NlpUned at SemEval-2025 Task 10: Beyond Training: A Taxonomy-Guided Approach to Role Classification Using LLMs

Alberto Caballero
UNED, Spain
acaballer382@alumno.uned.es

Alvaro Rodrigo
UNED, Spain
alvarory@lsi.uned.es

Roberto Centeno
UNED, Spain
rcenteno@lsi.uned.es

Abstract

Large Language Models (LLMs) have revolutionized natural language processing, demonstrating exceptional capabilities in understanding, reasoning, and generating human-like text. In this paper, we introduce an approach to automating the classification of roles in articles based on an entity and its location within the text. Instead of traditional model training, we leverage zero-shot and few-shot prompting, structuring LLM-based workflows with taxonomies and contextual cues to guide the classification process. By integrating signal processing and contextual enrichment, our approach enables the system to achieve competitive classification accuracy without fine-tuning or additional training. This underscores how context-aware LLMs can be effectively deployed in real-world applications where interpretability, adaptability, and ethical alignment are crucial—highlighting their potential for responsible AI deployment in large-scale human dynamics analysis.

1 Introduction

The SemEval-2025 Task 10 on Fine-Grained Role Classification (Subtask 1) (Piskorski et al., 2025) aims to assign one or more sub-roles to named entity mentions in news articles (Stefanovitch et al., 2025). The task provides a structured taxonomy covering three main role types, each of which is further divided into fine-grained subroles. As a multi-label, multi-class text-span classification challenge, this task holds significant potential for developing objective and transparent AI systems capable of classifying human behavior in news narratives (Gilardi et al., 2024). The ability to accurately identify and classify entities' roles in news stories can support media monitoring, misinformation detection, and conflict analysis while reducing human subjectivity in information processing (Berrondo-Otermin and Sarasa-Cabezuelo, 2023).

Our approach to this task centered on structuring the input and modeling strategy for optimal performance. The system first generates a summary of the news article to capture the broader context and then extracts the paragraph where the entity mention appears. Unlike a hierarchical approach that classifies a main role first and then predicts the subroles associated to that main role, the best-performing system uses a unified prompt that presents all possible subroles at once. This non-hierarchical approach allows the model to directly assign the most suitable fine-grained role without an intermediate classification step. The unified structure ensures that the model considers all possible subroles equally rather than being constrained by an initial high-level role decision, potentially capturing more nuanced role assignments[1].

Our results highlighted the critical role of signal processing and contextual representation in optimizing classification accuracy. Specifically, we found that the way information is structured—whether through hierarchical or unified approaches—significantly may impact performance (He et al., 2024). Contrary to expectations, breaking down the classification process into multiple stages and assigning explicit sub-role agents based on roles led to poorer results. Instead, models performed better when using a single-step decision-making process, selecting among all possible subroles simultaneously. This finding underscores the importance of input representation and classification strategy in complex multi-label tasks, where contextual nuances are key to accurately determining entity behavior.

[1]The full implementation, including architectures, prompts, and evaluation scripts, is publicly available at GitHub. We encourage further research and experimentation using our framework.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 296–301

2 Background

The SemEval25 task 10 focused on fine-grained role classification in news articles, where the goal was to assign one or more sub-roles to named entity mentions within a given article. The task is framed as a multi-label, multi-class text-span classification problem depending on the provided context.

The task relates to existing research in event extraction (Ahn, 2006), Named Entity Recognition (NER) (Lample et al., 2016), and stance detection (Mohammad et al., 2016). Unlike standard NER, which primarily focuses on entity categorization (e.g., person, organization, location), this task introduces a context-dependent role classification framework, requiring systems to infer entity intent and alignment from the article's discourse. Several studies have explored the role of large language models (LLMs) in legal and justice-related NLP tasks (Siino et al., 2025). Recent research has demonstrated LLMs' ability to analyze legal contracts (Jayakumar et al., 2023), assist in judicial decision-making (Chalkidis et al., 2021), and extract roles in court case narratives (Valvoda et al., 2024). Similarly, AI has been applied to bias detection in judicial decisions (Binns, 2018) and fact-checking in legal and political discourse (Thorne et al., 2018). These studies highlight the increasing relevance of AI systems in understanding narrative roles, power dynamics, and responsibility attribution, which aligns closely with the fine-grained role classification required in SemEval25.

3 System Overview

Our system tackles the task of fine-grained role classification of named entities in news articles. The objective is to classify each entity mention into corresponding sub-roles based on its portrayal in the text. This problem is challenging due to the subjectivity inherent in role perception, the contextual nature of entity portrayal, and the need for fine-grained semantic understanding. To address this, we explored three distinct classification strategies, each varying in preprocessing, hierarchical decision-making, and classification methodology. These approaches were evaluated based on their classification accuracy, robustness, and computational efficiency. Ultimately, the Single-Step Role and Sub-Role Classification approach without majority voting emerged as the best-performing system, achieving a performance of 0.33 in the reference metric, Exact Match Ratio (EMR) (Nazmi et al., 2020), in our experiments.

3.1 Data Preprocessing

The dataset comprises news articles containing multiple named entity mentions, each requiring role classification. Preprocessing plays a critical role in structuring the input data for classification. The key preprocessing steps include: Extracting entity mentions and their surrounding context to ensure relevant information is available for classification. Generating a concise summary of the article to capture the entity's role within the broader narrative. Identifying the exact location of the entity mention to isolate local context, providing a more context-aware classification process. These steps ensure that the model receives both global and local perspectives, allowing for a more nuanced classification of roles and sub-roles.

3.2 System Architectures

In this section, we will briefly describe the main components and features of the systems tested during the development stage. A common feature across all architectures is the use of the same underlying LLM (GPT-4o)[2] and the implemented prompting strategy, Chain-of-Thought (CoT) (Wei et al., 2022). Each approach was designed as a multi-step LLM-based nodal system, where different nodes receive different inputs and generate distinct outputs depending on their functions.

1. **Hierarchical Pipeline with Preprocessing:** This structured two-stage approach first determines the role of an entity before classifying its sub-role. The pipeline begins by generating an article summary, which provides a global understanding of the entity's involvement in the news. Simultaneously, the local context—the paragraph in which the entity appears—is extracted. Based on these inputs, a LLM-based node predicts the role of the entity. Once the main role is assigned, a role-specific sub-role classifier is invoked to refine the classification.

2. **Hierarchical Pipeline without Preprocessing:** A simplified variant of the hierarchical pipeline, this approach eliminates the article summarization step, relying only on the local context of an entity mention within the news

[2] https://platform.openai.com/docs/models#gpt-4o

article. The classification process remains hierarchical, first determining the main role before assigning a corresponding sub-role.

3. **Single-Step Sub-Role Classification with Preprocessing:** Unlike the hierarchical pipelines, this approach bypasses sequential classification and instead predicts an entity's role and sub-role in a single pass. The preprocessing steps remain the same, ensuring that a summary of the news is provided to the decision node. However, instead of first classifying the main role and then assigning a sub-role, the system jointly predicts both classifications in one step. This method eliminates the risk of cascading errors from hierarchical pipelines, where an incorrect role classification would automatically lead to an incorrect sub-role assignment. By handling role and sub-role prediction simultaneously, the model reduces decision fragmentation, making it less prone to error propagation. This approach achieved the best performance in our experiments.

3.3 Classification Strategies: Majority Voting vs. Single-Pass Inference

To further refine classification consistency, we experimented with two inference strategies: Majority Voting (MV) (Jain et al., 2022) and Non-Majority Voting (NMV).

1. **Majority Voting (MV):** In this strategy, each entity undergoes classification three times, and the most frequently predicted role/sub-role is selected as the final label. This method could help mitigate variability in LLM outputs and enhances stability in role assignments. However, it comes at the cost of higher computational requirements and increased token usage, as each entity needs to be classified multiple times.

2. **Non-Majority Voting (NMV):** A more computationally efficient approach, NMV classifies each entity only once, reducing inference time and resource consumption. However, the absence of redundancy makes it more susceptible to inconsistencies, as a single misclassification directly impacts the final role assignment.

4 Experimental Setup

The evaluation of our system is conducted under Subtask 1 of the SemEval 2025 task 10, which involves multiclass multi-label classification. Given the subjectivity of role perception and the context-dependent nature of entity portrayal, this classification task presents significant challenges. An entity may take on multiple sub-roles within a single article, requiring the model to make multi-label predictions while maintaining high precision.

4.1 Evaluation Measures

To assess system performance, the official evaluation metric for the task is the Exact Match Ratio (EMR), which quantifies the proportion of instances where the model's predicted labels exactly match the true labels. This strict metric ensures that partial correctness is not rewarded, emphasizing the need for precise sub-role assignments.

4.2 Challenges in Evaluation

Due to the multiclass multi-label nature of the task, prediction errors are highly penalized under EMR, making it a particularly difficult metric to optimize. Even a single incorrect sub-role prediction results in a zero score for that instance, which contrasts with more lenient multi-label classification metrics that reward partial correctness. Furthermore, the imbalance in sub-role distributions poses an additional challenge. Some sub-roles are more frequent than others, leading to potential bias in model predictions. To mitigate this, our system incorporates context-aware classification, ensuring that both global and local signals contribute to role assignments.

5 Results

Our best-performing system, the Single-Step Role and Sub-Role Classification without Majority Voting (NMV), achieved an EMR of 0.3277, securing 8th place in the SemEval 2025 shared task.

When compared to the top-ranked system (DUTIR), which achieved an EMR of 0.4128, our model exhibited an 8.5 percentual performance gap in EMR. However, an important distinction must be highlighted: our approach was based exclusively on architectural design, input preprocessing, and taxonomy framing, with no example-driven training or fine-tuning. Given this constraint, the results underscore the potential of prompt-based large language models (LLMs) for fine-grained entity role

classification without the need for domain-specific supervised learning.

5.1 Quantitative Analysis and Model Comparison

To assess the effectiveness of different modeling strategies, we conducted internal experiments on the development dataset, evaluating three distinct architectures:

- Model 1: Hierarchical Pipeline with Preprocessing
- Model 2: Hierarchical Pipeline without Preprocessing
- Model 3: Single-Step Sub-Role Classification with Preprocessing.

Each model was tested with the two prediction strategies, Majority Voting (MV) and Non-Majority Voting (NMV).

5.2 Experimental Results

In Table 1, we can see the performance of different models and architectural approaches implemented.

Model	MV	NMV
Model 1	0.26	0.30
Model 2	0.20	0.22
Model 3	0.32	0.33

Table 1: Comparison of classification performance in EMR for sub-role classification

From Table 1 we can extract the following insights:

- Single-Step Classification with Preprocessing (Model 3) outperformed hierarchical models in both voting strategies. Eliminating sequential dependencies and jointly predicting roles and sub-roles improved classification accuracy. The model reached 0.33 EMR with NMV, surpassing both hierarchical approaches.
- Non-Majority Voting (NMV) consistently outperformed Majority Voting (MV) across all architectures. Although MV was initially introduced to reduce variance in LLM outputs, but unlike expected it generated inconsistencies, possibly due to stochastic variations in sub-role assignment. This suggests that direct one-pass classification is more stable than aggregated predictions from multiple runs.
- Hierarchical models (Model 1 and Model 2) underperformed, particularly when preprocessing was removed. Model 2, which omitted article summarization, recorded the lowest EMR (0.20–0.22), highlighting the significance of incorporating global context for improved role classification. Given these experimental findings, we selected Model 3 (Single-Step Classification with NMV) as our final submission, as it demonstrated superior accuracy, computational efficiency, and robustness compared to hierarchical approaches. The decision was further reinforced by its consistency across development and test evaluations, confirming its effectiveness as a taxonomy-driven LLM-based classifier.

6 Conclusion

This study highlights the growing relevance of Large Language Models in the classification and assessment of human behaviour, particularly in news narratives and structured socio-political taxonomies. Our work in the SemEval-2025 Fine-Grained Role Classification task demonstrates that LLMs, when properly aligned with human-defined taxonomies, can effectively infer entity roles and sub-roles within complex textual contexts. By structuring input signals, leveraging contextual processing, and optimizing classification strategies, LLMs can be guided to generate taxonomically coherent and interpretable outputs, reinforcing their potential for media analysis, governance, and ethical AI applications. A key insight from our findings is the critical role of input structuring and system design in aligning LLM-based classification with human-defined taxonomies. Our results underscore that carefully designed input representations, contextual enrichment, and decision pipelines significantly impact performance—even in zero-shot settings where no explicit model training is conducted. This insight is particularly valuable for AI applications in legal, political, and ethical decision-making, where interpretability and alignment with human cognitive frameworks are essential, and data privacy could represent an obstacle to the creation of training datasets.

7 Acknowledgments

This work was supported by the Spanish Research Agency (Agencia Estatal de Investigación) DeepInfo (PID2021-127777OB-C22) project (MCIU/AEI/FEDER,UE), the CHIST-ERA HAMiSoN project grant CHIST-ERA-21-OSNEM-002 and by the AEI PCI2022-135026-2 project.

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Sydney, Australia. Association for Computational Linguistics.

Maialen Berrondo-Otermin and Antonio Sarasa-Cabezuelo. 2023. Application of artificial intelligence techniques to detect fake news: A review. *Electronics*, 12(24).

Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency*, pages 149–159. PMLR.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

Fabrizio Gilardi, Sabrina Di Lorenzo, Juri Ezzaini, Beryl Santa, Benjamin Streiff, Eric Zurfluh, and Emma Hoes. 2024. Disclosure of ai-generated news increases engagement but does not reduce aversion, despite positive quality ratings. *arXiv preprint arXiv:2409.03500*.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on llm performance? *arXiv preprint arXiv:2411.10541*.

Anmol Jain, Aishwary Kumar, and Seba Susan. 2022. Evaluating deep neural network ensembles by majority voting cum meta-learning scheme. In *Soft Computing and Signal Processing: Proceedings of 3rd ICSCSP 2020, Volume 2*, pages 29–37. Springer.

Thanmay Jayakumar, Fauzan Farooqui, and Luqman Farooqui. 2023. Large language models are legal but they are not: Making the case for a powerful legalllm. *arXiv preprint arXiv:2311.08890*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952, Portorož, Slovenia. European Language Resources Association (ELRA).

Shabnam Nazmi, Xuyang Yan, Abdollah Homaifar, and Emily Doucette. 2020. Evolving multi-label classification rules by exploiting high-order label correlations. *Neurocomputing*, 417:176–186.

Jakub Piskorski, Tarek Mahmoud, Nikolaos Nikolaidis, Ricardo Campos, Alípio Jorge, Dimitar Dimitrov, PurificaÃ§Ã£o Silvano, Roman Yangarber, Shivam Sharma, Tanmoy Chakraborty, Nuno Ricardo GuimarÃ£es, Elisa Sartori, Nicolas Stefanovitch, Zhuohan Xie, Preslav Nakov, and Giovanni Da San Martino. 2025. SemEval-2025 task 10: Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation*, SemEval 2025, Vienna, Austria.

Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. 2025. Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*, PP:1–1.

Nicolas Stefanovitch, Tarek Mahmoud, Nikolaos Nikolaidis, Jorge Alípio, Ricardo Campos, Dimitar Dimitrov, PurificaÃ§Ã£o Silvano, Shivam Sharma, Roman Yangarber, Nuno GuimarÃ£es, Elisa Sartori, Ana Filipa Pacheco, Cecília Ortiz, Cláudia Couto, Glória Reis de Oliveira, Ari GonÃ§alves, Ivan Koychev, Ivo Moravski, Nicolo Faggiani, Sopho Kharazi, Bonka Kotseva, Ion Androutsopoulos, John Pavlopoulos, Gayatri Oke, Kanupriya Pathak, Dhairya Suman, Sohini Mazumdar, Tanmoy Chakraborty, Zhuohan Xie, Denis Kvachev, Irina Gatsuk, Ksenia Semenova, Matilda Villanen, Aamos Waher, Daria Lyakhnovich, Giovanni Da San Martino, Preslav Nakov, and Jakub Piskorski. 2025. Multilingual Characterization and Extraction of Narratives from Online News: Annotation Guidelines. Technical Report JRC141322, Ispra (Italy).

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Josef Valvoda, Alec Thompson, Ryan Cotterell, and Simone Teufel. 2024. The ethics of automating legal actors. *Transactions of the Association for Computational Linguistics*, 12:700–720.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou,

et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.