Sent2Span: Span Detection for PICO Extraction in the Biomedical Text without Span Annotations

Shifeng Liu[1], Yifang Sun[2], Bing Li[3], Wei Wang[4], Florence T. Bourgeois[5,6], Adam G. Dunn[5,1]

[1]The University of Sydney, Australia [2]Northeastern University, China [3]IHPC, A*STAR, Singapore

[4]Hong Kong University of Science and Technology (Guangzhou), China

[5]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, United States

[6]Department of Pediatrics, Harvard Medical School, Boston, MA, United States

`{shifeng.liu,adam.dunn}@sydney.edu.au, sunyifang@cse.neu.edu.cn, weiwcs@ust.hk`

`li_bing@ihpc.a-star.edu.sg, florence.bourgeois@childrens.harvard.edu`

Abstract

The rapid growth in published clinical trials makes it difficult to maintain up-to-date systematic reviews, which require finding all relevant trials. This leads to policy and practice decisions based on out-of-date, incomplete, and biased subsets of available clinical evidence. Extracting and then normalising Population, Intervention, Comparator, and Outcome (PICO) information from clinical trial articles may be an effective way to automatically assign trials to systematic reviews and avoid searching and screening—the two most time-consuming systematic review processes. We propose and test a novel approach to PICO span detection. The major difference between our proposed method and previous approaches comes from detecting spans without needing annotated span data and using only crowdsourced sentence-level annotations. Experiments on two datasets show that PICO span detection results achieve much higher results for recall when compared to fully supervised methods with PICO sentence detection at least as good as human annotations. By removing the reliance on expert annotations for span detection, this work could be used in a human-machine pipeline for turning low-quality, crowdsourced, and sentence-level PICO annotations into structured information that can be used to quickly assign trials to relevant systematic reviews.

1 Introduction

Systematic reviews are a critical part of regulatory and clinical decision-making because they are designed to robustly make sense of all available evidence from primary research, especially clinical trials, accounting for study design quality and heterogeneity. Searching and screening for reports of clinical trials are time-consuming tasks that require specialised expertise but are a necessary component of systematic reviews. The rapid rate at which papers were published about COVID-19 (Wang and Lo, 2021) highlights the need for tools to improve the efficiency of systematic reviews.

A range of methods have been developed to help reduce the amount of human effort required to conduct systematic reviews (Tsafnat et al., 2014; Marshall and Wallace, 2019), but the methods developed to support the screening task are trained on data from a small number of systematic reviews and are not yet able to fully replace humans (O'Mara-Eves et al., 2015).

An alternative is to find ways to map all clinical trials to standardised representations of the populations, interventions, comparators, and outcomes (PICO) and aggregate information across studies that answer equivalent clinical questions. PICO extraction is a well-studied problem structure and was used as one of the examples in the development of SciBERT (Beltagy et al., 2019).

Many PICO extraction methods focus on annotating sentences that include PICO information. However, if the goal is to fully automate a process for augmenting a systematic review with new studies as their results become available, even a perfect annotation of just the sentence is not enough. An expert will still need to read the sentence and then extract and normalise the information representing the population, intervention, comparator, and primary outcomes (PICO) from those sentences. Full automation of the task requires the ability to identify the text spans that represent the PICO information.

Named entity recognition (NER) seeks to identify the types and boundaries of targeted spans in unstructured text data. Machine learning NER methods have been based on BiLSTM-CRF (Lample et al., 2016) and BERT (Devlin et al., 2019), and these require human annotated data for training. For other application domains where span annotation is designed for entities such as people, locations, and organisations, crowdsourcing of labels is somewhat easier because a broader range

1705

of people can annotate data without specialised training. Even when annotated by domain experts, there is still substantial inconsistency across annotators (Lee and Sun, 2019), though sentence level annotations tend to be more consistent (Zlabinger et al., 2020).

There is a gap in both the volume of available training data and demonstrated performance between NER in general domains such as news and biomedical applications compared to PICO extraction. There is a clear need for new approaches that can handle the more complicated and challenging token structure of biomedical entities including unusual synonyms and subordinate clauses that might include numerical information (*e.g.* drug dosage, or test result threshold); and can incorporate domain-specific knowledge in intelligent and useful ways.

In this paper, we address these challenges by proposing a novel PICO annotation task design. We propose a simplified requirement for annotation where we only need to know whether a sentence includes any PICO information. We use this coarser set of annotations to learn and infer PICO sentence types and span detections.

The pre-trained neural language model (Devlin et al., 2019; Lee et al., 2020; Peng et al., 2019) is firstly used as feature representation learning model for sentence classification to be fine-tuned, and identify whether a given sentence contains PICO spans. This makes the proposed approach also capable of inferring PICO sentences even without crowdsourced annotations.

To get the PICO spans, we apply a masked span prediction task to assist the inference process. The fine-tuned language model is then used as the task-specific knowledge provider for the PICO spans. Scored spans are then fed into an inference algorithm to produce the final detection results.

The contributions of this paper are as follows:

- We propose a span detection approach for PICO extraction that uses only low-quality, crowd-sourced, sentence-level annotations as inputs, which reduces the need for time-consuming annotations from experts.

- We evaluate a novel structure for identifying candidate PICO sentences and masked span inference together. The masked span inference task replaces input spans with pre-defined mask tokens and the language model is used to infer which spans contribute most to the PICO sentence classification results.

- We demonstrate results that substantially improve on recall in span detection to align with the use case in the systematic review application domain on two benchmark datasets.

2 Related Work

2.1 Neural Language Models

Pre-trained deep neural language models, such as ELMo (Peters et al., 2018), GPT (Radford et al., 2018), BERT (Lee et al., 2020) and its variants (Liu et al., 2019; Lan et al., 2020), have brought significant performance improvements in a wide range of NLP tasks, such as relation extraction (Alt et al., 2019), entity resolution (Li et al., 2021), and question answering (Lee et al., 2020). These language models generally benefit from large scale text corpora. A major advantage of these methods come from the way long dependency token relations are captured to produce contextualised representations.

To adapt language models for use in the biomedical domain, researchers took pre-trained language models and re-trained them with domain-specific corpora, including PubMed abstracts and full text articles. These were then applied to a diverse range of NLP tasks, such as named entity recognition (Lee et al., 2020; Peng et al., 2019) and document classification (Peng et al., 2019). In this paper, we make use of pre-trained neural language models as the backbone model to learn task-oriented information.

2.2 PICO Annotation and PICO Extraction

There are multiple uses cases associated with representing clinical trial reports (including article abstracts, registrations, and protocols) by the participant inclusion criteria including condition, the interventions used in each of the study arms, and the set of primary outcomes measured during the trial. Like named entity recognition (NER), PICO detection aims to identify certain spans in the text corresponding to the each of the categories: population, interventions, and outcomes. While is it possible to apply general NER methods for the PICO extraction task (Nguyen et al., 2017; Nye et al., 2018; Kang et al., 2021), there are several key differences that make general NER methods less effective. These differences include spans that often do not have distinguishing features such as capitalised tokens and PICO elements that are not limited to noun phrases.

Most PICO extraction methods are fully supervised and need annotated data, which requires expertise and can be time consuming. To acquire enough PICO annotations for training, researchers have developed methods that use crowdsourcing as an alternative (Nguyen et al., 2017; Nye et al., 2018). This results in low-quality annotations, especially for the boundaries of the spans. To improve the annotation quality, Zlabinger et al. (2020) simplified the annotation task from document level to sentence level and additionally guided workers with similar sentences that had already been annotated using an unsupervised semantic short-text similarity method. In this paper, we instead only require sentence-level crowdsourced annotations without a boundary information developing novel method to predict PICO spans without training data.

2.3 Span based Methods

Compared with token based methods such as BiLSTM-CRF (Lample et al., 2016; Ma and Hovy, 2016), span based methods treat the spans (*i.e.* consecutive tokens), as the targets. In one stream of research advances in span based methods, the aim is to extract the hidden representations of each token with the raw token sequence as input, then either use boundary token representations (Ouchi et al., 2018; Ebner et al., 2020) or aggregate all the token representations (Liu et al., 2020) as the span representation. All possible spans are enumerated, classified, and decoded. An alternative stream of span-based research aims to mask a span in the token sequence and recover the masked tokens with hidden representations (Joshi et al., 2020). Both research streams use supervised or self-supervised methods and high-quality annotations as training data. In this paper, we instead extract the information stored in the span using only sentence-level annotations derived from crowdsourcing.

3 Method

In this section, we define the task and then introduce the proposed approach with BLUE (Peng et al., 2019), a BERT (Devlin et al., 2019) structured neural language model in the biomedical domain, as our backbone model and the inference algorithm.

3.1 Task Definition

Generally, to construct a dataset for PICO span detection, the annotators are presented with the full text, *i.e.*, the entire abstract of a clinical trial report (Nguyen et al., 2017; Nye et al., 2018). To improve the quality of the annotation, Zlabinger et al. (2020) proposed a novel annotation task, asking annotators to annotate sentences instead of abstracts with retrieved expert annotated sentences as examples based on sentence similarity methods. Both annotation tasks require annotators to locate the boundaries of PICO spans.

However, getting agreed boundaries for PICO span annotation is challenging, especially for crowdsourcing annotators in the biomedical domain. To fuse the gap between PICO sentence prediction and PICO span detection, we formalise the annotated dataset and the task as follows. We represent the dataset as D with $|D|$ sentences and a sentence from the dataset as s with sentence annotations from $|C|$ annotators. We use $\langle i, j \rangle$ to denote the boundaries of a PICO span starting from the i-th token (inclusive) and ending before the j-th token (exclusive). The task is then to train a model M that is able to yield a PICO span $\langle i, j \rangle$ with the probability $P(i, j | M, s, C)$.

3.2 Sentence Classification Training

The pre-trained neural language model BLUE is trained on domain specific text (*e.g.* PubMed abstracts). This injects the neural language model with biomedical knowledge.

From the annotators, we have gathered whether a sentence contains PICO annotations or not for each sentence. These annotations represent task specific information and can be used to "teach" the model. Thus, we form this process as a sentence classification task and fine-tune the pre-trained neural language model BLUE to predict whether a given sentence is a PICO sentence.

For each sentence, we feed it into the pre-trained BLUE model and collect the contextualised representations $\mathbf{h} \in \mathbb{R}^h$ of `[CLS]` token. With this representation $\mathbf{h}$, we predict the probability of a PICO sentence using:

$$score(y \mid \mathbf{h}) = W\mathbf{h} + \mathbf{b}, \qquad (1)$$
$$p(y \mid \mathbf{h}) = \text{Softmax}(score(y \mid \mathbf{h})), \quad (2)$$

where $W \in \mathbb{R}^{h*h}$ and $\mathbf{b} \in \mathbb{R}^h$ are learnable parameters trained during the fine-tuning process, and y is a function of C annotations $y =$
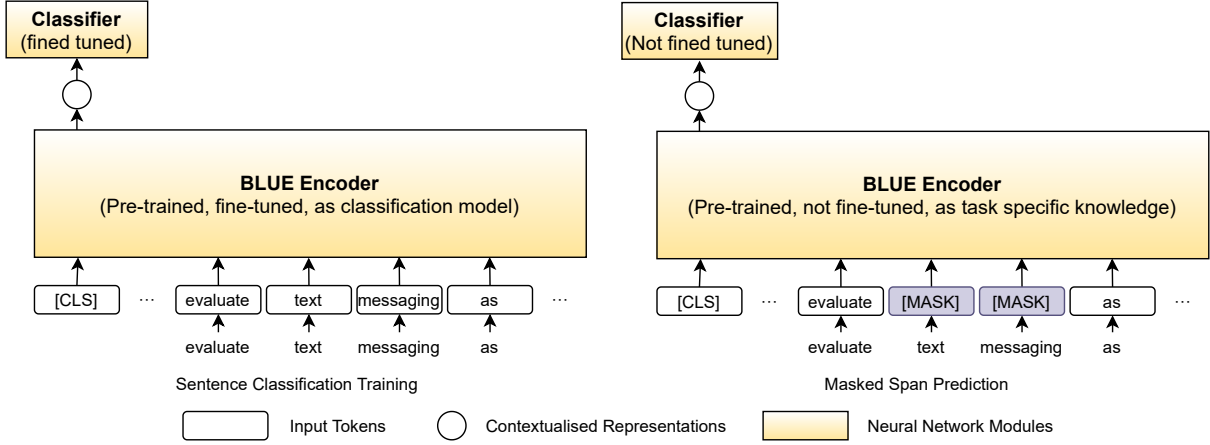
Figure 1: The proposed method has BLUE (Peng et al., 2019) as the backbone model, which is shared in the both the sentence classification (the model on the left) and masked span prediction (the model on the right). The BLUE encoder is firstly trained on the sentence prediction using the contextualised embedding of [CLS] token (the left part of the figure). Then the fine-tuned BLUE encoder predicts the score with span masked (the right part of the figure). The predicted scores of candidate spans along with the raw token sequence is collected for inference.

$f(C_1, \ldots, C_{|C|})$ (details of the functions are presented in Section 4.1). Compared with other state-of-the-art text classification methods (Huang et al., 2019; Zhang et al., 2020), our sentence classification method is relatively simple with just one extended module. The corresponding loss function we apply here is the cross entropy function:

$$\mathcal{L} = -\sum_{s \in D} y log(p(y \mid \mathbf{h})). \qquad (3)$$

3.3 Masked Span Prediction

Like how human beings locate PICO spans given known PICO sentences, we want the model to focus on the most indicative token spans in the sentences. A straightforward way is to directly use the last contextualised representations of the fine-tuned model to make predictions. However, such an approach is problematic because: first, the last contextualised representations contain information from the corresponding tokens and the surrounding tokens. Thus taking either the span boundary representations or inner span representations would introduce a faint amount of interference and lead to error-prone results. Second, these representations show the information in the token level rather than span level, which does not treat the span as a whole.

To address the aforementioned problems, we apply a Masked Span Prediction (MSP) task which is similar to the Masked Language Model training task applied in BERT (shown in Fig 1). In the MSP task, for span $\langle i, j \rangle$, we firstly mask the

original token sequence with [MASK] token. This masked token sequence is put into the fine-tuned neural language model to get the prediction score $score_{\langle i,j \rangle}$. The score $score_{\langle i,j \rangle}$ is then compared with the original score $score$ to infer the impact by the span $\langle i, j \rangle$, i.e., the contribution of the span in classifying the sentence as a PICO sentence. We define the contribution of a span $\langle i, j \rangle$ as $contribution_{\langle i,j \rangle} = score - score_{\langle i,j \rangle}$. The contribution $contribution_{\langle i,j \rangle}$ could be either positive or negative in classifying the sentence as a PICO sentence.

Given a sentence with N tokens, there are $O(N^2)$ candidate spans to be masked, which could generate an intractable number of spans. To reduce the number of candidate spans, we follow Ouchi et al. (2018) and Ebner et al. (2020) by limiting the number of tokens in a span up to M tokens. This would reduce the number of candidate spans to $\frac{M(2N-M+1)}{2}$, which is linear in the length of the sentence.

Algorithm 1: Nested Span Elimination

Input   : Eliminated Span Set RM, Span $\langle i, j \rangle$
Output : *Updated* Eliminated Span Set RM, Span Remove Flag F

1 $F \leftarrow False$;
2 **for** $p \leftarrow i + 1$ **to** $j - 1$ **do**
3   **if** $\langle i, p \rangle \in RM \land \langle p, j \rangle \in RM$ **then**
4          $RM \leftarrow RM \cup \langle i, j \rangle$;
5          $F \leftarrow True$;
6   break;
7   **end**
8 **end**
9 **return** RM, F

We observe that some spans with a negative contribution to PICO sentence classification have at least one nested span split where the split spans have negative contributions as well. To reduce the number of candidate spans for inference, we make use of this observation and apply with a bottom-up nested span elimination algorithm (Algorithm 1). Eliminated Span Set RM is initialised with single token spans of negative contributions to PICO sentence classification. For spans with more than two tokens, we search every split of the target span (Line 2). When the split spans are both in the existing eliminated span set RM (Line 3), we update RM with the target span (Line 4), claim this span can be eliminated (Line 5), and return the result. As we remove the spans based on model related set initialisation RM, we report the percentage of eliminated candidate span in Section 4.6.

3.4 Span Inference

Algorithm 2: Top-K Inference

Input	:Candidate Span Set U, the max select number K
Output	:Selective Span Set R

1 $R \leftarrow \emptyset$;
    /* descending sort U by contribution   */
2 $U \leftarrow sorted(U)$;
3 **for each** $span \in U$ **do**
    /* $noneoverlap(R, span)$: whether $span$ has overlapped tokens with any spans in the R */
4   **if** $noneoverlap(R, span)$ **then**
5         $R \leftarrow R \cup span$;
6   **if** $size(R) == K$ **then**
7   break;
8 **return** R

With all the reserved candidate spans and their contributions, we need to select spans that (1) have positive contributions to PICO sentence classification and (2) are not nested spans. Following previous work (Ouchi et al., 2018; Ebner et al., 2020), we establish a similar *argmax* inference method. However, different from the tasks in Ouchi et al. (2018) and Ebner et al. (2020), where each role is generally satisfied by exactly one span, there could be multiple spans appearing in one sentence in PICO span detection task. Thus we apply a top-K *argmax* inference method with pre-defined K for each PICO type. As shown in Algorithm 2, we firstly sort the candidate span set by span contribution (Line 2), and iteratively select spans (Line 3-

7). If the span does not overlap with selected spans (Line 4), we add it to the result set. The algorithm ends when we get K spans (Line 6) or traverse all the reserved candidate spans.

4 Experiments

In this section, we evaluate our method and compare it with supervised, semi-supervised methods, and crowdsourced annotations on two benchmark datasets. We also investigate the PICO sentence prediction results and the effect of candidate span reduction in the proposed method.

4.1 Dataset

We use two benchmark datasets for the PICO span detection task: EBM-NLP (Nye et al., 2018) dataset[1][2] and PICO-data (Nguyen et al., 2017) dataset[3]. The dataset statistics are shown in Table 1. All datasets are in English language.

The PICO dataset includes 3549, 500, and 191 abstracts for training, development, and test sets, respectively. This dataset is only annotated with Population (P) type.

Table 1: Dataset Statistic

Dataset	PICO Type	#Doc		
		Train	Dev	Test
PICO-data (Nguyen et al., 2017)	Population	3549	500	191
EBM-NLP (Nye et al., 2018)	Population	4282	500	188
	Intervention	4282	500	189
	Outcome	4170	500	190

The EBM-NLP dataset includes 4,993 PubMed abstracts annotated with Population (P), Intervention (I), and Outcome (O), respectively. Comparator (C) and Intervention (I) are combined together as Intervention (I). As each PICO type is annotated individually to avoid cognitive load, this makes the dataset contain three sub-datasets with a single PICO type. Without standard train-development split, we leave 500 abstracts from the training set as the development for each PICO type.

For all the training, development, and test sets in both datasets, they all collect the crowdsourced annotations, and the aggregated annotations. The test sets also have expert annotations from medical experts[4]. We additionally supply PICO sentence an-

---

[1]https://ebm-nlp.herokuapp.com/
[2]We use the latest version 2.0 in this paper.
[3]https://github.com/yinfeiy/PICO-data
[4]We note that some documents have no annotations and exclude them.

Table 2: Precision, Recall, and F_1 score for PICO Span Detection on the EBM-NLP dataset. Supervised and Semi-supervised methods are trained on the entire training set using the aggregated crowdsourced PICO span annotations. Weakly-supervised methods are trained on the entire training set *only* using PICO sentence annotations. Human annotation are the aggregated crowdsourced PICO span annotations on the test set. All methods are evaluated against expert span annotations on the test set. We highlight the recall as it is the most important metric when PICO detection is applied for systematic review processes. * indicates the results applying the crowdsourced sentence annotations.

Supervision Type	Method	Population			Intervention			Outcome		
		Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
Human Annotation	HMMCrowd(Nguyen et al., 2017)	0.72	0.76	0.70	0.64	0.80	0.68	0.50	0.81	0.59
Supervised	CRF(Lafferty et al., 2001)	0.55	0.51	0.53	0.65	0.21	0.32	0.83	0.17	0.29
Semi-Supervised	BiLSTM-CRF(Lample et al., 2016)	0.78	0.66	0.71	0.61	**0.70**	0.65	0.69	0.58	0.63
Weakly-Supervised	Sent2Span$_{agg}$	0.31	0.78	0.44	0.22	0.50	0.31	0.30	0.55	0.39
	Sent2Span$_{major}$	0.39	0.46	0.42	0.27	0.18	0.21	0.30	0.10	0.15
	Sent2Span$_{minor}$	0.30	**0.85**	0.45	0.23	0.51	0.31	0.27	**0.64**	0.39
Weakly-Supervised*	Sent2Span$_{agg}$	0.32	0.79	0.46	0.22	0.54	0.31	0.31	0.57	0.40
	Sent2Span$_{major}$	0.38	0.52	0.44	0.27	0.15	0.19	0.35	0.08	0.12
	Sent2Span$_{minor}$	0.30	**0.86**	0.45	0.23	0.51	0.31	0.28	**0.64**	0.39

Table 3: Precision, Recall, and F_1 score for PICO Span Detection on the PICO-data dataset. * indicates the results applying the crowdsourced sentence annotation.

Supervision Type	Method	Population		
		Precision	Recall	F_1
Human Annotation	HMMCrowd	0.73	0.75	0.74
Supervised	CRF	0.80	0.55	0.65
Semi-Supervised	BiLSTM-CRF	0.74	0.65	0.69
Weakly-Supervised	Sent2Span$_{agg}$	0.32	0.72	0.45
	Sent2Span$_{major}$	0.39	0.52	0.45
	Sent2Span$_{minor}$	0.29	**0.87**	0.43
Weakly-Supervised*	Sent2Span$_{agg}$	0.33	0.70	0.45
	Sent2Span$_{major}$	0.38	0.49	0.43
	Sent2Span$_{minor}$	0.29	**0.83**	0.44

notations by aggregating the existing crowdsourced annotations, indicating whether the sentence has at least one PICO span:

- *agg*: indicates whether a sentence received PICO annotations based on the aggregated annotations, to simulate the weighted annotator scenario

- *major*: indicates whether a sentence received PICO annotations from more than half annotators

- *minor*: indicates whether a sentence received at least one PICO annotation

On both datasets, we use spaCy [5] for tokenization and sentence split.

4.2 Hyperparameters

We use the BLUE (Peng et al., 2019) model trained on PubMed abstracts as the backbone model[6] with the BERT-Base structure and uncased tokenisation. The maximum sequence length is set to 512 tokens. The training batch size is 32 and the evaluation batch size is 64. For PICO sentence classification, we use Adam (Kingma and Ba, 2015) optimiser and set the peak learning rate to 2e-5. We train the models for 5 epochs and select the models with the best F_1 score on the development set.

For PICO Masked Span Prediction, the maximum span lengths M are set to be 20, 7, and 10 tokens [7], for Population, Intervention, and Outcome, respectively. The number of selected candidate spans K is set to 2 for all the PICO types. The maximum span lengths and the number of selected candidate spans are set based on the statistical information of the aggregated crowdsourced PICO spans on the development sets to cover at least 90% and 95% PICO spans, respectively. All the experiments are run on one NVIDIA V100 GPU. Our source code is available online[8].

---

[5] https://spacy.io/

[6] We also experimented with the one trained on PubMed abstracts and MIMIC-III and found similar results.

[7] This is on the raw string token sequence and this may vary after Bert tokenisation.

[8] Our implementation is publicly available at https://github.com/evidence-surveillance/sent2span

Table 4: Accuracy, Precision, Recall, and F_1 score of PICO sentence classification for crowdsourced annotations and proposed methods on the EBM-NLP dataset. Crowd_X refers to the aggregated crowdsourced annotations mentioned in Section 4.1. Sent2Span_X refers to proposed methods trained with different PICO sentence annotations mentioned in Section 4.1. The **bold-faced** scores represent the best results among crowdsourced annotations and proposed methods.

Method	Population				Intervention				Outcome			
	Accuracy	Precision	Recall	F_1	Accuracy	Precision	Recall	F_1	Accuracy	Precision	Recall	F_1
Crowd_{agg}	0.90	0.76	0.84	0.80	**0.90**	0.88	0.88	**0.88**	**0.84**	0.90	0.77	**0.83**
Crowd_{major}	0.88	0.90	0.57	0.70	0.70	**0.99**	0.32	0.48	0.61	**0.96**	0.22	0.36
Crowd_{minor}	0.73	0.47	**0.98**	0.64	0.83	0.73	**0.98**	0.83	0.78	0.71	0.97	0.82
Sent2Span_{agg}	**0.92**	0.84	0.84	**0.84**	0.88	0.91	0.80	0.85	0.81	0.89	0.70	0.79
Sent2Span_{major}	0.86	**0.91**	0.44	0.59	0.65	0.96	0.22	0.36	0.56	0.93	0.13	0.23
Sent2Span_{minor}	0.82	0.58	0.96	0.72	0.83	0.73	**0.98**	0.83	0.78	0.70	**0.98**	0.82

Table 5: Accuracy, Precision, Recall, and F_1 score of PICO sentence classification on the PICO-data dataset.

Method	Population			
	Accuracy	Precision	Recall	F_1
Crowd_{agg}	0.88	0.75	0.77	0.76
Crowd_{major}	0.86	**0.88**	0.51	0.64
Crowd_{minor}	0.78	0.54	0.90	0.68
Sent2Span_{agg}	**0.90**	0.82	0.81	**0.81**
Sent2Span_{major}	0.85	0.86	0.51	0.64
Sent2Span_{minor}	0.82	0.60	**0.96**	0.74

4.3 Comparison Methods and Evaluation Metrics

Following previous work (Nguyen et al., 2017; Nye et al., 2018), we use the token-wise precision, recall, and F_1 score of the output PICO spans against the expert annotations. Following previous work (Thomas et al., 2021), we focus on recall as it is the most important metric when PICO detection is applied for systematic review processes. We also report the accuracy, precision, recall and F_1 for the PICO sentence classification results against the expert annotations.

We compare Sent2Span against supervised method, semi-supervised method, and aggregated human annotations. All the methods are evaluated on the test set of expert annotations.

- HMMCrowd (Nguyen et al., 2017): HMM-Crowd extends Dawid-Skene model (Dawid and Skene, 1979) with a HMM component, and explicitly uses the sequential structure of spans. This model is directly applied on the crowdsourced annotations to get the aggregated annotations without training.

- Conditional Random Fields (CRF) (Lafferty

et al., 2001): The CRF model is fully supervised with a feature template, including the current, previous, and next words; part-of-speech tags; and character information such as whether a token contains digits, uppercase letters, symbols, etc. This model is trained with the aggregated crowdsourced span annotations on the training dataset.

- BiLSTM-CRF (Lample et al., 2016): BiLSTM-CRF model is a semi-supervised method with pre-trained word2vec embeddings trained on PubMed abstracts. This model is trained with the aggregated crowdsourced span annotations on the training dataset.

- Sent2Span_X: This is our proposed method with different sentence annotation generation functions X. X refers to the agg, $major$, and $minor$ mentioned in Section 4.1. This method is trained only with sentence annotations without any span annotations. This method is evaluated on the test datasets using both the predicted PICO sentence classification results and the crowdsourced PICO sentence annotations.

4.4 PICO Span Detection Results

Table 2 and Table 3 show the performance of different methods on the EBM-NLP and PICO-data datasets. Sent2Span_{minor} always shows the best recall with the best F_1 score in most case among the rest Sent2Span models. Compared with the rest methods, Sent2Span_{minor} achieves the best recall for the Population and Outcome types, and even better than the aggregated human annotation results for Population.

It surpasses the aggregated human annotation by 10% recall on Population on both datasets, even though the crowdsourcing annotators have been given annotation guidelines and certain examples. For Outcome, it achieves better recall compared to supervised and semi-supervised methods by around 6%, while these methods are trained with span-level annotations compared with the sentence-level annotation applied for Sent2Span. Though Sent2Span does not achieve the best recall for Intervention, it beats the supervised CRF methods by a large margin (0.51 vs 0.21).

Sent2Span does not exhibit a performance drop when applying sentence-level human annotation for inference. It shows better recall (*i.e.* Population on the PICO dataset) for all the proposed models. This indicates: (1) Sent2Span generalises well on the datasets; and (2) Sent2Span can be directly applied for PICO span detection.

4.5 PICO Sentence Classification Results

As the results are inferred from the PICO sentence classification results, it is worth examining the corresponding impact. We show the results of crowdsourced annotation and Sent2Span in Table 4 and Table 5 for both datasets.

From both tables, agg results have the best accuracy and F_1 score, $major$ results have the best precision, and $minor$ results have the best recall. Sent2Span_{minor} shows the better recall for Intervention and Outcome on the EBM-NLP dataset, and Population on the PICO dataset compared with crowdsourced annotations. This explains the equivalent and superior recall results in PICO span detection (see Table 2 and Table 3). Without losing potential PICO sentences, Sent2Span_{minor} selects the most PICO spans. And Sent2Span_{major} has the best precision with the worst recall.

4.6 Effect of Candidate Span Reduction

Table 6: Candidate Span Reduction with Algorithm 1. The number indicates the proportion of reduced candidate spans against all the candidate spans.

Method	PICO	EBM-NLP		
	Population	Population	Intervention	Outcome
Sent2Span_{agg}	0.15	0.20	0.16	0.21
Sent2Span_{major}	0.22	0.37	0.20	0.23
Sent2Span_{minor}	0.08	0.15	0.12	0.08

To demonstrate the effectiveness of the Nested

Table 7: Five examples of predicted PICO spans (Prediction) and their error types (boundary errors, overlap errors, false-positive errors, and false-negative errors). The illustration snippet contains 7 tokens, and tokens in blue are either annotated by experts or predicted by the methods (we assume the annotations and predictions have the same PICO type). FP and FN indicate false-positive and false-negative errors, respectively.

No.	Prediction	Gold	Error Type
1			Boundary
2			Boundary
3			Overlap
4			Overlap
5			FP and FN

Span Elimination Algorithm 1[9] in the Masked Span Prediction task, we count both the number of candidate spans with and without using the algorithm, and calculate the percentage of reduced candidate spans. The results are listed in Table 6 for both datasets.

At least 8% of candidate spans are eliminated and do not pass to the fine-tuned neural language model for inference, which saves inference time. In Comparison to models trained with different sentence-level annotations, Sent2Span_{major} has the most discarded candidate spans with more than 22% candidate spans eliminated cross both datasets. And Sent2Span_{minor} receives the least number of eliminated spans. This indicates that Sent2Span_{major} tends to ignore more spans and Sent2Span_{minor} builds more connections between the spans and PICO sentence classification results resulting in higher recall.

4.7 Error Analysis

We perform an error analysis using the test datasets for which there are expert annotations. Error types include boundary errors (BE), overlap errors (OE), false-positive errors (FP), and false-negative errors (FN) (Table 7).

The number of errors varies by type when comparing the proposed methods and the crowdsourced annotations (Table 8). Sent2Span_{major} gives the least number of BEs compared with the two other methods, at the cost of the largest number of FN errors and thus the lowest recall. All proposed

---

[9]We see no major difference in performance when we do not reduce candidate spans (less than 0.002 in precision, recall, and F1-score) compared with the performance using the algorithm.

Table 8: The number of errors from the proposed methods on the EBL-NLP and PICO-data datasets. Crowd indicates the aggregated crowdsourced annotations. BE indicates boundary error. OE indicates overlap error. FP and FN indicate false-positive and false-negative errors, respectively.

| Method | EBM-NLP | | | | | | | | | | | | PICO-data | | | |
| | Population | | | | Intervention | | | | Outcome | | | | Population | | | |
	BE	OE	FP	FN	BE	OE	FP	FN	BE	OE	FP	FN	BE	OE	FP	FN
Crowd	161	10	185	109	314	47	303	237	408	22	235	542	178	29	203	172
Sent2Span$_{agg}$	207	354	467	112	356	524	574	650	342	659	650	692	170	418	526	154
Sent2Span$_{major}$	100	198	182	366	111	152	168	1398	68	130	147	1590	121	295	324	355
Sent2Span$_{minor}$	201	415	1032	45	357	561	891	641	443	844	1532	422	274	486	1125	42

methods have more OEs when compared to crowd-sourced annotations, suggesting an area where future work would be valuable. In a post-hoc analysis of the OEs, we find that there are no nested predictions (for example, "training" is nested in "progressive muscle relexation training").

Sent2Span$_{minor}$ produces the most FP errors and the least FN errors across all datasets, and the difference is most pronounced for the Population type. Overall, Sent2Span$_{minor}$ produces a high recall, low precision tool, and this is related to the way K (the number of selected candidate spans) is set, introducing redundant selected spans in the final inference result. This approach is likely to be useful as part of certain pipelines where PICO extraction is used, but there is also room for further improvement in the span selection mechanism.

5 Discussion

The results of the experiments show that our proposed method could be a useful component of a pipeline for PICO detection and extraction, in use cases where costly expert annotation is limited and where the aim is to identify all relevant examples. The results show the approach compares favourably to existing supervised methods and achieves high recall for PICO span detection even without using any span annotations as training data.

Our proposed approach is likely to be useful in a range of other application domains. Many NLP tasks across NER (Sohrab and Miwa, 2018; Liu et al., 2020; Xia et al., 2019) and semantic role labelling (Ouchi et al., 2018) can be formulated as a span detection task. In cases where it is easier to acquire or estimate low-quality sentence-level annotations, and resource-intensive to acquire high-quality span-level annotations, our proposed approach may be appropriate.

Several opportunities exist for future work.

For the sake of simplicity, we used the BLUE model (Peng et al., 2019) with the BERT-base-uncased structure, rather than exploring models with BERT-large-cased structure. We also did not explore the use of other NLP tools such as part-of-speech tagging or dependency parsing. Integration of these methods in the proposed Sent2Span approach may improve the performance.

The Sent2Span method is designed to be part of a larger pipeline of techniques designed to support systematic review processes. Other than PICO detection, extraction, and representation, other methods have been proposed for identifying which trials should be included in systematic review updates (Surian et al., 2018). Future work in the space could include head-to-head comparisons of approaches for rapid systematic review updating that test for maximising the completeness of evidence identification and minimising human effort.

6 Conclusion

The Sent2Span method for PICO span detection we propose and test in this paper could be used to support new tasks in systematic review processes. The difference between Sent2Span and previous approaches to PICO detection include the use of only low-quality sentence-level annotations as training data and the results demonstrating achieve high recall in span detection, which is an important requirement for systematic review processes.

Acknowledgements

References

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *ACL (1)*, pages 1388–1398. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *EMNLP/IJCNLP (1)*, pages 3613–3618. Association for Computational Linguistics.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *ACL*, pages 8057–8077. Association for Computational Linguistics.

Lianzhe Huang, Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2019. Text level graph neural network for text classification. In *EMNLP/IJCNLP (1)*, pages 3442–3448. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguistics*, 8:64–77.

Tian Kang, Adler J. Perotte, Youlan Tang, Casey N. Ta, and Chunhua Weng. 2021. Umls-based data augmentation for natural language processing of clinical research literature. *J. Am. Medical Informatics Assoc.*, 28(4):812–823.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*. Morgan Kaufmann.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *HLT-NAACL*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *ICLR*. OpenReview.net.

Grace E. Lee and Aixin Sun. 2019. A study on agreement in PICO span annotations. In *SIGIR*, pages 1149–1152. ACM.

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. *Bioinform.*, 36(4):1234–1240.

Bing Li, Yukai Miao, Yaoshu Wang, Yifang Sun, and Wei Wang. 2021. Improving the efficiency and effectiveness for bert-based entity resolution. In *AAAI*, pages 13226–13233. AAAI Press.

Shifeng Liu, Yifang Sun, Bing Li, Wei Wang, and Xiang Zhao. 2020. HAMNER: headword amplified multi-span distantly supervised method for domain specific named entity recognition. In *AAAI*, pages 8401–8408. AAAI Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *ACL (1)*.

Iain J Marshall and Byron C Wallace. 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic reviews*, 8(1):1–10.

An Thanh Nguyen, Byron C. Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. 2017. Aggregating and predicting sequence labels from crowd annotations. In *ACL (1)*, pages 299–309. Association for Computational Linguistics.

Benjamin E. Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain James Marshall, Ani Nenkova, and Byron C. Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *ACL (1)*, pages 197–207. Association for Computational Linguistics.

Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto. 2018. A span selection model for semantic role labeling. In *EMNLP*, pages 1630–1642. Association for Computational Linguistics.

Alison O'Mara-Eves, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, 4(1):1–22.

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 *BioNLP@ACL*, pages 58–65. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT*, pages 2227–2237. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep exhaustive model for nested named entity recognition. In *EMNLP*.

Didi Surian, Adam G. Dunn, Liat Orenstein, Rabia Bashir, Enrico W. Coiera, and Florence T. Bourgeois. 2018. A shared latent space matrix factorisation method for recommending new trial evidence for systematic review updates. *J. Biomed. Informatics*, 79:32–40.

James Thomas, Steve McDonald, Anna Noel-Storr, Ian Shemilt, Julian Elliott, Chris Mavergames, and Iain J Marshall. 2021. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for cochrane reviews. *Journal of Clinical Epidemiology*, 133:140–151.

Guy Tsafnat, Paul Glasziou, Miew Keen Choong, Adam Dunn, Filippo Galgani, and Enrico Coiera. 2014. Systematic review automation technologies. *Systematic reviews*, 3(1):1–15.

Lucy Lu Wang and Kyle Lo. 2021. Text mining approaches for dealing with the rapidly expanding literature on COVID-19. *Briefings Bioinform.*, 22(2):781–799.

Congying Xia, Chenwei Zhang, Tao Yang, Yaliang Li, Nan Du, Xian Wu, Wei Fan, Fenglong Ma, and Philip S. Yu. 2019. Multi-grained named entity recognition. In *ACL (1)*.

Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang. 2020. Every document owns its structure: Inductive text classification via graph neural networks. In *ACL*, pages 334–339. Association for Computational Linguistics.

Markus Zlabinger, Marta Sabou, Sebastian Hofstätter, and Allan Hanbury. 2020. Effective crowd-annotation of participants, interventions, and outcomes in the text of clinical trial reports. In *EMNLP (Findings)*, pages 3064–3074. Association for Computational Linguistics.