A COMPARATIVE ANALYSIS OF RULE-BASED, MACHINE LEARNING-BASED AND DEEP LEARNING-BASED APPROACHES IN NAMED ENTITY RECOGNITION

ABSTRACT

In this essay, three Named Entity Recognition (NER) methodologies—rule-based methods, machine learning-based methods and deep learning-based methods—are examined side by side and evaluated. Their underlying ideas and methodologies are explored with an emphasis on their unique characteristics and potential applications. Additionally, we contrast them from both a theoretical and a practical standpoint, accounting for accuracy, efficiency, cost, and simplicity of implementation. Our essay suggests future directions of research within this field of study as well as identifies the optimal strategy to employ in the context of different situations.

INTRODUCTION

Our main objective of this essay is to compare at least two approaches of Named Entity Recognition from a theoretical as well as practical point of view. Named entity recognition (NER), also known as entity extraction, chunking, and identification, is a natural language processing (NLP) approach for extracting information from text. Named Entity Recognition entails recognising and categorising crucial textual information known as named entities. The major subjects of a piece of writing, such as people, locations, corporations, events, and goods, as well as themes, topics, times, monetary amounts, and percentages, are referred to as named entities. In our essay we are comparing rule-based methods, machine learning-based methods and deep learning-based methods assessing their abilities, shortcomings and effectiveness in Named Entity Recognition tasks. Due to linguistic diversity, domain-specificity, varied entity kinds, and the requirement for annotated training data, Named Entity Recognition is difficult. These features make it challenging to develop stable rules, adapt techniques to diverse domains, manage multiple object kinds, and acquire representative data with labels. It takes language understanding, machine learning skills, and domain-specific data to create effective Named Entity Recognition methods. Named Entity Recognition has a wide range of applications some of which are machine translation, information extraction, information retrieval. Natural Language Processing systems, such as chatbots, sentiment analysis tools, and search engines, rely heavily on Named Entity Recognition. It is utilised in healthcare, finance, human resources (HR), customer service, higher education, and social media analysis, among other fields.

DESCRIPTION OF THE THREE APPROACHES

Rule-based approach : Rule-based Named Entity Recognition (NER) is a method for detecting and categorising named entities in text that employs predetermined rules. Subject matter experts or linguists generally develop these rules, depending on language patterns, syntactic structures, or domain-specific expertise. The goal of rule-based NER is to find particular word patterns or combinations that signal the existence of named entities. A rule-

based NER system processes the text sequentially, with the rules being executed in a preset sequence. Each rule is made up of a pattern or a collection of requirements that must be met before a word or sequence of words is categorised as a named object. For example, a rule for identifying monetary values can state that it should start with a dollar sign followed by a numerical value such as $15. Another example for a rule could be that it's more probable that a person's name is a term that starts with a capital letter and ends with lowercase characters.

Machine Learning-based approach : Machine learning-based Named Entity Recognition (NER), models are taught to recognise and categorise named things in text by automatically deriving patterns and features from labelled training data. For capturing the statistical correlations between input data and named entity labels, these models make use of a variety of machine learning methods, including Conditional Random Fields (CRF) and Support Vector Machines (SVM). Giving the model a dataset with each word or string of words labelled with the appropriate named entity class is the first step in training the model. The model then learns to identify patterns in the data by removing pertinent information, including word embeddings, part-of-speech tags, or grammatical structures. The model can distinguish between named entities and other words in the text because of these features that provide crucial contextual information. After being trained, the model may be used to forecast the labels of named entities for fresh, unseen text. In order to make predictions based on the recognised patterns and features, the trained model must be fed the text. Each phrase or group of words is given a named entity label by the model, which identifies the kind of recognised entity, such as a person, organisation, or location.

Deep Learning-based approach : The identification and categorization of named entities is made possible by the cutting-edge technique known as Deep learning-based Named Entity Recognition (NER), which makes use of deep neural networks to automatically extract complex patterns and representations from text input. These models are excellent at collecting contextual information and sequential relationships, enabling precise predictions even for intricate language patterns. Deep learning-based NER commonly uses transformer models and recurrent neural networks (RNNs). By maintaining contextual information over lengthy sequences, RNNs, such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU), are skilled at processing sequential data. In order to forecast the named entity labels, they encode the input text into a fixed-length form and then send it through a classification layer. Due to their capacity to identify global relationships and parallelize computation, transformer models have grown significantly in favour in NLP applications like NER. Transformers successfully capture contextual linkages by modelling the interactions between words in the input text by using self-attention techniques. By pretraining on huge corpora and fine-tuning on particular NER datasets, models like BERT (Bidirectional Encoder Representations from Transformers) have attained state-of-the-art NER performance.

THEORETICAL COMPARISON

Rule-based approach : To recognise named entities, rule-based NER uses pre-defined rules created by linguists or subject-matter experts. Required features are language patterns,

syntactic constructions, and subject-matter expertise. It is more domain specific and offers limited flexibility because it depends on manually created rules that are unique to a domain or language. To update or modify the rules for new domains or languages involves time, knowledge and effort of experts. Sometimes it struggles with complex or unseen patterns. Since subject-matter specialists can examine and comprehend the rules this approach is highly interpretable. The rationale for entity classification is clear.

Machine Learning-based approach : Utilising Conditional Random Fields (CRF) and Support Vector Machines (SVM) to automatically learn patterns and features from labelled training data is a key component of the machine learning-based approach for Named Entity Recognition (NER). Because it learns patterns directly from the data, this method differs from rule-based NER in that it does not rely on predetermined rules. Because it can handle various language patterns and adapt to new domains without manually creating rules, machine learning-based NER offers better flexibility and adaptability. It can handle unseen data successfully by extracting features including word embeddings, part-of-speech tags, and contextual information. The model can learn and generalise for new cases to some extent with a significant amount of labelled training data. . However, a sizable quantity of labelled data is required for training, and the performance is greatly influenced by the quality and representativeness of the training dataset. Additionally, it might need to be retrained or adjusted for new domains. When compared to rule-based NER, interpretability is worse since the models learn complicated patterns that may be difficult to explain making it challenging to understand the rationale behind the entity classification.

Deep Learning-based approach : To automatically learn representations and identify complicated patterns from text input, deep learning-based NER uses deep neural networks, such as recurrent neural networks (RNNs) or transformer models. These models are excellent at capturing context and long-range relationships, which makes them suitable for NER tasks. It automatically collects features across the layers of neural networks. It is capable of dealing with intricate and subtle verbal patterns, even ones that may not be clearly outlined by rules. These models require little user intervention and are capable of capturing tiny contextual inputs and adapting to many domains. But in order to function at its best, deep learning-based NER often needs a lot of labelled training data. Deep learning training can be computationally expensive and demand a lot of computing power**.** Since deep learning-based NER relies on sophisticated representations that may be difficult to interpret, interpretability is a concern. When these models rely on complex network designs and extensive pre-training, it might be difficult to understand how they make decisions.

PRACTICAL COMPARISON

We will do our practical comparison between rule-based approach, machine learning-based approach and deep learning-based approach by comparing certain evaluation metrices like accuracy, precision, recall and f1 score on a dataset called CoNLL 2003 dataset which is a widely used benchmark dataset for Named Entity Recognition tasks. Let us see and compare the results to come to a conclusion.

Rule-Based Methods:

Accuracy: Rule-based methods on the CoNLL 2003 dataset achieved an accuracy in the range of 85% to 90% (Sang and Meulder, 2003).

Precision: Rule-based methods showed high precision, surpassing 90% on the CoNLL 2003 dataset (Tjong Kim Sang and De Meulder, 2003).

Recall: Rule-based methods showed recall ranging from 80% to 90% on the CoNLL 2003 dataset (Tjong Kim Sang and De Meulder, 2003).

F1 Score: The F1 score for rule-based methods on the CoNLL 2003 dataset ranged from 85% to 90% (Sang and Meulder, 2003).

Machine Learning-Based Methods:

Accuracy: Machine learning-based methods, like, Conditional Random Fields (CRFs), achieved an accuracy ranging from 88% to 92% on the CoNLL 2003 dataset (Finkel et al., 2005).

Precision: Machine learning-based methods showed precision ranging from 85% to 90% on the CoNLL 2003 dataset (Finkel et al., 2005).

Recall: Machine learning-based methods showed recall values ranging from 85% to 90% on the CoNLL 2003 dataset (Finkel et al., 2005).

F1 Score: The F1 score for machine learning-based methods on the CoNLL 2003 dataset ranged from 85% to 90% (Finkel et al., 2005).

Deep Learning-Based Methods:

Accuracy: Deep learning-based methods, like, Bidirectional LSTM-CRF models, achieved accuracy exceeding 90% on the CoNLL 2003 dataset (Lample et al., 2016).

Precision: Deep learning-based methods achieved precision above 90% on the CoNLL 2003 dataset (Lample et al., 2016).

Recall: Deep learning-based methods showed recall values above 90% on the CoNLL 2003 dataset (Lample et al., 2016).

F1 Score: The F1 score for deep learning-based methods on the CoNLL 2003 dataset exceeded 90% (Lample et al., 2016).

FURTHER WORK

There are many prospective directions for more study and advancements in Named Entity Recognition. To start with exploring and improving neural models could be done because they are capable to generate named entities in complex sentences. Making advantage of multilingual models is one strategy, another strategy is to employ transfer learning methods to challenge cross lingual NER where identifying named entities in languages other than the model's source language is the aim. We can explore development of hybrid models coming rule-based methods, machine learning-based methods and deep learning-based methods to improve accuracy and robustness of the model. Further research in rule-based NER might

concentrate on enhancing and increasing the rule sets for certain domains. This entails working with domain specialists to find and incorporate domain-specific patterns and linguistic subtleties into the rule-based method, hence improving its performance in specialised domains. Improved assessment measures are also required in order to fully represent the complexity and subtleties of NER. The effectiveness of NER models in real-world applications, where the cost of false positives and false negatives may vary, may not be adequately assessed by existing measures like accuracy, recall, and F1 score. To further comprehend the advantages and disadvantages of NER models, future research should examine the use of more complex measures, such as cost-sensitive assessment and error analysis.

CONCLUSION

In a nutshell the comparison of named entity recognition (NER) systems based on rule-based, machine learning-based, and deep learning reveals significant features and compromises. Theoretical study reveals that machine learning-based NER gives flexibility and adaptability by automatically learning patterns from labelled data, whereas rule-based NER offers interpretability and control through manually established rules. Deep learning-based NER are particularly effective in capturing intricate patterns and contextual data by using deep neural networks. Deep learning-based NER routinely outperforms rule-based NER and machine learning-based NER in practical evaluations using benchmark dataset, yielding better F1 scores on the CoNLL-2003 dataset. The selection of a NER technique, however, is based on the particular needs, the readily accessible labelled data, and the desired balance between performance and interpretability. The rule-based and machine learning-based approaches continue to be relevant in some domains where interpretability and control are prioritised, but overall, deep learning-based NER holds great promise for accurate and robust named entity recognition in a variety of applications.

REFERENCES

- https://www.analyticsvidhya.com/blog/2021/11/a-beginners-introduction-to-ner-named-entity-recognition/
- https://pub.towardsai.net/top-5-approaches-to-named-entity-recognition-ner-in-2022-38afdf022bf1
- https://www.techtarget.com/whatis/definition/named-entity-recognition-NER#:~:text=Named%20entity%20recognition%20(NER)%20is,text%20known%20as%20named%20entities.
- Chiu, J. P. C., & Nichols, E. (2016). Named Entity Recognition with Bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics, 4, 357-370.
- Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016) 'Neural Architectures for Named Entity Recognition', in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 260-270.
- Zhang, Y., Yang, Q., Zhang, S., & Huang, L. (2020). A Comparative Study of Sequence Labeling Models for Named Entity Recognition on Low-resource Languages.

Proceedings of the 28th International Conference on Computational Linguistics (COLING).

- Ma, X., & Hovy, E. (2016). End-to-End Sequence Labeling via Bi-directional LSTM-CNNs-CRF. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL).
- Sang, E. F., & Meulder, F. D. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL (pp. 142-147). Retrieved from http://www.aclweb.org/anthology/W03-0419
- Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 363-370). Retrieved from https://www.aclweb.org/anthology/P05-1045/
- Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL).
- Ratinov, L., & Roth, D. (2009). Design Challenges and Misconceptions in Named Entity Recognition. Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL).
- Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv preprint arXiv:1508.01991.
- Akbik, A., Blythe, D., & Vollgraf, R. (2019). Contextual String Embeddings for Sequence Labeling. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT).
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT).