An Annotated Dataset for Transformer-based Scholarly Information Extraction and Linguistic Linked Data Generation

Vayianos Pertsas°♦, Marialena Kasapaki°, Panos Constantopoulos°♦
°Athens University of Economics and Business, Department of Informatics
♦Athena R.C., Institute for the Management of Information Systems
vpertsas@aueb.gr, kasapakimariael@gmail.com, panosc@aueb.gr

Abstract

We present a manually curated and annotated, multidisciplinary dataset of 15,262 sentences from research articles (abstract and main text) that can be used for transformer-based extraction from scholarly publications of three types of entities: 1) research methods, named entities of variable length, 2) research goals, entities that appear as textual spans of variable length with mostly fixed lexico-syntactic-structure, and 3) research activities, entities that appear as textual spans of variable length with complex lexico-syntactic structure. We explore the capabilities of our dataset by using it for training/fine-tuning various ML and transformer-based models. We compare our finetuned models as well as LLM responses (chat-GPT 3.5) based on 10-shot learning, by measuring F1 scores in token-based, entity-based strict and entity-based partial evaluations across interdisciplinary and discipline-specific datasets in order to capture any possible differences in discipline-oriented writing styles. Results show that fine tuning of transformer-based models significantly outperforms the performance of few-shot learning of LLMs such as chat-GPT, highlighting the significance of annotation datasets in such tasks. Our dataset can also be used as a source for linguistic linked data by itself. We demonstrate this by presenting indicative queries in SPARQL, executed over such an RDF knowledge graph.

Keywords: Information Extraction from Text, Transformer-based Information Extraction, Scholarly Annotation Corpus, Linguistic Linked Data, RDF Knowledge Graph

1. Introduction

The steep increase of research publications in every major discipline (Bornmann et al., 2021) makes it increasingly difficult for experts to maintain an overview of their domain, increases the risk of missing new work or reinventing solutions, and makes it harder to relate ideas from different domains. To address this problem new "strategic reading" methodologies can be applied in order to transform the essence of knowledge encoded in textual form into structured format comprising concepts and relations that address the information needs of researchers, thus changing the ways in which they engage with literature (Renear & Palmer, 2009). This type of encoded information can alleviate the task of keeping up to date in a specific domain, while maintaining a bird's-eye-view over a discipline or across disciplines, something particularly useful in interdisciplinary fields. To this end, entities representing the encoded information need to be appropriately identified and extracted from text through the use of various NLP and ML methods. This task has been significantly alleviated by the recent advancements in Deep Learning, where the application of transformer-based models in various NLP tasks (Vaswani et al., 2017) enabled the extraction of semantically complex information from text, while at the same time increased the demand for large annotated datasets for fine-tuning the millions of parameters of those models.

Indeed, information extraction (IE) from scientific papers has attracted a lot of interest over the past years, as testified by the recent creation of various challenges on Scientific Information Extraction (ScienceIE). This constant challenge for new ML methods for ScienceIE calls for additional new datasets, capable of demonstrating and benchmarking the new capabilities of those methods.

In addition, despite the recent advancements in Large Language Models (LLMs) such as chat-GPT[1] and its remarkable ability to generate text that resembles human-like language, as demonstrated by numerous studies (Gao et al., 2023; Jimenez Gutierrez et al., 2022; X. Li et al., 2023; Ma et al., 2023; Qin et al., 2023; Qiu & Jin, 2024), when it comes to NLP tasks like IE and NER, these models underperform significantly compared to DL models that are finetuned in task specific annotated datasets, thus showcasing even more the significance of the latter in IE tasks.

In this paper we present such a manually curated dataset comprising of 15,262 sentences sampled from 3,500 research publications and 172 research subfields, that is specifically designed for extracting various types of entities of varied semantic complexities and lexico-syntactic characteristics. Specifically, we offer annotations for three different types of entitles: 1) research methods, named entities of variable length, 2) research goals, entities that appear as textual spans of variable length with mostly fixed lexico-syntactic-structure, and 3) research activities, entities that appear as textual spans of variable length with complex lexico-syntactic structure.

[1] https://chat.openai.com/chat

LDL: Linked Data in Linguistics @ LREC-COLING 2024, pages 84–93
20-25 May, 2024.

The concepts in this dataset are designed to be general enough so that they can be applied across disciplines and, at the same time, be capable of representing essential knowledge of "who has done what, why and how" in a research paper. Extracting such information can lead to creating RDF Knowledge Graphs capable of answering complex semantic queries like: *"find all papers that address a given problem"; "how was the problem solved"; "which methods are employed by whom in an activity addressing particular research goals",* etc. (Pertsas & Constantopoulos, 2023). This goes beyond the retrieval features of search engines widely used by researchers, such as Google Scholar[2], Scopus[3] or Semantic Scholar[4] that mostly leverage bibliographic metadata, while knowledge expressed in the actual text is exploited mostly by matching query terms to documents.

We explore the capabilities of our dataset along four dimensions: 1) *Classification Method*: we experiment with training/fine-tuning various ML and DL models as well as LLMs (chat-GPT 3.5) through prompting; 2) *Linguistic Characteristics*: we explore the performance of our methods across interdisciplinary and discipline-specific subsets in order to capture any possible differences in discipline-oriented writing styles as demonstrated in (Alluqmani & Shamir, 2018; Leong, 2024); 3) *Processing Granularity*: we test the effectiveness of classification at three levels of granularity: token-based, entity-based strict and entity-based partial. In addition, the included entities represent three levels of lexico-syntactic complexity: named entities of variable length, "non-named" entities (i.e. non real world objects that can't be denoted with proper names) that are of variable length with mostly fixed lexico-syntactic-structure and variable length with complex lexico-syntactic structure; 4) *Linguistic Linked Data Generation*: we demonstrate the capabilities of our dataset as a source for linguistic linked data, through semantically complex queries in SPARQL that can be executed over such an RDF Knowledge Graph.

The rest of the paper proceeds as follows: in Section 2 we present related work regarding the creation of datasets for Science IE; in Section 3 we present the characteristics of our dataset and describe the methodology for its creation; in Section 4 we demonstrate the capabilities of the dataset through various experiments with ML, DL transformer-based and LLM prompting methods; in Section 5 we discuss the performance of the dataset based on the evaluation experiments and demonstrate its capabilities as a source for linguistic linked data and in Section 6 we conclude the paper with insights for future work.

2. Related Work

Information extraction from scientific text constitutes an active research field where ML and DL models are trained/fine-tuned on annotated corpora designed for capturing specific knowledge according to the task at hand. Entity extraction is usually treated as a token classification or sequence labeling task where a classifier predicts whether each token belongs to the entity in question or not, based on the corresponding token-based annotations. In addition, recent advancements in LLMs have given rise to new methodologies regarding prompting techniques for interacting with these models based on few or even zero demonstrating examples in few / zero-shot learning (Brown et al., 2020; Das et al., 2022; Lu et al., 2022; Perez et al., 2021; X. Wei et al., 2023), while others implement chain-of-thought (CoT) reasoning (Ashok & Lipton, 2023; J. Wei et al., 2023) that can help in reasoning tasks such as solving mathematical problems, or works like (P. Li et al., 2023; Wang et al., 2023) that experiment with code generation. In our work, for comparison purposes, we include in our dataset experiments, a prompt template for LLMs (chat-GPT 3.5) that leverages both few-shot and code structure transformation.

Concerning the creation of datasets that can be used for IE, in domain specific fields like Biology and Bioinformatics, works like the BioText project (Rosario & Hearst, 2004) offer semantically annotated corpora, consisting of 3500 sentences drawn from MEDLINE abstracts labelled for *Disease* and *Treatment* and seven types of relation holding between them. In (Franzén et al., 2002; Kim et al., 2003) the Yapex and GENIA corpora offer annotated sentences with named entities of proteins and specific biological entities and events respectively. Regarding Medicine and Health Sciences, in (Roberts et al., 2009) the authors present a dataset from clinical texts, annotated with domain specific entities like *Condition, Investigation, Drug, Locus* etc. interrelated with relations: *has_target, has_type, location, modifies*. In (Borchert et al., 2022) the authors present a dataset of annotated named entities regarding Oncology (e.g. *Finding, Substance, Procedure*), which then evaluate using transformer-based models. In (Cheng et al., 2022) the authors present a manually annotated dataset from Japanese clinical reports with entities representing medical terms like *Diseases and Symptoms* and *Medicine*, as well as medical and temporal relations among them, which they evaluate using ML models. In Material Science, the authors of (Mullick et al., 2022) annotate a corpus with entities of type: *Code, Material, Method, Parameter* and *Structure* in order to train and evaluate their ML pipeline architecture.

In interdisciplinary ScienceIE projects, works like (Jain et al., 2020; Luan et al., 2018) present SciREC and SciREX, datasets from paper abstracts containing annotations of scientific entities (*Task, Method, Metric, Material, Other-ScientificTerm* and *Generic*). In (Qasemi, Zadeh & Schumann, 2016) a corpus of paper abstracts is manually annotated with terms classified into categories like *Method, Tool,*

[2] https://scholar.google.com/

[3] https://www.scopus.com/home.uri

[4] https://www.semanticscholar.org/

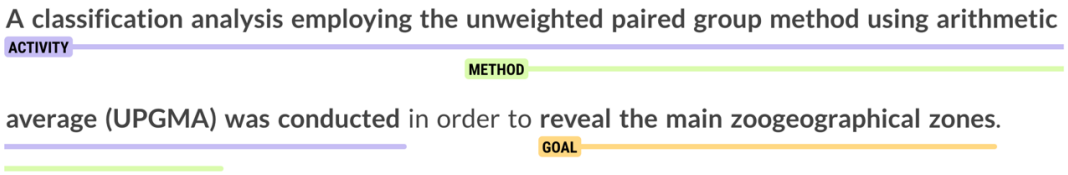

Figure 1: Example of Activity in passive voice, Method and Goal

Figure 2: Example of Activity in active voice, Method and Goal

Language Resource, Product, etc. In (Osenova et al., 2022) the authors present the Bulgarian Event corpus with annotations of named entities like *Locations, Events, Products*, etc. derived from the CIDOC-CRM Ontology and oriented mainly to Social Sciences and Humanities. In (Augenstein et al., 2017) the authors present a dataset with annotations of named entities like *Process, Task, Material* and relations like *hyponym-of* and *synonym-of*.

Compared to these works, we use a multidisciplinary dataset deriving from more than 170 research subfields in order to capture potential differences in writing styles among disciplines (Alluqmani & Shamir, 2018), since we use concepts that are general enough to be applied in any scientific field. In addition, to the best of our knowledge, our dataset is the first to contain entities of such lexico-syntactic complexity and variation in form and length. In this sense, it can be used for showcasing the capabilities of ML models in capturing various attributes of English language in a scholarly publication and not only those contained in a form of a named entity or an entity of relatively small length and fixed lexico-syntactic structure. The use of such semantically complex and -of highly variable length- entities, makes the problem of IE more challenging when it comes to employing prompting techniques for LLMs (as demonstrated in Section 4), thus showcasing the value of creating large, annotated datasets that can instead fine-tune DL transformer-based models with higher performance in such tasks.

3. Dataset Creation Methodology

For the creation of our dataset, we initially gathered a set of 25,681 papers spanning years 2000-2021 from JSTOR repository using the Constellate[5] portal. This initial material after various NLP processes for OCR Noise removal, text cleaning, tokenization and sentence segmentation, yielded in total 3,700,000 cleaned sentences. From those, we randomly sampled a total of 15,262 sentences deriving from 3,500 papers which, according to articles' metadata (fields: "publisher" and "tdmCategory") were published under 352 different publishers and derived from 172 different disciplines and subfields. The dataset is in English language since this is most commonly used in academia. The aim was to create a multidisciplinary corpus capturing as many different writing styles as possible.

The conceptual model behind the annotation schema is Scholarly Ontology (SO) (Pertsas & Constantopoulos, 2017), a domain-independent ontology of scholarly/scientific work. A specialization, in fact precursor, of SO already applied to the domain of Digital Humanities (that being an interdisciplinary field itself) is the NeDiMAH Methods Ontology (NeMO) (Constantopoulos et al., 2016). A brief overview of the definitions of SO concepts that were used in the annotation schema and guidelines is given below. For a full account see (Pertsas & Constantopoulos, 2017).

3.1 Annotation Schema

The Annotation schema used for the creation of this dataset was based on the following SO concepts and relations:

Activity: Instances of the Activity class represent research processes or steps thereof such as an experiment, a medical or social study, an archaeological excavation, etc. They usually manifest in text as spans of phrases in passive or active voice in first person singular or plural, according to the number of authors who are their actual participants.

Method: In contrast to activities, which are actual events carried out by actors, instances of the *Method* class denote procedures, such as an algorithm, a technique or a scheme that can be employed during an activity and describe how this was carried out. They are usually designated by single or multiple word terms, e.g. "ANOVA", "radio-carbon dating", etc., so their manifestations in text are mostly identified as named entities of variable length.

Goal: Goals represent the objectives of the activities and describe the intentional framework in which they were carried out. In addition, instances of the Goal class can represent general research goals of the paper that summarize the research objectives of all the activities described in it. In either case, they manifest in text as spans that declare purpose and are mostly introduced with purpose clauses like "for", "to" or "in order to".

[5] https://constellate.org/

Indicative examples of all the above textual manifestations of SO classes and relations can be seen in Figures 1 and 2.

3.2 Annotation Process

The annotation process was based on protocols described in (Roberts et al., 2009) and involved a trial phase during which three annotators, after appropriate training in the SO concepts, participated in 5 consecutive annotation trials covering in total 500 sentences from 300 papers. Each trial was followed by review of the entire batch by the group, discussion on the results and differences among annotations, re-adjustment of the annotation guidelines and evaluation of the inter-annotator agreement (IAA) using the Cohen's Kappa metric for IAA between annotator couples and Fleiss' Kappa for the group of three. We used the Prodigy[6] annotation tool for all the annotations and developed a Prodigy recipe for calculating the IAA scores.

After the trials, the best IAA scores reached 0.89 for *Activity*, 0.91 for *Method* and 0.92 for Goal, yielding sufficient agreement levels so that annotators could subsequently work on separate datasets. The entire annotator training process lasted approximately 25 hours.

As a general comment regarding the annotation of different types of entities, the most difficult type to agree upon was the Activity class. This can be attributed to the complexity of the lexico-syntactic structure of that particular entity type that produced differences among annotators, especially in the identification of boundaries in cases of very large lengths (compound phrases). On the other hand, Methods and Goals with clearer lexico-syntactic structures were easier to agree upon as can be seen from the higher agreement levels starting even from the first trial.

In addition to the annotation labels for the entities, the annotators used three "meta" labels for all the annotation sentences / spans: 1) *Accept*, where the annotator was confident for the annotation and the sentence/span is OK to be included in the dataset; 2) *Reject,* for the cases where the sentence/span was incomprehensible due to high noise from non-Unicode artifacts or non-English language and thus were to be excluded from the dataset; 3) *Ignore*, for the cases where the sentence/span was comprehensible but it wasn't clear if the annotation fulfils the specifications of the task at hand. The latter were agreed to be included in the dataset, since they can provide valuable material for other experiments, but not to be counted for the experiments mentioned in this paper since they were considered as prone to create outliers due to their ambiguity. Nevertheless, these cases were very few, counting less than 3% of the entire dataset.

When the annotation task was completed, the entire dataset was adjudicated by one annotator in order to maintain a constant annotation style throughout the entire dataset. Analytical results (group IAA) for each annotation trial and entity/relation type that show the progress in the agreement of the annotation tasks are presented in Table 1.

	Trial1	Trial2	Trial3	Trial4	Trial5
Activity	0.69	0.73	0.78	0.81	0.89
Method	0.71	0.78	0.84	0.89	0.91
Goal	0.81	0.86	0.92	0.90	0.92

Table 1: IAA scores per entity type for each annotation trial

3.3 Dataset Statistics

The annotation statistics of the final dataset, after adjudication, are shown in Table 2. In total, the dataset comprises 15,262 sentences and 517,499 tokens. At sentence level, the dataset contains 10,754 labeled sentences (i.e. sentences that contain at least one label). At span level (as a span we consider each individual textual chunk that is annotated as an entity) there are in total 19,173 entity labels (i.e., labels assigned to spans to denote them as activities, methods or goals). At token level (as tokens we consider individual lexical units like words, punctuation marks, etc.) the dataset contains in total 192,087 labeled tokens (i.e. annotation labels assigned to tokens, to denote them as part of a textual span representing an activity, goal and/or a method). Compared to other published benchmarks in ScienceIE tasks (Augenstein et al., 2017; Jain et al., 2020; Luan et al., 2018; Qasemi, Zadeh & Schumann, 2016) our dataset shows similar or higher numbers of annotations, which renders it a good source for ground truth in such experiments. The annotated dataset in jsonl format can be accessed from GitHub[7].

	Activity	Method	Goal	Total
Sent-level	6,610	6,028	4,029	10,754
Span-level	7,211	7,415	4,547	19,173
Token-level	126,702	14,036	51,349	192,087

Table 2: Dataset statistics for entity extraction

4. Experimental Setup

In order to evaluate the capabilities of the dataset in terms of how well it can fine-tune / train different types of ML models for performing the task at hand, we designed a total of 36 experiments measuring performance in entity extraction task.

4.1 Models and Methods

From the annotated dataset after random shuffling, we held out 20% for the evaluation set and the rest we split into training and development sets with the latter being 10% of the training set. We balanced our training sets but left unbalanced the evaluation sets so that we could measure performance in real case scenarios.

[6] https://prodi.gy/

[7] https://github.com/athenarc/ScholarlyIE-Datasets/

	Training/development			Total Evaluation set			H&B Subset			Humanities Subset		
	Act	Meth	Goal	Act	Meth	Goal	Act	Meth	Goal	Act	Meth	Goal
Sent	4,329	4,259	2,492	2,281	1,769	1,537	1,242	1072	699	1,008	658	889
Span	4,727	5,250	2,840	2,484	2,165	1,707	1,357	1,338	781	1,095	755	984
Token	84,469	9,716	32,237	42,233	4,320	19,112	24,577	2,706	9,665	15,944	1,489	10,237

Table 3: Number of annotated spans of the train/dev and eval subsets at sentence, span and token level.

From each of the following inputs included in "text" field, identify the textual spans representing entities that are defined as follows:

ACTIVITY: a research process like an experiment or a survey that is carried out by the author of the text.
GOAL: an objective of a research activity or a general research objective of the author of the text.
METHOD: the name of a research method denoting a procedure or a technique that was employed during an activity.

Your output should be in jsonl format containing the fields: "text" for the text in the input and "spans", a list of all the annotated spans each in a dictionary with the following fields: "start": character-based pointer to the start of the span, "end": character-based pointer to the end of the span, "token_start": token-based pointer to the start of the span, "token_end": character-based pointer to the end of the span, "label": the label of the annotated span, "span": the textual span of the annotation.

Below are some indicative examples that can be used as a guide.

Examples:
{"text":"A classification analysis employing the unweighted paired group method using arithmetic average (UPGMA) was conducted in order to reveal the main zoogeographical zones.", "spans":[{"start":0,"end":117,"token_start":0,"token_end":16,"label":"ACTIVITY","span":" A classification analysis employing the unweighted paired group method using arithmetic average (UPGMA) was conducted"}, {"start":40,"end":103,"token_start":5,"token_end":14,"label":"METHOD","span":" unweighted paired group method using arithmetic average (UPGMA)"}, {"start":130,"end":167,"token_start":20,"token_end":24,"label":"GOAL","span":" reveal the main zoogeographical zones "}]}
....
....

Input:
{"text":"All bryophyte species were collected and identified in the laboratory with the aid of a stereo microscope and a light microscope (Leica DMLB, Leica Microsystems SAS, Rueil Malmaison, France)."}
....
....

Figure 3: Indicative example of the prompt template. Each section is highlighted in different color.

In addition, in order to explore further possible differences in the writing styles of various disciplines, we created two subsets of the evaluation set: one with the sentences that were derived from papers in Humanities disciplines and another with sentences from papers in Health Sciences and Biology (H&B). Detailed statistics of the annotated entities contained in each subset are given in Table 3.

Regarding the entity extraction task, we used the above datasets to train / evaluate two different DL models for each entity: 1) a DL entity recognizer employing a Bert-base-NER transformer model that uses self-attention to process input sequences and generate contextualized representations of words in a sentence and 2) a DL entity recognizer employing a Roberta-base transformer, a variant of BERT, trained on a much larger dataset (10 times larger) and using a dynamic masking technique during training that helps the model learn more robust and generalizable representations of words. Both models came from the Hugging-Face library[8] and were used for vector representation in combination with a transition-based parser for the sequence labeling part. For the latter we used the development set for hyperparameter optimization (dropout=0.1, Adam optimizer -L2=0.01). All of the transformer models and the transition-based parsers were fine-tuned / trained on the same datasets. These are the models **A-BERT-base-NER**, **A-RoBERTa-base**, for the extraction of Activities, **M-BERT-base-NER** and **M-RoBERTa-base** for the extraction of Methods and **G-BERT-base-NER**, **G-RoBERTa-base** for the extraction of Goals.

In addition, for comparison reasons, we used the same dataset for training/evaluation of the spaCy default Named Entity Recognizer[9] consisting (at the time of writing this paper) of a CNN with Bloom Embeddings that utilize a stochastic approximation of traditional embeddings in order to provide unique vectors for a large number of words without explicitly storing a separate vector for each of them (Miranda et al., 2022). These are the models **A-CNN, M-CNN, G-CNN**.

Furthermore, we designed a prompt template that leverages k-shot learning and text-to-structure capabilities of chat-GPT (GPT 3.5), in order to recast the structured output in the form of code instead of natural language. More specifically, we used the development set for experimenting with various combinations in prompt, such as different number of included examples (k=3,5,10,20), inclusion or not of the actual entity spans and inclusion or not of the reasoning for each entity extraction. Responses of the LLM into various prompt types during development stage showed that: i) describing the type of output in combination with specific examples helps the LLM to understand how to perform the output transformation and the classification task; ii) Although the increase in the number of examples helps performance, the added computational (and budget) costs from the larger prompts need to be taken into account when setting the threshold for the number of included examples (in our case k=10 proved to be a fair threshold); iii) using only the reasoning field without any demonstrating examples didn't contribute

[8] https://huggingface.co/models

[9] https://spacy.io/api/entityrecognizer

	Humanities			Health & Biology			Total		
	Token	Partial	Strict	Token	Partial	Strict	Token	Partial	Strict
A-10-shot-GPT	48.53	42.66	12.37	69.64	44.91	15.17	56.17	45.32	13.25
A-CNN	64.99	61.12	47.21	71.46	65.71	50.51	68.15	63.06	48.36
A-Bert-base-NER	86.93	81.58	78.26	86.19	86.78	80.26	86.43	84.07	79.08
A-Roberta-base	**88.10**	**84.36**	**79.67**	**89.26**	**88.00**	**81.06**	**89.01**	**86.62**	**80.06**
G-10-shot-GPT	44.28	44.99	11.26	49.76	47.05	12.34	47.54	45.11	12.27
G-CNN	82.65	70.63	54.87	80.36	67.15	47.72	81.94	69.49	52.38
G-Bert-base-NER	86.99	80.68	71.63	87.12	78.61	66.29	86.98	79.97	69.51
G-Roberta-base	**87.03**	**81.45**	**73.01**	**88.84**	**82.20**	**70.11**	**88.59**	**80.62**	**72.79**
M-10-shot-GPT	40.03	33.96	18.87	43.89	34.03	19.19	43.11	34.31	19.74
M-CNN	74.41	72.86	64.85	76.33	74.49	66.95	75.54	73.75	65.84
M-Bert-base-NER	82.83	79.29	73.18	82.61	79.63	73.70	83.03	79.80	74.01
M-Roberta-base	**83.59**	**80.47**	**75.10**	**83.61**	**80.60**	**74.43**	**83.79**	**80.81**	**74.97**

Table 4: Evaluation results (F1 Scores). Prefixes A, G & M denote Activities, Goals & Methods respectively.

significantly to the overall performance increase (in comparison to adding more examples), as could be the case with other tasks like solving mathematical problems. Also it is to be noted that, similarly to (Fatemi & Hu, 2023), we experienced inconsistent performance across all experiments with variations in the output when the same input was repeated, even from a single account. Based on these observations, our proposed template consists of five sections: 1) description of the task at hand; 2) definitions of the entities, requested for extraction; 3) description of the requested output; 4) inclusion of 10 indicative examples for guidance; 5) input of the text to be annotated in the desired format. Using this template, the input is inserted as json lines (jsonl), each consisting of a dictionary containing the keys: "text" - with the actual text of the sentence and "spans" – a list of dictionaries, each containing the "label" denoting the type of the extracted entity, the entity span and pointers for the token-based and/or character-based entity boundaries, respectively. The LLM is enforced to recast the output in the same format, thus enabling easy integration with other workflows (through the Open AI API) and annotation tools such as Prodigy. The template is displayed in Figure 3. We used the same evaluation set in order to measure the performance of GPT 3.5 in the tasks at hand. These are the models **A-10-shot-GPT** for the extraction of Activities, **G-10-shot-GPT** for the extraction of Goals and **M-10-shot-GPT** for the extraction of Methods respectively.

4.2 Evaluation

The evaluation of Information Extraction methods involves comparing classifier results against a "gold standard" produced by human annotators. To this end, a confusion matrix is calculated based on the true positives (TP) -correctly classified predictions-, false positives (FP) -incorrectly classified predictions- true negatives (TN) -correctly non-classified predictions and false negatives (FN) -incorrectly non-classified predictions. Performance scores are then measured based on Precision (P), Recall (R) and F1 as usual.

For the entity extraction task, we conducted three types of evaluation experiments following the guidelines in (Segura-Bedmar et al., 2013) and using the nerevaluate 0.1.8[10] and the scikit-learn[11] python libraries: 1) *token-based*, where a true positive (TP) is a token correctly classified as part of a chunk representing the entity, etc.; 2) *entity based -partial matching*, where some overlap between the tagged entity and the "golden" entity is required, but counts as half compared to the exact matches and 3) *entity-based -strict matching*, where only exact boundaries of the entities are counted for the match. Detailed results for all the evaluation experiments (reported here as F1 scores per entity type, classification method, evaluation method and dataset) are shown in Table 4.

5. Discussion

As a general remark regarding all the evaluation experiments, overall performance suggests that the dataset can be used adequately for finetuning DL models like transformers.

5.1 Classification Method

Regarding the performance of each methodology, fine-tunned transformer-based models showed superior performance in comparison to the rest of the models.

Specifically, compared to the CNN, higher performance was expected since transformer-based models can capture far more language attributes from the textual context and thus "understand" better the individual characteristics even for syntactically complex entity types.

Performance of the LLM was also inferior, something expected since, as demonstrated in (Gao et al., 2023; Jimenez Gutierrez et al., 2022; X. Li et al., 2023; Ma et al., 2023; Qin et al., 2023; Qiu & Jin, 2024), when it

[10] https://pypi.org/project/nervaluate/

[11] https://scikit-learn.org/stable/

comes to NLP tasks like IE and NER, these models underperform significantly compared to DL models like BERT that are finetuned in task specific annotated datasets. This situation is expected to become worse when it comes to the extraction of entities with more complex lexico-syntactic structures than standard named entities and of variable length, as is the case in our dataset. This is demonstrated in particular by the low performance in the entity-strict evaluations, where probably due to the aforementioned reasons and the lack of massive training data that is available in fine-tunning methods, the LLM failed to capture the exact boundaries of the spans. Nevertheless, LLM's performance in partial- and token-based evaluations suggests their potential use in distance learning techniques, since they can easily yield massive (but noisy) annotations that could further be manually corrected, or filter candidate sentences for annotation, thus easing the total annotation cost in time and effort.

Regarding the fine-tunned transformer-based models, the difference in performance among the RoBERTa and the BERT models can be attributed to the fact that the former is pretrained on much larger datasets and in a more efficient way than the latter. The high performance of transformer-based models, with F1 reaching up to 89.26 in "lenient" token-based evaluation and up to 81.06 in strict entity-based evaluation, is also evidence of the adequacy and quality of the annotations in our dataset for fine-tuning/training.

5.2 Linguistic Characteristics

Regarding the variations in performance with respect to the different discipline-focused evaluation subsets, the biggest differences appear in the extraction of activities (F1=88.00 in H&B compared to F1=84.36 in Humanities subset). Apart from the difference in the number of labeled tokens between the two subsets, which could lead to lower performance, visual inspection of the errors showed that in Humanities disciplines (e.g. in Archeology, History, Paleontology, etc.) there are a lot of mentions of historical events which, being events themselves, have textual descriptions that bear similar lexico-syntactic structures with those of research activities. Such cases, especially in passive voice with missing agent, are more difficult to discern. A similar situation arises in certain cases of research goals extraction, especially when these are goals of those "misclassified activities".

Based on visual inspection of more than 1000 sentences from the evaluation set and their comparison the rest of the dataset, the aforementioned cases could be considered as "extreme scenarios" of the dataset, since in these situations, the semantics for discerning a textual span representing a general activity or a goal (that are irrelevant of the research described in the paper) are not enough for the classifier to be able to make the correct prediction. Nevertheless, these errors could probably be resolved with heuristics that analyze only specific sections of the paper (e.g. excluding related work, background, historical references sections, etc.).

5.3 Processing Granularity

Analyzing the results of each entity type, showed that the highest performance was achieved in Activity extraction. This can be attributed to the differences in the number of labeled tokens for each entity that follows the overall differences in performance. So, the extraction of Methods -having the fewest labeled tokens per sentence on average-, despite being the simplest of all, in terms of lexico-syntactic structure, yielded lower performance compared to Goals which, in turn, fared slightly lower than Activities.

Regarding the extraction of instances of the Goal class, analysis showed that, despite the fewer labeled tokens compared to activities spans, the overall good performance could be attributed to the fact that textual manifestations of goals have a concrete and consistent lexico-syntactic representation that allows for easier generalization of the corresponding DL models. Errors mainly occurred in cases of textual spans representing purpose that was not attributed to the author of the paper and thus should not be classified as a research goal according to SO definitions (e.g.: "The consortium's survey of East Los Angeles was one of the first holistic efforts to document historic and cultural resources in the community.").

Similar performance was also observed in the recognition of Methods. Analysis showed that the errors mainly occurred in cases of named entities other than methods, which, however, appear in similar textual contexts. For example, consider the sentence: "In May 2005 two of us traveled to the Angolan provinces of Namibe and Bengo, where we employed a geographic information system (GIS) to model the potential distribution of new species.". Here the tool: "geographic information system (GIS)" is erroneously annotated as a method by the classifier, probably due to the similar lexical form or the textual context of the sentence.

Regarding the extraction of textual spans referring to the Activity class, errors were observed in some instances of the Activity class in passive voice, not recognized as such by the classifier. For instance, in the sentence: "In this study carbon isotope discrimination was performed to assess the growing conditions of fossil cereal grains", the classifier failed to recognize the activity span. These errors could be attributed to the inclusion of negative training samples (i.e., cases of sentences in passive voice referring to historical events or activities not performed by the authors and thus not being annotated as activities) in the training set.

Regarding the variation in performance across different evaluation experiments (token-based, entity-based-partial, and entity-based-strict evaluations) it can be seen that the exact boundaries of the entity are difficult to capture even for the highest performing

models. Analysis indicates that such errors mostly occur in cases where one type of entity overlaps with another. E.g., "As a consequence of different growth behavior of trees in the juvenile phase, two different methods to estimate the juvenile rings **were used**.". Here, the boundaries of the enclosed entity were incorrectly detected, just like the tokens "were used" (in bold) were erroneously recognized as part of the goal span, although they are part of the overlapping activity. Other cases of erroneous boundary detection involved the inclusion of punctuation marks immediately following the entity inside the textual span. E.g., "To fulfill this purpose, we analyzed cranial discrete traits from this population.". Especially concerning the Method class, such cases also involved the inclusion of information inside parentheses or brackets adjacent to the entities, probably due to the similarity in form with cases where the acronym of a method inside parentheses follows the method name (e.g., "Evaluation of Logistic Regression **(P, R, F1)** yielded good performance results…", see also Fig. 1 for another example).

5.4 Linguistic Linked Data Generation

Apart from fine tuning transformer-based models for information extraction, this dataset can be used directly as a source for linguistic linked data by itself. Specifically, using the methodologies described in (Pertsas & Constantopoulos, 2023) the dataset can be transformed into an RDF Knowledge Graph (KG) adhering to Linked Data Standards. Such a KG can offer structured semantic views of the content of publications, which enhance our capability for comprehensive exploration of research work. This can be demonstrated through semantically complex queries executed over the KG. Indicative such queries, expressed in SPARQL are presented below:

Query 1: Retrieve all researchers that participate in activities or have research objectives that deal with linguistic analysis.

```
SELECT DISTINCT ?p_label
WHERE {
 ?p rdfs:label ?p_label
 ?p so:hasGoal / rdfs:label ?g_label
 ?p so:participatesIn / rdfs:label ?a_label
 filter contains(lcase(?g_label,?a_label),
 "linguistic analysis").}
```

Here, through the use of property chains in SPARQL and the *filter contains* SPARQL expression, all the methods employed in activities that have objectives with labels (i.e. textual spans) that contain the words "linguistic analysis" can be retrieved.

Query 2: For a specific paper (e.g. "Paper1") retrieve all the research activities, conducted by the authors, along with their objectives and the methods they employed.

```
SELECT ?m_label ?a_label ?g_label
WHERE {
 ?a so:isDocumentedIn so:Paper1.
 ?a rdfs:label ?a_label.
 ?g so:isDocumentedIn so:Paper1.
 ?g rdfs:label ?g_label.
 ?m so:isDocumentedIn so:Paper1.
 ?m rdfs:label ?m_label. }
```

Here, the overall activity reported in a paper is decomposed into a series of activities denoting "what" the authors have done, associated with the methods they employed, and the goals they were trying to accomplish. Through this way, basic questions of "what", "how" and "why" regarding information described in a research publication can be answered. Using such queries, the reader has access to an enhanced "bird's-eye" view of what is described in a paper before actually reading it. Additional information regarding the authors, their research interests or the abstract can also be retrieved using the appropriate SO classes and relations.

6. Conclusion

In this paper we presented a manually curated dataset of 15,262 sentences in English, derived from 3,500 research articles (abstract and main text) and 172 different disciplines and subfields. The dataset contains in total 23,562 labels for three types of entities: 1) research methods, named entities of variable length, 2) research goals, entities that appear in text as textual spans of variable length with mostly fixed lexico-syntactic-structure, and 3) research activities, entities that appear as textual spans of variable length with complex lexico-syntactic structure.

We explored the capabilities of our datasets along four dimensions: 1) *Classification Method*: we experimented with training/fine-tuning various ML and DL models as well as LLMs (chat-GPT 3.5) through prompting; 2) *Linguistic Characteristics*: we explored the performance of our methods across interdisciplinary and discipline-specific subsets in order to capture any possible differences in discipline-oriented writing styles; 3) *Processing Granularity*: we tested the effectiveness of classification at three levels of granularity: token-based, entity-based strict and entity-based partial. In addition, the included entities represent three levels of lexico-syntactic complexity: named entities of variable length, "non-named" entities of variable length with mostly fixed lexico-syntactic-structure and variable length with complex lexico-syntactic structure; 4) *Linguistic Linked Data Generation:* we explored the capabilities of our dataset as a potential source for linguistic linked data through the use of SPARQL queries that can be executed over an RDF KG that can be created from it.

Evaluation scores showed high performance in all the experiments, especially with transformer-based models, showcasing the capabilities of our dataset in fine-tuning / training transformer models that can achieve very high results in entity extraction reaching up to F1=89.26 in "lenient" token-based evaluation and up to F1=81.06 in strict entity-based evaluation, even for entities of complex lexico-syntactic structure and variable length like the ones of research activities.

Future work includes expansion of our dataset with annotation of other types of entities and relations of the Scholarly Ontology concerning research publications. Specifically, we intend to provide annotations as well as trained DL models for the relations among SO entities, such as *employs(Activity,Method)*,*hasObjective(Activity,Goal*) for interrelating the extracted activities with their corresponding methods and goals respectively, thus enhancing the produced linguistic linked data.

In addition, we intend to produce annotations for the research findings, arguments that describe various experiment results and interrelate them with their associated research activities that provide the supporting evidence or premise for those findings.

7. Bibliographical References

Alluqmani, A., & Shamir, L. (2018). Writing styles in different scientific disciplines: A data science approach. *Scientometrics*, *115*. https://doi.org/10.1007/s11192-018-2688-8

Ashok, D., & Lipton, Z. C. (2023). *PromptNER: Prompting For Named Entity Recognition* (arXiv:2305.15444). arXiv. http://arxiv.org/abs/2305.15444

Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, *8*(1), 224. https://doi.org/10.1057/s41599-021-00903-w

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., & Henighan, T. (2020). Language Models are Few-Shot Learners. *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 1877–1901.

Constantopoulos, P., Hughes, L. M., Dallas, C., Pertsas, V., & Christodoulou, T. (2016). Contextualized Integration of Digital Humanities Research: Using the NeMO Ontology of Digital Humanities Methods. *Digital Humanities 2016*, 161–163.

Das, S. S. S., Katiyar, A., Passonneau, R., & Zhang, R. (2022). CONTaiNER: Few-Shot Named Entity Recognition via Contrastive Learning. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6338–6353. https://doi.org/10.18653/v1/2022.acl-long.439

Fatemi, S., & Hu, Y. (2023). *A Comparative Analysis of Fine-Tuned LLMs and Few-Shot Learning of LLMs for Financial Sentiment Analysis* (arXiv:2312.08725). arXiv. http://arxiv.org/abs/2312.08725

Gao, J., Zhao, H., Yu, C., & Xu, R. (2023). *Exploring the Feasibility of ChatGPT for Event Extraction* (arXiv:2303.03836). arXiv. http://arxiv.org/abs/2303.03836

Jimenez Gutierrez, B., McNeal, N., Washington, C., Chen, Y., Li, L., Sun, H., & Su, Y. (2022). Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4497–4512. https://doi.org/10.18653/v1/2022.findings-emnlp.329

Leong, A. P. (2024). Marked Themes in academic writing: A comparative look at the sciences and humanities. *Text & Talk*, *0*(0). https://doi.org/10.1515/text-2022-0188

Li, P., Sun, T., Tang, Q., Yan, H., Wu, Y., Huang, X., & Qiu, X. (2023). CodeIE: Large Code Generation Models are Better Few-Shot Information Extractors. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15339–15353. https://doi.org/10.18653/v1/2023.acl-long.855

Li, X., Chan, S., Zhu, X., Pei, Y., Ma, Z., Liu, X., & Shah, S. (2023). *Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks* (arXiv:2305.05862). arXiv. http://arxiv.org/abs/2305.05862

Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022). Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8086–8098. https://doi.org/10.18653/v1/2022.acl-long.556

Ma, Y., Cao, Y., Hong, Y., & Sun, A. (2023). Large Language Model Is Not a Good Few-shot Information Extractor, but a Good Reranker for Hard Samples! *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10572–10601. https://doi.org/10.18653/v1/2023.findings-emnlp.710

Miranda, L. J., Kádár, Á., Boyd, A., Van Landeghem, S., Søgaard, A., & Honnibal, M. (2022). *Multi hash embeddings in spaCy* (arXiv:2212.09255). arXiv. http://arxiv.org/abs/2212.09255

Perez, E., Kiela, D., & Cho, K. (2021, May). *True Few-Shot Learning with Language Models*. https://doi.org/10.48550/arXiv.2105.11447

Pertsas, V., & Constantopoulos, P. (2017). Scholarly Ontology: Modelling scholarly practices. *International Journal on Digital Libraries*, *18*(3), 173–190. https://doi.org/10.1007/s00799-016-0169-3

Pertsas, V., & Constantopoulos, P. (2023). Ontology-Driven Extraction of Contextualized Information from Research Publications: *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 108–118. https://doi.org/10.5220/0012254100003598

QasemiZadeh, B., & Schumann, A.-K. (2016). The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1862–1868.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). *Is ChatGPT a General-Purpose Natural Language Processing Task Solver?* (arXiv:2302.06476). arXiv. http://arxiv.org/abs/2302.06476

Qiu, Y., & Jin, Y. (2024). ChatGPT and finetuned BERT: A comparative study for developing intelligent design support systems. *Intelligent Systems with Applications*, *21*, 200308. https://doi.org/10.1016/j.iswa.2023.200308

Renear, A. H., & Palmer, C. L. (2009). Strategic Reading, Ontologies, and the Future of Scientific Publishing. *Science*, *325*(5942), 828–832. https://doi.org/10.1126/science.1157784

Segura-Bedmar, I., Martinez, P., & Zazo, M. H. (2013). SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, *2*, 341–350.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.

Wang, X., Li, S., & Ji, H. (2023). Code4Struct: Code Generation for Few-Shot Event Structure Prediction. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3640–3663. https://doi.org/10.18653/v1/2023.acl-long.202

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Findings of the Association for Computational Linguistics: ACL 2023*, 6519–6534. https://doi.org/10.18653/v1/2023.findings-acl.408

Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., Jiang, Y., & Han, W. (2023). *Zero-Shot Information Extraction via Chatting with ChatGPT* (arXiv:2302.10205). arXiv. http://arxiv.org/abs/2302.10205

8. Language Resource References

Augenstein, I., Das, M., Riedel, S., Vikraman, L., & McCallum, A. (2017). SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 546–555. https://doi.org/10.18653/v1/S17-2091

Borchert, F., Lohr, C., Modersohn, L., Witt, J., Langer, T., Follmann, M., Gietzelt, M., Arnrich, B., Hahn, U., & Schapranow, M.-P. (2022). GGPONC 2.0—The German Clinical Guideline Corpus for Oncology: Curation Workflow, Annotation Policy, Baseline NER Taggers. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, 3650–3660.

Cheng, F., Yada, S., Tanaka, R., Aramaki, E., & Kurohashi, S. (2022). JaMIE: A Pipeline Japanese Medical Information Extraction System with Novel Relation Annotation. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, 3724–3731.

Franzén, K., Eriksson, G., Olsson, F., Asker, L., Lidén, P., & Cöster, J. (2002). Protein names and how to find them. *International Journal of Medical Informatics*, *67*(1), 49–61. https://doi.org/10.1016/S1386-5056(02)00052-7

Jain, S., Van Zuylen, M., Hajishirzi, H., & Beltagy, I. (2020). SciREX: A Challenge Dataset for Document-Level Information Extraction. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7506–7516. https://doi.org/10.18653/v1/2020.acl-main.670

Kim, J.-D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). GENIA corpus—A semantically annotated corpus for bio-textmining. *Bioinformatics*, *19*(suppl_1), i180–i182. https://doi.org/10.1093/bioinformatics/btg1023

Luan, Y., He, L., Ostendorf, M., & Hajishirzi, H. (2018). Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3219–3232. https://doi.org/10.18653/v1/D18-1360

Mullick, A., Pal, S., Nayak, T., Lee, S.-C., Bhattacharjee, S., & Goyal, P. (2022). Using Sentence-level Classification Helps Entity Extraction from Material Science Literature. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, 4540–4545.

Osenova, P., Simov, K., Marinova, I., & Berbatova, M. (2022). The Bulgarian Event Corpus: Overview and Initial NER Experiments. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 3491--3499. https://aclanthology.org/2022.lrec-1.374

Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., & Setzer, A. (2009). Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*, *42*(5), 950–966. https://doi.org/10.1016/j.jbi.2008.12.013

Rosario, B., & Hearst, M. A. (2004). Classifying semantic relations in bioscience texts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, 430-es. https://doi.org/10.3115/1218955.1219010