or an expression that unmistakably recognizes one thing from a lot of different things that have comparable qualities. In the articulation named element, the word named limits the extent of substances that have one or numerous unbending designators that represents a referent. Typically, Rigid designators incorporate legit names, however it relies upon area of intrigue that may allude the reference word for object in space as named substances. For instance, in sub-atomic science and bio-informatics, substances of intrigue are qualities and quality items.

Basra Jehangir, Saravanan Radhakrishnan, Rahul Agarwal published the paper “Named entity Recognition” Currently, the data is typically presented in its raw form (unstructured, native language, confusing), from various sectors of our economy, government, and private and public lives so summarizing, searching, drawing conclusions, and doing statistical analysis are all challenging tasks for humans. We perform various NLP operations on the text to complete the abovementioned tasks. This makes NLP crucial in today’s data processing. One of the operations is Named Entity Recognition.

Nadeesha Perera, Matthias Dehmer, Frank Emmert-Streib published the paper “Named entity Recognition” The number of scientific publications in the literature is steadily growing, containing our knowledge in the biomedical, health, and clinical sciences. Since there is currently no automatic archiving of the obtained results, much of this information remains buried in textual details not readily available for further usage or analysis. For this reason, natural language processing (NLP) and text mining methods are used for information extraction from such publications. In this paper, we review practices for Named Entity Recognition (NER) and Relation Detection (RD), allowing, e.g., to identify interactions between proteins and drugs or genes and diseases. This information can be integrated into networks to summarize large-scale details on a particular biomedical or clinical problem, which is then amenable for easy data management and further analysis. Furthermore, we survey novel deep learning methods that have recently been introduced for such tasks.

Zhen sun , xinfu Li published the paper “Named entity recognition” can deeply explore semantic features and enhance the ability of vector representation of text data. This paper proposes a named entity recognition method based on multi-head attention to aim at the problem of fuzzy lexical boundary in Chinese named entity recognition. Firstly, Word2vec is used to extract word vectors, HMM is used to extract boundary vectors, ALBERT is used to extract character vectors, the Feedforward-attention mechanism is used to fuse the three vectors, and then the fused vectors representation is used to remove features by BiLSTM. Then multi-head attention is used to mine the potential word information in the text features. Finally, the text label classification results are output after the conditional random field screening. Through the verification of WeiboNER, MSRA, and CLUENER2020 datasets, the results show that the proposed algorithm can effectively improve the performance of named entity recognition.

Block Diagram

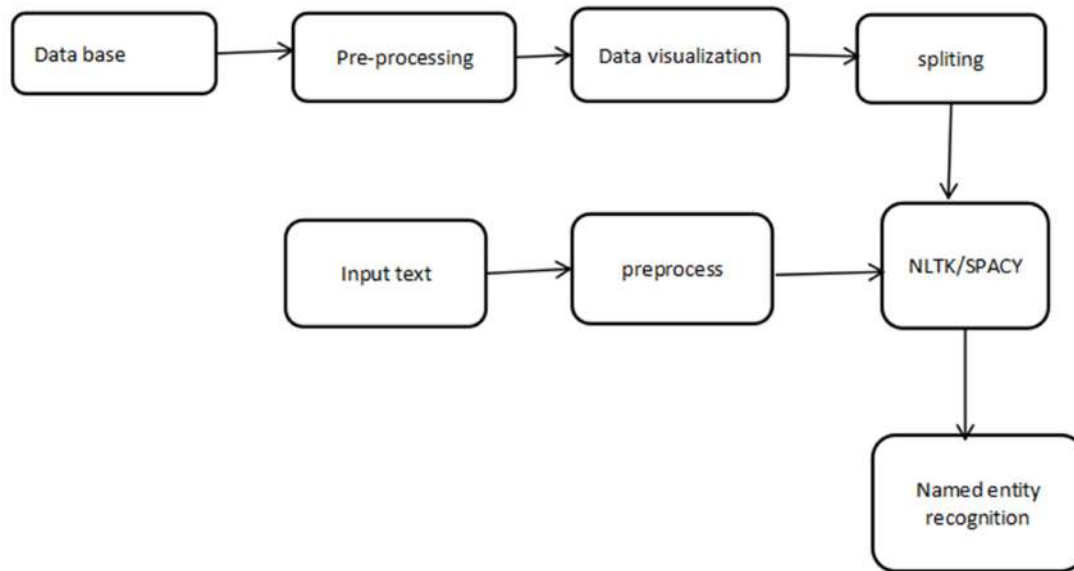


Fig 2.3.1: Block Diagram of Existing System

Proposed System

Algorithms are well-suited to classifying, processing and making predictions based on time series data. Evaluation metrics like Precision, Recall, and F1-score would gauge model performance, with testing on unseen data for generalization. The deployment of this model could be as part of an NLP pipeline or a standalone application.

Continuous improvement would be achieved through periodic model updates with new data to enhance accuracy and keep up with evolving language patterns.

III. ANALYSIS

The "Analysis of Named Entity Recognition" project is a comprehensive exploration of the intricacies and methodologies employed in the field of Named Entity Recognition (NER). NER plays a pivotal role in Natural Language Processing (NLP), involving the identification and classification of entities like names of individuals, organizations, locations, dates, and specific terms within a given text. The project aims to dissect various aspects of NER, ranging from traditional rule-based approaches to modern machine learning techniques. This includes an in-depth examination of algorithms, exploration of diverse datasets, evaluation metrics, and an investigation into challenges and limitations associated with NER systems. The project also scrutinizes recent advancements, such as the integration of contextual embeddings and transformer-based models, and their impact on the state-

of-the-art in NER. By analyzing practical applications across different domains, the project seeks to contribute valuable insights to researchers and practitioners, fostering a deeper understanding of the evolving landscape of NER in natural language processing.

IV. DESIGN

The "Design of Named Entity Recognition Project" is motivated by the imperative to address the evolving complexities of linguistic nuances and diverse contextual variations inherent in vast and varied datasets. This undertaking goes beyond the conventional boundaries of entity recognition, extending into the incorporation of cutting-edge methodologies that harness the power of machine learning and artificial intelligence. The project's architecture is meticulously crafted to accommodate the dynamic nature of language, enabling the system to adapt to emerging patterns and trends. It embraces a hybrid approach, combining the structured rule-based techniques with the ability of machine learning models to capture intricate contextual dependencies.

In the design phase, careful attention is given to data preprocessing, ensuring that the input text is transformed into a format conducive to effective entity recognition. Feature extraction techniques, such as word embeddings and contextual embeddings, are explored to enhance the system's understanding of the semantics and relationships within the text. The project also places a strong emphasis on model training, validation, and fine-tuning, utilizing a diverse and well-annotated dataset to bolster the system's learning capabilities.

Moreover, the design incorporates measures for post-processing refinement, addressing potential challenges in entity boundary cases and refining the system's outputs for improved precision. Scalability and efficiency are integral considerations, allowing the designed NER system to handle both small-scale and large-scale datasets seamlessly.

Ultimately, the ambition of the "Design of Named Entity Recognition Project" extends beyond the mere identification of entities; it seeks to pave the way for a sophisticated and adaptable solution capable of navigating the intricacies of language, thereby contributing to the advancement of natural language processing and its myriad applications. Through a meticulous and thoughtful design process, this project endeavors to provide a foundation for NER systems that not only meet current challenges but also remain flexible and resilient in the face of evolving linguistic landscapes.

Class Diagram

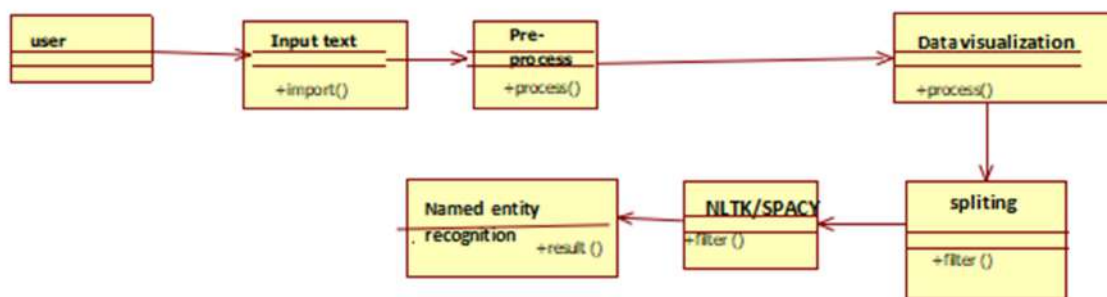


Fig 4.2.2: Class Diagram

Activity Diagram

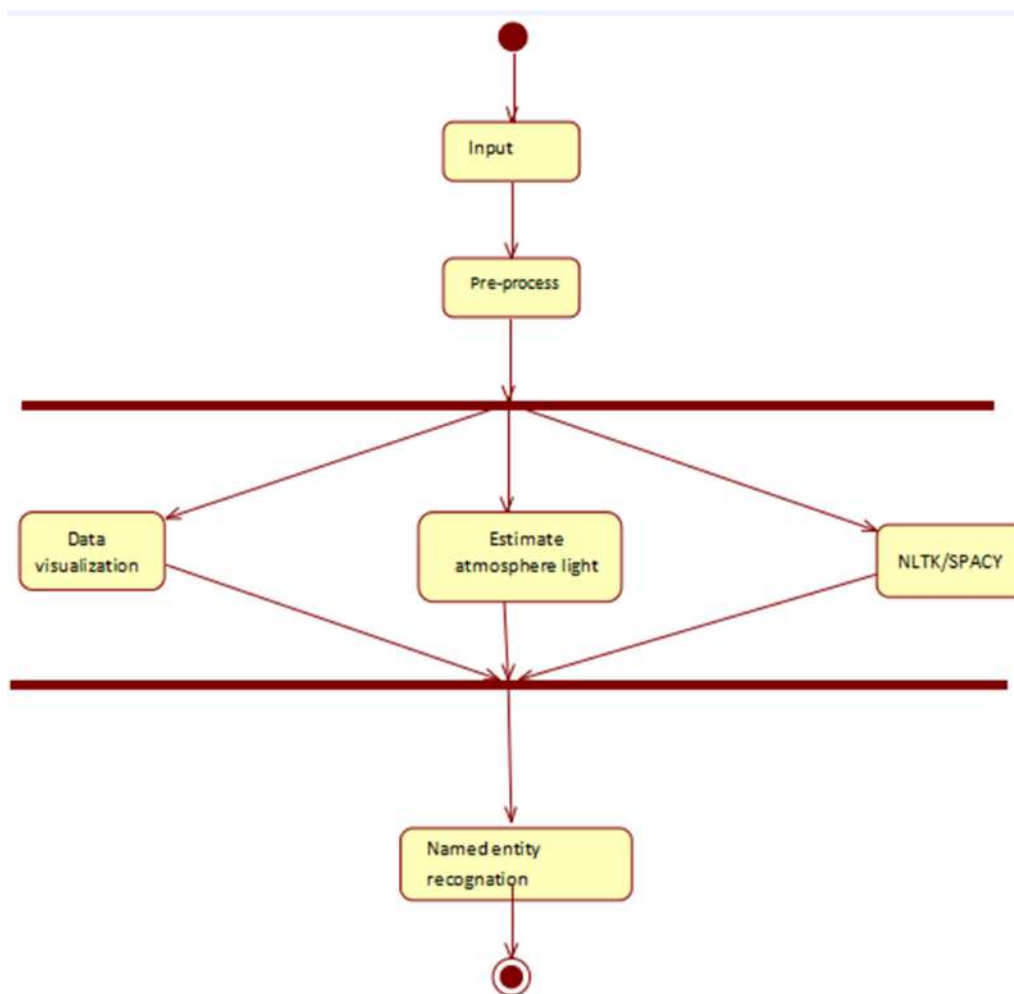


Fig 4.2.5: Activity Diagram

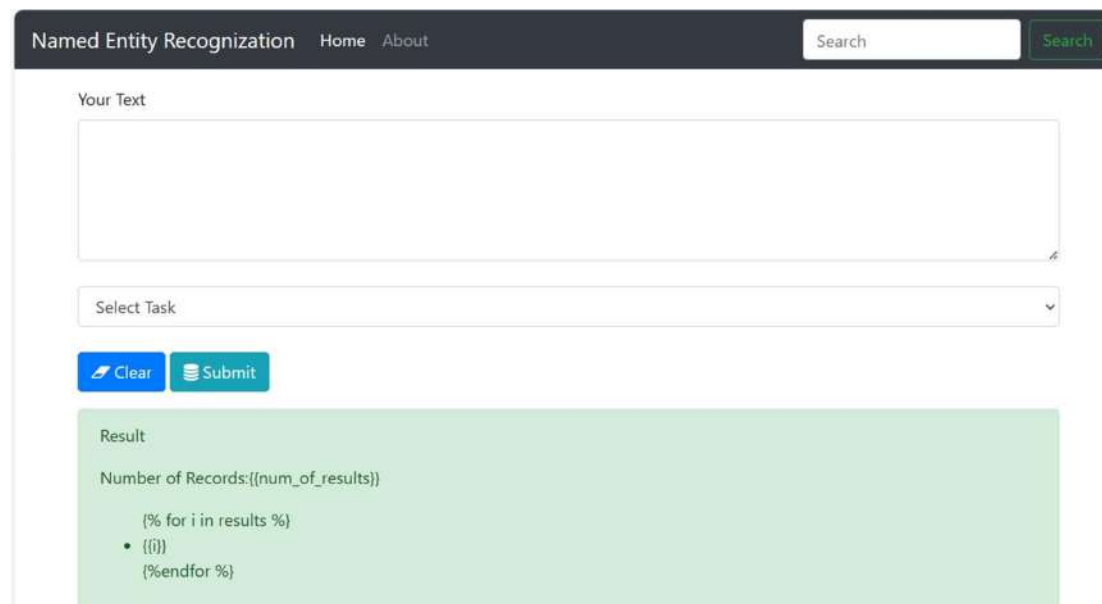
Activity diagrams are graphical representations of Workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

V. IMPLEMENTATION AND RESULTS

The chapter tells about the implementation part of the Entity Recognition. In the codebase, you'll find key components like data preprocessing, where the input text is tokenized, cleaned, and transformed into a suitable format for training. The heart of the project lies in selecting and training an appropriate NER model, such as spaCy's pre-trained models or a custom-built solution using machine learning frameworks like scikit-learn or TensorFlow.

OUTPUT SCREENS

Main page:



The screenshot shows a web application titled "Named Entity Recognition". It has a dark header bar with "Home" and "About" links. A search bar with a "Search" button is on the right. Below the header, there is a large text input area labeled "Your Text". Below the input area is a dropdown menu labeled "Select Task". At the bottom of the input section are two buttons: "Clear" and "Submit". Below these buttons is a green box labeled "Result" containing the following text: "Number of Records:{{num_of_results}}", "{% for i in results %}", a bulleted list item "• {{i}}", and "{%endfor %}".

Fig 5.4.1: Output Template

This is the basic output of the code in this we use to give the input and select entity and by clicking the submit button we will get output.

Named Entity Extractor Home About

Search

Your Text

Prime Minister Justin Trudeau of Canada promised a fresh approach to politics, one that was based on openness, decency and liberalism. Now he is embroiled in a scandal involving accusations of back-room deal-making and bullying tactics, all to support a Canadian company accused of bribing the Libyan government when it was run by the dictator Muammar el-Qaddafi. Canadian newspapers are filled with outrage and opposition parties are calling for a resignation. Elections are still seven months away, but some members of Mr. Trudeau's own governing party fear the scandal has armed opposition parties with rich campaign fodder against its leader, who promised "sunny ways" in politics. "This is a huge, huge blow to Justin Trudeau's

Person

Clear Submit

Result

Number of Records:9

- Justin Trudeau
- Muammar el-Qaddafi
- Trudeau
- Justin Trudeau's
- Justin Trudeau
- Shachi Kurl
- Jody Wilson-Raybould
- Trudeau
- Wilson-Raybould

Fig 5.4.2: Output Template of Person

In the above Figure we have given the input and selected the entity **PERSON** and it displays the output of the person names.

Named Entity Extractor Home About

Search

Your Text

Prime Minister Justin Trudeau of Canada promised a fresh approach to politics, one that was based on openness, decency and liberalism. Now he is embroiled in a scandal involving accusations of back-room deal-making and bullying tactics, all to support a Canadian company accused of bribing the Libyan government when it was run by the dictator Muammar el-Qaddafi. Canadian newspapers are filled with outrage and opposition parties are calling for a resignation. Elections are still seven months away, but some members of Mr. Trudeau's own governing party fear the scandal has armed opposition parties

Money

Clear Submit

Result

Number of Records:2

- 47.7 million Canadian dollars
- 129.8 million Canadian dollars

Fig 5.4.3: Output Template of Money

In the above Figure we have given the input and selected the entity **MONEY** and it displays the output of the Money.

Named Entity Extractor Home About

Search

Your Text

Prime Minister Justin Trudeau of Canada promised a fresh approach to politics, one that was based on openness, decency and liberalism. Now he is embroiled in a scandal involving accusations of back-room deal-making and bullying tactics, all to support a Canadian company accused of bribing the Libyan government when it was run by the dictator Muammar el-Qaddafi. Canadian newspapers are filled with outrage and opposition parties are calling for a resignation. Elections are still seven months away, but some members of Mr. Trudeau's own governing party fear the scandal has armed opposition parties

Organization

Clear Submit

Result

Number of Records:3

- the Angus Reid Institute
- SNC-Lavalin
- Liberal Party's

Fig 5.4.4: Ouput Template of an Organization

In the above Figure we have given the input and selected the entity **ORGANIZATION** and it displays the output of the Organization names.

Named Entity Extractor Home About

Search

Your Text

Prime Minister Justin Trudeau of Canada promised a fresh approach to politics, one that was based on openness, decency and liberalism. Now he is embroiled in a scandal involving accusations of back-room deal-making and bullying tactics, all to support a Canadian company accused of bribing the Libyan government when it was run by the dictator Muammar el-Qaddafi. Canadian newspapers are filled with outrage and opposition parties are calling for a resignation. Elections are still seven months away, but some members of Mr. Trudeau's own governing party fear the scandal has armed opposition parties with rich campaign fodder against its leader, who promised "sunny ways" in politics. "This is a huge, huge blow to Justin Trudeau's

Geopolitical

Clear Submit

Result

Number of Records:5

- Canada
- Vancouver
- Quebec
- Libya
- Canada

Fig 5.4.5: Output Template of Geopolitical

In the above Figure we have given the input and selected the entity **GEOPOLITICAL** and it displays the output of the Locations.

VI. CONCLUSION

In this paper, we have tried to give the information about NER techniques, tools and algorithms in history, state-of-the-art current and few future working. In this article helps the new researchers to gain information about named entities issues and solutions. In this survey, introduce of Named Entity Recognition and also compared techniques, tools and algorithms. This paper provides briefly review of learning based systems, rule-based systems and hybrid NER systems all this information is available in tabular form and also talk about these systems in detail. In this article we compared the different techniques which are Spacy, StanfordNLP, TensorFlow and ApacheOpenNLP in news corpus. The Spacy give good results and less predication time as compares the other techniques. The evaluation measures on the base of training accuracy, model size, time prediction, training loss data and Fmeasure discussed in detail. At the rest of paper some future directions are also be provided so that this NER research field will explore continue.

REFERENCES

- [1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” *Advances in neural information processing systems*, pp. 3111–3119, 2013.
- [2] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global Vectors for Word Representation,” *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014.
- [3] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching Word Vectors with Subword Information,” *Transactions of the Association for Computational Linguistics*, vol. 5, no. 2307-387X, pp. 135–146, 7 2017.
- [4] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, “Deep learning with word embeddings improves biomedical named entity recognition,” *Bioinformatics*, vol. 33, no. 14, pp. i37–i48, 7 2017.
- [5] A. M. Dai and Q. V. Le, “Semi-supervised Sequence Learning,” 11 2015.
- [6] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” *arXiv preprint arXiv:1802.05365*, 2 2018.
- [7] J. Howard and S. Ruder, “Universal Language Model Fine-tuning for Text Classification,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 328–339.
- [8] A. Radford, “Improving Language Understanding by Generative Pre-Training,” in URL [https://s3-us-west-2.amazonaws.com/openaiassets/researchcovers/languageunsupervised/language understanding paper. pdf,](https://s3-us-west-2.amazonaws.com/openaiassets/researchcovers/languageunsupervised/language%20understanding%20paper.pdf) 2018.
- [9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” In *Advances in neural information processing systems*, pp. 5998–6008, 6 2017.

- [10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in arXiv preprint arXiv:1810.04805, 10 2018.
- [11] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual String Embeddings for Sequence Labeling," in COLING, 2018.
- [12] T. Eftimov, B. Korous Seljak, and P. Korošec, "A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations," PLoS ONE, vol. 12, no. 6, 2017.
- [13] T.-V. T. Nguyen, A. Moschitti, and G. Riccardi, "Kernel-based reranking for named-entity extraction," in Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics (ACL), 2010, pp. 901–909.
- [14] M. Collins, "Ranking algorithms for named-entity extraction," in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics (ACL), 2001, pp. 489–496.
- [15] J. Bjorne and T. Salakoski, "Generalizing Biomedical Event Extraction," in Proceedings of BioNLP Shared Task 2011 Workshop, 2011, pp. 183–191.
- [16] H. Isozaki and H. Kazawa, "Efficient support vector classifiers for named entity recognition." Association for Computational Linguistics (ACL), 2002, pp. 1–7.
- [17] R. Patra and S. K. Saha, "A kernel-based approach for biomedical named entity recognition," The Scientific World Journal, vol. 2013, 2013.
- [18] D. Li, L. Huang, H. Ji, and J. Han, "Biomedical Event Extraction Based on Knowledge-driven Tree-LSTM," in Proceedings of NAACLHLT 2019. Association for Computational Linguistics, 2019, pp. 1421–1430.
- [19] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, no. btz682, 2019.
- [20] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, "Publicly Available Clinical BERT Embeddings," in Proceedings of the 2nd Clinical Natural Language Processing Workshop, 4 2019, p. 72–78.

[21] M. Basaldella and N. Collier, “BioReddit: Word Embeddings for UserGenerated Biomedical NLP,” in Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019). Association for Computational Linguistics (ACL), 11 2019, pp. 34–38.

[22] L. Akhtyamova and J. Cardiff, “LM-based Word Embeddings Improve Biomedical Named Entity Recognition: a Detailed Analysis,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Bioinformatics) [to be published in June 2020]. Springer Verlag.