NLPSharedTasks: A Corpus of Shared Task Overview Papers in Natural Language Processing Domains

Anna Martin[1], **Ted Pedersen**[2], **and Jennifer D'Souza**[3]
[1]University of Minnesota, Minneapolis, MN 55455*
mart5877@umn.edu
[2]University of Minnesota, Duluth, MN 55812
tpederse@d.umn.edu
[3]TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
jennifer.dsouza@dtib.eu

Abstract

As the rate of scientific output continues to grow, it is increasingly important to be able to develop systems to improve interfaces between researchers and scholarly papers. Training models to extract scientific information from the full texts of scholarly documents is important for improving how we structure and access scientific information. However, there are few annotated corpora that provide full paper texts. This paper presents the NLPSharedTasks corpus, a new resource of 254 full text Shared Task Overview papers in NLP domains with annotated task descriptions. We calculated strict and relaxed inter-annotator agreement scores, achieving Cohen's kappa coefficients of 0.44 and 0.95, respectively. Lastly, we performed a sentence classification task over the dataset, in order to generate a neural baseline for future research and to provide an example of how to preprocess unbalanced datasets of full scientific texts. We achieved an F1 score of 0.75 using SciBERT, fine-tuned and tested on a rebalanced version of the dataset.

1 Introduction

Scholarly Document Processing (SDP) research is concerned with developing methods for improving the retrieval and organization of information from academic papers. This interest is partly driven by the rapid growth rate of scientific publications, which Larsen and von Ins (2010) estimate to be between 2.7 and 13.5 percent between 1997 and 2006. Some disciplines are expanding even more rapidly. Dhawan et al. (2020) examined the global output of machine learning research between 2009 and 2018 and estimated a growth rate of roughly 28 percent per year, while Li et al. (2020) suggest an average annual growth rate of 152.9 percent in the deep learning domain between 2013 and 2019.

Because of the rapid expansion of scientific literature, it is beneficial to use natural language processing (NLP) and information extraction (IE) techniques to structure scientific and bibliometric data into machine-actionable forms. One method is to automatically identify scientific and bibliometric entites and relations from scholarly literature and organize them into knowledge graphs, which can be used to improve access to scholarly documents by enhancing Digital Libraries (Ammar et al., 2018, Auer et al., 2020).

One scientific entity type relevant to NLP and Machine Learning domains is TASK. Machine Learning and NLP tasks can be useful to extract, as they are a unit of information relevant to understanding research trends and constructing leaderboards (Hou et al., 2019). We are particularly interested in the utility of augmenting scholarly digital library resources with automatically extracted task descriptions such that a reader could quickly understand the NLP task described in the paper at hand. We find that NLP shared task workshop overview papers are a rich resource for training a model to extract such task descriptions.

Our contribution towards information extraction (IE) from scientific articles is a new gold-standard corpus of task description annotations from Shared Task Overview papers. This corpus provides an interesting IE situation for two reasons. First, the full texts are provided for each paper in the corpus rather than individual sentences or paragraphs. Second, the annotation goal was to extract a single span of text from each paper rather than any number of qualifying phrases. The benefit of this kind of annotation strategy is that it provides test data that is close to the "real world" data that downstream applications might encounter, such as a digital library tool tasked with extracting the task descriptions from NLP papers. This IE scenario is also difficult, since extracting a single span from full paper texts results in an extremely

*The work presented in this paper was performed while the first author was affiliated with the University of Minnesota, Duluth.

Proceedings of the 1st Workshop on Information Extraction from Scientific Publications, pages 105–120

unbalanced dataset. For this reason, we describe in detail the data preparation and preprocessing steps we performed for the sentence classification task we ran over the NLPSharedTasks corpus. The original NLPSharedTasks corpus, preprocessed dataset, and experimental code is available at https://github.com/anmartin94/martin-masters-thesis-2022.

2 Related Work

Numerous corpora for scientific information extraction have been hand-annotated by experts in computational linguistics and NLP domains. Many of these corpora provide parts of scientific papers, such as paragraphs (Augenstein et al., 2017), abstracts (Gábor et al., 2018, Gábor et al., 2016, QasemiZadeh and Schumann, 2016, Luan et al., 2018), and sentences (Hou et al., 2021).

The SemEval-2021 Task 11 (NLP Contribution Graph) (D'Souza et al., 2021) provided the corpus that serves as the main source of inspiration for our annotation project. The NLPContributionGraph corpus comprises 442 scholarly papers in NLP domains, with 12 different types of information annotated at three levels of granularity (D'Souza and Auer, 2020). It is similar to our work in that full paper texts and sentence-level annotations are provided, but the annotation scheme allowed for multiple spans to be extracted for each entity type, rather than a single sequence from each paper. Additionally, the NLPContributionGraph annotation scheme includes a TASKS information unit, which was applied to 277 triples found across approximately 69 sentences in eight papers.

The differences between the NLPContributionGraph and NLPSharedTask annotation schemes relate to the different intended downstream tasks. The information extracted by D'Souza et al., 2021 is designed to populate a research knowledge graph with a variety of types of scientific information, while the information extracted in NLPSharedTasks is intended to convey to a human reader the task described in the Shared Task Overview paper.

3 Corpus Selection

The resource we drew from was the annual research workshop SemEval and similar initiatives. These venues host shared tasks that approach a wide variety of semantic problems and provide a rich resource for understanding the state of the art in semantic analysis. We assembled our task description corpus by searching the ACL Anthology for shared task description papers, including all SemEval task description papers from the year 2001 to 2021, all CoNLL[1] shared tasks 2000-2020, and selected shared tasks from a variety of other conferences and workshops (see Table 1). The dataset was developed in two stages. The first stage selected only Shared Task Overview papers associated with the SemEval workshop from 2001 to 2020, yielding 165 papers. During the second stage we added 89 papers to the dataset. These papers were found by searching the ACL Anthology for Shared Task Overview papers published at non-SemEval workshops hosted by the venues described in Table 1. Additionally, this second batch of papers contained the newly published set of papers from SemEval 2021. The final dataset contains a total of 254 shared task description papers between the years 2000 and 2021 and encompasses twenty natural language processing research topics that we identified (see Figure 1).

Venue	Frequency
SemEval Workshop	176 (69%)
CoNLL Conference	21 (8%)
ACL Conference	18 (7%)
EMNLP Conference	12 (5%)
NAACL Conference	8 (3%)
EACL Conference	7 (3%)
IJCNLP Conference (2017)	5 (2%)
BioNLP Workshop (2011)	3 (1%)
AACL Conference	2 (<1%)
∗SEM Workshop	2 (<1%)
Total	254

Table 1: The conferences and workshops that hosted the Shared Tasks in our corpus and the number of papers from each venue.

4 Annotation Methodology

The aim of this annotation project was to develop a gold standard corpus of shared task overview papers with annotations of shared task descriptions. We define "shared task description" as a span of text containing information on an NLP or computational linguistics task to be performed by participating systems. This information must describe in brief what is to be done to accomplish the task, and may also contain details on the dataset the task is performed over.

[1] https://www.conll.org/

Set #	Strict Score	Relaxed Score
Set 1	.3830	.6401
Set 2	.4374	.9488

Table 2: This table presents the inter-annotator agreement scores measured with Cohen's kappa coefficient. Strict scores were calculated by comparing the exact spans of text. The relaxed scores were calculated by including the full sentence(s) containing the span. The difference between rows 2 and 3 is due to guideline revisions. The annotators often chose sequences that overlapped but were not exactly the same, resulting in the difference between columns 2 and 3.

The first annotator extracted a task description sequence[2] from every paper in the corpus, generated a set of guidelines for the second annotator to follow, and created two representative sets of twenty papers each. Intra-annotator agreement was determined using Cohen's kappa coefficient. A strict score and a relaxed score were calculated for each dataset, where the strict score compared the exact sequence spans and the relaxed score counted overlapping annotations as matches. After the first subset was annotated by the second annotator, the guidelines were refined by the first annotator to address ambiguities before releasing the second set to annotator 2.

4.1 Guidelines

The final version of annotation guidelines performs two functions: it defines Task Description and describes various subtypes and task-description scenarios including "full task description", "partial task description", and "multiple subtasks description"; and it provides two sets of rules, one explaining how the task description sequence boundaries should be determined, and another detailing how ambiguous annotation situations might be resolved. See Appendix A.6 for more information on the annotation process.

5 Annotation Results

We calculated the inter-annotator agreement between annotator 1 and annotator 2 using Cohen's kappa coefficient (Cohen, 1960). The strict Cohen's kappa coefficient for the first subset was 0.383, and the relaxed Cohen's kappa coefficient was 0.6401, indicating fair to substantial agreement (Viera et al., 2005). After we made revisions and clarifications to the annotation guidelines, we annotated the second subset, and achieved a strict score of 0.4373 and a relaxed score of 0.9488, indicating moderate to almost perfect agreement. The difference between the strict and relaxed scores indicates that, though the annotators often spans from the same sentence context, mutually choosing equivalent sequences is somewhat difficult. For example, from the following sentence

> The 2020 iteration of our task is similar to CoNLL-SIGMORPHON 2017 (Cotterell et al., 2017) and 2018 (Cotterell et al., 2018) in that participants are required to design a model that learns to generate inflected forms from a lemma and a set of morphosyntactic features that derive the desired target form.
> -*SIGMORPHON 2020 Shared Task 0: Typologically Diverse Morphological Inflection*, Vylomova et al. (2020),

annotator 1 extracted "design a model that learns to generate inflected forms from a lemma and a set of morphosyntactic features that derive the desired target form", and annotator 2 extracted "*participants are required to* design a model that learns to generate inflected forms from a lemma and a set of morphosyntactic features that derive the desired target form".

6 Corpus Statistics

One benefit of annotating shared task overview papers published over a long period of time is that this resource could potentially be used to study NLP research progress and trends. For this reason, we provide some basic statistics on the content of the papers included (see Section 6.1). We also provide data on the extracted task descriptions in Section 6.2, as such information may be useful to others for building task description extraction systems.

6.1 Characteristics of Shared Task Overview Papers

The 254 shared task overview papers collected for this dataset encompass a wide variety of research topics. We identified 20 distinct topics (see Figure 1), and found that the frequency of publications

[2] on occasion, the first annotator extracted two sequences if the texts were extremely similar. Following is such an example: "Given a set of documents and a set of target entities, the task consisted of building a timeline for each entity, by detecting, anchoring in time and ordering the events involving that entity" and "Given a set of documents and a set of target entities, the task consists of building a timeline related to each entity, i.e., detecting, anchoring in time, and ordering the events in which the target entity is involved".

	mean$^{+}_{-}$ std	max	min
word count	$29^{+}_{-}21.8$	126	3
sentences in span	$1.17^{+}_{-}.48$	4	1

Table 3: Mean word count and sentences per task description

included in our corpus increases between the year 2000 and 2021 (see Figure 2). Another interesting characteristic of these shared tasks is that not all tasks are novel; it is fairly common for tasks to be re-run for several years. This allows participants to improve benchmarks by building on previous work, and allows task organizers to add to the complexity of the task. Approximately 65 papers in the corpus describe rerun tasks.

6.2 Task Description Characteristics

One of the most consistent patterns observed is that task descriptions tend to appear under the same limited set of section headers (Figure 3). While they are most commonly found in the abstract, they also frequently appear in introduction sections. Unsurprisingly, sections with titles such as "Task Description" or "Task Overview" often contain task descriptions suitable for our project. Rarely, papers may not contain a good task description until the conclusion or discussion section. Furthermore, there were thirteen papers that did not contain a task description in the body, but had a title that was sufficient. Consequently, the first quadrant of full paper texts contain a higher concentration of task descriptions, as seen in Figure 4. This pattern persists within sections as well; more than half of the task descriptions were found in the first half of the containing section (see Figure 5).

A complicating aspect of this corpus in terms of information extraction and text classification is the varying lengths of task descriptions. This low-homogeneity can make it more difficult to train traditional classifiers, but is important because it provides a more "real-world" environment. The extracted sequences span between 1 and 4 sentences, and contain between 3 and 126 tokens (Table 3).

7 Dataset Preparation

Scholarly papers are often stored as PDFs, which are not very machine-actionable[3]. For this reason the full text for each paper had to be extracted and stored in a different format. We processed paper PDFs into XML encoded files using GROBID (Lopez, 2009), then extracted the text data into plain text files. GROBID is not always completely accurate, so we manually compared each text file with the original PDF. Two papers had to be manually typed because the PDF files could not be processed by GROBID. For the majority of papers with tables, the table output from GROBID had to be manually removed.

To prepare our corpus for a sentence classification task, we randomly divided the 254 papers into a training set of 228 papers and a test set of 26 papers. The resulting training set contains 259 positive samples and 41,493 negative samples, and test set contains 34 positive and 4725 negative samples. This is an extremely unbalanced dataset, where less than 1% (0.63%) of the total sentences are positive samples. The reason for this is the annotation goal was to extract a single candidate per paper. However, extra steps must be taken to change the balance enough so that machine classifiers are able to learn how to identify task descriptions.

7.1 Leveraging Paper Context and Hierarchical Structure

Scholarly papers tend to have a predictable structure. Task description overview papers usually start with an abstract and introduction, which tend to be followed by task description and dataset preparation sections, before describing the system solutions and reporting results. There are patterns within sections as well; for example, sections that contain a task description often contain the sequence near the beginning of the section. For this reason, we added positional data as features to the dataset following the example of (Liu et al., 2021).

We added a section header feature to the dataset by iterating through the plain text files and capturing the header for each section. Each sentence's position was quantified with four values: the sentence index relative to its section; the sentence index relative to the full paper; the quadrant of the sentence's section; and the quadrant of the paper that the sentence is found in.

We ran experiments with and without the header feature and positional features and ultimately found that the additional features did not improve model performance (see Table 6 to compare results).

In addition to extracting positional information

[3]Some journals and archives such as arXiv (https://arxiv.org/) provide LaTeX source code for papers in addition to PDFs.

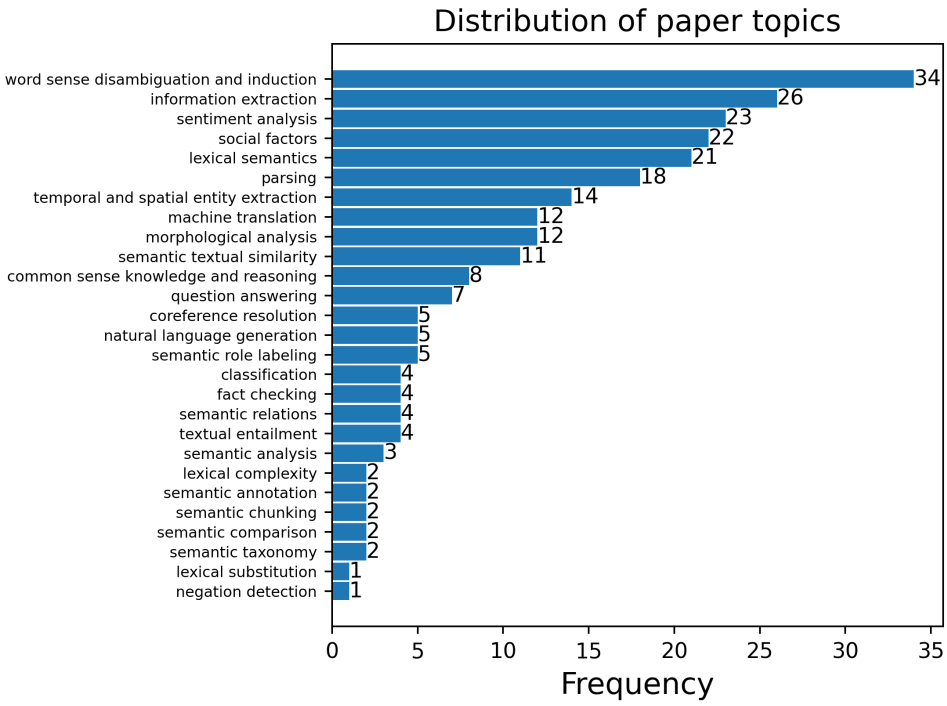

Figure 1: The distribution of paper topics. There were situations where a Shared Task encompassed more than one topic. In this situation, we chose the more specific topic. For example, note that the topic **classification** appears to only contain five papers. There are more classification tasks found in the corpus, but they were assigned other desciptors such as **sentiment analysis** and **social factors**.

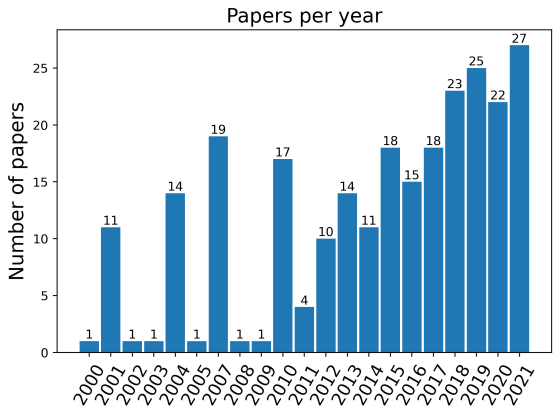

Figure 2: The distribution of publication dates. Note that the years 2000, 2002, 2003, 2005, 2008, and 2009 appear to be outliers. This is because most of the corpus (69.3%) was taken from the SemEval workshops, which were not held in those years.

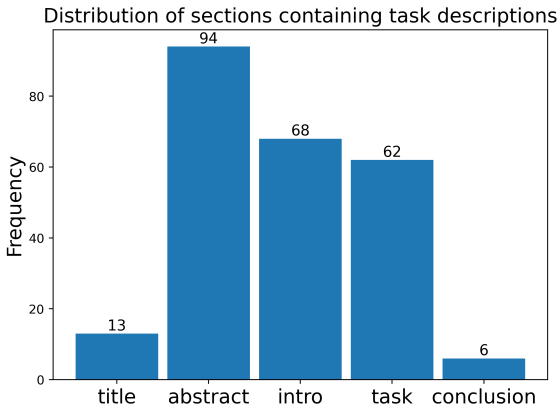

Figure 3: Distribution of sections containing task descriptions

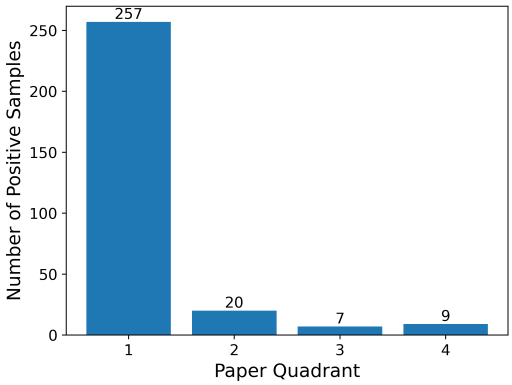

Figure 4: Distribution of task descriptions across paper quadrants

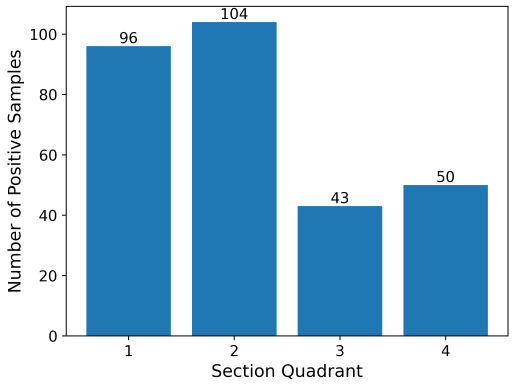

Figure 5: Distribution of task descriptions across section quadrants

for each sentence as features, we also removed any section for each publication that did not provide a task description. A paper with the task description in the introduction, for example, would only have its introduction included in our dataset. This improved the balance between positive and negative samples by increased the proportion of task descriptions to non-task descriptions. It also addressed the following problem: because the goal was to extract a single sequence from each paper, some papers have negative samples that would actually qualify as task descriptions if a better candidate had not been found. Reducing each paper to a single section eliminated some of those perplexing sentences. The resulting training set contains 259 positive samples and 2,304 negative samples, and the resulting test set contains 34 positive samples and 293 negative samples. After reducing the dataset, 11.28% of the total data is positive, which is more manageable than the previous 0.64%.

One problem with manually removing samples from the dataset based on which sections contain task descriptions is that the reduced test set is less "real world". In a non-experimental setting, the machine reader should be able to extract a task description from a whole paper, since it does not know ahead of time which section contains the task description. To address this issue, we tested our model on three versions of the test set. The first was manually reduced the same way as the training set. The second had sections automatically removed by fine-tuning a BERT model on section headers seen in the training set. This model was then applied to the test set to classify section headers as either likely or unlikely to contain a task description. This is a more fair test set because one could apply this classifier to any unseen papers to filter out paper sections. The third set is the full test set without any data removed.

8 Sentence Classification Experiments

Despite the fact that task descriptions are defined as sequences that can be longer or shorter than a single sentence, we designed a sentence classification task because we achieved much higher inter-annotator agreement scores when we compared the chosen sentences spanned by the sequences rather than the exact sequences (see Section 5). We fine-tuned the cased and uncased base versions of BERT (Devlin et al., 2019) and SciBERT (Beltagy et al., 2019) on every hyperparameter combination in Table 4.

hyperparameter	settings
epochs	2, 3, 4
batch size	16, 32
learning rate	2e-5, 3e-5, 5e-5

Table 4: The hyperparameter options are based on the fine-tuning recommendations made by Devlin et al. (2019).

8.1 Training Loop

Each hyperparameter and BERT model combination was fine-tuned on two versions of the training dataset, ten times each. The first version of the dataset contained the contextual features described in Section 7.1, and the second version contained only the text data. In between runs, the data was shuffled and a new validation set containing 10% of the training data was selected. The precision, recall, and F1 score was recorded for each training run. Then the mean scores and standard deviation were calculated for each classifier-encoding pair.

8.2 Baseline

We calculated a baseline based on common vocabulary and positional patterns. We analyzed common word patterns in the training set by tokenizing each sample, removing English stop words, and looking at the 10, 15, 20, 25, and 30 most frequent words in the positive and negative samples from the training set. The most common words for the positive and negative samples are identical, but the density of common words per sentence differs. The density of common words is greater in task description sentences: see Table 5 for the mean common word density per sentence for task descriptions and non-task descriptions. In calculating the baseline, we used a threshold density value of > 0.03 as one of the criteria for classifying a sentence as a task description, with the common word list containing 20 words.

We also experimented with the use of positional information seen in Figures 4 and 5 in calculating our baseline. We found that restricting positive classifications to the first halves of each paper section yielded the highest baseline scores. However, setting a threshold for the total paper quadrants lowered the scores.

The highest baseline scores were calculated by classifying sentences as task descriptions when the density of common words was greater than 0.03 and the sentence was found in the first half of its section. The F1, precision, and recall baseline scores are .4000, .2687, and .7826, respectively.

	Common word density	
N	Task	Non-task
10	0.0518	0.0281
15	0.0647	0.0314
20	0.0762	0.0349
25	0.0848	0.0435

Table 5: The mean density of N most common words among task description sentences and non-task description sentences. Density is calculated by dividing the number of common words in the sentence by the total number of words in the sentence.

8.3 BERT Training Results

The precision, recall, and F1 scores for the best model and hyperparameter combination are shown in Table 6. Scores are reported for both the dataset with additional contextual features and the dataset containing sentences alone.

The highest performing model scored better on the dataset comprising sentence data only without additional features. The cased scibert model earned an average F1 score of 0.72 on the simple dataset and an average F1 score of 0.69 on the dataset containing contextual features. However, the other three models all returned higher mean scores when trained on the dataset containing contextual features. The mean F1 score across all four models trained on the contextual dataset is 0.7, while the mean score across all four models trained on the simple dataset is 0.68. Notice also that the standard deviations are somewhat high, indicating a not insignificant spread around the mean. From this data it is unclear whether one variant of the dataset is better than the other.

8.4 Test Results

Tests were run using the cased SciBERT model fine-tuned on the simple dataset over four epochs with a batch size of 32 and a learning rate of 5e-05 (the model with the highest training results). Three versions of the test dataset were used in order to determine how well our system would perform given data of varying levels of preprocessing. The three versions of the test data are:

1. The dataset manually reduced in the same way that the training data is reduced. Only sections that contain a task description are included;

model	epochs	batch size	learning rate	metric	score
Training results using data annotated with positional features					
bert-cased	4	16	2e-05	Precision Recall F1	0.69 ± 0.1 0.73 ± 0.1 0.71 ± 0.09
scibert_uncased	3	16	3e-05	Precision Recall F1	0.69 ± 0.03 0.73 ± 0.12 0.71 ± 0.06
Training results using text data only					
bert-uncased	3	32	5e-05	Precision Recall F1	0.63 ± 0.08 0.7 ± 0.1 0.66 ± 0.08
scibert_cased	4	32	5e-05	Precision Recall F1	0.73 ± 0.11 0.71 ± 0.07 $\mathbf{0.72 \pm 0.08}$
Baseline					
baseline	-	-	-	Precision Recall F1	0.27 0.78 0.40

Table 6: Mean training results and standard deviations for BERT and SciBERT classifiers across ten runs. Only the results for the best hyperparameter and model combinations are reported here.

		Predicted Labels		
True Labels	Manually reduced test set		+	-	Sum
		+	24 (7.34%)	10 (3.06%)	34 (10.40%)
		-	6 (1.83%)	287 (87.77%)	293 (89.60%)
		Sum	30 (9.17%)	297 (90.83%)	Total=327
	Automatically reduced test set		+	-	Sum
		+	21 (1.76%)	8 (0.67%)	29 (2.43%)
		-	63 (5.27%)	1104 (92.31%)	1167 (97.58%)
		Sum	84 (7.03%)	1112 (92.98%)	Total=1196
	Full test set		+	-	Sum
		+	25 (0.53%)	9 (0.19%)	34 (0.72%)
		-	128 (2.69%)	4597 (96.60%)	4725 (99.29%)
		Sum	153 (3.22%)	4606 (96.79%)	Total=4759

Table 7: The confusion matrices for the test results on the manually reduced, automatically reduced, and full (non-reduced) test sets. The sums of the positive and negative labels are displayed for the predicted labels and the true labels, as well as the total number of samples in the respective test set. Occasionally the percentages don't sum to 100%; this occurs due to rounding.

test dataset	precision | recall | F1
manually reduced	0.80 | 0.71 | 0.75
automatically reduced	0.25 | 0.72 | 0.37
full test set	0.16 | 0.74 | 0.27

Table 8: Test results for each version of the test dataset

2. The dataset automatically reduced by learning which section headers are likely to appear over a section containing a task description. Only sections that have a high probability of containing a task description are included;

3. The full dataset without any sections removed from any papers.

Figure 7 shows the resulting confusion matrices for each version of the test dataset. The scores reflect the variation in proportion of positive to negative samples; the most balanced dataset is associated with the highest F1 score (0.75) and the least balanced is associated with the lowest (0.27).

Surprisingly, the F1 score for the manually reduced dataset (0.75) is higher than the mean training result (0.72). This is surprising because the hyperparameter settings were chosen based only on the training data; the test data was unseen during the process of hyperparameter selection. However, 0.75 is within one standard deviation of the mean training result (standard deviation = ± 0.08). The dataset used to train the model used to classify the test set was bigger than the dataset used during training experiments because 10% of it did not need to be set aside for validation. It is possible that, due to the relatively small amount of positive samples, that increasing the training data by a small amount could be enough to improve results on during testing.

8.5 Error Analysis

Many of the errors made by our system reflect the situations that were difficult or ambiguous for the human annotators. Papers with subtasks, joint tasks, and multiple tracks were particularly hard. There were two papers with subtasks in the test set for which the system failed to classify any sentences as task descriptions; one paper that describes multiple tracks for which the system wrongly chose multiple sentences (one for each track); and a paper describing four joint tasks for which the system found all but one of the four task descriptions[4].

There were six instances where, when faced with more than one good task description candidate, the system either chose both or chose the wrong one. One interesting pattern is that the false positives are often adjacent to true positives extracted by the system. While these false positives may be lacking in detail on their own, some of them work quite well as auxiliary sentences to the true positives.

Our system struggled in two cases to recognize short task description phrases embedded in broader, more generic statements. This indicates that taking a span-based approach to Task Description extraction could be more effective than sentence classification. See Appendix B for more examples.

9 Conclusion

Our primary contribution is the creation of a new Scholarly Document Processing corpus that provides full paper texts rather than short, curated contexts, and a method for reducing and rebalancing the dataset for an information extraction task. Corpora such as NLPSharedTasks can be used in scholarly information extraction systems to automatically identify and display fine grained scientific information to users of digital libraries. Our most significant finding is the importance of the data preparation and preprocessing decisions. These choices about how to build and filter the datasets had a much greater impact on the results than the hyperparameter settings.

A future annotation project could be conducted that is generally based on our rules but is more lenient in terms of the sentences to be extracted. Instead of focusing on conciseness, this project would prioritize obtaining as much information as is required to produce a more thorough account of the shared task. This resource might subsequently be utilized as the basis for an extractive task summary effort. A span-based information extraction task could be designed over our corpus to extract the original annotated sequences rather than full sentences. Sentence classification could be used as a preprocessing step to narrow down the search space.

[4]The guidelines instructed the annotators to only extract subtask descriptions if they appeared in consecutive sentences, did not allow annotators to extract track descriptions, and permitted annotators to choose multiple task descriptions for joint task papers even if the spans were discontiguous.

References

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans - Louisiana. Association for Computational Linguistics.

S. Auer, A. Oelen, Muhammad Haris, M. Stocker, Jennifer D'Souza, K. Farfar, Lars Vogt, Manuel Prinz, Vitalis Wiens, and M. Y. Jaradeh. 2020. Improving access to scientific literature with knowledge graphs. *Bibliothek Forschung und Praxis*, 44:516 – 529.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

S. M. Dhawan, B. M. Gupta, and N. K. Singh. 2020. Global machine-learning research: a scientometric assessment of global literature during 2009-18. *World Digit. Libr.*, 13:105–120.

Jennifer D'Souza, Sören Auer, and Ted Pedersen. 2021. SemEval-2021 task 11: NLPContributionGraph - structuring scholarly NLP contributions for a research knowledge graph. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 364–376, Online. Association for Computational Linguistics.

Jennifer D'Souza and S. Auer. 2020. Nlpcontributions: An annotation scheme for machine reading of scholarly contributions in natural language processing literature. *ArXiv*, abs/2006.12870.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.

Kata Gábor, Haïfa Zargayouna, Davide Buscaldi, Isabelle Tellier, and Thierry Charnois. 2016. Semantic annotation of the ACL Anthology corpus for the automatic analysis of scientific literature. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3694–3701, Portorož, Slovenia. European Language Resources Association (ELRA).

Jacek Haneczok, Guillaume Jacquet, Jakub Piskorski, and Nicolas Stefanovitch. 2021. Fine-grained event classification in news-like text snippets - shared task 2, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 179–192, Online. Association for Computational Linguistics.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2019. Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5203–5213, Florence, Italy. Association for Computational Linguistics.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2021. TDMSci: A specialized corpus for scientific literature entity tagging of tasks datasets and metrics. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 707–714, Online. Association for Computational Linguistics.

Peder Olesen Larsen and Markus von Ins. 2010. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, 84:575 – 603.

Yang Li, Zeshui Xu, Xinxin Wang, and Xizhao Wang. 2020. A bibliometric analysis on deep learning dur-

ing 2007–2019. *International Journal of Machine Learning and Cybernetics*, pages 1–20.

Haoyang Liu, M. Janina Sarol, and Halil Kilicoglu. 2021. UIUC_BioNLP at SemEval-2021 task 11: A cascade of neural models for structuring scholarly NLP contributions. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 377–386, Online. Association for Computational Linguistics.

Patrice Lopez. 2009. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries*, pages 473–474, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Behrang QasemiZadeh and Anne-Kathrin Schumann. 2016. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1862–1868, Portorož, Slovenia. European Language Resources Association (ELRA).

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. SemEval-2012 task 1: English lexical simplification. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 347–355, Montréal, Canada. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

A Data Statement

Provided in this is the Data Statement for our corpus NLPSharedTasks, version 1, following Bender and Friedman (2018).

A.1 Curation Rationale

Our corpus contains the full texts of 254 Shared Task Overview papers published in the ACL Anthology between the year 2000 and 2021. The criteria for inclusion are:

- The paper was written by the organizers of a Shared Task
- The paper provides a description of the Shared Task, including details on the dataset the task is performed over, the task to be implemented by participating systems, and an overview of participating systems
- The Shared Task described in the paper was hosted by some research workshop in the domain of computational linguistics or natural language processing (NLP)

These criteria ensure that the papers included in the corpus are likely to contain a Shared Task Description. The ACL Anthology was chosen as the source because it provides a catalog that is easy to browse for qualifying candidates for inclusion. Furthermore, choosing a single anthology to draw from provided some consistency of paper style and organization. The starting year (2000) was chosen because the formatting of papers describing earlier initiatives was too dissimilar.

A.2 Language Variety

The papers included in NLPSharedTasks are in English as used in scientific communication in linguistics, computer science, and natural language processing domains.

A.3 Speaker Demographic

The demographics of the paper authors are unknown. The speakers are likely researchers and students of computational linguistics and natural language processing.

A.4 Annotator Demographic

The annotation was performed by two English-speaking annotators well versed in a broad range of NLP topics. Annotator 1 is a graduate student in computer science with a B.S. in computer science, and annotator 2 is a post doctoral researcher in data science with a PhD in computer science. Both annotators had shared task experience, annotator 1 as a participant and annotator 2 as an organizer of SemEval 2021: NLPContributionsGraph (D'Souza et al., 2021). Neither annotator was compensated.

A.5 Speech Situation

The papers included in NLPSharedTasks were written between 2000 and 2021 in research settings. The speech included in these papers is written and is assumed to be scripted and edited, as well as peer-reviewed. In the case of multiple authors, it is unknown whether interaction was either synchronous or asynchronous. The intended audience of the papers included in NLPSharedTasks is researchers and practitioners of computational linguistics and natural language processing.

A.6 Text Characteristics

The genre of the texts included in NLPSharedTasks can be described as written scientific communication in computational linguistics domains and other fields. As such, scientific vocabulary is used throughout that is specific to these domains and the documents are structured in a formal way. Texts are structured with sections under headers including *Title*, *Abstract*, *Introduction*, *Related Work*, *Task Description*, *Results*, and *Conclusion*, among others.

We define a task description as a span of text containing information on the task that must be performed by participating systems. The annotation goal was to extract sequences of text that efficiently describe the Shared Task such that a human reader can understand the task outside of the context of the full paper. Encountering a variety of ways of describing tasks, we developed three sub-definitions: *full task description*, *partial task description*, and *multiple subtasks description*, where a *full task description* contains information on the input data and a brief description of what the participating system must accomplish with the input data, a *partial task description* only describes the task to be performed by participating systems without mention of the data to be used, and a *multiple subtasks description* is a sequence of text that covers multiple subtasks in a single continuous sequence (such a task description is permitted even if the content spans multiple sentences). See Table 9.

Type	Example	Frequency
Full	"<TASK>Given a short context, a target word in English, and several substitutes for the target word that are deemed adequate for that context, the goal of the English Simplification task at SemEval-2012 is to rank these substitutes according to how "simple" they are, allowing ties</TASK>." From *SemEval-2012 Task 1: English Lexical Simplification*, (Specia et al., 2012).	127
Partial	"We describe the CoNLL-2000 shared task: <TASK> dividing text into syntactically related non-overlapping groups of words, so-called text chunking</TASK>." From *Introduction to the CoNLL-2000 Shared Task Chunking*, (Tjong Kim Sang and Buchholz, 2000).	104
Subtask	"The task is <TASK>divided into three subtasks: (a) classification of text snippets reporting sociopolitical events (25 classes) for which vast amount of training data exists, although exhibiting slightly different structure and style vis-a-vis test data, (b) enhancement to a generalized zero-shot learning problem (Chao et al., 2016), where 3 additional event types were introduced in advance, but without any training data ('unseen' classes), and (c) further extension, which introduced 2 additional event types</TASK>, announced shortly prior to the evaluation phase." From *Fine-grained Event Classification in News-like Text Snippets - Shared Task 2, CASE 2021*, (Haneczok et al., 2021).	13
NULL	N/A	12

Table 9: Number of full, partial, subtask, and null task descriptions in 254 shared task overview papers with examples. The full task description contains a description of the input ("Given a short context, target word in English, and several substitutes for the target word"), and a description of what participating systems must do ("rank these substitutes according to how "simple" they are, allowing ties"). In contrast, the partial task description only contains a description of what participating systems must do ("dividing text into syntactically related non-overlapping groups of words, so-called text chunking").

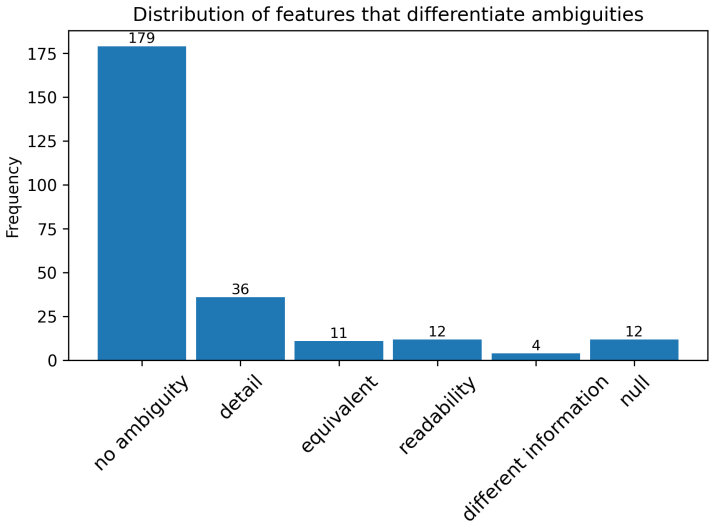

Figure 6: Distribution of features that help choose between two or more candidate task descriptions

Option 1	Option 2	Discussion
automatically assessing humor in edited news headlines	build systems for rating a humorous effect that is caused by small changes in text	We chose option 2 because it contains more **detail**.
quantify the degree of prototypicality of a target pair by measuring the relational similarity between it and pairs that are given as defining examples of a particular relation	rate word pairs by the degree to which they are prototypical members of a given relation class	This is a difficult example because initially option 1 seems better because it appears to have more detail. However, the second option has better **clarity**, and is more specific because of the phrase "word pairs" instead of "target pairs".
annotate instances of nouns, verbs, and adjectives using WordNet 3.1	label each instance with one or more senses, weighting each by their applicability	Both of these phrases provide **different pieces of information** about the task. Because these sentences are adjacent, the guidelines permit extracting the full sequence of text including both phrases and the text in between them.
given a set of documents and a set of target entities, the task consisted of building a timeline for each entity, by detecting, anchoring in time and ordering the events involving that entity	given a set of documents and a set of target entities, the task consists of building a timeline related to each entity, i.e. detecting, anchoring in time, and ordering the events in which the target entity is involved	Both phrases are equally good candidates and are **equivalent** in meaning. Either may be chosen.

Table 10: Examples of ambiguous annotation scenarios where it may be difficult to choose between two candidates

There are a number of situations that caused ambiguity during the annotation process. Certain kinds of sentences may appear at first glance to contain Task Descriptions, but actually served a different role. For example, task descriptions will often mention the research area, but a sequence that only describes the general research area is insufficient if it does not contain specific information on the task to be performed, as in the following example:

> "*Sensiting inflectionality*: Estonian task for SENSEVAL-2"
> **Discussion**: "Sensiting inflectionality" describes the research area, but is insufficient to describe the shared task to be performed.

One other pitfall we observed is the fact that sometimes paper authors use language when describing the aim, goal, or "task" of the task organizers or dataset annotators that makes it seem like they are describing the task to be performed by participating systems. A description of the organizers' aim or the dataset creation task would not be extracted as a task description according to our guidelines. For example:

> Aiming to *catalyze the development of models for predicting LE*, we organized the shared task described in this paper.
> **Discussion**: "catalyze the development of models for predicting LE" sounds like it could be a task description. The surrounding context shows us that it actually is describing the aim of the task organizers ("Aiming to... we organized the shared task").

Another source of ambiguity for the annotators is the presence of sub tasks, joint tasks, and multi-track or multi-language tasks. Developing a machine reader to determine how many subtasks are described in the paper and to extract a task description for each one from potentially disparate parts of the paper would not be trivial. For this reason, we do not annotate subtask descriptions unless they appear in consecutive sequences of text.

Another ambiguous situation is the scenario where there are two or more candidate task descriptions that are all decent choices. These ambiguities could be resolved by choosing the option that had either more **detail** or better **clarity**; choosing the sequence that works best out of context when the options contain complementary but **different information**; or choosing any candidate when the sequences are truly **equivalent**. The frequencies of each of these choices in the dataset can be seen in Figure 6, and examples of ambiguous cases can be seen in Table 10.

Lastly, sometimes a paper does not contain a sequence of text that sufficiently describes the task out of context. In any situation where a task description cannot be found, we use a portion of the title of the paper if the title contained a phrase describing the task. If no task description could be found in the body of the paper and the title did not sufficiently describe the task, then that paper would not receive an annotation. There were twelve such cases in the entire corpus.

A.7 Corpus Access

NLPSharedTasks corpus is available on GitHub and is licensed under a Creative Commons Attribution 4.0 International License.

B Error Analysis

Table 11 on the following page presents examples and analysis of errors made by our system on the test set.

Error Type	Sample In Context	Discussion
False Negative	Unsupervised Word Sense Induction and Discrimination (WSID, also known as corpus-based unsupervised systems) has followed this line of thinking, and tries to ***induce word senses directly from the corpus***.	This sentence may have been difficult for the system to classify because the actual task description span is relatively short compared to the overall sentence context.
False Negative	***Nine sub-tasks were included, covering problems in time expression identification, event expression identification and temporal relation identification***.	Papers with subtasks were difficult for the system. The system did not extract a single sentence from the paper containing this example.
Partial False Negative	***This task required participating systems to annotate instances of nouns, verb, and adjectives using Word-Net 3.1 (Fellbaum, 1998), which was selected due to its fine-grained senses. Participants could label each instance with one or more senses, weighting each by their applicability***.	Annotators were permitted to select sequences of text that spanned multiple sentences, if the additional text provided important details. Our system successfully classified the first sentence in this example as a task description, but missed the second sentence.
False Positive & False Negative	We present a counterfactual recognition (CR) task, the task of determining whether a given statement conveys counterfactual thinking or not, and further analyzing the causal relations indicated by counterfactual statements. In our counterfactual recognition task, we aim to ***model counterfactual semantics and reasoning in natural language.***	Some of the errors were also difficult cases for human annotators. In this example, the system selected the first sentence rather than the second. However, the annotator chose to prioritize readability over detail in this case.
Partial False Positive	This task seeks to evaluate the capability of systems for predicting dimensional sentiments of Chinese words and phrases. For a given word or phrase, participants were asked to ***provide a real-valued score from 1 to 9 for both the valence and arousal dimensions, respectively indicating the degree from most negative to most positive for valence, and from most calm to most excited for arousal***.	The system classified both of these sentences as task descriptions, although the annotator only chose a span from the second sentence.

Table 11: Examples of errors made by our system. The bolded and italicized spans of text are the original sequences identified by human annotators as task descriptions.