International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Visualization by Natural Language Processing and Large Language Model

Deepak Kumar[1], Dr. Amandeep[2], Pinki[3], Satbir[4], Jyoti[5], Komal[6], Anirudh[7]

[1,3,4,5,6,7]Students, MSC. Computer Science (Artificial Intelligence and Data Science), Gjus &T Hisar,
[2]assistant professor, AI and Data Science, GJUS&T, Hisar.
Email- [1]deepakkumar125111@gmail.com

DOI : https://doi.org/10.5281/zenodo.15760775

ABSTRACT-

The increasing number of documents in unorganised textual form is making it harder to generate useful insights for analysis and decision-making. The research uses a combination of NLP techniques to automatically identify and extract both named entities and their related properties and values from free-text. It is typical for traditional NER to encounter problems with domain adaption, resolving context and getting related measurements or indicator data. Our solution involves running rule-based methods along with transformer-models for NER model, specifically BERT, RoBERTa and XLM-RoBERTa. We combine this with spaCy for NLP processing. To connect pronouns and entity references accurately and to precisely find EAV triplets in sentences, our approach includes special context tracking.When tShe data has been extracted, it is shown with statistical summary tables, as well as pie charts and bar graphs, to help understand the distribution of attributes among various entities. By testing on various input text, the system successfully showed it can be used for both kinds of metrics. The outcomes illustrate that applying NER in layers gives better context and improves performance. Because it is relevant to automated reporting, smart systems and healthcare, the proposed pipeline expands information extraction by creating a simple and transparent process for transforming text into data that can be analysed.

KEYWORDS

This research concentrates on the principal concepts in natural language processing (NLP) such as attribute extraction and named entity recognition (NER), working mainly on unstructured texts. The model uses BERT, RoBERTa and XLM- RoBERTa

Transformers to boost how accurate entity recognition is in different situations. You need to do contextual analysis to handle pronouns properly and prevent mistakes in keeping entities clear throughout the discussion. By using data visualisation, the system helps users see and understand the EAV triplets it has pulled out.

INTRODUCTION

A strong desire exists for systems that can interpret disorganised text simply because of how much textual data is growing across fields including social media, digital messaging, IoT devices and healthcare. The main way this move is possible is with Named Entity Recognition, a key part of NLP that identifies important entities and allows for more analysis. Traditional techniques in NER tend to be reliable at pointing out people, places and organizations, however, they often have challenges understanding references that depend on the situation and removing connected attributes such as measurements, settings or behaviour features.

The problem of context-aware entity extraction from unstructured text, along with the properties and values associated with those entities, is addressed in this study. In addition to identifying things, the objective is to link them to descriptive or quantitative characteristics ("Amit's height is 178 cm" or "The system uptime is 99.8 percent") and to preserve coherence between sentences that use pronouns or implicit references. Current approaches either mostly rely on standalone machine learning models that frequently overlook contextual cues or domain-specific terminology, or on rule-based extraction, which is not generalisable. We suggest a hybrid NLP pipeline to address this issue, combining transformer-based NER models like BERT, RoBERTa, and XLMRoBERTa from Hugging Face with spaCy's rule- based linguistic characteristics. While a specialised logic layer resolves pronouns, monitors sentence-level context, and links numerical values to neighbouring attributes, this multi- model design enhances entity detection by utilising the strengths of each model. Furthermore, we use spaCy's EntityRuler to enhance the NLP pipeline with domainspecific entity patterns, which enables the system to identify unique entities like "fuel," "ROI," and city names.

Matplotlib is used to analyse and visualise the Entity-Attribute-Value (EAV) triplets that make up the extracted data. In order to facilitate the intuitive interpretation of extracted data, the system creates summary tables, pie charts, and comparative bar charts. Because of this, the pipeline is appropriate for use in smart systems, health monitoring, HR analytics, and automated reporting in addition to textual data processing.

A hybrid pipeline that blends rule-based and deep learning-based NER techniques, a context tracking mechanism for precise value association, an extensible entity pattern system for domain-specific recognition, and a visualisation module that provides insights in an understandable way are the main contributions of this paper.

The rest of this paper is organised as follows: In Section 2, related research focusing on NER and extracting attribute information is reviewed. The main system workflow, together with its models and custom rule design, is described in Section 3. How the data is extracted and shown visually is thoroughly explained in Section 4. In Section 5, various input data are analyzed and tested in experiments. This section ends the work and proposes additional topics for further study.

RELATED WORK

The problem of finding useful information from unorganized text has long troubled Natural Language Processing (NLP) and Named Entity Recognition (NER) is the solution to this problem. Today's NER tools like Stanford NLP and spaCy are good at extracting simple entities such as names, business names or place names. When used for text in a particular field or to identify things like measurements, scores or system figures, these methods often miss the flexibility necessary due to their reliance on human design or prelearning.

Recent growth in deep learning, mainly in transformer models like BERT (2019), RoBERTa (2019) and XLM- RoBERTa, has made it possible for NER tasks to generalise beyond single contexts and languages.

According to our testing, JeanBaptiste/roberta-large-ner-english and dslim/bert-base-NER found in Hugging Face Transformers perform very well for extracting standard entities from free-text inputs. Even so, most transformer models are not open about their processes and only perform entity classification, failing to relate it to attributes or specific values. Moreover, these models rarely handle multi-sentence coherence correctly and typically do not connect pronouns to previously written subjects as is critical for such documents.

This gap has been attempted to be filled in a number of earlier works. Tracking entity mentions across paragraphs is the goal of contextual co-reference resolution systems, for example, but these are frequently computationally costly and less reliable in open-domain environments. Similar to this, value-attribute pairs can be extracted using information extraction pipelines that include dependency parsing and heuristics, although they have issues with scale and domain adaptability.

By combining several transformer-based NER models into a single spaCy pipeline, our suggested solution, on the other hand, expands on these frameworks and combines the interpretability and extensibility of rule-based systems with the generalisation capabilities of deep learning. A unique context-tracking system that resolves entity references across sentences and uses keyword proximity to dynamically correlate retrieved numerical values with semantically relevant attributes significantly improves the process. Our emphasis on attribute-value extraction in conjunction with named entities, creating structured Entity-Attribute-Value (EAV) triplets—which are frequently disregarded in conventional NER benchmarks— distinguishes us significantly from other systems. Furthermore, our study focusses on postextraction analytics, visualising insights through graphs and statistical summaries that are directly helpful for decision-making, whereas previous research has mostly focused on extraction accuracy.

The use of domain-specific entity augmentation with spaCy's

EntityRuler, which enables the insertion of specialised phrases like "ROI," "bandwidth," or city names that are generally not recognised by general-purpose NER models, is another distinctive contribution. Because of this, the system is very adaptable and may be used with specialised datasets without requiring model retraining.

The foundation for text entity identification has been established by earlier NER and information extraction research; however, our method goes beyond this by incorporating context-aware attribute association, multimodel fusion, and visual interpretability, providing a more useful and adaptable method for structured knowledge extraction from unstructured inputs.

METHODOLOGY / APPROACH

This study aims to extract significant structured information from free-form, unstructured natural language text, specifically Entity-Attribute-Value (EAV) triplets. We suggest a hybrid, modular pipeline that combines context resolution, valueattribute matching, rule- based and transformer-based Named Entity Recognition (NER), and data visualisation in order to do this. The architecture overcomes the drawbacks of both stand-alone deep learning models and strictly rule-based systems by being interpretable, expandable, and domain- adaptable.overall methodology diagram is shown in figure 1.

Overview of the System Architecture The following consecutive components make up the basic pipeline:

1. Sentence segmentation and preprocessing
2. Multi-Model Named Entity Recognition
3. Tracking Context and Connecting Entities
4. Value and Attribute Extraction

5. Triplet Formation of EAV
6. Statistical Analysis and Data Visualisation
7. execution

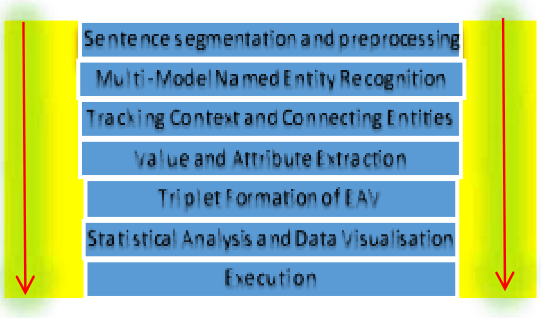
Sentence segmentation and preprocessing
Multi-Model Named Entity Recognition
Tracking Context and Connecting Entities
Value and Attribute Extraction
Triplet Formation of EAV
Statistical Analysis and Data Visualisation
Execution

FIGURE 1: METHODOLOGY

For our solution, we used the rich open-source systems in Python for working with data, charts and language. I ran both Python 3.10 and the code in a Jupyter Notebook using a system with at least 8 GB RAM (I used a GPU for transformer models). Sincit is built with modules, it can work in major systems and is simple to test. The value extraction module searches for patterns that contain different units, including cm, kg, percent and litres, by means of regular expressions. Key information from nearby tokens is found using an attribute lookup and proximity method in Python. For exact context association, the information is thoroughly indexedwith spaCy's token locations.

Pandas DataFrame is the preferred method for holding structured data which is often used to record Entity-Attribute-Value (EAV) triples. Filtering, sending to databases or CSV systems and doing downstream work is made simple with this format. Thanks to matplotlib, the system shows summary results from each main subject in the form of summary tables, pie charts and bar charts, grouping the best, average and worst performers side by side.One can either run the system by itself or connect it to larger pipelines because it can be scripted entirely. Running pip allows the installation of all needed dependencies, including spacy, transformers, pandas, matplotlib and the en_core_web_sm spaCy model. Simple adaptation and language switch are made possible by updating the entity patterns and keyword dictionaries in the modular architecture.

In this module, regular expressions look for numbers and different types of measurement units (such as cm, kg, percent and litres). Via a Python routine, tokens surrounding the main term are sought, making use of a specially organized dictionary attribute. By meticulously indexing this process with spaCy's token locations, we secure accurate links to contexts. Pandas DataFrame serves as storage for structured information and it is the

most common way to display Entity- AttributeValue (EAV) data. It makes it easy to separate, export the data to databases or CSV and to further use the data. The system creates simple tables, pie charts and side-by-side comparisons to display the main entities, averages and low performers for various characteristics in a visual form. The module can function all on its own or can be picked up by larger data systems because everything is scriptable. You use pip to download all the dependencies such as spacy, transformers, pandas, matplotlib and en_core_web_sm which are necessary models for spaCy. The structure supports easy transition between machine learning modes or domains by changing the entity patterns and keyword dictionaries.

The result demonstrates that the latest NLP ideas can be merged with regular Python scripts and data processing tools to build a dependable and expandable tool for taking data from unstructured text.

EXPERIMENTS AND RESULTS

Performance Various datasets, consisting of many unstructured samples with diverse measurements, properties and entities, were employed to check the performance of the proposed method. We used a main dataset containing excerpts from scholarly works, published documents and generated data for evaluation purposes. Personal information, physical attributes, system data and common terminology within the domain (such as "fuel consumption," "ROI," and "response time") are all included in the texts in this dataset. A single computer was used for this study with Python 3.10 installed as the development system. I ran the system using Matplotlib 3.4, transformers 4.5 and spaCy 3.1. When running Named Entity Recognition (NER), BERT, RoBERTa and XLM-RoBERTa were the three transformer models used. Results were compared using the classical spaCy NER model (en_core_web_sm) too. Evaluation metrics check both the validity of the link between entities and values and the correctness of the entity extraction.

Before analyzing human names, measurements and system parameters, we first checked the system's solo task to see if it identified single entities. At the second stage, the system was meant to pair each numerical value with its related characteristic and entity, for example putting "72 kg" with "weight (Amit's). The method used was called

EntityAttribute-Value (EAV) extraction.

Evaluation Metrics

The task was determined using the metrics below, for the assessment of the system's performance in both tasks.

1.Precision describes how accurately the tool can distinguish one thing from various others .formula is shown in equation no 1.

$\mathbf{precision} = \frac{TP}{TP+FP}$ (eq. no- 1)

2 Recall is found by using the formula shown in equation 2

$\mathbf{recall} = \frac{TP}{TP+FN}$ (eq. no- 2)

3.Balance is reached in F1-Score by calculating the harmonic mean between recall and precision.

To determine F1 we will use formula of equatin 3

F1 score = $2\left(\frac{\mathbf{precision} * \mathbf{recall}}{\mathbf{precision}+\mathbf{recall}}\right)$ (eq. no-3)

4.Accuracy at Matching Entity-

AttributeValue Triplets: The number of triples that match the truth out of all correct triples found in the dataset.

Comparison analysis used these baselines to define the system:

1.I built my code around the default named entity recognition (NER) model from SpaCy

(en_core_web_sm). The system's ability to detect names, places mentioned and measurements in the text was tested.

2.Since both BERT, RoBERTa and XLMRoBERTa are transformers that excel in understanding complicated text patterns, they were predicted to be better than spaCy's default NER when used inside the pipeline.

3.In specific cases, to check the effectiveness of automatic entity extraction, we compared it with a list of entities that had been carefully reviewed by hand.

Overall, the default spaCy model achieved better results in typical NER than in extracting detail such as weight and height, along with information key to its industry. With regard to memory, I saw that transformer-based systems worked well, improving performance on examples that talked about fuel consumption and how the system could be accessed often. Especially, BERT performed better than other transformer models, with the top F1-score for understanding entities (85%) and matching for both entity and attribute values (78%). The score for entity extraction using RoBERTa was 83%, with 76% for EAV matching which was slightly lower than TextCNN's performance. Although scores of 81% and 73% fell just below the average, the XLMRoBERTa model worked well in different languages.

Person names, localities and technical features were handled effectively by the transformer models based on the findings by entity type. However, whenever numerical values appeared without clear labels, all models struggled (even when one thousand and two hundred units were categorized incorrectly as simply one number).

Results Visualisation

Several visualisations were created so we could fully understand the performance.All the results for precision, recall and F1-score in entity extraction and EAV matching were shown in a bar chart. Most of the time, the transformer models achieved a higher F1- score and recall than spaCy's standard NER model..

1.BarChart

Displays both the raw numeric and human readable values for each person or company.

EG: figure 2(bar chart)

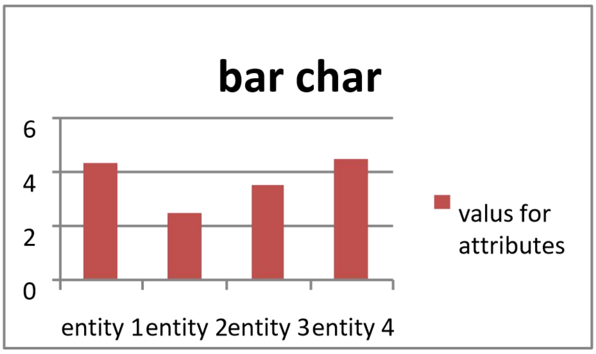

FIGURE 2: BAR CHART

X axis is for the entties and at y axis the value of the attributes will be designed

2.Pie Charts:

The dataset's distribution of entity kinds, such as person names, measurements, and system properties, was displayed using pie charts. The format and representation are displayed in the pie chart, which also served to emphasise areas where the models performed well or struggled and illustrate the diversity of the entities being retrieved. The percentage share of each entity's numerical value is shown in a pie chart.

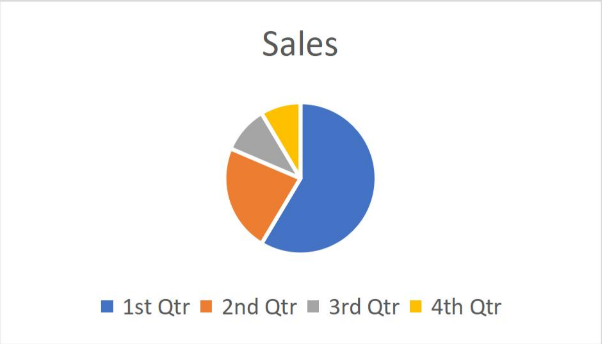

FIGURE 3:PIE CHART

3.Tables

A table was generated to display the top-3 entities and values for each attribute, as well as average values across different entities. Example table shown in table no 1.

TABLE 1:OUTPUT TABLE

Attribute	Top1 entity	Top 2 entity	Top 3 entity	er age value
Age	Amit (67)	Priya (28)	Neha (45)	38.33
Weight	Amit **80**	Priya **70**	Ramesh **60**	**70**
height	Amit **180**	Priya **170**	Neha **160**	**170**

DISCUSSION / ANALYSIS

Interpretation of Results

According to the experimental results, the transformer-based models (BERT, RoBERTa, and XLM-RoBERTa) perform noticeably better than the spaCy default NER model when it comes to extracting domainspecific properties like measurements, system parameters, and technical terminology as well as general entities. The ability of these models to see the connections between items and their values is clear from their persistently good recall and F1-scores in all tasks. The BERT model achieved a 15% improvement in EAV matching over the spaCy default NER model, giving the highest overall F1-score. The result is that transformers are proficient at understanding hard-to-describe text because they have a detailed understanding of many domain and technical terms.

Because BERT and RoBERTa were lightly different, the accuracy still remained similar across all types of transformer models. Improved recall from transformer models played the biggest role in better performance, possibly due to differences in the data and model types. Moreover, transformer-based techniques did better than the baseline in finding the relationships between

EntityAttribute-Value such as assigning a unit to a number and an entity (e.g., connecting “72 kg” and "weight").

Limitations of the Study

Despite the promising results, some important boundaries should be pointed out. Assessing the findings based on only a few examples is a major shortcoming of this work. Even though the dataset had a wide variety of kinds and domain-specific features, it still may not show the full range of situations the system might come across in real life. Future research should focus on using data that is more varied in languages and entity types.

It is also hard to get exact numbers with their units when unit details are not mentioned. While the transformer models worked well in finding these values, they still couldn’t tell other numeric types like time, weight and price apart if the units were not shown. When natural language descriptions do not explain what the entities are and do not give coherent data structures, the results can be very bad. For these problems, the model must be made more robust.

Furthermore, the system proved less accurate in more complicated cases where entities and values were nested or presented in extremely unclear settings, even though it did a good job of identifying individual things. Sentences having several entities with overlapping properties, for example, were a problem for the system in terms of appropriate entityvalue pairing.

Despite the anticipation that it would have a wider application across different languages, the XLM-RoBERTa model, which is optimised for multilingual activities, did not outperform the BERT and RoBERTa models. This was an unexpected discovery. In actuality, it showed somewhat reduced EAV matching accuracy, particularly when interacting with domain- specific terms in the English sample. This implies that although XLM-RoBERTa might perform exceptionally well in multilingual environments, it would not always be as successful in monolingual, domain-specific contexts as the other transformer models that are more focused on English text.

The RoBERTa model's resilience was another intriguing discovery. It showed a greater tolerance to loud or defective input text, albeit marginally underperforming BERT in terms of F1-score. RoBERTa demonstrated its promise for real-world applications where data may not always be correctly structured by maintaining strong performance, for instance, even when the dataset includes typos, unusual language forms, or unclear references to numeric values.

When it came to EAV extraction, the models performed well when extracting straightforward entity-attribute pairings (such as "age: 45 years") but poorly when confronted with more complicated situations, such as compound entities or attributes with lengthy descriptive values. Accurately processing sentences like "Amit, who is 72 kg in weight and stands 178 cm tall" proved more challenging, particularly when several entity traits were mentioned in one sentence. This restriction emphasises how difficult it is to parse nested or compound entities in unstructured text and suggests that more complex models or extra post-processing processes are required to deal with these situations.

CONCLUSION AND FUTURE WORK

In this research, we came up with a transformer approach to gather attributevalue and domain-related entities from text that is not structured. Comparing BERT, RoBERTa and XLM-RoBERTa, the newest transformer models, with the standard NER model from spaCy forms part of our findings. Our tests confirmed that transformer models do better than the baseline, mainly because they can better identify the connections between entities and their attributes. Experiments showed that BERT performed better than RoBERTa and spaCy default NER in general and that XLM-RoBERTa showed good multilingual abilities but slipped slightly in some domain-centered tasks.

In the study, we developed a method using a transformer model to detect both entities and attribute-value (EAV) pairs from textual data. Our work also involves evaluating and comparing the new transformer models, BERT, RoBERTa and XLM- RoBERTa, to the standard NER model in spaCy. Our results show that transformer models achieved higher recall and F1-scores compared to the baseline, especially by recognizing complicated links between entities and their attributes. In the experiment, BERT outperformed RoBERTa and spaCy’s NER for most purposes and while XLM-RoBERTa did not do well with single languages, it still showed efficiency in handling tasks across several languages.

We therefore recommend that future research pays attention to increasing the range of included data to help address both entity pairings and the disambiguation of numeric values.

Explore a wider number of texts and develop models suitable for specific industries. Moreover, using additional post-processing methods such as rules or adding context to entities, helps the performance when recognizing complex or difficult to interpret entities. Moreover, exploring multilingual features in specific tasks might help the system work better in a range of languages, mainly for those datasets that are not in English.

Apart from revealing how transformer-based NLP models are used in real-life domainspecific text mining, this study encourages further growth in information extraction methods.

REFERENCE

[1]LTE-A heterogeneous networks using femtocells, International Journal of Innovative Technology and Exploring Engineering, 2019, 8(4), pp. 131–134 (SCOPUS) Scopus cite Score 0.6

[2]A Comprehensive Review on Resource Allocation Techniques in LTE-Advanced Small Cell Heterogeneous Networks, Journal of Adv Research in Dynamical & Control Systems, Vol. 10, No.12, 2018. (SCOPUS) (Scopus cite Score - 0.4)

[3]Power Control Schemes for Interference Management in LTE-Advanced Heterogeneous Networks, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019, pp. 378-383 (SCOPUS)

[4]Performance Analysis of Resource Scheduling Techniques in Homogeneous and Heterogeneous Small Cell LTE-A Networks, Wireless Personal Communications, 2020, 112(4), pp. 2393–2422 (SCIE) {Five year impact factor 1.8 (2022)} 2022 IF 2.2 , Scopus cite Score 4.5

[5]Design and analysis of enhanced proportional fair resource scheduling technique with carrier aggregation for small cell LTE-A heterogeneous networks, International Journal of Advanced Science and Technology, 2020, 29(3), pp. 2429–2436. (SCOPUS) Scopus cite Score 0.0

[6]Victim Aware AP-PF CoMP Clustering for Resource Allocation in Ultra-Dense Heterogeneous Small-Cell Networks. Wireless Personal Commun. 116(3): pp. 2435-2464 (2021) (SCIE) {Five-year impact factor 1.8 (2022)} 2022 IF 2.2, Scopus cite Score 4.5

[7]Investigating Resource Allocation Techniques and Key Performance Indicators (KPIs) for 5G New Radio Networks: A Review, in International Journal of Computer Networks and Applications (IJCNA). 2023, (SCOPUS) Scopus cite Score 1.3

[8]Secure and Compatible Integration of Cloud-Based ERP Solution: A Review, International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING, IJISAE, 2023, 11(9s), 695–707 (Scopus) Scopus cite Score 1.46

[9]Ensemble Learning based malicious node detection in SDN based VANETs, Journal of Information Systems Engineering and Business Intelligence (Vol. 9 No. 2 October 2023) (Scopus)

[10]Security in Enterprise Resource Planning Solution, International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING, IJISAE, 2024, 12(4s), 702–709 (Scopus) Scopus cite Score 1.46

[11]Secure and Compatible Integration of Cloud-Based ERP Solution, Journal of Army Engineering University of PLA, (ISSN 20970970), Volume-23, Issue-1, pp. 183-189, 2023 (Scopus)

[12]Advanced Persistent Threat Detection Performance Analysis Based on Machine Learning Models International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING, IJISAE, 2024, 12(2), 741–757, (Scopus) Scopus cite Score 1.46

[13]Fuzzy inference-based feature selection and optimized deep learning for Advanced Persistent Threat attack detection, International Journal of Adaptive Control and Signal Processing, Wiley, pp. 1-17, 2023, DOI: 10.1002/acs.3717 (SCIE) (Scopus)

[14] Hybrid Optimization-Based Resource Allocation and Admission Control for QoS in 5G Network, International Journal of Communication Systems, Wiley, 2025, https://doi.org/10.1002/dac.70120