# CATKoDR: Hybrid Context-Awareness Model Architecture for Natural Language Processing

Nour Matta
*LIST3N*
*University of Technology of Troyes*
Troyes, France
nour.matta@utt.fr

Nada Matta
*LIST3N*
*University of Technology of Troyes*
Troyes, France
nada.matta@utt.fr

Agata Marcante
*R&D*
*Namkin*
Troyes, France
a.marcante@namkin.fr

Nicolas Declercq
*Namkin*
Troyes, France
n.declercq@namkin.fr

***Abstract*—Context plays a crucial role in understanding and representing the meaning of words. In previous work [1], we proposed a context awareness approach to enhance knowledge discovery from textual data. In this paper, we will present the system architecture to enable a hybrid approach to extract knowledge and its context from text.**

***Keywords— Text Mining, Semantic Network, Event Ordering, Knowledge Representation, Context-Awareness***

## I. INTRODUCTION

Knowledge Discovery is an important branch of Artificial Intelligence. Our main focus is textual knowledge discovery and how to enhance the understanding of the information extracted. Knowledge is defined as the meaning assigned to a symbol based on a specific reference [2]. When analyzing relationships between textual concepts, some relations and definitions are context dependent; the relationship "dog" "is a" "product", is a relation considered correct only in the context of a pet shop where dogs are sold. In our work, we aim to extract information from texts and create an episodic memory [3] of each text while keeping track of the context to enable a contextualized learning process. In this paper, we will present the architecture of our CATKoDR (**C**ontext-**A**ware **T**ext **Kno**wledge **D**iscovery and Representation) that will enable the extraction along with the pipeline.

## II. RELATED WORK

The context awareness field was introduced as means of enabling system the ability to adapt to a changing environment [4] and considering the context in the processing. In the text analysis, context plays an important role in eliminating ambiguity, improving coreference resolution and indicating referent, and detecting conversational implicature and intentions [5]. There are two different conceptualizations of context and context used in NLP. The first considers the context of the target word or phrase as its surroundings [6]. The second is relative to knowledge extraction and the use of ontologies. The use of predefined ontologies to orient and target the information, and the domain of the ontology would play the role of the context. It is needless to say that there is a bidirectional relationship between ontologies and natural language processing [7]. Ontologies can be populated using text analysis and texts can be used to extract and build ontologies.

When using different texts, the categorization of concepts may not identify changes in concept definitions, evolution analysis of the context and its elements. A need to track contextual dependency such as the information temporality to limit the inferences based on their context is identified. Therefore the need for a context awareness approach that provides contextual information about each processed text. In the following section, we will review the context elements and their semantic network.

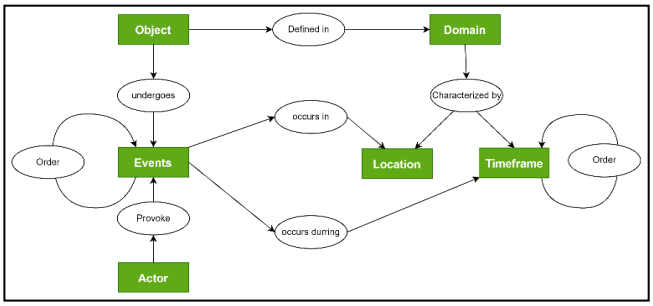
## III. CONTEXT ELEMENTS

In previous work [8], we presented the different context entities that we will be extracting and the semantic network. The first element is objects, they are extracted using hypernym-hyponym relation extraction. This enables the extraction of all types of entities mentioned along with their characteristic. Actors and Locations are extracted using a BERT model trained to extract entities from sentences based on their role in relation to the verb in accordance with VerbNet [9]. The domain is an external concept that can be provided by the user or through topic modeling in future work. The timeframe is a concept we introduced to enable gathering events that happen within the same timeframe to improve flashback and flash-forward analysis, but also consider direct speech timeframes and the publication date [10]. This approach reduces the number of temporal relationships extracted between different events by segmenting text into multiple narrative timeframes. Consecutive direct speeches are gathered in spoken timeframes and processed according to the narrative timeframe in which they appear. An event ordering model, CAEVO [11], is used to extract events, time expressions, and their temporal relations.

Figure 1 presents the semantic network available between the context entities. We can notice that the domain is characterized by the timeframes, locations, and objects mentioned. Objects are defined in a specific way in the domain and have subObjects. The main reason is that objects represent the extracted ontologies within a text and how they are defined through hypernym extraction. Events occur in a location during a timeframe and there is a temporal

relationship between events that will order them. Actors are entities that provoke events. Finally, timeframes are also ordered. This representation enables the creation of episodic knowledge. In the learning process to generate knowledge, an event coreference can be used to analyze the evolution of the same event in multiple texts over time, or the evolution of concepts over time. In the following section, we will present the pipeline of the context elements extraction.

Fig. 1. Context sementic Network.




## IV. CATKoDR Extraction Pipeline

The first step of the extraction process is the timeframe extraction. The text is segmented into multiple timeframes and the temporal relationship between the different timeframes is extracted. The next step is either object extraction or event ordering. For object extraction, we extract hypernym relationships using Hearst pattern extraction and structural base extraction. Objects keep track of the timeframe they appear in, but can be processed with the timeframe to analyze how the object evolves within a text. But can also be processed without consideration of the timeframe and be used for comparing concepts from one text to another. As for the event extraction, the event ordering model extracts the event, time expression, and the temporal relationship between them. The last step provided in the pipeline is the location and actor extraction but can be modified to extract all types of roles entities can play in events. In the following section, we provide the model architecture and the intended next step of our approach to improve contextual knowledge discovery from texts.

## V. Context Representation

Figure 2 presents the intended CATKoDR model architecture that aims a learning process on the extracted information. The pipeline provided previously is represented as the context entities extraction. The extracted elements will be stored in a NoSQL database that will be considered as the episodic knowledge base database. We will apply an event coreference to provide the identification of mention of the same event. An Entity Linking model will be also used to identify the same entities and allow the use of an index per entity to be able to query data that evolves specific entities. The semantic and synonymy analysis has the role of representing similar synonyms and concepts using one specific word to improve the analysis and also identify how these concepts changed over time. The last element is the similarity analysis to cluster different texts based on their graphical representation taking into account events, objects, actors, entities, and locations. This architecture has not been fully implemented but will be built in future work.

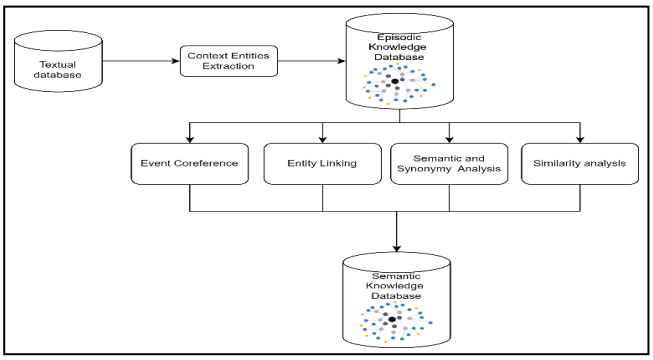



Fig. 2. CATKoDR Architecture.

## VI. Conclusion

Finally, we presented our CATKoDR model, context-aware textual knowledge discovery and representation model. From a text mining perspective, our goal is to provide a model that enables a learning process using each text we represented as episodic memory. This paper provided the context elements extraction pipeline along with the architecture of the extraction model. Our model uses semantic networks and knowledge representation from the symbolic AI to model the extracted information and provide a tool to build learning models on data available in natural language. CATKoDR uses a mix of methods, between pattern matching and machine learning to extract and populate the semantic network. It is also intended to apply more connectionist AI models to generate knowledge and build semantic memory making the CATKoDR a hybrid AI model.

## References


[1] N. Matta, N. Matta, A. Marcante, and N. Declercq, 'Structuring Context Entities for Knowledge Discovery', in 2022 International Conference on Computational Science and Computational Intelligence, Dec. 2022.

[2] B. Bachimont, A. Isaac, and R. Troncy, 'Semantic Commitment for Designing Ontologies: A Proposal', in *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*.

[3] J. Richard, 'Les Activités Mentales. Comprendre, Raisonner, Trouver des Solutions', *Rev. Philos. Fr. Etranger*, vol. 182, no. 4, pp. 491–491, 1992

[4] B. Schilit, N. Adams, and R. Want, 'Context-Aware Computing Applications', in *1994 First Workshop on Mobile Computing Systems and Applications*, Dec. 1994, pp. 85–90. doi: 10.1109/WMCSA.1994.16.

[5] S. Lichao, 'The Role of Context in Discourse Analysis', *J. Lang. Teach. Res.*, vol. 1, Nov. 2010, doi: 10.4304/jltr.1.6.876-879.

[6] A. Lenci, 'Distributional Models of Word Meaning', *Annu. Rev. Linguist.*, vol. 4, no. 1, pp. 151–171, 2018.

[7] A. Lenci, 'The life cycle of knowledge', Feb. 2023.

[8] N. Matta, N. Matta, E. Giret, and N. Declercq, 'Enhancing Textual Knowledge Discovery using a Context-Awareness Approach', in 2021 International Conference on Computational Science and Computational Intelligence, Dec. 2021, pp. 233–237.

[9] P. Shi and J. Lin, 'Simple BERT Models for Relation Extraction and Semantic Role Labeling'. arXiv, Apr. 10, 2019.

[10] N. Matta, N. Matta, N. Declercq, and A. Marcante, 'Semantic Patterns to Structure TimeFrames in Text', presented at the INTELLI 2022, The Eleventh International Conference on Intelligent Systems and Applications, May 2022, pp. 16–23.

[11] N. Chambers, T. Cassidy, B. McDowell, and S. Bethard, 'Dense Event Ordering with a Multi-Pass Architecture', *Trans. Assoc. Comput. Linguist.*, vol. 2, pp. 273–284, Oct. 2014, doi: 10.1162/tacl_a_00182.