A Method Based on NER and CRF for Extracting Named Entities from Text and Textual Representations of Chemical Reactions

Anna Vasileva[1] and Natalia Evstifeeva[2]
[1]*Department of Engineering Cybernetics, National Research Technological University MISIS, Leninsky Prospekt 4, 119049 Moscow, Russia*
[2]*Department of Engineering of Technological Equipment, National Research Technological University MISIS, Leninsky Prospekt 4, 119049 Moscow, Russia*
vasilevaaa.ao@gmail.com, evstifeeva@mail.ru

Keywords: Named Entity Recognition, Conditional Random Fields, Ontology-Based Information Extraction, Knowledge Graph Construction, Scientific Text Processing.

Abstract: The rapid growth of scientific and technological publications has increased the demand for automated methods capable of extracting structured knowledge from unstructured textual data. This problem is particularly relevant for chemical and technological texts, where essential information is often represented through chemical reactions, physical formulas, and domain-specific terminology that standard natural language processing techniques handle poorly. This paper proposes a hybrid information extraction method that combines Named Entity Recognition (NER), a rule-based MiningNOUN algorithm, and Conditional Random Fields (CRF) to improve the identification of entities and relationships in domain-specific scientific texts. The proposed approach integrates statistical and ontological principles, enabling the recognition of substances, processes, physical quantities, and formally structured expressions that are typically missed by baseline NER models. The method was evaluated on a corpus of chemical and technological texts describing experimental procedures and reaction processes. The results show that the combined NER + MiningNOUN + CRF configuration significantly increases the coverage of extracted entities compared to a standard NER pipeline, allowing the system to capture information expressed in both natural language and formal notation. The extracted entities and relations are integrated into an ontological knowledge graph compliant with RDF/OWL standards and further applied within a Retrieval-Augmented Generation (RAG) architecture. The proposed method supports the development of more reliable knowledge graphs for intelligent scientific data processing and can be adapted to other technical domains with complex symbolic representations.

1 INTRODUCTION

In recent years, the volume of scientific information presented in the form of natural language texts has grown significantly. However, automating the analysis of such texts and extracting formalized knowledge from them remains a challenging task. This research is particularly important in the context of chemical and technological texts, where a significant portion of the data is presented in the form of formulas, physical symbols, and mathematical expressions. These elements contain key information about processes, but their extraction and structuring are difficult using standard natural language processing methods.

The issue of automating the processing of data containing mathematical models of technological processes in the current state of the information field in the form of texts in natural language is currently a task with an achievable solution. In this regard, the most promising approach remains the use of RAG architecture to develop an information system based on a Large Language Model (LLM) in combinationwith a knowledge graph [1]. Such an architecture and set of components, on the one hand, provides effective integration of technological solutions for intelligent processing of data in the form of textual information on the natural language, and on the other hand, forms deep knowledge taking into account the specifics of the subject area based on the use of an ontological model as a knowledge graph in the form of formalism.

Modern LLMs are trained on big data volumes from various subject areas. However, their depth of

knowledge in highly specialized subject areas is still insufficient. Retraining Large Language Models (LLMs) is expensive. The RAG architecture offers a cost-effective alternative by combining existing LLM knowledge with structured data from a knowledge graph, enabling deeper and more accurate information integration. The use of RAG architecture makes it possible to avoid the process of direct LLM retraining and also reduces the problem of hallucination [2].

The goal of this research is to develop a method for automatically extracting entities and relationships from textual information to formalize a knowledge graph, with further use in RAG architecture. Unlike existing solutions, the proposed approach integrates the capabilities of NER, the MiningNOUN algorithm, and CRF, ensuring the formation of functional and hierarchical relationships for describing technological and chemical processes. At the same time, standard NER algorithms show limited results when processing text records of chemical reactions. They are unable to correctly identify chemical compounds, formulas, and expressions that have formal syntax and deviate from natural language. Therefore, an extended approach is required that is adapted to the structural, contextual, and semantic dependencies of the task and subject area.

2 LITERATURE OVERVIEW

Recently, the problem of automatically extracting chemical reactions from scientific texts has been gaining momentum. In the paper "Extracting chemical reactions from text using Snorkel", the authors use the Snorkel framework, based on weakly supervised learning, to extract reaction relations from annotated abstracts [3]. They achieve an accuracy of up to 84% with low recall (for a highly specialized corpus) and demonstrate scalability to large text corpora with a high degree of class imbalance.

Another approach presents a modular tool that extracts information about chemical reactions from text, tables, and illustrations [4]. The model combines neural components for different subtasks (molecule identification, reaction identification) and then combines the results using algorithms that take chemical constraints into account. The authors demonstrate an F1 score of approximately 69.5% for reaction schemes and an accuracy of 64.3% when compared to the Reaxys database.

A recent study, Suitability of large language models for extraction of high-quality chemical reaction dataset from patent literature, investigated the possibility of using LLM (GPT-3.5, Llama 2, etc.) to extract reaction data from patent documents. The authors show that an automated approach can add 26% new entries to chemical reaction databases and identify errors in previously manually compiled databases [5].

In addition to the chemical field, considerable attention is paid to general NER and relation extraction tasks. The article "Joint entity recognition and relation extraction as a multi-head selection problem" proposes a neural network model that simultaneously solves the tasks of entity recognition (using a CRF layer) and relation extraction, demonstrating the advantages of joint training over separate approaches [6].

An ontology-driven approach is also important. The paper "Ontology-Driven Extraction of Contextualized Information" explores a two-stage architecture where entities are first extracted and then linked to metadata (article, authors, etc.), after which everything is integrated into an RDF knowledge graph [7].

An analysis of existing research shows that modern methods of extracting information from scientific texts, in particular NER and CRF models, have achieved significant success in the tasks of annotation and term extraction. However, most approaches are focused on general text and do not provide effective processing of highly specialized data containing chemical reactions, physical formulas, and complex hierarchical relationships between entities.

The use of large language models (LLMs) demonstrates the potential for improving the quality of analysis. However, problems remain unresolved regarding the integration of extracted knowledge into formalized graph structures and ensuring their connectivity [8].

Thus, a method is needed that combines the advantages of structuring with simultaneous semantic fine-tuning that considers the contextual features of the analyzed information, ensuring both the extraction of entities and the construction of hierarchies between them. This paper proposes a hybrid method based on NER, the MiningNOUN algorithm, and CRF, integrated into the RAG architecture to form coherent and generalized ontological representations.

3 METHODOLOGY

3.1 System Architecture

The effectiveness of solving the problem of increasing knowledge depth using RAG architecture directly depends, among other things, on the

knowledge graph, both in terms of its structure and its content. This work uses a knowledge graph constructed in accordance with the ontological model of the subject area. The quality of the knowledge graph's content is determined not only by the algorithm used to form it, but also by the set of textual information expertly compiled in the form of a text corpus. The initial, basic corpus of textual information on the subject area is selected by an expert and contains, in addition to a textual description of the technological process, textual records of chemical reactions, physical formulas, and mathematical calculations. The addition of textual information and the expansion of the knowledge graph should occur without the direct participation of experts, but with verification and evaluation through a decision-making system. The main task in the process of evaluating textual information is to form a textual representation of the technological process that is as objective, complete, consistent, and accurate as possible, considering the possibility of integration.

The developed information system is presented in general terms in Figure 1. In addition to processing textual information about the subject area to construct a knowledge graph as an ontological model, ready-made ontological models from open sources are also used. They are also automatically evaluated in the decision-making system for compliance with international standards and generalized, generally accepted requirements for the presentation of data and knowledge for dictionaries and ontological models in the subject area.

Thus, the modular architecture of the system ensures the continuous integration of new data and automatic updating of the knowledge graph without the participation of an expert.

3.2 Ontological Model and Knowledge Graph Representation

In accordance with the international standard ISO 25964, the ontological model takes into account the requirements for term structures, hierarchical relationships, and compatibility standards according to the Simple Knowledge Organization System (SKOS). SKOS-compliant thesauri are those that have a precisely formalized knowledge organization model developed by the W3 consortium for the Semantic Web and based on its technologies.

In terms of data representation format, the developed ontological model is formed according to the structure and formal requirements in accordance with the Resource Description Framework (RDF). In addition, the developed knowledge graph of the ontological model fully complies with the Web Ontology Language (OWL). This adds the possibility of integrating the developed knowledge graph according to the protocol included in the W3C stack for the Semantic Web as a universally recognized international formalism for ontological models [9].

Like most graph models for representing knowledge, the developed structure breaks down into single triples that are semantically equivalent to the subject-predicate-object (SPO) structure, where predicate typing is limited to several types: functional type 1, functional type 2, hierarchical, attributive, and modifying. An example of a basic triplet is the construction (Substance → participatesIn → Reaction), where the nodes correspond to entities and the predicate reflects the type of their connection [10].

The resulting ontological model forms the basis for subsequent stages of extracting named entities and building hierarchical connections between them.

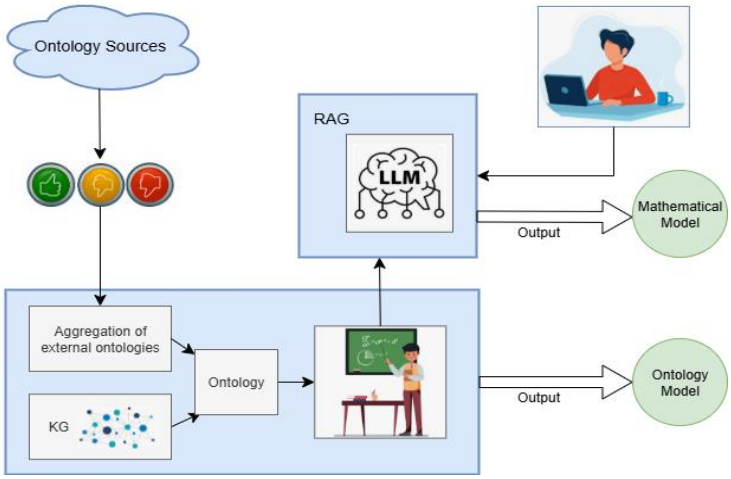

Figure 1: General architecture of the RAG-based knowledge graph system.

3.3 Hybrid NER + MiningNOUN

To identify entities and establish relationships between them, a hybrid approach is used, based on a combination of the classic NER algorithm and the developed MiningNOUN method, focused on the analysis of nouns and their contextual dependencies [11]. Let us consider the method of obtaining a hierarchical predicate type and establishing relationships by identifying a generalizing class or defining and adding it as an instance to an already named existing class, as well as defining a set of vertices as entities.

The algorithm for forming a set of nodes for a knowledge graph defines sets of entities more precisely than objects and subjects, which will be connected in pairs by a specific type of predicate, thus establishing a connection in the knowledge graph between the nodes. Two parallel algorithms are used to select a set of objects and subjects. The basic algorithm is Named-entity recognition (NER) [11]. NER is a classifier in the form of a light and fast-working neural network. The standard NER model correctly classifies only a limited set of categories, such as PERSON, ORG, LOC, DATE, and QUANTITY. However, when analyzing scientific and technical texts, significant terms (e.g., substances, parameters, physical properties) do not fall into these classes, which require additional processing. During the development of an algorithm based on NER, it was found that it generally copes well with the text source set for entity extraction, as shown in Figure 2.

Copper sulfate reacts with zinc to form zinc sulfate and copper. The reaction occurs at a temperature of 80°C in an aqueous solution. The produced copper mass is approximately 2.3 grams per experiment.

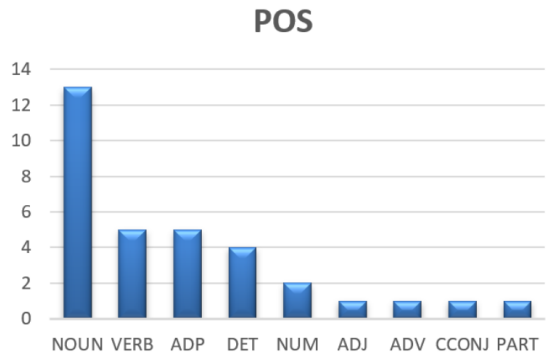

Figure 2: Classification of object types extracted from the text corpus.

As shown in Table 1, in the case of a corpus of texts on a given subject area and a description of the technological process of copper production, many semantically significant objects and subjects fall into the NOUN class. To compensate for the loss of significant objects, the MiningNOUN algorithm has been developed, which is integrated with NER and performs additional classification processing. MiningNOUN is based on a series of sequential evaluation algorithms, each chain of which forms an evaluation coefficient:

1) Resolving ambiguities associated with replacing pronouns with nouns that they were originally replaced with in the text;
2) Identifying unclassified tokens with nodes of the ontological model that was constructed earlier;
3) Detection of tokens from the NOUN class using a thesaurus;
4) Execution of chains of production rules that mimic grammatical constructions as templates of triplets with predicates reflecting functional and hierarchical relationships from general to specific or from specific to general.

Table 1: Composition of the NOUN class in the domain-specific corpus.

POS	COUNT	Example
NOUN	13	Copper, sulfate, zinc, reaction, temperature, solution, mass, grams, experiment
VERB	5	reacts, form, occurs, produced, is
ADP	5	with, at, of, in, per
DET	4	The, a, an

As a result of implementing the MiningNOUN algorithm, it was possible to refine the semantic tags for terms characteristic of chemical and technological texts. The algorithm assigns more accurate ontological classes to nouns that are not recognized by the standard NER model, as shown in Table 2.

Table 2: Examples of semantic refinement using the MiningNOUN algorithm.

Original token	NER Label	MiningNOUN Class
Copper	-	MATERIAL
Reaction	-	PROCESS
Temperature	-	PARAMETER
Mass	-	PHYSICAL_QUANTITY

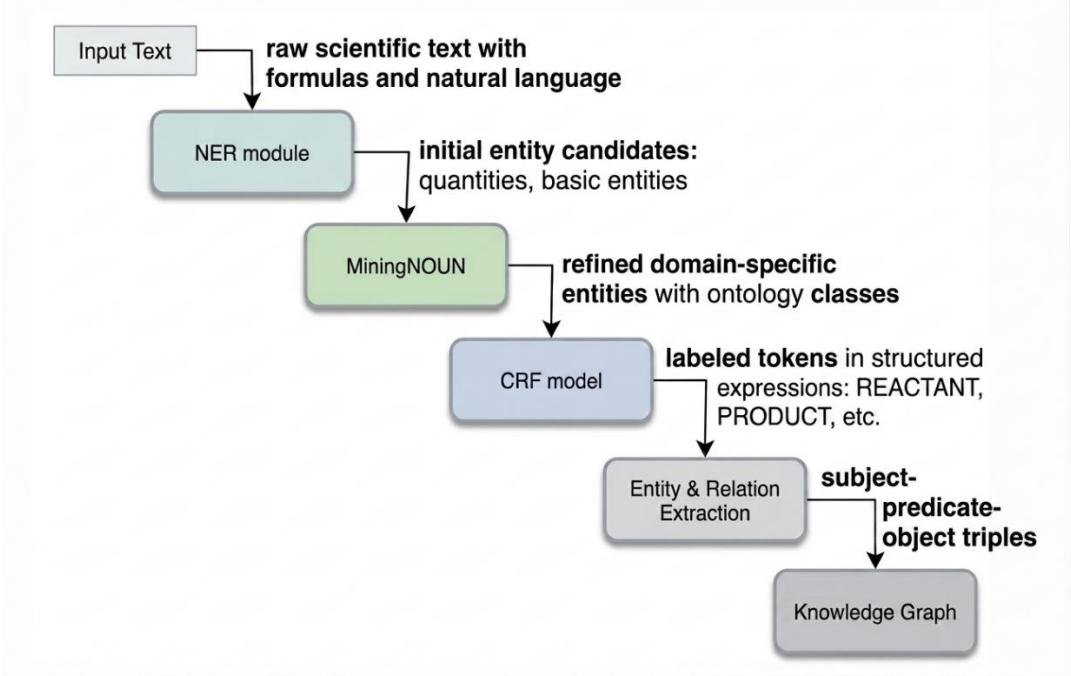

Figure 3: Interaction between the NER module, MiningNOUN module, and CRF model in the proposed hybrid architecture.

To better illustrate the interaction between the components of the proposed method, Figure 3 presents the overall processing pipeline. The system combines rule-based and statistical approaches, where different components perform complementary roles in entity extraction.

In this architecture, the NER module performs initial entity recognition from natural language text. The MiningNOUN module, implemented as a rule-based component, refines and enriches the extracted entities by incorporating domain-specific knowledge and ontological structures.

The CRF model is applied at the final stage to process structured expressions such as chemical reactions and formulas. Unlike the MiningNOUN module, which relies on predefined rules, the CRF component is a probabilistic sequence labeling model that captures dependencies between tokens in complex expressions.

This distinction explains the different roles of the MiningNOUN module and the CRF model within the system, as they address complementary aspects of the information extraction task.

3.4 CRF for Formulas

When identifying objects and subjects based on textual representations of chemical reactions, physical formulas, and mathematical calculations, the standard NER algorithm demonstrates a high level of misclassification. Classic models do not handle such contexts well due to the complex internal structure of expressions and the absence of linguistic features characteristic of natural text [12], [13]. To improve the efficiency of intelligent processing of such specialized data, an algorithm based on Conditional Random Fields (CRF) is used, which can be adapted to the task at hand. Its task is to identify structural components from text records of chemical reactions and formulas and assign them roles according to the classes of the knowledge graph vertices [14], [15].

The CRF model is trained on a labeled corpus that includes pairs of “formula - semantic role”. For each formula token, a set of features is formed:

1) Character type (letter, number, operator, index);
2) Token position in the expression;
3) Context (surrounding characters and symbols);
4) Belonging to the dictionary of chemical elements, physical quantities, or mathematical functions.

In the absence of labeled training data sets, CRF can be implemented as a set of rules and templates that imitate a probabilistic model, allowing for partially deterministic classification.

Compared to the Hidden Markov Model (HMM), where the current state only depends on the previous one, CRF calculates probabilities based on the whole sequence of hidden states, which makes it easier to recognize related parts of expressions [16]-[19]. This

is especially important when analyzing multi-level reactions and formulas, where a single symbol can have different meanings depending on the context.

As a result of the CRF module's work, each word in the text receives a corresponding label reflecting its role in the chemical expression. As shown in Figure 4, in the sentence "Hydrogen reacts with oxygen," both reagents (Hydrogen and oxygen) are classified as B-REACTANT, while function words (reacts, with) are marked as O (outside entity). This approach makes it possible to model the relationships between tokens and form structured representations of knowledge in the form of triples (subject-predicate-object) that can be integrated into an ontological model.

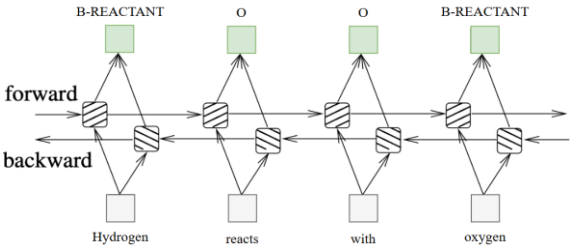

Figure 4: Example of CRF-based token classification in a chemical sentence.

3.5 Dataset and Evaluation Setup

To evaluate the proposed method, a dataset was constructed consisting of 850 sentences collected from scientific publications in the field of chemical engineering. The corpus includes descriptions of experimental procedures, chemical reactions, and physical processes, containing both natural language and structured expressions such as formulas.

The dataset was manually annotated to identify entities belonging to the following categories: MATERIAL, PROCESS, PARAMETER, and PHYSICAL_QUANTITY. In addition, tokens related to chemical reactions were labeled according to their semantic roles (e.g., REACTANT, PRODUCT, OPERATOR).

The dataset was divided into training and evaluation subsets, where approximately 70% of the data was used for model configuration and rule tuning, and 30% was used for evaluation.

4 RESULTS

To demonstrate the effective performance of the NER+MiningNOUN algorithm, a set of sentences from scientific publications in the chemical industry was compiled. This text contains numerical values representing temperature or time, formulas, and other chemical terms. After testing it on the standard SpaCy model "en_core_web_sm," we can see once again that the model is not designed to work with scientific text. The algorithm correctly recognized only basic numerical parameters (time, quantity, temperature), while chemical compounds and physical parameters were not classified. This confirms the need to refine the base model by integrating the MiningNOUN module and specialized analysis of formal expressions (see Fig. 5).

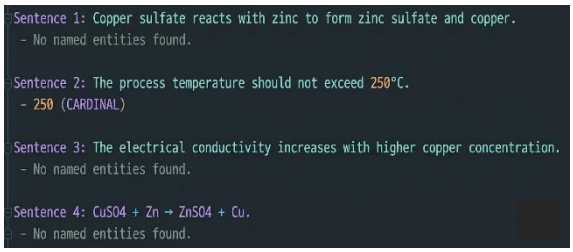

Figure 5: Classification of object types extracted from the text corpus.

The addition of the developed MiningNOUN module, which uses subject area ontology, has significantly expanded the number of recognized entities. In particular, elements like copper or zinc, chemical compounds (sulfate, solution), and physical parameters (temperature, pressure, energy, density) were correctly classified.

A manually compiled dictionary of subject area terms was used as the basic ontological model. In the future, integration with open ontologies such as ChEBI and QUDT is planned, which will ensure compatibility with the RDF/OWL format.

According to the experiment, the resulting copper mass was 12 grams.

- 12 grams (QUANTITY) [NER]
- copper (MATERIAL) [MiningNOUN]
- mass (PHYSICAL_QUANTITY) [MiningNOUN]

Figure 6: Example of token classification using NER and MiningNOUN.

To demonstrate how the proposed MiningNOUN algorithm works, let us consider an example of analyzing a fragment of the corpus describing experimental data (see Fig. 6). The basic NER algorithm correctly identifies quantitative expressions (e.g., "12 grams") but does not identify other important entities such as copper and mass, as they do not belong to the predefined categories of the standard model.

The MiningNOUN module performs additional processing and classifies these tokens, correlating them with the ontological classes MATERIAL and PHYSICAL_QUANTITY. This achieves more complete coverage of the subject area and refines the semantic structure of the text.

At the final stage, a subsystem for analyzing structural expressions was implemented, based on the Conditional Random Fields (CRF) model, adapted to the specifics of the subject area. Its task is to process text fragments containing chemical reactions, physical formulas, and mathematical equations to identify structural elements and their semantic roles.

The implemented CRF module is integrated into the overall system alongside NER and MiningNOUN. While the first two blocks are responsible for extracting entities from descriptive texts, the CRF module processes expressions with a formal structure, where classic NER algorithms produce a high percentage of misclassifications. To this end, formula analysis is implemented using regular expressions that mimic CRF behavior and allow variables, reagents, and operators to be identified.

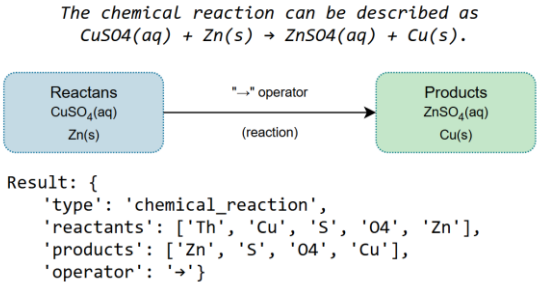

Figure 7: Example of CRF-based parsing of a chemical reaction.

The results of the module's work showed that adding the CRF component significantly increases the completeness of information extraction. Textual representations of chemical reactions and mathematical formulas that were not previously recognized by NER are now correctly classified and integrated into the knowledge graph. An example of the analysis is shown in Figure 7, which shows the selection of structural elements and their comparison with the classes of the ontological model.

Thus, the inclusion of the CRF module made it possible to eliminate gaps in the extraction of entities associated with formal expressions and ensured the completeness of the graph representation of the technological process.

Table 3: Evaluation results of the proposed method.

Method	Precision	Recall	F1-score
NER	0.68	0.54	0.61
NER + MiningNOUN	0.81	0.72	0.76
NER + MiningNOUN + CRF	0.88	0.79	0.83

To quantitatively evaluate the performance of the proposed method, standard evaluation metrics including Precision, Recall, and F1-score were used.

The results presented in Table 3 demonstrate that the baseline NER model shows relatively low recall due to its inability to recognize domain-specific entities. The integration of the MiningNOUN module significantly improves both precision and recall by enriching the set of extracted entities using ontological rules.

The addition of the CRF model further enhances performance, particularly in the analysis of structured expressions such as chemical reactions and formulas. As a result, the combined NER + MiningNOUN + CRF configuration achieves the highest F1-score among the evaluated methods.

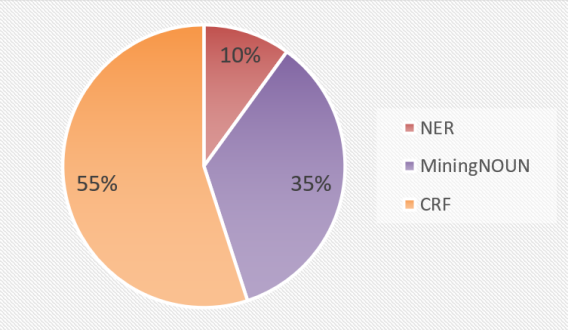

Figure 8: Distribution of extracted entities across NER, MiningNOUN, and CRF methods.

Figure 8 shows the distribution of extracted entities among the three configurations of the proposed system: the basic NER algorithm, the hybrid NER + MiningNOUN, and the extended NER + MiningNOUN + CRF. The values for the diagram were calculated as the average number of extracted entities per sentence.

Thus, the pie chart shows that the contribution of the MiningNOUN and CRF modules accounts for over 80% of the total number of entities found, significantly expanding the semantic coverage of the knowledge graph.

5 CONCLUSIONS

The paper proposes a hybrid method for extracting structured knowledge from scientific and technological texts based on a combination of NER, MiningNOUN, and CRF algorithms. The developed approach demonstrated a significant improvement in the quality of entity extraction compared to basic NER, especially when processing terms related to chemical and physical processes.

The integration of the MiningNOUN module made it possible to additionally identify nouns and match them with ontology classes, while the application of the CRF model ensured the correct analysis of chemical and mathematical expressions, which increased the accuracy of the semantic interpretation of the text.

The results of the experiments confirmed that the use of the combined NER + MiningNOUN + CRF system increases the completeness and accuracy of entity extraction, contributing to the enrichment of the knowledge graph in RAG architecture. In the future, we plan to expand the ontological base using external sources (e.g., ChEBI and QUDT), as well as train the CRF model on annotated corpora containing texts from specialized subject areas to improve the accuracy of entity role classification in complex scientific expressions.

REFERENCES

[1] L. Ehrlinger and W. Wöß, “Towards a definition of knowledge graphs,” in Proc. SEMANTiCS 2016 Conf., 2016.

[2] A. Hogan et al., “Knowledge graphs,” in ACM Computing Surveys, vol. 54, no. 4, pp. 1-37, 2021, [Online]. Available: https://doi.org/10.1145/3447772.

[3] S. L. Dixon, K. R. M. Mackay, and A. A. Butler, “Extracting chemical reactions from text using Snorkel,” in BMC Bioinformatics, vol. 21, no. 1, pp. 1-14, 2020, [Online]. Available: https://doi.org/10.1186/s12859-020-03542-1.

[4] Y. Chen, M. Sun, and J. Zhao, “OpenChemIE: An information extraction toolkit for chemistry literature,” arXiv:2404.01462, 2024.

[5] S. I. Sanabria and T. N. Hart, “Suitability of large language models for extraction of high-quality chemical reaction dataset from patent literature,” in Journal of Cheminformatics, vol. 16, no. 1, pp. 1-12, 2024, [Online]. Available: https://doi.org/10.1186/s13321-024-00928-8.

[6] D. Bekoulis, J. Deleu, T. Demeester, and C. Develder, “Joint entity recognition and relation extraction as a multi-head selection problem,” arXiv:1804.07847, 2018, doi: 10.48550/arXiv.1804.07847.

[7] J. Ferreira, R. Martins, and M. Araújo, “Ontology-driven extraction of contextualized information,” in Proc. ICAART 2023, vol. 3, pp. 642-649, 2023.

[8] İ. Karadeniz and A. Özgür, “Linking entities through an ontology using word embeddings and syntactic re-ranking,” in BMC Bioinformatics, vol. 20, no. 1, p. 156, 2019, [Online]. Available: https://doi.org/10.1186/s12859-019-2678-8.

[9] M. Y. Jaradeh et al., “Information extraction pipelines for knowledge graphs,” in Knowledge and Information Systems, 2023, [Online]. Available: https://doi.org/10.1007/s10115-022-01826-x.

[10] Q. Qiu et al., “Integrating NLP and ontology matching into a unified system for automated information extraction from geological hazard reports,” in Journal of Earth Science, vol. 34, no. 5, pp. 1433-1446, 2023, [Online]. Available: https://doi.org/10.1007/s12583-022-1716-z.

[11] Z. Han and J. Wang, “Knowledge enhanced graph inference network based entity-relation extraction and knowledge graph construction for industrial domain,” in Frontiers of Engineering Management, vol. 11, no. 1, pp. 143-158, 2024, [Online]. Available: https://doi.org/10.1007/s42524-023-0273-1.

[12] K. Kozaki et al., “Role representation model using OWL and SWRL,” in Proc. Workshop on Roles and Relationships in Object-Oriented Programming, Multiagent Systems, and Ontologies, 2007, pp. 39-46.

[13] L. Massel, T. Vorozhtsova, and N. I. Pjatkova, “Ontology engineering to support strategic decision-making in the energy sector,” in Ontology of Designing, vol. 7, pp. 66-76, 2017, [Online]. Available: https://doi.org/10.18287/2223-9537-2017-7-1-66-76.

[14] S. Yu, “Application of artificial intelligence methods in knowledge graphs,” in Applied and Computational Engineering, vol. 106, pp. 52-58, 2024, [Online]. Available: https://doi.org/10.54254/2755-2721/106/20241287.

[15] D. Vrandecic and M. Krötzsch, “Wikidata: A free collaborative knowledge base,” in Communications of the ACM, vol. 57, no. 10, pp. 78-85, 2014, [Online]. Available: https://doi.org/10.1145/2629489.

[16] M. Nickel et al., “A review of relational machine learning for knowledge graphs,” in Proceedings of the IEEE, vol. 104, no. 1, pp. 11-33, 2016, [Online]. Available: https://doi.org/10.1109/JPROC.2015.2483592.

[17] R. Speer, J. Chin, and C. Havasi, “ConceptNet 5.5: An open multilingual graph of general knowledge,” in Proc. AAAI Conf. Artificial Intelligence, 2017, [Online]. Available: https://doi.org/10.1609/aaai.v31i1.11164.

[18] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” arXiv:1908.10084, 2019.

[19] F. Souza, R. Nogueira, and R. Lotufo, “BERTimbau: Pretrained BERT models for Brazilian Portuguese NLP,” arXiv:1909.10649, 2019.