A Dataset for N-ary Relation Extraction of Drug Combinations

Aryeh Tiktinsky[*,1] **Vijay Viswanathan**[*,2] **Danna Niezni**[3] **Dana Meron Azagury**[3]
Yosi Shamay[3] **Hillel Taub-Tabib**[1] **Tom Hope**[1,4] **Yoav Goldberg**[1,5]
[1]Allen Institute for AI [2]Language Technologies Institute, Carnegie Mellon University
[3]Faculty of Biomedical Engineering, Technion – Israel Institute of Technology
[4]School of Computer Science & Engineering, The Hebrew University of Jerusalem
[5]Computer Science Department, Bar Ilan University
aryeht@allenai.org vijayv@andrew.cmu.edu

Abstract

Combination therapies have become the standard of care for diseases such as cancer, tuberculosis, malaria and HIV. However, the combinatorial set of available multi-drug treatments creates a challenge in identifying effective combination therapies available in a situation. To assist medical professionals in identifying beneficial drug-combinations, we construct an expert-annotated dataset for extracting information about the efficacy of drug combinations from the scientific literature. Beyond its practical utility, the dataset also presents a unique NLP challenge, as the first relation extraction dataset consisting of variable-length relations. Furthermore, the relations in this dataset predominantly require language understanding beyond the sentence level, adding to the challenge of this task. We provide a promising baseline model and identify clear areas for further improvement. We release our dataset,[1] code,[2] and baseline models[3] publicly to encourage the NLP community to participate in this task.

1 Introduction

"*So far, many monotherapies have been tested, but have been shown to have limited efficacy against* ***COVID-19****. By contrast,* ***combinational*** *therapies are emerging as a useful tool to treat SARS-CoV-2 infection.*" (Ianevski et al., 2021).

Indeed, combining two or more drugs together has proven to be useful for treatments of various medical conditions, including cancer (DeVita et al., 1975; Carew et al., 2008; Shuhendler et al., 2010), AIDS (Bartlett et al., 2006), malaria (Eastman and Fidock, 2009), tuberculosis (Bhusal et al., 2005), hypertension (Rochlani et al., 2017) and COVID-19 (Ianevski et al., 2020).

In this work, we examine the clinically significant and challenging NLP task of extracting known drug combinations from the scientific literature. We present an expert-annotated dataset and baseline models for this new task. Our dataset contains 1600 manually annotated abstracts, each mentioning between 2 and 15 drugs. 840 of these abstracts describe one or more positive drug combinations, varying in size from 2 to 11 drugs. The remaining 760 abstracts either contain mentions of drugs not used in combination, or discuss combinations of drugs that do not give a combined positive effect.

For the clinical setting, solving the drug combination identification task can help researchers suggest and validate complex treatment plans. For example, when searching for effective treatments for cancer, knowing which drugs interact synergistically with a first line treatment allows researchers to suggest new treatment plans that can subsequently be validated in-vivo and become a standard protocol (Wasserman et al., 2001; Katzir et al., 2019; Ianevski et al., 2020; Niezni et al., 2022).

From an NLP perspective, the drug combination identification task and dataset pushes the boundaries of relation extraction (RE) research, by introducing a relation extraction task with several challenging characteristics:

Variable-length n-ary relations Most work on relation extraction is centered on *binary relations* (e.g. Li et al. (2016), see full listing in §5), or on *n-ary relations with a fixed* n (e.g. Peng et al. (2017)). In contrast, the drug combination task involves *variable-length n-ary relations*: different passages discuss combinations of different numbers of drugs. For each subset of drugs mentioned in a passage, the model must predict if they are used together in a combination therapy and whether this drug combination is effective.

* Equal contribution.

[1]https://huggingface.co/datasets/allenai/drug-combo-extraction

[2]https://github.com/allenai/drug-combo-extraction

[3]https://huggingface.co/allenai/drug-combo-classifier-pubmedbert-dapt

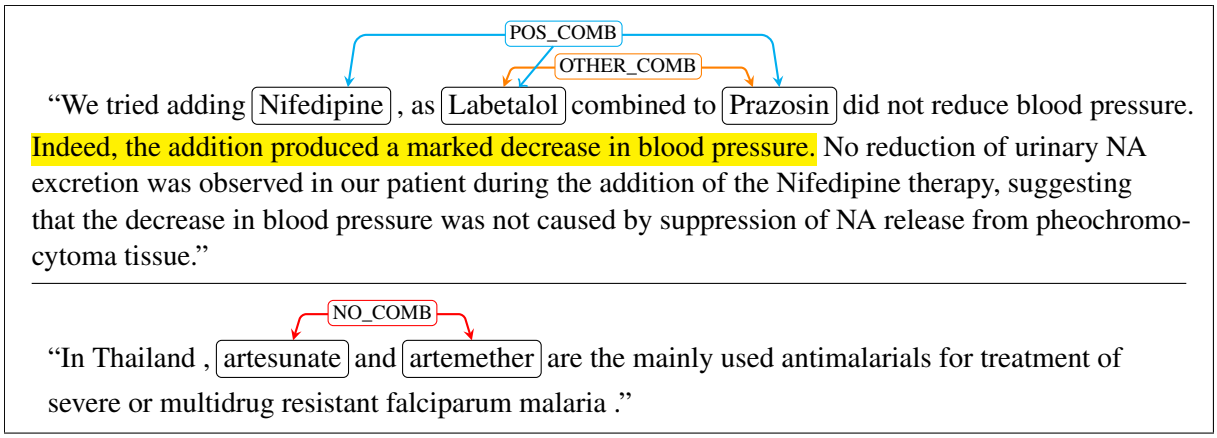

Figure 1: Examples of our label scheme. The top example contains two relations: a binary OTHER_COMB relation and a ternary POS_COMB relation. The evidence required to annotate the latter relation is found in a different sentence (highlighted). In the bottom example, each drug is described as a separate treatment rather than a combination therapy.

No type hints As noted by Rosenman et al. (2020) and Sabo et al. (2021), in many relation extraction benchmarks (Han et al., 2018; Sabo et al., 2021; Zhang et al., 2017), the argument types serve as an effective clue. However, argument types do not apply naturally to the drug combination task, in which all possible relation arguments are entities of the same type (drugs) and we need to identify specific subsets of them.

Long range dependencies The information describing the efficacy of a combination is often spread-out across multiple sentences. Indeed, our annotators reported that for 67% of the instances, the label could not be determined based on a single sentence, requiring reasoning with a larger textual context. Interestingly, our experiments show that our models *are not* helped by the availability of longer context, showing the limitations of current standard modeling approaches. This suggests our dataset can be a test-bed for models that attempt to incorporate longer context.

Challenging inferences As we show in our qualitative analysis (§4.2), instances in this dataset require processing a range of phenomena, including coordination, numerical reasoning, and world knowledge.

We hope that by releasing this dataset we will encourage NLP researchers to engage in this important clinical task, while also pushing the boundaries of relation extraction.

2 The Drug Combinations Dataset

A set of drugs in a biomedical abstract are classified to one of the following labels:

Positive combination (POS_COMB): the sentence indicates the drugs are used in combination, and the passage suggests that the combination has additive, synergistic, or otherwise beneficial effects which warrant further study.

Non-positive combination (OTHER_COMB): the sentence indicates the drugs are used in combination, but there is no evidence in the passage that the effect is positive (it is either negative or undetermined).[4]

Not a combination (NO_COMB): the sentence does not state that the given drugs are used in combination, even if a combination is indicated somewhere else in the wider context. An example is given in the lower half of Figure 1, where each of the drugs Artesunate and Artemether is given in isolation, and no combination is reported.

Our primary interest is to identify sets of drugs that are positive combinations.

2.1 Relevant Context Size for Classifying Drug Combinations

When formulating the extraction task and designing our data collection methodology, we first analyzed the locality of the phenomenon: to what extent are drug combinations are expressed in a single sentence, or is a larger context is needed? We sampled 275 abstracts that contained known

[4] We also experimented with another label for combinations that are discouraged (antagonistic, harmful or not effective). The agreement for this label was low, leading us to keep it as a subset of OTHER_COMB.

drug combinations according to DrugComboDB.[5] Analysis showed that 51% of these abstracts mentioned attempted drug combinations. In 97% of the abstracts containing drug combinations, all participating drugs in the attempted combination could be located within a single sentence in the abstract (for an example, see the OTHER_COMB relation in Figure 1). However, establishing the efficacy of the combination frequently required a larger context (such as the context accompanying the POS_COMB relation in Figure 1).

2.2 Task Definition

We define each instance in the Drug Combination Extraction (DCE) task to consist of a sentence, drug mentions within the sentence, and an enclosing context (e.g. paragraph or abstract).

The output of the task is a set of relations, each consisting of a set of participating drug spans and a relation label (POS_COMB or OTHER_COMB). Each subset of drug mentions not included in the output set is implicitly considered to have relation label NO_COMB.

More formally, DCE is the task of labeling an instance $X = \{C, i, D\}$ with a set of relation instances R, where $C = (S_1, ...S_n)$ is an ordered list of context sentences (e.g. all the sentences in an abstract or paragraph), $1 \leq i \leq n$ is an index of a target sentence $S_i = (w_1, ..., w_{n(i)})$ with $n(i)$ words, and $D = \{(d_{\mathbf{1}start}, d_{\mathbf{1}end}), ..., (d_{\mathbf{m}start}, d_{\mathbf{m}end})\}$ is a set of $m >= 2$ spans of drug mentions in S. The output is a set $R = \{(c_i, y_i)\}$ where $c_i \in \mathcal{P}(D)$ is a drug combination from $\mathcal{P}(D)$, the set of all possible drug combinations, and $y_i \in \{\text{POS_COMB}, \text{OTHER_COMB}\}$ is a combination label.

2.3 Evaluation Metric

We consider two settings: "Exact Match", a strict version which considers identifying exact drug combinations, and "Partial Match", a more relaxed version which assigns partial credits to correctly identified subsets.

We use standard precision, recall and F1 metrics for both settings. For the partial-match case, we replace the binary 0 or 1 score for a given combination with a refined score: $shared_drugs/total_drugs$. If there are multiple partial matches with gold relations, we take the one with maximum overlap. We compute **recall** as $identified_relations/all_gold_relations$, and **precision** as $correct_relations/identified_relations$.

We consider two metrics, the averaged Positive Combination F1 score which compares POS_COMB to the rest, and the averaged Any Combination F1 score which counts correct predictions for any combination label (POS or OTHER) as opposed to NO_COMB. The latter is an easier task, but still valuable for identifying drug combinations irrespective of their efficacy.

2.4 Collecting Data for Annotation

To collect data for annotation we curated a list of 2411 drugs from DrugBank [6] and sampled from PubMed a set of sentences which mention 2 or more drugs. Analysis of the first 50 sentences from this sample showed that only $8/50$ of the sentences included mentions of drug combinations. This meant that annotating the full sample will be costly, and will result in a dataset that's highly skewed toward relatively trivial NO_COMB instances.

We therefore repeated this experiment, sampling sentences whose PubMed abstract included a trigger phrase.[7] 48% of 50 sampled sentences included mentions of drug combinations. Evaluating the coverage of the trigger list against a new sample of abstracts with known drug combinations showed that 90% of these new abstracts included one of the trigger words. This suggests our trigger list is useful for fetching label-balanced data, without prohibitively restricting coverage and diversity.

Accordingly, we collected the majority of instances for annotation, 90%, using a basic search for sentences that contain at least two different drugs and whose abstract contains one of the trigger phrases. To overcome the lexical restrictions imposed by our trigger list, we sample the remaining 10% of instances using distant supervision: fetch-

[5] We used `Syner&Antag_voting.csv` taken from `http://drugcombdb.denglab.org/download/` and ranked according to the *Voting* metric.

[6] Curation included downloading a premade drug list from DrugBank's website, while removing non pharmacological intervention such as Vitamins and Supplements. The later we got from the FDA orange book.

[7] Triggers were selected by manually identifying words and phrases which frequently appear in abstracts mentioning drug combinations. These are phrases like "combination", "followed by", "prior to", etc. (see full list in Appendix A.3). The triggers are recall oriented, so while a presence of a trigger increases the chances that an abstract mentions a drug combination, it is definitely not clearly indicative. Importantly, since we're dealing with a wide context, the presence of a trigger in an abstract which includes multiple drugs does not mean the trigger is related to the drugs.

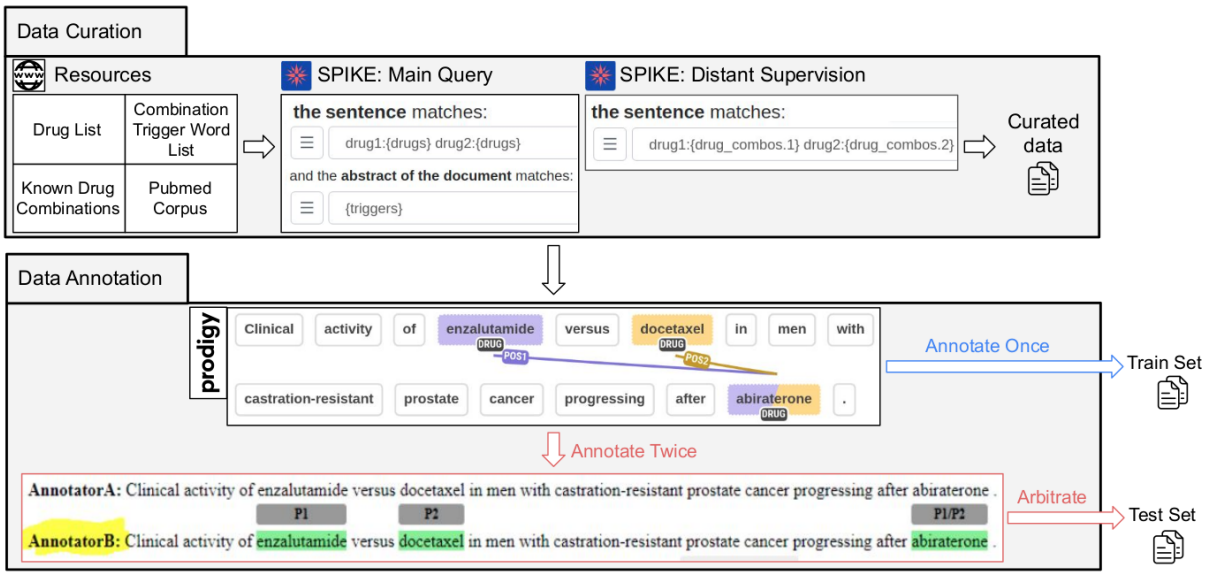

Figure 2: Illustration of the data construction process. First we construct the required knowledge resources. Then, we collect data using SPIKE –an extractive search tool– over the PubMed database. The train and test sets were annotated using Prodigy over the curated data. For test data, we collected two annotations for each sample, and then had a domain expert resolve annotation disagreements.

ing sentences containing pairs of drugs known to be synergistic according DrugComboDB, but whose abstract does not include one of our trigger phrases. All data collecting queries were performed using the SPIKE Extractive Search tool (Shlain et al., 2020; Taub-Tabib et al., 2020). The process is illustrated at the top of Figure 2.

2.5 The Annotation Process

Seven graduate students in biomedical engineering took part in the annotation task. The students all completed a course in combination therapies for cancer and were supervised by a principled researcher with expertise in this field.

We provided the participants with annotation guidelines which specified how the annotation process should be carried out (see Appendix A.1) and conducted an initial meeting where we reviewed the guidelines with the group and discussed some of the examples together.

Each of the participants had access to a separate instance of the Prodigy annotation tool (Montani and Honnibal, 2018), pre-loaded with the candidate annotation instances. Once a session starts, the instances (containing of a sentence with marked drug entities, and its context) appear in a sequential manner, with no time limit. For each instance we instructed the annotators to mark all subsets of drugs that participated in a combination, and for each subset to indicate its label (POS_COMB or OTHER_COMB). Moreover, we instructed them to indicate whether the context was needed in order to determine the positive efficacy of the relation.

Metric	Partial Match	Exact Match
Avg. Any Combination F1	88.9	86.1
Avg. Positive Combination F1	83.4	79.6

Table 1: Agreement scores using our adaptation of F1 score to allow for partial-match.

Despite the considerable time required for expert annotation, we collected annotations for 1634 passages. Among these, 272 were assigned to at least two annotators. After further arbitration by the lead researcher, these were used for the test set. The process is illustrated in the bottom part of Figure 2.

2.6 Inter-annotator Agreement

During the course of the task we calculated inter-annotator agreement multiple times to identify cases of disagreement and provide feedback to annotators. Each time, a set of 25 instances were randomly selected and assigned to all annotators. Agreement was calculated based on a pairwise F1 measure (with some modifications as described in §2.3) and averaged over all pairs of annotators (see discussion of alternative metrics in Appendix A.2).

Final agreement numbers, in Table 1, are satisfactory (Aroyo and Welty, 2013; Araki et al., 2018).

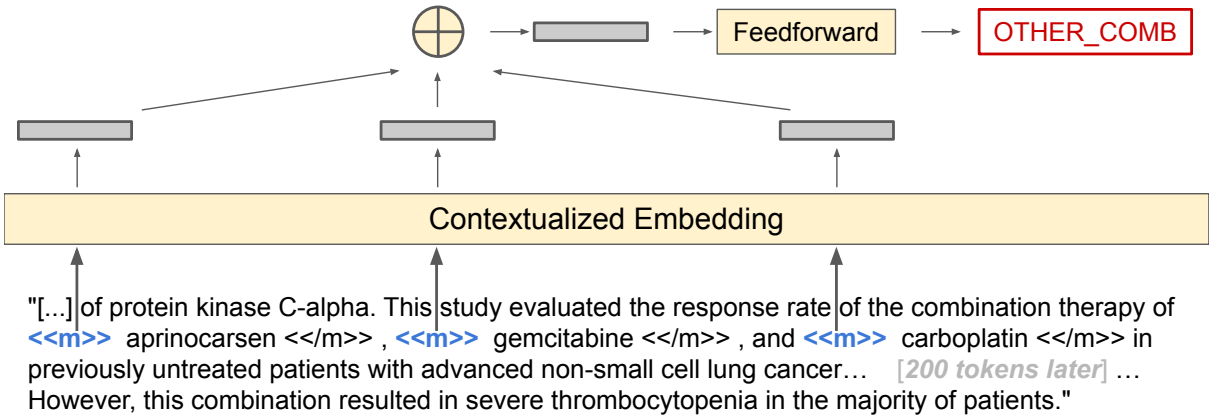

Figure 3: Our baseline architecture, adapted from the PURE model (Zhong and Chen, 2021)

2.7 Resulting Dataset

The dataset consists of 1634 annotated abstracts,[8] split into 1362 train and 272 test instances. These include 1248 relations; 838 are POS_COMB and 410 are OTHER_COMB (with the same label ratio in the train and test sets). 591 sentences contain no drug combination, 877 contain one relation (either POS_COMB or OTHER_COMB), and 166 contain two or more different combinations. Among annotated relations, 900 are binary, 226 are 3-ary, 69 are 4-ary, and 53 are 5-ary or more.

For each instance in the resulting dataset we include the context-required indication provided by the annotators. In 835 out of 1248 relations the annotator marked the context as needed which is 67% of the time, showing the importance of the context in the DCE task.

3 Experiments

3.1 Baseline Model Architecture

We establish a baseline model to measure the difficulty of our dataset and reveal areas for improvement. For our underlying baseline model architecture, we adopt the PURE architecture from Zhong and Chen (2021), which is state-of-the-art on several relation classification benchmarks, including the SciERC binary scientific RE dataset (Luan et al., 2018). The PURE architecture, designed for 2-ary and 3-ary relation extraction, consists of three components. First, special "entity marker" tokens are inserted around all entities in a candidate relation. Next, these marker tokens are encoded with a contextualized embedding model. Finally, the entity marker embeddings are concatenated and fed to a feedforward layer for prediction.

Unlike the original PURE architecture, we consider the more challenging case of extracting relations of variable arity. To support this setting, we *average* the entity marker tokens in a relation rather than concatenate. The final baseline model architecture is shown in Figure 3. For the contextual embedding component of this architecture, we experiment with four different pretrained scientific language understanding models (SciBERT (Beltagy et al., 2019), BlueBERT (Peng et al., 2019), PubmedBERT (Gu et al., 2020), and BioBERT (Lee et al., 2020)). During training, we only finetune the final *BERT layer. We train each model architecture for 10 epochs on a single NVIDIA Tesla T4 GPU with 15GB of GPU memory, which takes roughly 7 hours to train for each model.

To our knowledge, there are no other models designed for variable-length N-ary relation extraction, so we consider no other baselines.

3.2 Domain-Adaptive Pretraining

Our baseline model architecture relies heavily on a pretrained contextual embedding model to provide discriminative features to the relation classifier. Gururangan et al. (2020) showed that continued domain-adaptive pretraining almost always leads to significantly improved downstream task performance. Following this paradigm, we performed continued domain-adaptive pretraining ("DAPT") on our contextual embedding models.

We acquired in-domain pretraining data using the same procedure used to collect data for annotation: running a SPIKE query against PubMed to find abstracts containing multiple drug names

[8] This is a similar size to existing human-labeled biomedical relation extraction datasets, such as BioCreative V CDR (Li et al., 2016), which has 1500 abstracts annotated, BioCreative VI (Krallinger et al., 2017), which has 2432 abstracts, and DDI (Herrero-Zazo et al., 2013), which has 714 abstracts.

Model	Positive Combination F1		Any Combination F1	
	Exact Match	Partial Match	Exact Match	Partial Match
Human-Level	79.6	83.4	86.1	88.9
Rule-based	31.8	45.6	39.1	57.4
SciBERT	44.6 (± 4.6)	55.0 (± 5.9)	50.2 (± 1.9)	63.6 (± 2.7)
w/ DAPT	54.8 (± 3.2)	63.6 (± 2.0)	61.8 (± 2.7)	72.8 (± 2.1)
BlueBERT	41.2 (± 4.8)	51.7 (± 6.0)	47.3 (± 4.2)	59.9 (± 6.2)
w/ DAPT	56.6 (± 2.3)	63.5 (± 3.1)	64.2 (± 2.6)	74.7 (± 2.7)
PubmedBERT	50.7 (± 5.5)	59.6 (± 5.8)	55.9 (± 3.2)	66.7 (± 3.8)
w/ DAPT	**61.8** (± 5.1)	**67.7** (± 4.8)	**69.4** (± 1.7)	**77.5** (± 2.2)
BioBERT	45.4 (± 3.7)	55.8 (± 2.2)	46.7 (± 3.6)	58.3 (± 5.1)
w/ DAPT	56.0 (± 6.5)	63.5 (± 7.5)	65.6 (± 1.8)	75.7 (± 2.2)

Table 2: Comparing different foundation models (with and without continued domain-adaptive pretraining) on Exact-Match and Partial-Match relation extraction metrics. Mean score from 4 different random seeds is reported, and standard deviation is computed across seeds.

and a “trigger phrase" (from the list in Appendix A.3). This query resulted in 190K unique abstracts. We do not include any paragraphs from our annotated dataset. We then perform domain-adaptive training against this dataset using the `Hugging Face Transformers` library. We train for 10 epochs using a learning rate of 5e-4, finetuning all *BERT layers and using the same optimization parameters specified by Gururangan et al. (2020). This pretraining took roughly 8 hours using four 15GB NVIDIA Tesla T4 GPUs.

3.3 Relation Prediction

To apply the model to drug combination extraction, we reduce the RE task to an RC task by considering all subsets of drug combinations in a sentence, treating each one as a separate classification input, and combining the predictions.

This poses two challenges: there may be a large number of candidate relations for a given document, and each relation is classified independently despite the combinatorial structure. To handle these issues, we use a greedy heuristic of choosing the smallest set of disjoint relations whose union covers as many drug entities as possible in the sentence. We do this iteratively: at each step, we choose the largest predicted relation that does not contain any drugs found in the relations chosen at previous iterations.

This greedy heuristic favors large (high arity) relations. Nonetheless, we empirically find this method is helpful for extracting high-precision drug combinations from our model architecture.

3.4 Rule-based baseline

To further validate that the trigger words do not introduce bias to our task, we consider an additional baseline based on the following rule: if a trigger word is found in the same sentence with multiple drugs, this set of drugs is tagged as POS_COMB.

4 Results

4.1 Effect of Pretrained LMs and Domain-Adaptive Pretraining

We show results of our baseline model architectures in Table 2. For each model, we report the mean and standard deviation of each metric over four identical models trained with different seeds.[9] Among the four base scientific language understanding models in our experiments, we observe PubmedBERT to be the strongest on every metric. We additionally find that domain-adaptive pretraining provides significantly improvements for every base model, consistently giving 5-10 points of improvement on Positive Combination F1 score. The value of domain-adaptive pretraining supports our observation that encoding domain knowledge is critical to solving this new task.

The rule-based approach underperformed all learned models (30 F1 points under our strongest model, PubmedBERT-DAPT). This shows this task cannot be reduced to keyword identification.

4.2 Qualitative Error Analysis

We identify classes of challenges that make this task difficult, both in terms of human annotation and machine prediction.

Coordination Ambiguity: A known linguistic challenge is the ambiguity that stems from vague coordination. In cases where explicit combination words (e.g. combination, plus, together with, etc)

[9] Seeds used are 2021, 2022, 2023, and 2024

Model	Positive Combination F1		Any Combination F1	
	Exact Match	Partial Match	Exact Match	Partial Match
No Extra-sentential Context	63.4 (± 0.6)	68.5 (± 1.1)	69.7 (± 1.3)	76.8 (± 1.7)
1 Sentence of Context	63.9 (± 2.3)	69.4 (± 3.5)	71.9 (± 1.1)	78.6 (± 1.8)
2 Sentences of Context	61.9 (± 9.0)	67.6 (± 9.2)	70.1 (± 2.3)	77.9 (± 1.3)
3 Sentences of Context	65.2 (± 2.3)	72.4 (± 1.3)	70.8 (± 1.7)	78.7 (± 1.2)

Table 3: The effect of extra-sentential context on model performance. n sentences are included on each side of the relation-bearing sentence. Mean and standard deviation of each metric are reported over 4 different random seeds.

are not used, it may be unclear whether two drugs are being used together or separately. For example in *"These findings may help clinicians identify patients for whom acamprosate **and** naltrexone may be most beneficial"* it is unclear if *acamprosate* and *naltrexone* are being described in combination or as independent treatments, leading to either a POS label for the former or NO_COMB for the latter.

Numerical and Relative Reasoning: In some cases, the effect of a treatment is described in relative or numerical terms, rather than an absolute claim. Consider the example, *"The infection rate in the control group was 3.5% and in the treated group 0.5%."*. Here, the reader must compare the control vs experimental groups and deduce that the experimental outcome is positive, because the treatment yields a lower infection rate.

Domain Knowledge: Similarly, classifying relations in this dataset may require an understanding of domain knowledge. In *"Growth inhibition and apoptosis were significantly higher in BxPC-3, HPAC, and PANC-1 cells treated with **celecoxib and erlotinib** than cells treated with either **celecoxib** or **erlotinib**"*, one must understand that having higher values of *Growth inhibition and apoptosis* in specific cells is a positive outcome, in order to classify this combination as positive.

Context related Complications: The following are kinds of complications found when the evidence lies in the wider part of the context.

Coreference: Anaphoric or coreferential reasoning is sometimes needed to understand the efficacy of the combination e.g. *"it was demonstrated that **they** could be combined with acceptable toxicity."*.

Contradicting Evidence: the reader often must infer a conclusion given opposing claims within a given abstract. This can happen as combinations can be referred as e.g. *toxic but effective*.

Long Distance: The target sentence can be far—up to 41 sentences apart—from the evidence sentence, making it difficult for even humans to process.

4.3 Quantitative Error Analysis

To probe this task, we analyze the performance of our strongest model—the one using a PubmedBERT base model tuned with domain-adaptive pretraining—along different partitions of test data. We trained with four random seeds and perform comparisons using a paired multi-bootstrap hypothesis test where bootstrap samples are generated by sampling hierarchically over the four random seeds and the subsets of the test set (Sellam et al., 2021). We use 1000 bootstrap samples in each test.

4.3.1 Do models leverage context effectively?

Each relation in our dataset consists of entities contained within a single sentence, but labeling the relation frequently requires extra-sentential context to make a decision. In our dataset, annotators record whether or not each relation requires paragraph-level context to label, reporting that 67% of drug combinations required context to annotate.

To understand the extent that models make use of paragraph-level context, we trained and evaluated our PubmedBERT-based model using varying amounts of extra-sentential context around the sentence containing drug entities. In Table 3, we see that adding context provides nearly identical performance to training a model with no extra-sentential context at all, with differences rarely exceeding one standard deviation of F1 score.

However, we see increased variability in "Positive Combination F1" performance when extra-sentential context is used. To explain this, recall from §2.1 that determining the *efficacy* of a drug combination often requires paragraph-level context for annotators, while identifying *any combination* usually requires no context. From qualitative analysis of attention maps, we observed that our models are not able to consistently identify the salient parts of paragraph-level context, potentially causing instability with larger inputs.

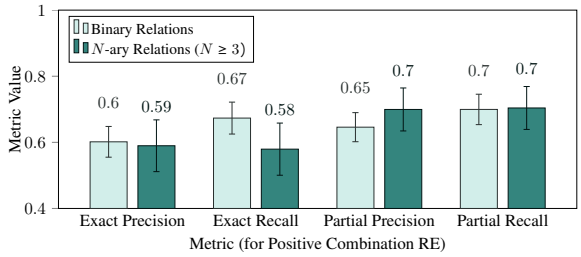

Figure 4: Comparing models performance on binary vs higher-order N-ary relations, averaged over 4 seeds of the PubmedBERT-DAPT model. No consistent significant differences were observed; p-values for these comparisons are 0.456, 0.149, 0.240, and 0.276.

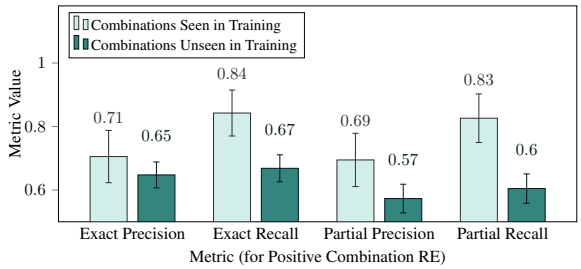

Figure 5: Comparing relation extraction on test set drug combinations that are observed in the training set or not, using the PubmedBERT-DAPT model. Paired multi-bootstrap test p-values for these four comparisons are 0.262, 0.025, 0.103, and 0.009, respectively.

These results suggest ample room for improvement in extracting document-level evidence. This makes our dataset a potentially useful benchmark for document-level language understanding.

4.3.2 Binary vs. higher-arity relations

Given that our dataset is the first relation extraction dataset with variable-arity relations, do higher-order relations pose a particular challenge for our models? To answer this, we separate all predicted and ground truth relations for the test set into binary relations and higher-arity relations. We then report precision among each subset of predicted relations and recall among each subset of ground truth relations. We perform this experiment across four different model seeds, and report results in aggregate using a paired multi-bootstrap procedure. In the results in Figure 4, we see no consistent significant differences between models of different arities, suggesting that our technique of computing relation representations by averaging entity representations scales well to higher-order relations.

4.3.3 Generalizing to new drug combinations

How well can relation extraction models classify drug combinations not seen during training? Similar to the setup in §4.3.2, we divide all predicted and ground truth relations for the test set into the set of drug combinations which are also annotated in our training set, and the set that have not been. In our dataset, over 80% of annotated test set relations are not found in the training set.

In Figure 5, performance is consistently better for relations observed in the training set than for unseen relations, by a margin of 10-15 points. Recall, in particular, is significantly worse for relations unseen during training (at 95% confidence), and precision is potentially also worse. Considering that unseen drug combinations are practically more valuable than already-known combinations, improving generalization to new combinations is a critical area of improvement for this task.

5 Related Work

The DDI dataset (Herrero-Zazo et al., 2013) is the only work to our knowledge that annotates drug interactions for text mining. However, it fundamentally differs from our dataset in the type of annotations provided: the DDI annotates the type of discourse context in which a drug combination is mentioned, without providing explicit information about combination efficacy. In contrast, our dataset is focused on semantically classifying the efficacy of drug combinations as stated in text.

Other RE datasets exist in the biomedical field (Peng et al., 2017; Li et al., 2016; Wu et al., 2019; Krallinger et al., 2017), but do not focus on drug combinations. Similarly, several RE datasets tackle the N-arity problem in the scientific domain (Peng et al., 2017; Jain et al., 2020; Kardas et al., 2020; Hou et al., 2019), and in the non-scientific domain (Akimoto et al., 2019; Nguyen et al., 2016), however, **all of them consider a fixed choice of** N.

6 Conclusions

We present a new resource for drug combination and efficacy identification. We establish baseline models that achieve promising results but reveal clear areas for improvement. Beyond the immediate, application-ready value of this task, this task poses unique relation extraction challenges as the first dataset containing variable-arity relations. We also highlight challenges with document-level representation learning and incorporating domain knowledge. We encourage others to partici-

pate in this task. Our processed dataset[10] and best baseline model[11] are available on Hugging Face, and our model training code is available to the public at https://github.com/allenai/drug-combo-extraction.

Acknowledgements

This project has received funding in part from the European Research Council (ERC) under the European Union's Horizon2020 research and innovation programme, grant agreement 802774 (iEXTRACT), and in part from the NSF Convergence Accelerator Award #2132318. We would also like to thank our annotators from the Shamay lab at the Faculty of Biomedical Engineering, Technion, including Shaked Launer-Wachs, Yuval Harris, Maytal Avrashami, Hagit Sason-Bauer and Yakir Amrusi

References

Kosuke Akimoto, Takuya Hiraoka, Kunihiko Sadamasa, and Mathias Niepert. 2019. Cross-sentence n-ary relation extraction using lower-arity universal schemas. 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 6225–6231, Hong Kong, China. Association for Computational Linguistics.

Jun Araki, Lamana Mulaffer, Arun Pandian, Yukari Yamakawa, Kemal Oflazer, and Teruko Mitamura. 2018. Interoperable annotation of events and event relations across domains. In *Proceedings 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 10–20.

Lora Aroyo and Chris Welty. 2013. Measuring crowd truth for medical relation extraction. In *2013 AAAI Fall Symposium Series*.

John A Bartlett, Michael J Fath, Ralph Demasi, Ashwaq Hermes, Joseph Quinn, Elsa Mondou, and Franck Rousseau. 2006. An updated systematic overview of triple combination therapy in antiretroviral-naive hiv-infected adults. *Aids*, 20(16):2051–2064.

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.

Y Bhusal, CM Shiohira, and N Yamane. 2005. Determination of in vitro synergy when three antimicrobial agents are combined against mycobacterium tuberculosis. *International journal of antimicrobial agents*, 26(4):292–297.

Jennifer S Carew, Francis J Giles, and Steffan T Nawrocki. 2008. Histone deacetylase inhibitors: mechanisms of cell death and promise in combination cancer therapy. *Cancer letters*, 269(1):7–17.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

VT DeVita, RC Young, and GP Canellos. 1975. Combination versus single agent chemotherapy: a review of the basis for selection of drug treatment of cancer. *Cancer*, 35(1):98—110.

Richard T Eastman and David A Fidock. 2009. Artemisinin-based combination therapies: a vital tool in efforts to eliminate malaria. *Nature Reviews Microbiology*, 7(12):864–874.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.

[10] https://huggingface.co/datasets/allenai/drug-combo-extraction

[11] https://huggingface.co/allenai/drug-combo-classifier-pubmedbert-dapt

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.

Aleksandr Ianevski, Rouan Yao, Svetlana Biza, Eva Zusinaite, Andres Mannik, Gaily Kivi, Anu Planken, Kristiina Kurg, Eva-Maria Tombak, Mart Ustav, et al. 2020. Identification and tracking of antiviral drug combinations. *Viruses*, 12(10):1178.

Aleksandr Ianevski, Rouan Yao, Hilde Lysvand, Gunnveig Grødeland, Nicolas Legrand, Valentyn Oksenych, Eva Zusinaite, Tanel Tenson, Magnar Bjørås, and Denis E. Kainov. 2021. Nafamostat–interferon- combination suppresses sars-cov-2 infection in vitro and in vivo by cooperatively targeting host tmprss2. *Viruses*, 13(9).

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. Scirex: A challenge dataset for document-level information extraction.

Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. 2020. AxCell: Automatic extraction of results from machine learning papers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8580–8594, Online. Association for Computational Linguistics.

Itay Katzir, Murat Cokol, Bree B Aldridge, and Uri Alon. 2019. Prediction of ultra-high-order antibiotic combinations based on pairwise interactions. *PLoS computational biology*, 15(1):e1006774.

Martin Krallinger, Obdulia Rabal, Saber Ahmad Akhondi, Martín Pérez Pérez, Jesus Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio Baso López, Umesh K. Nandal, Erin M. van Buel, Anjana Chandrasekhar, Marleen Rodenburg, Astrid Lægreid, Marius A. Doornenbal, Julen Oyarzábal, Anália Lourenço, and Alfonso Valencia. 2017. Overview of the BioCreative VI chemical-protein interaction track.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jiao Li, Yueping Sun, Robin Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn Mattingly, Thomas Wiegers, and Zhiyong lu. 2016. Biocreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

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.

Ines Montani and Matthew Honnibal. 2018. Prodigy: A new annotation tool for radically efficient machine teaching. *Artificial Intelligence*, to appear.

Kiem-Hieu Nguyen, Xavier Tannier, Olivier Ferret, and Romaric Besançon. 2016. A dataset for open event extraction in English. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1939–1943, Portorož, Slovenia. European Language Resources Association (ELRA).

Danna Niezni, Hillel Taub-Tabib, Yuval Harris, Hagit Sason-Bauer, Yakir Amrusi, Dana Azagury, Maytal Avrashami, Shaked Launer-Wachs, Jon Borchardt, M Kusold, Aryeh Tiktinsky, Tom Hope, Yoav Goldberg, and Yosi Shamay. 2022. Extending the boundaries of cancer therapeutic complexity with literature data mining. *bioRxiv preprint*.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65.

Yogita Rochlani, Mohammed Hasan Khan, Maciej Banach, and Wilbert S Aronow. 2017. Are two drugs better than one? a review of combination therapies for hypertension. *Expert opinion on pharmacotherapy*, 18(4):377–386.

Shachar Rosenman, Alon Jacovi, and Yoav Goldberg. 2020. Exposing shallow heuristics of relation extraction models with challenge data.

Ofer Sabo, Yanai Elazar, Yoav Goldberg, and Ido Dagan. 2021. Revisiting few-shot relation classification: Evaluation data and classification schemes.

Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, Ian Tenney, and Ellie Pavlick. 2021. The multiberts: Bert reproductions for robustness analysis. *ArXiv*, abs/2106.16163.

Micah Shlain, Hillel Taub-Tabib, Shoval Sadde, and Yoav Goldberg. 2020. Syntactic search by example. In *ACL*.

Adam J Shuhendler, Richard Y Cheung, Janet Manias, Allegra Connor, Andrew M Rauth, and Xiao Yu Wu. 2010. A novel doxorubicin-mitomycin c co-encapsulated nanoparticle formulation exhibits anti-cancer synergy in multidrug resistant human breast cancer cells. *Breast cancer research and treatment*, 119(2):255–269.

Hillel Taub-Tabib, Micah Shlain, Shoval Sadde, Dan Lahav, Matan Eyal, Yaara Cohen, and Yoav Goldberg. 2020. Interactive extractive search over biomedical corpora. *arXiv preprint arXiv:2006.04148*.

Ernesto Wasserman, William Sutherland, and Esteban Cvitkovic. 2001. Irinotecan plus oxaliplatin: a promising combination for advanced colorectal cancer. *Clinical colorectal cancer*, 1(3):149–153.

Ye Wu, Ruibang Luo, Henry CM Leung, Hing-Fung Ting, and Tak-Wah Lam. 2019. Renet: A deep learning approach for extracting gene-disease associations from literature. In *International Conference on Research in Computational Molecular Biology*, pages 272–284. Springer.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *North American Association for Computational Linguistics (NAACL)*.

A Appendices

A.1 Annotation Guidelines

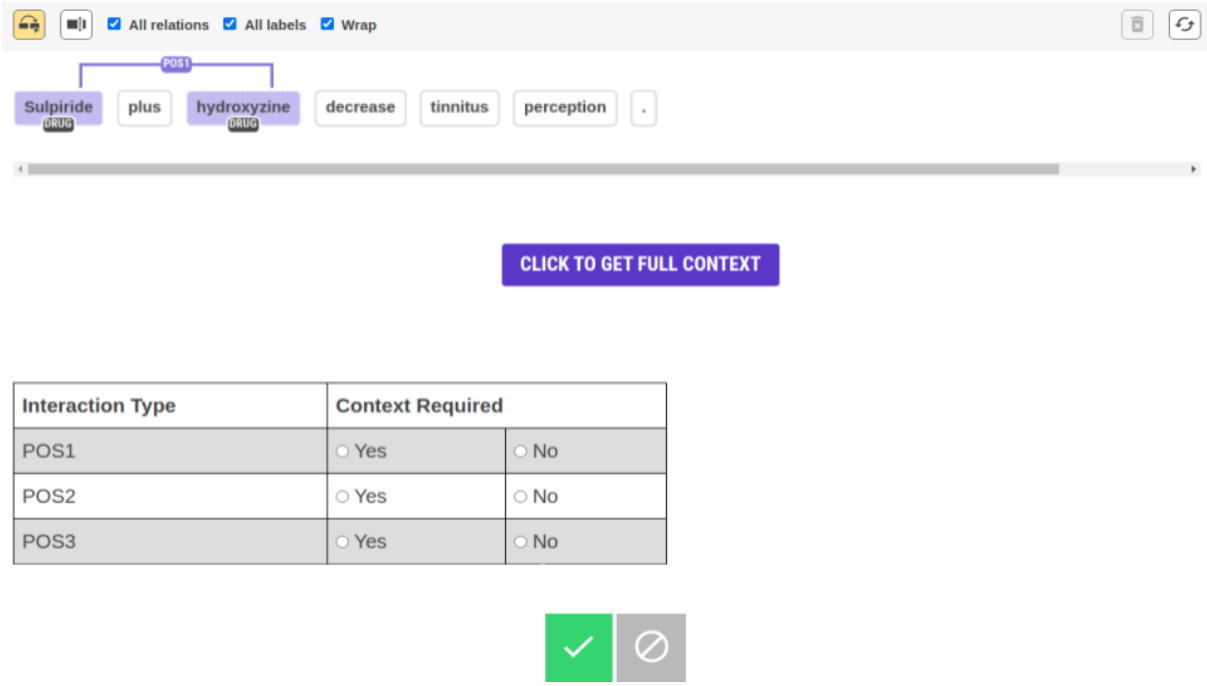

Figure 6: Annotation instance in the Prodigy environment. The screen is constructed of the sentence where they should mark relations, a button to show the full context and a selection per relation to indicate the necessity of the context.

For this task we recruited 7 annotators all studying for advanced degrees in biomedical engineering. The annotators were payed by their advisor, an amount that is standard for annotation projects in their country of residence. All participating annotators were provided with annotation guidelines. The guidelines specified how the annotation process should be carried out and provided definitions and examples for the different labels used. As the task progressed, the guidelines were also expanded to include discussion of frequently encountered issues.

For a given instance, such as presented in the top of Figure 6 the annotator needs to first recognize any missing drugs and mark them, and then label any interactions they find among the drugs. In case they need to consult a wider context they can press on a 'show more context' button and a text box with the wider context will appear. This context can be again hidden by clicking the same button if needed. Lastly, in the bottom of the sample page, we present a table with questions regarding the necessity of using the context.

Then the annotator should decide if they need to ignore the current sample or to complete the current instance and accept it, by pressing the accept and ignore buttons.

The annotators are instructed as follows. They should read the sentence carefully, and try to answer a two phase question to themselves. First, if the drugs are mentioned in any form of combination or they should be given separately. Second, if indeed the annotator recognized the drugs as a combination can they determine the efficacy of the combination by the sole sentence.

In case they can not determine the efficacy they are instructed to press on the 'get more context' button and read the entire context in order to determine what is the correct efficacy. If after reading the context they can still not determine the efficacy then the label of the interaction should be OTHER_COMB (aside from negative label experimentation mentioned in Footnote 4). Otherwise it should be POS_COMB. In case that they recognized that there is no combination between the drugs in the sentence then they should not use any label and simply accept the current instance. Then they should answer the context related questions for the POS_COMB label in order to signal if the context was needed.

While reading the sentence if the annotators find unmarked drugs they can mark them before continuing

to the interaction-labeling phase and treat them the same as the other drugs, but, it is not required to mark a word as drug in order to use it in an interaction. If a drug is marked in a wrong manner they should try and fix it, e.g. the span of the drug is incorrect.

In order to achieve more consistent and accurate annotations, they are also instructed to annotate all the interactions that they can find in a given sentence. They should always use the *accept* button even if there are no interactions in the sentence. Only in cases where they want to skip a sentence (e.g. when there is an inherent problem with it) or leave it for a future discussion they should use the *ignore* button. An interaction can occur between more than two drugs, if so they should notice that they don't need each pair from this group to have a marked interaction, as long as they all connect to the same graph. e.g. "Drugs A, B and C are synergistic." connecting A to B and B to C is sufficient, no need to connect drug A to drug C. Each interaction should be marked with a different tag (POS_COMB1, POS_COMB2..., OTHER_COMB1, OTHER_COMB2...).

A.2 Evaluation Metric Discussion

For measuring the agreement, we chose to use our adaptation of F1 score and not other common metrics such as Cohen's Kappa (Cohen, 1960) or one of its variations (e.g. Feliss's Kappa (Fleiss, 1971) and Krippendorf's Alpha (Hayes and Krippendorff, 2007)). These metrics expect a setup where the *relation* candidates are already marked and the task is only to label them – a labeling task and not an extraction task. This causes two problems, one is that they inherently do not need to handle a partial match. So if for example there are three drugs in a sentence, the first annotator annotated a relation between drugs A and B, while a second annotator annotated the same relation between drugs A, B and C. So we will either underestimate or overestimate their agreement score if we considered this a mismatch or a match respectively. Moreover, their calculations depend on the *hypothetical agreement by chance* normalization factor, but this will not reflect the difficulty of random choosing in our setup as they ignore the size of the combinatorial set of relation candidates we can possibly have.

A.3 Trigger List

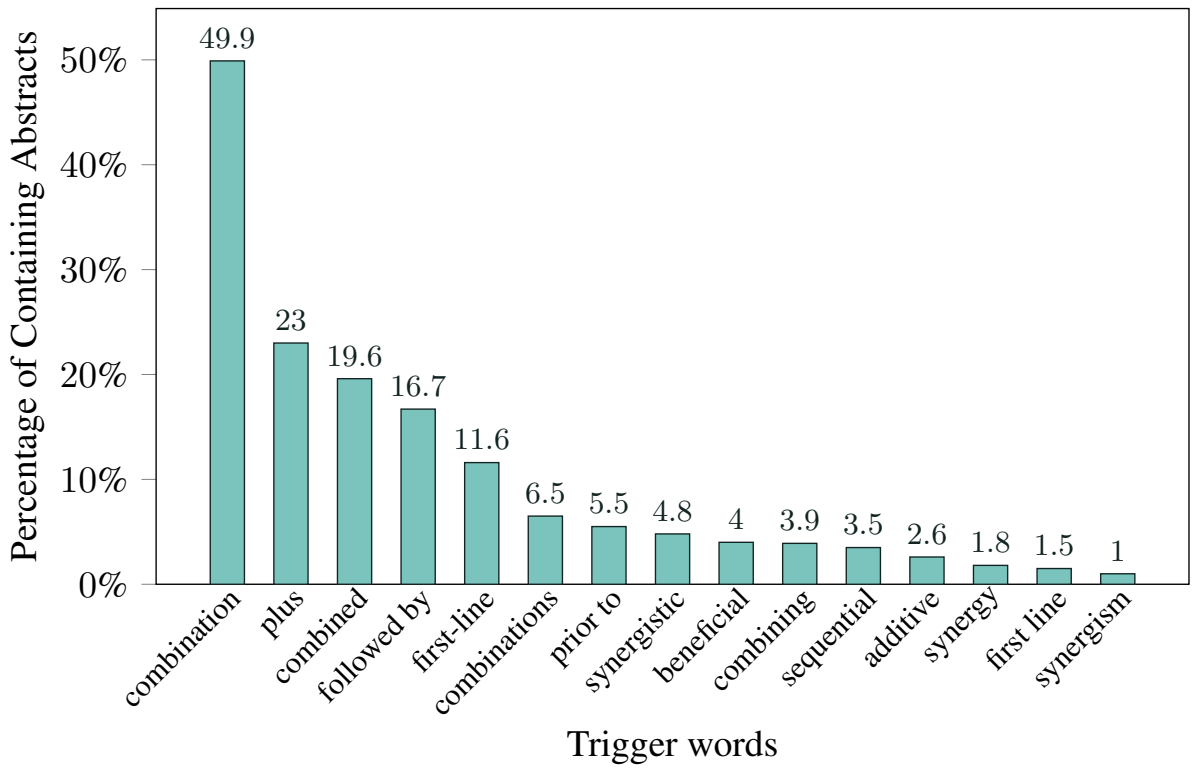

Figure 7: Abstracts percentage including each trigger word (1634 abstracts included; 44 words in the full word list; Words <1% were neglected from the figure.

In Figure 7 we show the triggers that we used in the Spike queries. We show the percentage of abstracts that included each trigger (others under 1%: *conjunction, two-drug, first choice, additivity, combinational, synergetic, simultaneously with, supra-additive, five-drug, combinatory, over-additive, timed-sequential, co-blister, super-additive, synergisms, synergic, synergistical, less-than-additive, greater-than-additive, additivesynergistic, supraadditive, superadditive, overadditive, subadditive, first-choice, 2-drug, sub-additive, more-than-additive, 3-drug*).