

# Effective Crowd-Annotation of Participants, Interventions, and Outcomes in the Text of Clinical Trial Reports

**Markus Zlabinger**

TU Wien, Vienna, Austria

markus.zlabinger@tuwien.ac.at

**Marta Sabou**

TU Wien, Vienna, Austria

marta.sabou@tuwien.ac.at

**Sebastian Hofstätter**

TU Wien, Vienna, Austria

s.hofstaetter@tuwien.ac.at

**Allan Hanbury**

TU Wien, Vienna, Austria

allan.hanbury@tuwien.ac.at

## Abstract

The search for Participants, Interventions, and Outcomes (PIO) in clinical trial reports is a critical task in Evidence Based Medicine. For an automatic PIO extraction, high-quality corpora are needed. Obtaining such a corpus from crowdworkers, however, has been shown to be ineffective since (i) workers usually lack domain-specific expertise to conduct the task with sufficient quality, and (ii) the standard approach of annotating entire abstracts of trial reports as one task-instance (i.e. HIT) leads to an uneven distribution in task effort. In this paper, we switch from entire abstract to sentence annotation, referred to as the SENBASE approach. We build upon SENBASE in SENSUPPORT, where we compensate the lack of domain-specific expertise of crowdworkers by showing for each task-instance similar sentences that are already annotated by experts. Such *tailored task-instance examples* are retrieved via unsupervised semantic short-text similarity (SSTS) method – and we evaluate nine methods to find an effective solution for SENSUPPORT. We compute the Cohen’s Kappa agreement between crowd-annotations and gold standard annotations and show that (i) both sentence-based approaches outperform a BASELINE approach where entire abstracts are annotated; (ii) supporting annotators with tailored task-instance examples is the best performing approach with Kappa agreements of 0.78/0.75/0.69 for P, I, and O respectively.

## 1 Introduction

Evidence Based Medicine is the practice of decision making based on the best available scientific information. Finding such information rapidly is essential, especially in the current pandemic crisis where thousands of medical articles about COVID-19 are published weekly (Škorić et al., 2020). To

make the search process time-efficient, the PICO model enables specific search for: **Participants** (e.g. “patients with headache”), **Interventions** (“ibuprofen”), **Comparisons** (“placebo”), and **Outcomes** (“pain reduction”) (Huang et al., 2006). To allow a search for structured PICO information in trial reports, a prior automatic extraction is necessary.

The effectiveness of an automatic PICO extraction depends on the quality of manually annotated corpora. As an alternative to scarce and expensive expert annotators, Nye et al. (2018) hired crowdworkers from the Mechanical Turk platform (MTurk) to annotate Participants, Interventions, and Outcomes (PIO<sup>1</sup>) in clinical trial reports. The crowdworkers, however, reached low agreements to expert annotations, potentially affected by (i) a lack of domain-specific expertise of the crowdworkers, and (ii) an uneven task length distribution. **The lack of domain-specific expertise** makes it difficult for crowdworkers to understand the terminology and jargon that prevails in medical literature (Kim et al., 2011; Wallace, 2018). As a result, workers experience medical tasks as cognitively overwhelming, with the side effect of a decreased label quality (Finnerty et al., 2013).

**An uneven task length distribution** makes the effort to complete individual task-instances<sup>2</sup> unevenly distributed thus enticing workers to “cherry pick” short task-instances or rush longer ones (Cheng et al., 2015; Feyisetan et al., 2017). In the task design of Nye et al. (2018), *entire abstracts* of clinical trial reports were annotated. These abstracts contain on average 268 words with a high standard deviation of 89, resulting in an uneven task length distribution.

In this paper, we address these problems by

<sup>1</sup>The I and C were unified as Intervention

<sup>2</sup>Referred to as HIT on the Mechanical Turk platform

proposing two novel PIO task designs:

**SENBASE:** The uneven task length is addressed by shifting from annotating abstracts to a *sentence-based* annotation. This makes the effort to complete individual task-instances more evenly distributed: Sentences have with an average word length of 25 and a standard deviation of 13 a 85% reduced word length variety compared to abstracts.

**SENSUPPORT:** This task design builds upon SENBASE by additionally compensating the lack of domain-specific expertise of crowdworkers. A common strategy to train crowdworkers for a task is to provide a few examples that illustrate how the task should be performed. In addition to such *static examples*, we propose to show for each task-instance similar sentences that are already annotated by experts. Such *tailored task-instance examples* are retrieved from a set of expert annotations—usually available to evaluate the performance of non-expert annotators (Daniel et al., 2018)—via an unsupervised semantic similarity method.

During our search for an effective, unsupervised semantic short-text similarity (SSTS) method for SENSUPPORT, we observed a lack of comparative evaluations for biomedical tasks. Therefore, to address this gap, we perform a comparative evaluation of nine SSTS methods, ranging from traditional count-based methods (e.g. TFIDF) to recent text embedding methods (e.g. Sen-BERT (Reimers and Gurevych, 2019)). The results on two biomedical benchmark corpora show the high effectiveness of the BioSent2Vec model, which we utilize to retrieve the tailored task-instance examples in SENSUPPORT.

We evaluate the sentence-based approaches SENBASE/SENSUPPORT and the abstract-based BASELINE of Nye et al. (2018) by comparing their collected MTurk annotations to gold standard annotations. We find that the highest label quality is obtained with SENSUPPORT with Cohen’s Kappa agreements of 0.78/0.75/0.69 for P/I/O. We show further that annotations obtained via the sentence-based approaches lead to substantially higher Kappa agreements than annotations from the BASELINE approach, especially for the labeling of Interventions and Outcomes. The largest source of disagreement in the BASELINE approach is caused by crowdworkers overlooking entire text phrases that *should* be annotated – whereas, in the sentence-based approaches, crowdworkers tend to

label text phrases that *should not* be annotated.

The contributions of this paper are:

- We propose and evaluate two novel task designs for the collection of high-quality PIO annotations from crowdworkers.
- We evaluate nine unsupervised semantic short-text similarity (SSTS) methods based on two biomedical corpora to identify an effective method for SENSUPPORT. The obtained results are also useful to other researchers who work on related biomedical IR tasks, like ad-hoc search or question answering.

We discuss related work in Section 2. The PIO task designs are presented in Section 3. We evaluate the unsupervised SSTS methods in Section 4 and the PIO task designs in Section 5.

## 2 Related Work

### 2.1 PICO Annotation

The traditional PICO annotation task design was to collect coarse-grained annotations of whether a given sentence contains PICO or not (Demner-Fushman and Lin, 2007; Kim et al., 2011). Only recently, the trend has moved from coarse-grained binary annotation to a fine-grained *text span annotation*. The fine-grained annotation, however, is more difficult and makes the accurate annotation of PICO labels a challenging task (Nye et al., 2018; Zlabinger et al., 2018).

The core strategy of Nye et al. (2018) to obtain decent quality text span annotations from non-expert crowdworkers was to collect several redundant annotations, which were then aggregated to a meta-annotation of higher quality. As additional measure to reduce the task’s complexity, the Intervention and the Comparison were not differentiated by Nye et al. (2018), resulting in the PIO task. While these two measures lead to annotations of higher quality, it remained unclear whether a more effective task design could further improve the label quality of (i) individuals and (ii) aggregated annotations. In this paper, we investigate this research gap by performing a comparative evaluation of two novel PIO task designs and the task design of Nye et al. (2018).

### 2.2 Crowdsourcing Task Design

The *lack of domain specific experience* of crowdworkers has been primarily addressed by training

through examples. In a large-scale study of micro-tasks, it was shown that the availability of examples had a clear effect of reducing disagreement in collected annotations (Jain et al., 2017). Furthermore, Doroudi et al. (2016) show that training workers based on examples annotated by experts is highly effective compared to various other training strategies. Liu et al. (2016) propose an annotation task design called Gated Instructions to improve the quality of crowdworkers. In the Gated Instructions approach, annotators are trained by tutorials, feedback is provided throughout the annotation task, and the annotation process is continuously monitored. Singla et al. (2014) further advance the process of providing examples: For an image labeling task, a machine learning approach was utilized to dynamically select relevant examples from an expert-authored set based on the progress of each worker. SENSUPPORT is based on a similar principle, but adopts an unsupervised text similarity method to find relevant expert examples as opposed to creating an internal machine learner model.

Several studies have shown that task complexity (e.g., in terms of *task length distribution*) affects the performance of crowdsourcing: a task’s cognitive complexity was shown to affect both accuracy and completion time (Finnerty et al., 2013); breaking up large tasks into smaller tasks increased output quality and worker experience for the task types arithmetic, sorting and transcription (Cheng et al., 2015); experiments related to Named-Entity Recognition (NER) of tweets found that the length and number of entities in a tweet influenced the quality of the crowd-annotations: A better quality was obtained for shorter tweets with fewer entity mentions (Feyisetan et al., 2017). As best practice in corpus annotation (Sabou et al., 2014), it is advisable to keep the text that is annotated reasonably short, without compromising the context. Sentences provide sufficient context for most NLP tasks (except for tasks like long-distance anaphora discovery Poesio et al., 2013).

In this paper, we provide further insights into the field of biomedical data acquisition by conducting thorough experimentation for the annotation of Participants, Interventions, and Outcomes. The acquisition of labeled data in this specific domain is challenging since annotators require medical expertise to understand the jargon and terminology that prevails in the biomedical literature. Therefore, experimental findings reported in related studies but

in different domains can often not be generalized to the biomedical domain.

### 2.3 Unsupervised Short-Text Similarity

The two standard biomedical corpora for the evaluation of semantic short-text similarity (SSTS) methods are BIOSSES (Soğancı oğlu et al., 2017) and MedSTS (Wang et al., 2018). Studies conducted on these two corpora usually evaluate the effectiveness of *supervised* similarity methods (Antunes et al., 2020; Liu et al., 2019); however, we are interested in the effectiveness of *unsupervised* methods for SENSUPPORT. Although results for individual unsupervised methods are reported (e.g. Chen et al., 2019; Tawfik and Spruit, 2020), no comprehensive evaluation exists. We address this research gap and evaluate nine unsupervised methods based on BIOSSES and MedSTS.

## 3 Task Designs for PIO Annotation

In this section, we first describe the BASELINE task design of Nye et al. (2018), followed by our proposed task designs SENBASE and SENSUPPORT.

### 3.1 BASELINE

In the task design of Nye et al. (2018), the *entire abstract* of a clinical trial report is presented to annotators who are asked to label the PIO entities. The annotation of P, I, and O is conducted as three individual sub-tasks to reduce the cognitive overload needed to switch between the three labels. For each sub-task, annotation guidelines are crafted to prepare the workers. The guidelines consist of a few *static examples*, which illustrate how the task should be performed, and annotation instructions, which describe what text phrases should or should not be annotated as PIO.

### 3.2 SENBASE

The annotation of entire abstracts leads to an uneven distribution of task effort to complete individual task instances. We illustrate this problem in Table 2, where we compare the word counts of abstracts to sentences. The table shows that the annotation of sentences leads to a better distribution in task effort, indicated by the substantially lower std. dev. of 13 for sentences compared to abstracts.

Based on this analysis, we propose a new task design, SENBASE, in which we switch from abstract to sentence annotation. Specifically, we split each abstract into individual sentences, in which

P	Task Instance	<b>Thirty-nine subjects</b> completed the study and were included in the data analysis.
	Tailored Example	<b>Ninety-three subjects</b> were randomly assigned.
I	Task Instance	<b>QYJDR</b> is an effective formula in treatment of EMs-related infertility.
	Tailored Example	<b>Eltrombopag</b> is an oral thrombopoietin receptor agonist for the treatment of thrombocytopenia.
O	Task Instance	There were no serious <b>adverse events</b> .
	Tailored Example	<b>Adverse events</b> did not significantly differ in the 2 groups.

Table 1: Task-instances with tailored examples for [Participants](#), [Interventions](#), and [Outcomes](#). The **bold text spans** should be annotated by the crowdworkers.

annotators label the PIO entities – or mark a checkbox if no PIO entity could be identified. Similar to the BASELINE (i) the task is divided into three individual sub-tasks for PIO, and (ii) the annotators are trained with the same annotation guidelines, available as an appendix of [Nye et al. \(2018\)](#).

	# Words			
	Min.	Max.	Avg.	Stdev.
Abstract	57	562	268	89
Sentence	5	105	25	13

Table 2: Analysis of the word counts of abstracts versus sentences. We measure the word count based on tokenized text excluding punctuation. Data basis of this analysis is the EBM-NLP corpus, described in Sec. 5.1.

Although the annotation of sentences improves the distribution in task effort, sentences might appear out-of-context. This means that two consecutively annotated sentences could stem from two different abstracts. The inability to preserve a certain task instance order is typical for crowdsourcing platforms, since workers can usually (i) skip individual task instances and (ii) start/stop working on task instances arbitrarily. The lack of context can be problematic since the context is essential, e.g., to identify the meaning of an abbreviation that was defined in an earlier sentence. To address the lack of context in SENBASE, we give workers access to the entire abstract via an expandable window.

### 3.3 SENSUPPORT

This approach builds upon SENBASE by additionally addressing the lack of domain-specific expertise of crowdworkers. The common approach to train crowdworkers for difficult tasks is to provide a few examples that illustrate how the task should be performed. Providing examples is essential for a successful task-design ([Daniel et al., 2018](#)), however, examples are usually defined *statically* over an entire task and might not be helpful at individ-

ual task instances. To improve the effectiveness of examples, we propose the SENSUPPORT task design in which annotators are supported by *tailored task-instance examples*.

The tailored task-instance examples (see some examples in Table 1) are retrieved from a set of sentences that are already annotated by medical experts. Note that expert annotations are usually available since they are crucial to measure the performance of non-expert annotators ([Snow et al., 2008](#); [Doroudi et al., 2016](#)). We propose to split expert annotations into: (i) a test set that is used to measure the performance of non-expert annotators and (ii) a training set from which the tailored examples are retrieved. To identify an effective unsupervised sentence similarity method for the retrieval of task-instance examples in SENSUPPORT, we evaluate nine methods.

We note that the SENSUPPORT approach was first described in our preliminary study of [Zlabinger et al. \(2020\)](#). We extend our preliminary study in this paper as follows: First, we report baseline results for the case where no task-instance examples are shown (i.e. the SENBASE approach). Second, we perform additional experiments by analyzing the types of errors that annotators commonly make in each annotation approach. Finally, we conduct a comparative evaluation to identify an effective method for the retrieval of task-instance examples.

## 4 Evaluation of Similarity Methods

In this section, we evaluate the effectiveness of nine unsupervised semantic short-text similarity (SSTS) methods. Each method computes a similarity score  $\text{sim}(t, k) \in \mathbb{R}$  between two short-texts  $t, k$ . For methods that produce a vector representation of a text, we compute the similarity  $\text{sim}(t, k)$  between the vectors  $\mathbf{v}_t, \mathbf{v}_k \in \mathbb{R}^n$  using the cosine similarity.

*Word Count Based.* These methods compute the similarity between two texts based on their words in common. We evaluate TFIDF-weighted word

vectors and the Levenshtein Distance defined as the number of edits to transform the word sequence of a text into the other. (Manning et al., 2008)

**Aggregated Word Embeddings.** A word embedding  $e_w \in \mathbb{R}^n$  is an  $n$  dimensional vector representation of a word  $w$ . To obtain a text embedding  $v_t$  from individual word embeddings, an aggregation step is necessary. We evaluate three aggregation strategies: the average of the word embeddings (AVG), a TFIDF-weighted average (WAVG) (Le and Mikolov, 2014), and an aggregation via the Smooth Inverse Frequency (SIF) (Arora et al., 2017). In SIF, the relative word frequency is used to obtain aggregated text representations from which the second principal component is removed.

**Text Embeddings.** Methods from this category infer a contextualized text embedding  $v_t$  for an input text  $t$ . We evaluate: Sentence BERT (SenBERT) (Reimers and Gurevych, 2019), which uses a siamese network to create BERT-based text representations; the transformer-based Universal Sentence Encoder (USE) (Cer et al., 2018); InferSent (Conneau et al., 2017), which uses word embeddings and a combination of LSTMs and hierarchical CNNs to produce universal sentence representations; and finally, Sent2Vec (Pagliardini et al., 2018), which computes sentence representations by combining the Continuous Bag of Words Model (CBOW) and character n-gram embeddings.

#### 4.1 Experiment Setup

We compare the similarity methods on two biomedical sentence-to-sentence similarity corpora.

**BIOSSES** (Soğancı oğlu et al., 2017): This corpus contains 100 sentence pairs with labeled similarity scores from 0 (no relation) to 4 (high relation). The sentences are sampled from biomedical research papers.

**MedSTS** (Wang et al., 2018): This corpus contains 1,068 sentence pairs annotated from 0 (no relation) to 5 (high relation). The sentences are sampled from anonymized electronic health records of patients of the Mayo Clinic.

We evaluate the methods by computing the Pearson correlation coefficient between the ground truth labels and the score computed by the unsupervised methods. The Pearson correlation is the standard metric reported for these two corpora. The evaluation is conducted on all samples of each corpus, as

opposed to a training/test split which is not needed for the evaluation of unsupervised methods.

The methods based on word or text embeddings require a language model trained on large amounts of text data. The pretrained models used in our experiments are summarized in Table 3. All described models are freely available and more details on the models can be found in the referenced papers including download links, hyper-parameter settings, and descriptions of the text corpora used for training. Note that we preferably select models that are pretrained on biomedical data. For the universal methods USE and InferSent, there is no specific biomedical model available. In the appendix of this paper, we describe and evaluate additional pretrained models that are not presented in the paper due to (i) the page limitation and (ii) the inferior effectiveness compared to the presented models.

We differentiate between three preprocessing functions (i) *Identity* where no preprocessing is conducted, (ii) *Lower* where text is lowercased, and (iii) *LowerStop* where text is lowercased and stopwords are removed. We use the English stopword list of the NLTK Python library<sup>3</sup>. For the tokenization needed for the methods Levenshtein, TFIDF, AVG, WAVG, and SIF, we use the *word\_tokenize* function of the NLTK library. Finally, the hyperparameter  $a$  of the SIF method is set to  $10^{-3}$ , as suggested in Arora et al. (2017).

#### 4.2 Evaluation of Effectiveness

The evaluation results in Table 4 show the high effectiveness of methods that use the BioSent2Vec or BioWord2Vec model. The common denominator of these models is the pretraining on biomedical research papers (i.e. PubMed) and clinical notes (i.e. the MIMIC III corpus), which is similar to the underlying data source of BIOSSES and MedSTS.

Apart from the pretraining, the method also has a substantial impact on the obtained results. The SenBERT method, although pretrained on biomedical publications, is rather ineffective, even outperformed by TFIDF-weighted word vectors. Similarly ineffective are the universal methods USE and InferSent. These findings align with other studies that report that transformer-based text representations are highly effective as input for supervised learning, but less effective in an unsupervised setting (Reimers and Gurevych, 2019; Tawfik and Spruit, 2020).

<sup>3</sup><https://www.nltk.org/> (version 3.5)

Embedding	Model	Training Data	Used by Method
Word	BioWord2Vec (Zhang et al., 2019)	PubMed abstracts, MIMIC III corpus (Johnson et al., 2016)	AVG, WAVG, SIF
Text	BioBERT (Lee et al., 2019)	PubMed abstracts	SenBERT
	BioSent2Vec (Chen et al., 2019)	PubMed abstracts, MIMIC III corpus (Johnson et al., 2016)	Sent2Vec
	USE 4.0 (Cer et al., 2018)	Wikipedia, web news, online forums, SNLI corpus (Bowman et al., 2015)	USE
	InferSent 2.0 (Conneau et al., 2017)	SNLI corpus (Bowman et al., 2015)	InferSent

Table 3: Overview of the pretrained models utilized in our evaluation.

Category	Method	Model	Preprocessing	MedSTS	BIOSSES	Avg.
Word count	TFIDF	-	Lower	<u>0.74</u>	0.70	0.72
	TFIDF	-	LowerStop	<u>0.74</u>	<u>0.73</u>	<u>0.74</u>
	Levenstein	-	Lower	0.55	0.64	0.60
	Levenstein	-	LowerStop	0.64	0.69	0.66
Word embedding	AVG	BioWord2Vec 2019	Lower	0.61	0.72	0.66
	AVG		LowerStop	0.72	<b>0.77</b>	0.75
	WAVG		Lower	0.73	0.75	0.74
	WAVG		LowerStop	0.76	<b>0.77</b>	0.76
	SIF		Lower	0.79	0.75	<u>0.77</u>
	SIF		LowerStop	0.78	0.76	<u>0.77</u>
Text embedding	SenBERT	BioBERT 2019	Identity	0.78	0.58	0.68
	USE	USE 4.0 2018	Identity	0.66	0.72	0.69
	InferSent	InferSent 2.0 2017	Identity	0.49	0.65	0.57
	Sent2Vec	BioSent2Vec 2019	Lower	<b>0.81</b>	0.74	0.78
	Sent2Vec	BioSent2Vec 2019	LowerStop	<b>0.81</b>	<b>0.77</b>	<b>0.79</b>

Table 4: Pearson correlation between the ground truth labels and the unsupervised semantic similarity methods. For each corpus, we highlight the overall best result **bold** and the best result per category by underline.

The effect of preprocessing shows that stopword removal is usually beneficial, especially for Levenshtein and AVG since these two methods do not have an incorporated mechanism for weighting word importance. Notice that we did not report exhaustive preprocessing results for all methods since certain methods expect (i) a specific preprocessing to be effective (e.g. lowercasing for TFIDF), or (ii) the raw unprocessed text as input, as it is the case for SenBERT, USE, and InferSent.

Based on the conducted evaluation of unsupervised similarity methods, we use the BioSent2Vec model with lowercasing and stopword removal for the retrieval of similar examples in SENSUPPORT.

## 5 Experiments on PIO Task Designs

In this section, we describe our experimental evaluation of BASELINE, SENBASE, and SENSUPPORT.

### 5.1 Experiment Setup

As data source for our experiments, we use the EBM-NLP corpus (Nye et al., 2018) consisting of 191 clinical trial report abstracts. Each clinical

report is annotated by three medical experts. We aggregate the three expert labels by a majority vote to derive a final gold standard label.

We divide the 191 trial reports into a test set consisting of 41 abstracts and a training set consisting of 150 abstracts. To split the abstracts into sentences for SENBASE and SENSUPPORT, we use the CoreNLP library (Manning et al., 2014), resulting in a total of 423 sentences for the test set and 1,636 sentences for the training set.

The sentences of the training set are used for the retrieval of tailored task-instance examples for the SENSUPPORT approach. We retrieve the top-3 most similar sentences for each sentence in the test set and show them as tailored task-instance examples to the crowdworkers.

The samples in the test set are used to compare the three annotation approaches. The annotations for the BASELINE approach are downloaded from <https://github.com/bepnye/EBM-NLP>, which were published in the scope of Nye et al. (2018). The annotations for SENBASE and SENSUPPORT are specifically collected for this

Design	#Workers	#Redundant	HIT
BASELINE	403	8 - 17	abstract
SENBASE	38	3	sentence
SENSUPPORT	31	3	sentence

Table 5: Overview of the compared annotation sets.

study. Therefore, we implement both approaches<sup>4</sup> and follow the same annotation setup of Nye et al. (2018), namely: annotations are collected from the MTurk platform; workers require a minimum approval rate of 90% on previous tasks to participate; spammers are removed in a small-scale test run; and finally, the payment per HIT is set to \$0.06 per sentence (which we reduced from \$0.30 to reflect the reduced effort needed to complete a HIT).

An overview of the three annotation sets is given in Table 5. For SENBASE and SENSUPPORT, we collect 3 redundant annotations per sentence, resulting in  $423 \times 3 = 1,269$  HITs in each PIO sub-task. In the BASELINE, more unique workers contributed compared to SENBASE and SENSUPPORT because more redundant annotations of 8-17 (average 11 and std. dev. 1.7) were collected.

## 5.2 Agreement of Individual Crowdworkers

We measure the label quality between individual crowdworkers and the gold standard annotations by computing the inter-annotator agreement in terms of Cohen’s Kappa (McHugh, 2012), a standard metric for the label quality in annotation projects. The results in Figure 1 show a clear improvement of Kappa scores of the sentence-based task designs, compared to the abstract-based task design BASELINE. Substantially higher agreements are reached for the sub-tasks Interventions and Outcomes. Notable is the outlier of the SENSUPPORT approach for the annotation of Interventions, denoted by a dot. This one worker reached a distinctly lower agreement to the gold standard than the other workers of the SENSUPPORT approach.

The results of SENBASE compared to SENSUPPORT show that the utilization of tailored task instance examples further increases the Kappa agreement, especially for the annotation of Interventions. This additional improvement was obtained at no additional costs since we pay \$0.06 per HIT in both sentence-based approaches.

The analysis of individual workers has the dis-

<sup>4</sup>The annotation interfaces are illustrated in the appendix.

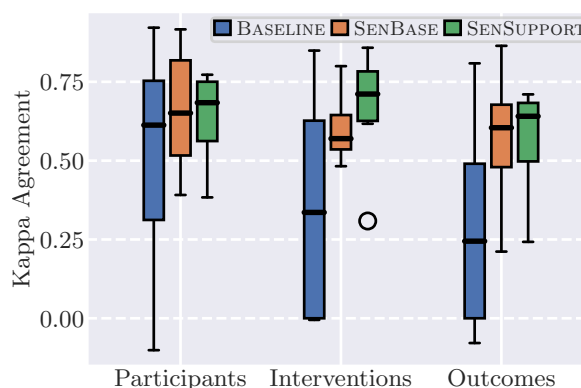


Figure 1: Kappa agreements between annotations from individual workers and the gold standard.

advantage that workers who labeled only a few task-instances are less reliable than workers who labeled several task-instances. We addressed this problem by limiting the presented analysis to workers who labeled at least 5% of the test set. All workers are considered in the analysis of aggregated annotations, described next.

## 5.3 Agreement of Aggregated Annotations

Here, we analyze the label quality of meta-annotations that are aggregated from multiple redundant annotations. We consider two aggregation methods: (i) majority voting (MV) where individual workers are weighted equally and (ii) Dawid-Skene<sup>5</sup> (DS) where the reliability of individual workers is automatically computed and used for a weighted aggregation (Dawid and Skene, 1979).

We measure the quality of aggregated annotations by computing the Kappa agreement to the gold standard annotations. We compute the aggregations for the sentence-based approaches based on the 3 available redundant annotations. Since there are 8-17 redundant annotations available for the BASELINE approach, we (i) select 3 random annotations, (ii) aggregate them, and (iii) compute the agreement to the gold standard. Since the random selection in (i) can be affected by a lucky/unlucky seed, we repeat (i-iii) 20 times and compute a robust final agreement by averaging the 20 individual Kappa scores.

The results for 3 aggregated annotations show that the highest agreements to the gold standard are reached by the SENSUPPORT approach, followed by SENBASE (Table 6). Especially, for Interventions and Outcomes, the sentence-based ap-

<sup>5</sup>We use the implementation from [https://github.com/dallascard/dawid\\_skene](https://github.com/dallascard/dawid_skene)

proaches significantly outperform the BASELINE approach. Note that aggregation via DS is only effective for the BASELINE annotations. This is expected since weighted aggregation methods rely on a certain noise level of the underlying annotations, which was high in the BASELINE (see Figure 1).

The results when all 8-17 annotations of the BASELINE approach are aggregated are indicated by  $MV_{ALL}$  and  $DS_{ALL}$  in Table 6. As expected, the Kappa agreements substantially improve compared to the aggregation of only 3 annotations. However, for I and O, 8-17 aggregated annotations still reach substantially lower agreements to the gold standard than only 3 aggregated annotations of the SENSUPPORT approach, caused by the low quality of the underlying annotations (Figure 1). Only for P, the  $DS_{ALL}$  agreement of 0.867 significantly improves over all other aggregations of  $MV_3$  or  $DS_3$ . We investigated this result and found that DS picks up the signal from two workers of the BASELINE approach who reach exceptionally high agreements to the gold standard of 0.83 and 0.84 – and who both annotated a majority of the abstracts in the test set (34/41 and 41/41).

	Cohen’s Kappa ( $\kappa$ )		
	P	I	O
BASELINE $_{MV3}$	0.702	0.455	0.352
SENBASE $_{MV3}$	0.715	0.675 <sup>a</sup>	0.655 <sup>a</sup>
SENSUPPORT $_{MV3}$	0.780 <sup>ab</sup>	<b>0.757<sup>ab</sup></b>	<b>0.694<sup>ab</sup></b>
BASELINE $_{DS3}$	0.729	0.579	0.458
SENBASE $_{DS3}$	0.726	0.674 <sup>a</sup>	0.654 <sup>a</sup>
SENSUPPORT $_{DS3}$	0.776 <sup>a</sup>	0.756 <sup>ab</sup>	<b>0.694<sup>ab</sup></b>
BASELINE $_{MV_{ALL}}$	0.760	0.476	0.343
BASELINE $_{DS_{ALL}}$	<b>0.867</b>	0.633	0.677

Table 6: Kappa agreements between aggregated annotations of each approach and the gold standard. We show significant improvements for both categories  $MV_3$  and  $DS_3$  where *a* refers to BASELINE and *b* to SENBASE (two-sided, paired t-test:  $p < 0.05$ ).

#### 5.4 Analysis of Agreement Types

We switch from analyzing Kappa agreements to analyzing which types of agreement appear between the gold standard and the non-expert annotators. The analysis of agreement types gives additional insights in the labeling behavior of annotators (Lee and Sun, 2019). We differentiate between four agreement types, summarized in Table 7. We differentiate between cases where a text-span annotation of a crowdworker and the gold standard (i)

*disagree* entirely (Miss and Redundant) or (ii) they *agree* exactly (Exact) or at least partially (Partial)




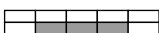
Type	Example
Exact	
Partial	
Miss	
Redundant	

Table 7: Overview of the differentiated agreement types. The examples show annotations between crowdworkers (gray) and the gold standard (yellow).

The analysis results in Figure 2a show a substantial difference between the types Miss and Redundant, when comparing the sentence-based approaches to the abstract-based approach BASELINE. In the BASELINE approach, we see a high frequency of Miss and fewer cases of Redundant in all PIO sub-tasks. On the other hand, in SENBASE and SENSUPPORT, we see a high frequency of Redundant cases and much fewer cases of Miss. This result shows that (i) crowdworkers who annotate entire abstracts frequently overlook text phrases that *should* be annotated and (ii) crowdworkers who annotate sentences tend to label text phrases that *should not* be annotated.

The analysis results in Figure 2b shows how often annotators exactly or at least partially agree with the gold standard annotations. We find, aligned with our previous results, that SENSUPPORT is the most effective approach, followed by SENBASE and BASELINE. The frequency of Exact cases is constantly higher in the sentence-based approaches compared to the BASELINE, especially for I and O. This shows that crowdworkers of the sentence-based approaches are more likely to fully agree with the gold standard than crowdworkers of the BASELINE approach.

## 6 Conclusion & Future Work

We presented two novel task designs for crowdsourcing PIO annotations in clinical trial report abstracts. Specifically, we propose to switch from annotating entire abstracts of clinical trial reports to the annotation of sentences (SENBASE), and to additionally support non-expert annotators with tailored task-instance examples (SENSUPPORT).

The task-instance examples were retrieved from a set of expert annotations using the BioSent2Vec



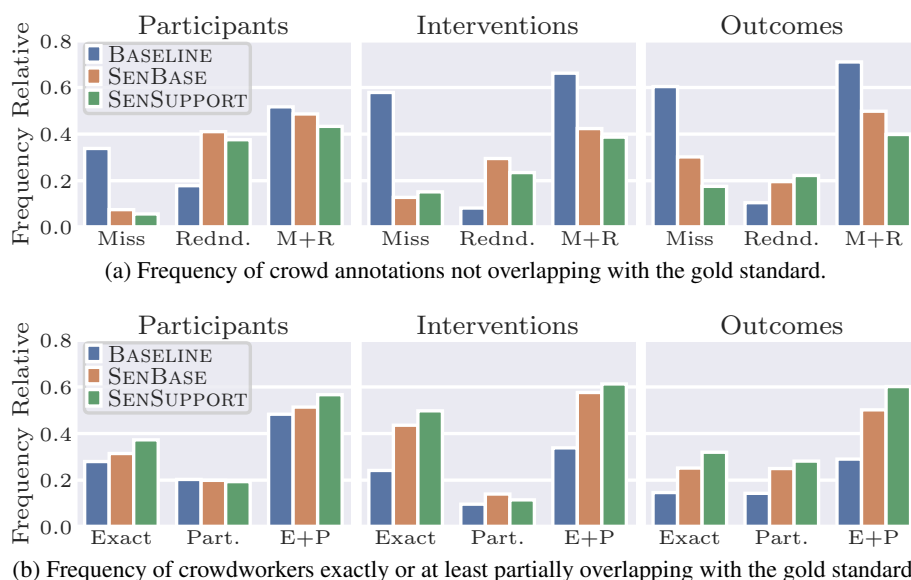


Figure 2: Relative frequency of the different agreement types between crowdworkers and the gold standard annotations. The combined result of Miss+Redundant and Exact+Partial is referred to as M+R and E+P respectively.

model, which was found to be effective by an empirical evaluation of nine unsupervised SSTS methods based on two biomedical corpora.

We evaluated the sentence-based annotation approaches SENBASE and SENSUPPORT, and the abstract-based approach BASELINE by comparing crowd-annotations of each approach to a set of gold standard annotations. We found that the highest Kappa agreement to the gold standard is reached by annotations of the SENSUPPORT approach. Therefore, whenever expert annotations can be spared, they should be utilized as tailored task-instance examples. Furthermore, we showed that annotations from the sentence-based approaches SENBASE/SENSUPPORT substantially outperform annotations from the BASELINE approach, especially for Interventions and Outcomes.

Finally, we conducted a pairwise comparison of the token overlap of annotations of either approach with the gold standard and find that crowdworkers using the sentence-based approaches are prone to annotate text phrases that *should not* be annotated, whereas workers using the abstract-based approach are prone to overlook text phrases that *should* be annotated.

The core limitation of the SENSUPPORT approach is the availability of reference samples from which the tailored task-instance examples are retrieved. Obtaining reference samples is usually expensive since expert annotators need to be employed. Therefore, in future work, we aim to identify the minimum number of reference samples that

is needed to still preserve a high annotation quality of crowdworkers. More specifically, we aim to combine the selection of reference samples with an active learning approach. By applying active learning, the informativeness of samples can be computed and used as a deciding factor in the selection of a few reference samples that are effective as task-instance examples.

Another promising future research direction is the application of task-simplification approaches for different tasks and domains. In this study, we showed that a simple shift from annotating sentences rather than entire abstracts can significantly increase the annotation quality obtained from crowdworkers. Similar improvements might be possible in other challenging tasks/domains, such as the legal or patent domain.

The code of our experiments on the various unsupervised SSTS methods, including the implementation of each method and the computed scores, is available at <https://github.com/Markus-Zlabinger/ssts>. The collected annotations of SENBASE and SENSUPPORT are available at <https://github.com/Markus-Zlabinger/pico-annotation>.

## Acknowledgment

Marta Sabou was funded by the HOnEst Austrian Science Fund (FWF) project:V 745-N.

## References

- Rui Antunes, João Figueira Silva, and Sérgio Matos. 2020. Evaluating semantic textual similarity in clinical sentences using deep learning and sentence embeddings. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing, SAC '20*, pages 662–669, Brno, Czech Republic. Association for Computing Machinery.
- Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of ICLR*.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.
- Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. [Universal Sentence Encoder](#). *arXiv:1803.11175 [cs]*.
- Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. BioSentVec: Creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5. IEEE.
- Justin Cheng, Jaime Teevan, Shamsi T. Iqbal, and Michael S. Bernstein. 2015. Break It Down: A Comparison of Macro- and Microtasks. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*.
- Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.
- Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Computing Surveys (CSUR)*, 51(1):7:1–7:40.
- Alexander Philip Dawid and Allan M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.
- Dina Demner-Fushman and Jimmy Lin. 2007. Answering Clinical Questions with Knowledge-Based and Statistical Techniques. *Computational Linguistics*, 33(1):63–103.
- Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. 2016. [Toward a Learning Science for Complex Crowdsourcing Tasks](#). In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16*, page 2623–2634, New York, NY, USA. Association for Computing Machinery.
- Oluwaseyi Feyisetan, Elena Simperl, Markus Luczak-Roesch, Ramine Tinati, and Nigel Shadbolt. 2017. [An extended study of content and crowdsourcing-related performance factors in named entity annotation](#). *Semantic Web*, 9(3):355–379.
- Ailbhe Finnerty, Pavel Kucherbaev, Stefano Tranzillini, and Gregorio Convertino. 2013. [Keep it simple: Reward and task design in crowdsourcing](#). In *ACM International Conference Proceeding Series*, pages 1–4, New York, New York, USA. Association for Computing Machinery.
- Xiaoli Huang, Jimmy Lin, and Dina Demner-Fushman. 2006. Evaluation of PICO as a Knowledge Representation for Clinical Questions. *AMIA Annual Symposium Proceedings*, 2006:359–363.
- Ayush Jain, Akash Das Sarma, Aditya Parameswaran, and Jennifer Widom. 2017. [Understanding workers, developing effective tasks, and enhancing marketplace dynamics: A study of a large crowdsourcing marketplace](#). In *Proceedings of the VLDB Endowment*, volume 10, pages 829–840.
- Alistair EW Johnson, Tom J. Pollard, Lu Shen, H. Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.
- Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support Evidence Based Medicine. *BMC Bioinformatics*, 12(2):S5.
- Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, ICML'14*, pages II–1188–II–1196, Beijing, China. JMLR.org.
- Grace E. Lee and Aixin Sun. 2019. A Study on Agreement in PICO Span Annotations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'19*, pages 1149–1152, Paris, France. ACM Press.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. [BioBERT: A pre-trained biomedical language representation model for biomedical text mining](#). *Bioinformatics*.

- Angli Liu, Stephen Soderland, Jonathan Bragg, Christopher H. Lin, Xiao Ling, and Daniel S. Weld. 2016. Effective Crowd Annotation for Relation Extraction. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 897–906, San Diego, California. Association for Computational Linguistics.
- Sijia Liu, Yanshan Wang, and Hongfang Liu. 2019. Selected articles from the BioCreative/OHNLN challenge 2018. *BMC Medical Informatics and Decision Making*, 19(10):262.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge university press.
- Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.
- Mary L. McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282.
- Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.
- Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*.
- Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. [Phrase Detectives: Utilizing Collective Intelligence for Internet-Scale Language Resource Creation](#). *ACM Trans. Interact. Intell. Syst.*, 3(1).
- Nils Reimers and Iryna Gurevych. 2019. SentenceBERT: Sentence Embeddings using Siamese BERT-Networks. 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 3980–3990, Hong Kong, China. Association for Computational Linguistics.
- Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pages 859–866.
- Adish Singla, Ilija Bogunovic, Gábor Bartók, Amin Karbasi, and Andreas Krause. 2014. Near-optimally teaching the crowd to classify. In *Proceedings of the 31st International Conference on Machine Learning - Volume 32, ICML'14*, page II–154–II–162. JMLR.org.
- Lea Škorić, Anton Glasnović, and Jelka Petrak. 2020. A publishing pandemic during the COVID-19 pandemic: How challenging can it become? *Croatian Medical Journal*, 61(2):79–81.
- Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics.
- Gizem Soğancı oğlu, Hakime Öztürk, and Arzucan Özgür. 2017. BIOSSES: A semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58.
- Noha S. Tawfik and Marco R. Spruit. 2020. Evaluating sentence representations for biomedical text: Methods and experimental results. *Journal of Biomedical Informatics*, 104:103396.
- Byron C. Wallace. 2018. Automating Biomedical Evidence Synthesis: Recent Work and Directions Forward. In *BIRNDL@ SIGIR*, pages 6–9.
- Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang D. Liu. 2018. MedSTS: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*.
- Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, 6(1):1–9.
- Markus Zlabinger, Linda Andersson, Allan Hanbury, Michael Andersson, Vanessa Quasnik, and Jon Brassey. 2018. Medical Entity Corpus with PICO elements and Sentiment Analysis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*.
- Markus Zlabinger, Marta Sabou, Sebastian Hofstätter, Mete Sertkan, and Allan Hanbury. 2020. DEXA: Supporting Non-Expert Annotators with Dynamic Examples from Experts. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20*, pages 2109–2112, New York, NY, USA. Association for Computing Machinery.