# Modular Self-Supervision for Document-Level Relation Extraction

**Sheng Zhang**[1]   **Cliff Wong**[1]   **Naoto Usuyama**[1]   **Sarthak Jain**[2*]
**Tristan Naumann**[1]   **Hoifung Poon**[1]
[1]Microsoft Research, Redmond, WA
[2]Khoury College of Computer Sciences, Northeastern University
`{shezhan,clwon,naotous,tristan,hoifung}@microsoft.com`
`jain.sar@husky.neu.edu`

## Abstract

Extracting relations across large text spans has been relatively underexplored in NLP, but it is particularly important for high-value domains such as biomedicine, where obtaining high recall of the latest findings is crucial for practical applications. Compared to conventional information extraction confined to short text spans, document-level relation extraction faces additional challenges in both inference and learning. Given longer text spans, state-of-the-art neural architectures are less effective and task-specific self-supervision such as distant supervision becomes very noisy. In this paper, we propose decomposing document-level relation extraction into relation detection and argument resolution, taking inspiration from Davidsonian semantics. This enables us to incorporate explicit discourse modeling and leverage modular self-supervision for each sub-problem, which is less noise-prone and can be further refined end-to-end via variational EM. We conduct a thorough evaluation in biomedical machine reading for precision oncology, where cross-paragraph relation mentions are prevalent. Our method outperforms prior state of the art, such as multi-scale learning and graph neural networks, by over 20 absolute F1 points. The gain is particularly pronounced among the most challenging relation instances whose arguments never co-occur in a paragraph.

## 1 Introduction

Prior work on information extraction tends to focus on binary relations within sentences. However, practical applications often require extracting complex relations across large text spans. This is especially important in high-value domains such as biomedicine, where obtaining high recall of the latest findings is crucial. For example, Figure 1 shows a ternary (drug, gene, mutation) relation signifying that a tumor with MAP2K1 mutation



*"... The patient's peripheral blood indices are shown over time relative to the first dose of the MEK inhibitor cobimetinib ..."*

(... 17 sentences spanning 2 paragraphs ...)

*"... MAP2K1 mutations appeared later with p.K57T expanding to become the dominant clone ..."*

(... 10 sentences spanning 3 paragraphs ...)

*"... Expression of the MAP2K1 mutants in a lymphoid cell line showed that while each mutation was able to activate ERK in the presence of vemurafenib, all mutations remained sensitive to MEK inhibitors ..."*

Figure 1: An example document-level drug-gene-mutation relation. *cobimetinib* never co-occurs with *MAP2K1* or *K57T* in any paragraph.

K57T is sensitive to cobimetinib, yet the entities never co-occur in any single paragraph. Such precision oncology knowledge is key for determining personalized treatment for cancer patients, but it is scattered among a vast biomedical literature of more than 30 million papers, with over 1 million being added each year[1].

Recently, there has been increasing interest in cross-sentence relation extraction, but most existing work still focuses on short text spans. Quirk and Poon (2017) and Peng et al. (2017) restrict extraction to three consecutive sentences and Verga et al. (2018) to abstracts. DocRED (Yao et al., 2019), a popular document-level relation extraction dataset, consists of Wikipedia introduction sections, each with only eight sentences on average. Further, half of the relation instances reside in a single sentence, all effectively in a single paragraph.

To the best of our knowledge, Jia et al. (2019) is the first to consider relation extraction in full-text articles, which is considerably more challenging. They use the CKB dataset (Patterson et al., 2016) for evaluation where each document contains, on

---

*Work done as an intern at Microsoft Research.
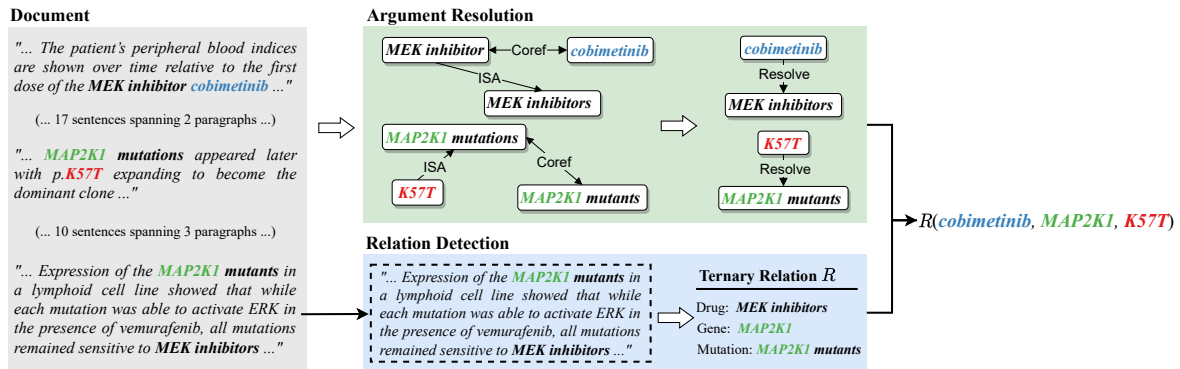
[1]`http://www.ncbi.nlm.nih.gov/pubmed`

Figure 2: Decomposing document-level $n$-ary relation extraction into relation detection and argument resolution.

average, 174 sentences, spanning 39 paragraphs. Additionally, while prior work focuses mainly on binary relations, Jia et al. (2019) follows Peng et al. (2017) to extract ternary relations. However, while Jia et al. (2019) admits relation instances for which the three arguments never co-occur in a paragraph, it still requires that the two arguments for each binary subrelation co-occur at least once in a paragraph, leaving one fifth of findings out of reach for their method.

All this prior work considers document-level relation extraction as a single monolithic problem, which presents major challenges in both inference and learning. Despite recent progress, there are still significant challenges in modeling long text spans using state-of-the-art neural architectures, such as LSTM and transformer. Moreover, direct supervision is scarce and task-specific self-supervision, such as distance supervision, becomes extremely noisy when applied beyond short text spans.

In this paper, we explore an alternative paradigm by decomposing document-level relation extraction into local relation detection and global reasoning over argument resolution. Specifically, we represent $n$-ary relation using Davidsonian semantics and combine paragraph-level relation classification with discourse-level argument resolution using global reasoning rules (e.g., transitivity over argument resolution). Each component problem resides in short text spans and their corresponding self-supervision is much less error-prone. Our approach takes inspiration from modular neural networks (Andreas et al., 2016) and neural logic programming (Rocktäschel and Riedel, 2017) in decomposing a complex task into local neural learning and global structured integration. However, instead of learning from end-to-end direct supervision, we admit modular self-supervision for the

component problems, which is more readily available. Our method can thus be viewed as applying deep probabilistic logic (Wang and Poon, 2018) to combine modular self-supervision and joint inference with global reasoning rules.

This modular approach enables us to not only handle long text spans such as full-text articles like Jia et al. (2019), but also expand extraction to the significant portion of cross-paragraph relations that are out of reach to all prior methods. We conduct a thorough evaluation in biomedical machine reading for precision oncology, where such cross-paragraph relations are especially prevalent. Our method outperforms prior state of the art such as multiscale learning (Jia et al., 2019) and graph neural networks (Zeng et al., 2020) by over 20 absolute F1 points. The gain is particularly pronounced among the most challenging relations whose arguments never co-occur in a paragraph.

## 2  Document-Level Relation Extraction

Let $E_1, \ldots, E_n$ be entities that co-occur in a document $D$. Relation extraction amounts to classifying whether a relation $R$ holds for $E_1, \ldots, E_n$ in $D$. For example, in Figure 2, $R$ is the relation of precision cancer drug response, and $E_1, E_2, E_3$ represent drug cobimetinib, gene MAP2K1, mutation K57T, respectively. The relation mention spans multiple paragraphs and dozens of sentences. Direct extraction is challenging and ignores the elaborate underlying linguistic phenomena. The drug-response relation is explicitly mentioned in the last paragraph, though it is between "*MEK inhibitors*" and "*MAP2K1 mutations*". Meanwhile, the top paragraph states the ISA relation between "*cobimetinib*" and "*MEK inhibitors*", as apparent from the apposition. From the middle paragraph one can infer the ISA relation between "*K57T*"

5292

and "*MAP2K1 mutations*". Finally, "*MAP2K1 mutants*" in the last paragraph can be resolved with "*MAP2K1 mutations*" in the middle based on semantic similarity. Combining these, we can conclude that the drug-response relation holds for (cobimetinib, MAP2K1, K57T) in this document (Figure 2), even though cobimetinib never co-occurs with MAP2K1 or K57T in any paragraph.

Formally, we represent $n$-ary relation extraction by neo-Davidsonian semantics (Parsons, 1990):

$$R_D(E_1, \cdots, E_n) \doteq \exists T \in D \; \exists r.$$
$$[R_T(r) \wedge A_1(r, E_1) \wedge \cdots \wedge A_n(r, E_n)]$$

Here, $T$ is a text span in $D$, $r$ is a reified event variable introduced to represent relation $R$, and the arguments are represented by binary relations with the event variable. The distributed nature of this representation makes it suitable for arbitrary $n$-ary relations and does not require drastic changes when arguments are missing or when new arguments are added. Given this representation, document-level relation extraction is naturally decomposed into local relation detection (e.g., classifying if $R_T(r)$ holds for some paragraph $T$) and global argument resolution (e.g., classifying $A_i(r, E_i)$).

Entity-level argument resolution can be reduced to mention-level argument resolution $A(r, E) \doteq \exists e. [\texttt{Mention}(e, E) \wedge A(r, e)]$, where $\texttt{Mention}(e, E)$ signifies that $e$ is an entity mention of $E$. Additionally, the transitivity rule applies:

$$A(r, e) \wedge \texttt{Resolve}(r, e, e') \implies A(r, e')$$

Here, $\texttt{Resolve}(r, e, e')$ signifies that the mentions $e, e'$ are interchangeable in the context of relation mention $r$. For brevity, in this paper we drop the relation context $r$ and simply consider $\texttt{Resolve}(e, e')$. If $e, e'$ are coreferent or semantically equivalent, as in (MAP2K1 mutations, MAP2K1 mutants), $\texttt{Resolve}$ obviously holds. More generally, $\texttt{ISA}$ (e.g., K57T and MAP2K1 mutations) and $\texttt{PartOf}$ (e.g., a mutation and a cell line containing it) may also signify resolution:

$$\texttt{Coref}(e, e') \implies \texttt{Resolve}(e, e')$$
$$\texttt{ISA}(e, e') \implies \texttt{Resolve}(e, e')$$
$$\texttt{PartOf}(e, e') \implies \texttt{Resolve}(e, e')$$

Also, transitivity generally holds for $\texttt{Resolve}$:

$$\texttt{Resolve}(e, e') \wedge \texttt{Resolve}(e', e'')$$
$$\implies \texttt{Resolve}(e, e'')$$

| Modular Self-Supervision | | |
|---|---|---|
| **Relation Detection** | Distant Supervision | CIVC, GDKD, OncoKB |
| **Argument Resolution** | Data Programming | Identical mentions, apposition |
| | Joint Inference | Transitivity |

Table 1: Modular self-supervision for relation detection and argument resolution.

## 3 Modular Self-Supervision

Our problem formulation makes it natural to introduce modular self-supervision for relation detection and argument resolution (Table 1).

**Relation Detection** The goal is to train a classifier for $R_T(r)$. In this paper, we consider paragraphs as candidates for $T$ and use distant supervision (Mintz et al., 2009) for self-supervision. Specifically, knowledge bases (KBs) with known relation instances for $R$ are used to annotate examples from unlabeled text. Co-occurring mention tuples of known relations are annotated as positive examples and those not known to have relations are sampled as negative examples. These examples are then used to train a paragraph-level relation classifier. Here, we leverage the fact that paragraph-level distant supervision is much less noise-prone, but document-level relation mentions still observe similar textual patterns as paragraph-level ones, as can be seen in Figure 2.

**Argument Resolution** The goal is to train a classifier for $\texttt{Resolve}(e, e')$ based on local context for entity mentions $e, e'$. As stated in the prior section, $\texttt{Resolve}$ is strictly more general than coreference and may involve $\texttt{ISA}$ and $\texttt{PartOf}$ relations. For self-supervision, we introduce data programming rules that capture identical mentions and appositives. These are used as seed self-supervision to annotate high-precision resolution instances. In turn, additional instances in the same document can be generated by applying the transitivity rule. E.g., in Figure 2, by deriving $\texttt{Resolve}$(cobimetinib, MEK inhibitor) in the top paragraph based on the apposition, we may annotate additional $\texttt{Resolve}$ instances between "*cobimetinib*" and "*MEK inhibitors*" in the bottom paragraph. As in distant supervision, there will be noise, but on balance, such joint inference helps learn more general resolution patterns.
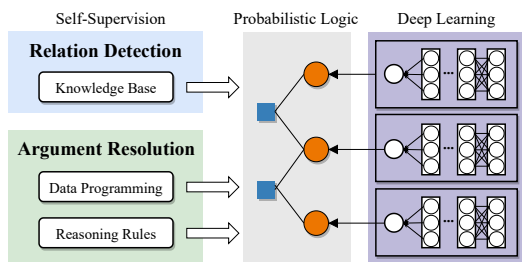
Figure 3: Our approach for document-level relation extraction applies deep probabilistic logic to incorporate modular self-supervision and joint inference for relation detection and argument resolution.

## 4 Our Model

Figure 3 shows our document-level relation extraction system, which uses deep probabilistic logic (Wang and Poon, 2018) to incorporate modular self-supervision and joint inference.

**Prediction Module**   The prediction module comprises transformer-based neural models (Vaswani et al., 2017) for local relation detection and global argument resolution. For relation detection, let $(m_1, \cdots, m_n)$ be a candidate co-occurring mention tuple in a paragraph $T$. We input $T$ to a transformer-based relation classifier, with mentions $m_i$ dummified.[2] For argument resolution, let $(m, m')$ be a candidate mention pair. We compute the contextual representation using a transformer model for both mentions and classify the pair using a comparison network. The input concatenates contextual representations of the entities as well as their element-wise multiplication. For the detailed neural architectures, see Appendix A.

**Supervision Module**   As described in the previous section, the supervision module incorporates the relation KBs and resolution data programming rules as seed self-supervision, as well as reasoning rules such as resolution transitivity for joint inference. Note that these self-supervision rules can be noisy, but for simplicity we still treat them as hard constraints. Deep probabilistic logic offers a principled way to soften them and model their noisiness, which can be investigated in future work.

**Learning**   The prediction and supervision modules define a joint probabilistic distribution

$$P(K, Y|X) \propto \prod_{v \in K} \Phi_v(X, Y) \cdot \prod_i \Psi(X_i, Y_i)$$

---
[2]Alternatively, we can add entity markers for each mention.

| | CKB | CKB$^{\text{HARD}}$ | DocRED |
|---|---|---|---|
| **Documents** | 430 | 391 | 1000 |
| **Relations** | 1904 | 332 | 12323 |
| **Candidates** | 17744 | 12122 | 198395 |
| **Document-level statistics (mean)** | | | |
| Words | 5480.1 | 5576.2 | 200.7 |
| Sentences | 170.6 | 173.6 | 8.1 |
| Paragraphs | 38.9 | 39.3 | 1.0 |
| **Minimal span for relations (mean)** | | | |
| Words | 161.8 | 717.7 | 30.9 |
| Sentences | 5.7 | 22.6 | 2.2 |
| Paragraphs | 1.7 | 4.8 | 1.0 |

Table 2: Comparison of statistics among CKB, CKB$^{\text{HARD}}$, and DocRED (validation; test annotations are not publicly available). DocRED comprises short Wikipedia introduction sections, whereas CKB features full-text articles. CKB also features relations spanning much longer text, especially in the CKB$^{\text{HARD}}$ subset.

Here, $K$ represents the self-supervision and $(X_i, Y_i)$ the input-output pairs of relation detection and argument resolution. $\Phi, \Psi$ are the supervision and prediction modules, respectively. Learning is done via variational EM. In the E-step, we compute a variational approximation $q(Y) \propto P(Y|K, X)$ using loopy belief propagation, based on current $\Phi, \Psi$. In the M-step, we treat $q(Y)$ as the probabilistic labels and refine parameters of $\Phi, \Psi$. As aforementioned, we treat the self-supervision in $\Phi$ as hard constraints, so the M-step simplifies to fine-tuning the transformer-based models for relation detection and argument resolution, treating $q(Y)$ as probabilistic labels.

**Inference**   After learning, given a test document and candidate entities and mentions, it is straightforward to run the neural modules for relation detection and argument resolution. Additionally, we would incorporate joint inference for argument resolution as in self-supervision (e.g., transitivity) using loopy belief propagation.

## 5 Experiments

In this section, we study how our modular self-supervision approach performs in document-level relation extraction. A popular dataset is DocRED (Yao et al., 2019), which features Wikipedia introduction sections and general-domain relations. However, upon close inspection, DocRED does not have many truly long-range relation instances. As Table 2 shows, each DocRED document contains only eight sentences in average, most within a single paragraph. About half of relations (49%) can

|       | Documents | Positive ex. | Negative ex. |
|-------|-----------|--------------|--------------|
| Train | 5563      | 5019         | 12367        |
| Dev   | 2400      | 5539         | 6582         |

Table 3: Distant supervision for relation detection.

be extracted from a single sentence, per evidence annotation. Consequently, there is very little room to explore the more challenging scenario where relations span multiple paragraphs in large text spans. In fact, Huang et al. (2021) finds that over 95% instances in DocRED require no more than three sentences as supporting evidence, and 87% requires no more than two sentences. Ye et al. (2020) shows that a simple BERT-based system (a special case of our approach with just local relation detection) yields 60.06% F1 on DocRED test (Table 3 in their paper), very close to the state-of-the-art results of 62.76% by GAIN (Zeng et al., 2020).

We thus focus on biomedical machine reading, where there is a pressing need for comprehensive extraction of the latest findings from full-text articles, and cross-paragraph relation mentions are prevalent. Following Peng et al. (2017); Jia et al. (2019), we consider the problem of extracting precision oncology knowledge from PubMed Central full-text articles, which is critical for molecular tumor boards and other precision health applications. Concretely, the goal is to extract drug-gene-mutation relations as shown in Figure 1: given a drug, gene, mutation, and document in which they are mentioned, determine whether the document asserts that the mutation in the gene affects response to the drug.

## 5.1 Datasets

**Self-Supervision** For training and development, we use unlabeled documents from the PubMed Central Open Access Subset (PMC-OA)[3]. For relation detection, we derive distant supervision from three knowledge bases (KBs) with manually-curated drug-gene-mutation relations: CIVIC[4], GDKD (Dienstmann et al., 2015), OncoKB (Chakravarty et al., 2017). We randomly split the generated examples into training and development sets and ensure no overlap of documents. Table 3 summarizes their statistics. For argument resolution, we use the global reasoning rules such as transitivity, as well as data programming rules capturing two anaphoric

phenomena: identical mentions, apposition.

**Evaluation** Following Jia et al. (2019), we use CKB CORE™ from the Clinical Knowledgebase (CKB; Patterson et al. 2016)[5] as our gold-standard test set. CKB contains high-quality document-level annotations of drug-gene-mutation interactions, which are manually curated from PubMed articles by The Jackson Laboratory (JAX), an NCI-designated cancer center. CKB has minimal overlap with the three KBs used in training and development. To avoid contamination, we remove CKB entries whose documents are used in our training and development. See Table 2 for statistics. Note that compared to the version used in Jia et al. (2019), the latest dataset (accessed in Oct. 2020) contains substantially more relations from recent findings. For about one fifth of annotated relations (17.4%), the key entities such as drug and mutation never co-occur in the same paragraph. These relations are out of scope in Jia et al. (2019). We denote this subset as CKB[HARD], which comprises particularly challenging instances requiring cross-paragraph discourse modeling.

## 5.2 Systems

We implemented our modular self-supervision method (`Modular`) using PyTorch (Paszke et al., 2019). We conducted variational EM for eight iterations, which appear to work well in preliminary experiments. In the M-step, we incorporate early stopping to identify the best checkpoint based on the development performance for fine-tuning the relation detection and argument resolution neural modules. We initialized the encoding layers in both modules with PubMedBERT (Gu et al., 2021), which has been pretrained from scratch on PubMed articles and demonstrated superior performance in a wide range of biomedical NLP applications.

We follow Wang and Poon (2018); Jia et al. (2019) to conduct standard data preprocessing and entity linking. We used the AdamW optimizer (Loshchilov and Hutter, 2019). For training, we set the mini-batch size to 32 and the learning rate 5e-5 with 100 warm-up steps and 0.01 weight decay. The drop-out rate is 0.1 for transformer-based encoders, and 0.5 for other layers. The hidden size is 765 for transformer-based encoders, and 128 for all other feed-forward networks. We generate checkpoints at every 4096 steps. Three random seeds are tried in our experiments: [7, 12, 17].

---

[3] `www.ncbi.nlm.nih.gov/pmc/`
[4] `civicdb.org/home`

[5] `ckbhome.jax.org`

| Test Set | System | Precision (%) | Recall (%) | F1 (%) |
|---|---|---|---|---|
| CKB | AllPositive | 10.7 | 100.0 | 19.4 |
| | Multiscale (Jia et al., 2019) | 21.2 ±0.5 | 59.3 ±0.2 | 31.2 ±0.5 |
| | GAIN (Zeng et al., 2020) | 42.3 ±2.0 | 33.5 ±1.5 | 37.3 ±0.2 |
| | Modular (Ours) | 49.9 ±0.3 | 71.5 ±0.3 | **58.8** ±0.1 |
| CKB$^{\text{HARD}}$ | AllPositive | 2.7 | 100.0 | 5.3 |
| | Multiscale (Jia et al., 2019) | 4.7 ±0.2 | 42.8 ±0.9 | 8.5 ±0.3 |
| | GAIN (Zeng et al., 2020) | 7.9 ±0.4 | 13.7 ±1.4 | 10.0 ±0.4 |
| | Modular (Ours) | 22.6 ±0.1 | 46.1 ±0.4 | **30.3** ±0.1 |

Table 4: Comparison of test results on CKB and CKB$^{\text{HARD}}$. Relations in CKB$^{\text{HARD}}$ are particularly challenging as key entity pairs such as drug and mutation never co-occur in a paragraph. All systems were trained using the same three KBs for distant supervision, with no overlap with CKB. We report results from three random runs.

For self-supervised relation detection, following Jia et al. (2019), we further decompose it into classifying drug-mutation relations and then augmenting them with high-precision gene-mutation associations. As stated in Section 3, at training time, only named entities such as drug, gene, mutation are considered, whereas at inference time, in principle any co-occurring noun phrases within a paragraph would be considered (see Figure 2, bottom paragraph). In practice, however, this would incur too much computation, most of which wasted on irrelevant candidates. Therefore, we employ the following heuristics to leverage argument resolution results for filtering candidates: In argument resolution, we focus on resolving candidate mentions with drugs, genes, mutations. We also stipulate that a candidate mention must contain within it some relevant biomedical entity mentions (e.g., cell lines, genes, etc., as in "MEK inhibitors" that contains gene reference "MEK"). In relation detection, we only consider candidate mentions that are classified as resolving with entities among drugs, genes, mutations, based on current prediction module.

We compare Modular with the following baselines: AllPositive is a recall-friendly baseline that always predicts positive; Multiscale (Jia et al., 2019) is a state-of-the-art approach that combines local mention-level representations into an entity-level representation over the entire document; GAIN (Zeng et al., 2020) is another state-of-the-art approach that constructs mention-level graphs and applies graph convolutional network (Kipf and Welling, 2017) to model interdependencies among intra- and inter-sentence mentions, attaining top performance on DocRED.

For fair comparison, we replaced the encoders in Multiscale and GAIN with the state-of-the-art PubMedBERT as in our approach, which helped improve the performance (Appendix B). The orig-

inal Multiscale encodes each paragraph separately using LSTM, so it's straightforward to replace that with PubMedBERT. GAIN, on the other hand, encodes the entire input text all at once. This is feasible in DocRED, where each "document" is actually a Wikipedia introduction section, thus more like a paragraph (average only eight sentences long). But it doesn't work in CKB, where each document is a full-text article. Even the minimal text span covering given entities is often too long to encode using a transformer. Therefore, we ran the encoder on individual paragraphs. Note that the original version of Multiscale can't make prediction for any instances where key entity pairs such as drug and mutation never co-occur in a paragraph. We implemented a natural extension that would generate local mention-level representations even for singleton mentions (i.e., only one relevant entity shows up in a paragraph).

## 5.3 Main Results

Table 4 compares various approaches for document-level relation extraction on CKB. Our modular self-supervision approach (Modular) substantially outperforms all other methods, gaining more than 20 absolute F1 points compared to prior state of the art such as multiscale learning and graph neural networks. This demonstrates the superiority in leveraging less noise-prone modular self-supervision as well as fine-grained discourse modeling in argument resolution. Note that the results for Multiscale are different than that in Jia et al. (2019) as we used the latest CKB which contains considerably more cross-paragraph relations.

Compared to multiscale learning (Multiscale), the graph-neural-network approach (GAIN) attains significantly better precision, as it incorporates more elaborate graph-based reasoning among entities across sentences. However, this comes with

|  | Prec. | Recall | F1 |
|---|---|---|---|
| `Modular` (Ours) | 22.6 | 46.1 | **30.3** |
| ▷ Replace relation detection with `AllPositive` | 13.7 | 66.1 | 22.7 |
| ▷ Replace argument resolution with Multi-Sieve Pass | 11.0 | 42.8 | 17.5 |
| ▷ Replace argument resolution with SpanBERT Coref | 60.2 | 1.5 | 2.9 |

Table 5: Ablation study on CKB$^{\text{HARD}}$.

substantial expense at recall. Our approach outperforms both substantially in precision and recall.

On the most challenging subset CKB$^{\text{HARD}}$, the contrast is particularly pronounced, as all other systems could only attain single-digit precision. The graph-neural-network approach also suffers heavily in recall. Our approach attains much better precision and recall, and more than triples the F1.

## 5.4 Ablation Study

To understand the impact of our modular self-supervision, we conducted an additional ablation study on CKB$^{\text{HARD}}$. See Table 5 for results.

To assess the limitation of our current argument resolution module, we replace self-supervised relation detection with a baseline that always predicts positive for candidate tuples whose components have been resolved with some drug, gene, mutation entities. This yields a maximum recall of 66.1%, which means that about a third of the especially hard cases of cross-paragraph relations are still out of reach for our method. In some cases, this is because the only hint at the relation resides in figures or appendix, which are currently not in scope for extraction. In other cases, the argument resolution fails to make correct resolution with the corresponding entities. We leave further investigation and improvement to future work. With self-supervised relation detection, our full model improves both F1 and precision for the end extraction on CKB$^{\text{HARD}}$.

Next, we investigate the impact of our self-supervision for argument resolution by replacing it with state-of-the-art coreference systems: Multi-Sieve Pass (Raghunathan et al., 2010; Lee et al., 2011), and SpanBERT Coreference (Joshi et al., 2020). Multi-Sieve Pass is a rule-based system that incorporates a series of resolution rules with increasing recall but lower precision. SpanBERT Coreference is a state-of-the-art transformer-based system fine-tuned on OntoNotes (Pradhan et al., 2012), an annotated corpus with diverse text. Both result in significant performance drop. Using Multi-

Sieve results in substantial drop in precision, indicating that coreference heuristics suitable for general domains are less effective in biomedicine. SpanBERT, on the other hand, suffers catastrophic drop in recall. This suggests that argument resolution for document-level relation extraction may involve more general anaphoric phenomena such as `ISA` and `PartOf`, which are out of scope in standard coreference annotations. Remarkably, bootstrapping from the simple data programming rules of identical mentions and apposition, our self-supervised module is able to perform much better argument resolution than these state-of-the-art systems for document-level relation extraction.

| Test Data | Pre-trained Encoder | Prec. | Recall | F1 |
|---|---|---|---|---|
| CKB | PubMedBERT | 49.9 | 71.5 | **58.8** |
|  | BERT | 49.5 | 68.6 | 57.5 |
| CKB$^{\text{HARD}}$ | PubMedBERT | 22.6 | 46.1 | **30.3** |
|  | BERT | 21.8 | 39.2 | 28.1 |

Table 6: Domain-specific pretraining improves test performance over general-domain pretraining.

Given that our evaluation is in the biomedical domain, it is natural to initialize our self-supervised neural modules with PubMedBERT (Gu et al., 2021). Table 6 shows that this is indeed advantageous, with domain-specific pretraining attains significant gain over general-domain pretraining.

## 5.5 Discussion

**Interpretability** We envision that machine reading is used not as standalone automation, but as assisted curation to help expert curators attain significant speed-up (Peng et al., 2017). For extraction within short text spans, human experts can validate the results by simply reading through the provenance text. For document-level relation extraction, as in Jia et al. (2019), this can be challenging, as the intervening text span is long and validation may require a significant amount of reading that is not much faster than curation from scratch. Our modular approach not only enables us to tackle the harder cases of cross-paragraph relations, but also yields a natural explanation for an extraction result with the local relation and chains of argument resolution, all of which can be quickly validated by curators. We leave studying the impact on assisted curation to future work.

**Error Analysis** We focus our error analysis on CKB$^{\text{HARD}}$, which is particularly challenging. An im-

mediate opportunity lies in significant recall miss by argument resolution. As shown in the prior subsection, the maximum recall for our current argument resolution module is 66%. From preliminary sample analysis, there are three main types of errors. Some relation instances are hinted in figures, tables, or supplements, which are beyond the current scope of extraction. In other cases, the relation statement is vague and scattered, and relation detection requires inference piecing together multiple evidences. However, the bulk of recall errors simply stem from argument resolution failures. Likewise, we found that in the majority of precision errors, relation detection appears to make the right call, but argument resolution is mistaken.
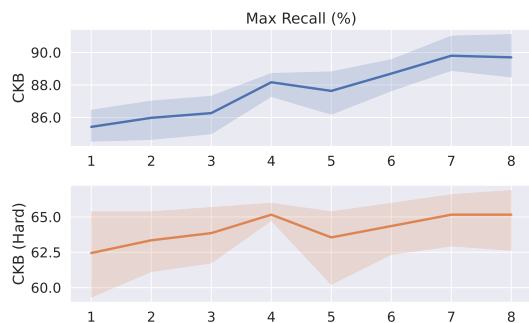


Figure 4: Argument resolution improves during learning as global reasoning rules augment seed self-supervision to raise maximum recall.

**Variational EM**   One direction to improve argument resolution is to augment the self-supervision used in the resolution module. Figure 4 shows that argument resolution did improve during learning, thanks to the global reasoning rules, at least in terms of expanding maximum recall. However, we notice that our current mention filtering rules may be overly strict, which limit the room for growth in recall. Additionally, we treat the reasoning rules such as transitivity as hard constraints, whereas in practice they can be noisy. (E.g., qualifiers like "some MAP2K1 mutations" or negation are currently not considered in resolution.)

## 6   Related Work

**Document-Level Relation Extraction**   Due to the significant challenges in modeling long text spans and obtaining high-quality supervision signals, document-level relation extraction has been relatively underexplored (Surdeanu and Ji, 2014). Prior work often focuses on simple extensions of sentence-level extraction (e.g., by incorporat-

ing coreference annotations or considering special cases when document-level relations reduces to sentence-level attribute classification) (Wick et al., 2006; Gerber and Chai, 2010; Swampillai and Stevenson, 2011; Yoshikawa et al., 2011; Koch et al., 2014; Yang and Mitchell, 2016). Recently, cross-sentence relation extraction has seen increasing interest (Li et al., 2016; Quirk and Poon, 2017; Peng et al., 2017; Verga et al., 2018; Christopoulou et al., 2019; Wu et al., 2019; Yao et al., 2019), but most efforts are still limited to short text spans, such as consecutive sentences or abstracts. A notable exception is Jia et al. (2019), which considers full-text articles that comprise hundreds of sentences. However, they still model local text units in isolation and can't effectively handle relations whose arguments never co-occur in a paragraph. In contrast, we provide the first attempt to systematically explore cross-paragraph relation extraction. Most prior work focuses on binary relations. We instead follow Peng et al. (2017); Jia et al. (2019) to study general $n$-ary relation extraction, using precision oncology treatment as a case study.

**Discourse Modeling**   Given the focus of standard information extraction on short text spans, discourse modeling has not featured prominently in prior work. An exception is coreference resolution, though the focus tends to be improving sentence-level extraction, as in Koch et al. (2014). Here, we show that document-level relation extraction often requires modeling more general anaphoric phenomena. As discussed in the experiment section, many remaining errors lie in argument resolution, which offers an exciting opportunity to study discourse modeling for an important end application.

**Self-Supervision**   Task-specific self-supervision alleviates the annotation bottleneck by leveraging freely available domain knowledge (as in distant supervision (Craven and Kumlien, 1999; Mintz et al., 2009)) and expert-derived labeling rules (as in data programming (Ratner et al., 2017)). Unfortunately, such self-supervision becomes extremely noisy when applied to full-text documents, prompting many prior efforts to focus on short text spans (Quirk and Poon, 2017; Peng et al., 2017; Verga et al., 2018; Yao et al., 2019). We instead decompose end-to-end document-level extraction into relation detection and argument resolution modules, for each of which we leverage modular self-supervision that is much less error-prone.

**Neural-Symbolic NLP** In the past few decades, the dominant paradigm in NLP has swung from logical approaches (rule-based or relational systems) to statistical and neural approaches. However, given the prevalence of linguistic structures and domain knowledge, there has been increasing interest for synergizing the contrasting paradigms to improve inference and learning. Neural logic programming replaces logical operators with neural representations to leverage domain-specific constraints with end-to-end differentiable learning (Rocktäschel and Riedel, 2017). Similarly, modular neural networks integrate component neural learning along a structured scaffold (e.g., syntactic parse of a sentence for visual question-answering) (Andreas et al., 2016). On the other hand, deep probabilistic logic (Wang and Poon, 2018) combines probabilistic logic with neural networks to incorporate diverse self-supervision for deep learning. We take inspiration from modular neural networks and neural logic programming, and use deep probabilistic logic to combine relation detection and argument resolution using global reasoning rules for document-level relation extraction.

# 7 Conclusion

We propose to decompose document-level relation extraction into local relation detection and global argument resolution, and apply modular self-supervision and discourse modeling using deep probabilistic logic. On the challenging problem of biomedical machine reading, where cross-paragraph relations are prevalent, our approach substantially outperforms prior state of the art such as multiscale learning and graph neural networks, gaining over 20 absolute F1 points. Future directions include: improving discourse modeling for argument resolution; studying the impact on assisted curation; applications to other domains.

## Acknowledgements

## References

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 39–48. IEEE Computer Society.

Debyani Chakravarty, Jianjiong Gao, Sarah Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E Rudolph, Rona Yaeger, Tara Soumerai, Moriah H Nissan, et al. 2017. Oncokb: a precision oncology knowledge base. *JCO precision oncology*, 1:1–16.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. 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 4925–4936, Hong Kong, China. Association for Computational Linguistics.

Mark Craven and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 77–86.

Rodrigo Dienstmann, In Sock Jang, Brian Bot, Stephen Friend, and Justin Guinney. 2015. Database of genomic biomarkers for cancer drugs and clinical targetability in solid tumors. *Cancer discovery*, 5(2):118–123.

Matthew Gerber and Joyce Chai. 2010. Beyond NomBank: A study of implicit arguments for nominal predicates. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1583–1592, Uppsala, Sweden. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *Special Issue on Computational Methods for Biomedical Natural Language Processing, ACM Transactions on Computing for Health*.

Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. 2021. Three sentences are all you need: Local path enhanced document relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 998–1004, Online. Association for Computational Linguistics.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3693–3704, Minneapolis, Minnesota. 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. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Mitchell Koch, John Gilmer, Stephen Soderland, and Daniel S. Weld. 2014. Type-aware distantly supervised relation extraction with linked arguments. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1891–1901, Doha, Qatar. Association for Computational Linguistics.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, Portland, Oregon, USA. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Terence Parsons. 1990. *Events in the Semantics of English: A Study in Subatomic Semantics*. MIT Press.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Sara E Patterson, Rangjiao Liu, Cara M Statz, Daniel Durkin, Anuradha Lakshminarayana, and Susan M Mockus. 2016. The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. *Human genomics*, 10(1):1–13.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph LSTMs. *Transactions of the Association for Computational Linguistics*, 5:101–115.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1171–1182, Valencia, Spain. Association for Computational Linguistics.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501, Cambridge, MA. Association for Computational Linguistics.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. *Proc. VLDB Endow.*, 11(3):269–282.

Tim Rocktäschel and Sebastian Riedel. 2017. End-to-end differentiable proving. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3788–3800.

Mihai Surdeanu and Heng Ji. 2014. Overview of the english slot filling track at the tac2014 knowledge base population evaluation. In *Proc. Text Analysis Conference (TAC2014)*.

Kumutha Swampillai and Mark Stevenson. 2011. Extracting relations within and across sentences. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 25–32, Hissar, Bulgaria. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 872–884, New Orleans, Louisiana. Association for Computational Linguistics.

Hai Wang and Hoifung Poon. 2018. Deep probabilistic logic: A unifying framework for indirect supervision. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1891–1902, Brussels, Belgium. Association for Computational Linguistics.

Michael Wick, Aron Culotta, and Andrew McCallum. 2006. Learning field compatibilities to extract database records from unstructured text. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 603–611, Sydney, Australia. Association for Computational Linguistics.

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

Bishan Yang and Tom M. Mitchell. 2016. Joint extraction of events and entities within a document context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 289–299, San Diego, California. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential Reasoning Learning for Language Representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186, Online. Association for Computational Linguistics.

Katsumasa Yoshikawa, Sebastian Riedel, Tsutomu Hirao, Masayuki Asahara, and Yuji Matsumoto. 2011. Coreference based event-argument relation extraction on biomedical text. *Journal of Biomedical Semantics*, 2(5):1–14.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640, Online. Association for Computational Linguistics.

## A  Neural Architectures



$P(R(\textit{MEK inhibitors, MAP2K1, MAP2K1 mutants}))$

**Relation Classifier**

$h_{\text{[CLS]}}$ $h_{\text{Expression}}$ ··· ··· $h_{\text{[SEP]}}$

**Pre-trained Transformer Encoder**

[CLS]Expression of the [Gene][Mutation] in a lymphoid cell ... all mutations remained sensitive to [Drug][SEP]

(a) Relation Detection

$P(\text{Resolve}(\textit{K57T, MAP2K1}))$

**Resolution Classifier**

··· $h_{\text{MAP2K1 mutations}}$ ··· ··· $h_{\text{K57T}}$ ···

**Pre-trained Transformer Encoder**

[CLS]... MAP2K1 mutations appeared later with p.K57T expanding to become the dominant clone ...[SEP]
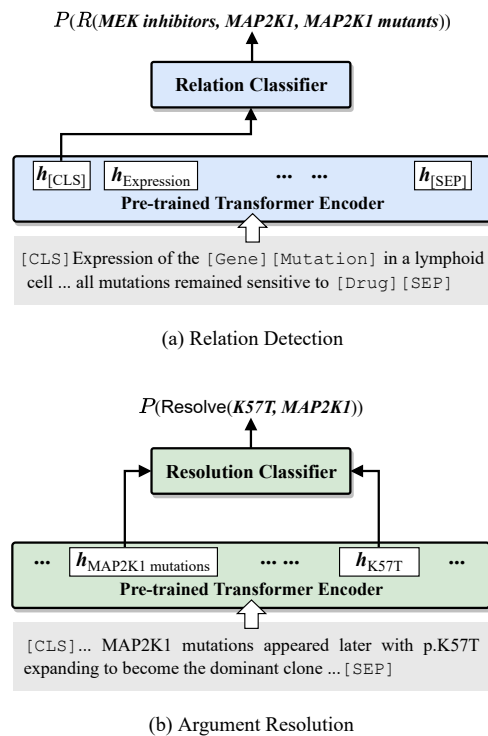
(b) Argument Resolution

Figure 5: Neural architectures for relation detection and argument resolution.

Figure 5 shows the neural architectures for relation detection and argument resolution. For relation detection, we input a paragraph to a transformer-based encoder, with mentions dummified. The hidden state $h_{\text{[CLS]}}$ in the last layer is then passed to a simple feed-forward classifier defined as below:

$$\text{classifier}(h) = \text{sigmoid}(\text{FFNN}_1(h))$$

where $\text{FFNN}_1$ is a two-layer feed-forward network using ReLU as activation functions.

For argument resolution, given a candidate mention pair $(m_i, m_j)$ and their context, we first compute their contextual representations using a transformer-based encoder. If a mention span contains multiple tokens, we use average pooling to combine their contextual representations (hidden states in the last layer). The pair of mention representations $(h_{m_i}, h_{m_j})$ are then passed to a classifier defined as below:

$$\text{classifier}(h_{m_i}, h_{m_j}) = \text{sigmoid}(s(h_{m_i}, h_{m_j}))$$

where $s(x, y)$ is a scoring function similar to those

| System | AUC | Prec. | Recall | F1 |
|---|---|---|---|---|
| **Base versions** | | | | |
| SENTLEVEL | | | | |
| Jia et al. (2019) | 22.4 | 39.3 | 34.7 | 36.9 |
| Our reproduction | 22.4 | 38.9 | 35.5 | 37.1 |
| w. PubMedBERT | 24.4 | 43.6 | 35.1 | 38.9 |
| PARALEVEL | | | | |
| Jia et al. (2019) | 33.1 | 36.5 | 44.6 | 40.1 |
| Our reproduction | 32.8 | 35.6 | 44.3 | 39.5 |
| w. PubMedBERT | 33.2 | 49.4 | 39.5 | 43.9 |
| DOCLEVEL | | | | |
| Jia et al. (2019) | 36.7 | 45.4 | 38.5 | 41.7 |
| Our reproduction | 37.0 | 43.3 | 41.9 | 42.6 |
| w. PubMedBERT | 34.9 | 47.7 | 41.6 | 44.5 |
| MULTISCALE | | | | |
| Jia et al. (2019) | 37.3 | 41.8 | 43.4 | 42.5 |
| Our reproduction | 36.9 | 38.5 | 46.2 | 42.0 |
| w. PubMedBERT | 35.9 | 43.4 | 46.3 | 44.8 |
| **+ Noisy-Or** | | | | |
| SENTLEVEL | | | | |
| Jia et al. (2019) | 25.3 | 39.3 | 35.3 | 37.2 |
| Our reproduction | 25.3 | 38.4 | 35.9 | 37.1 |
| w. PubMedBERT | 26.0 | 43.6 | 35.2 | 39.0 |
| PARALEVEL | | | | |
| Jia et al. (2019) | 35.6 | 44.3 | 40.6 | 42.4 |
| Our reproduction | 35.5 | 45.5 | 39.2 | 42.1 |
| w. PubMedBERT | 37.3 | 51.6 | 38.8 | 44.2 |
| DOCLEVEL | | | | |
| Jia et al. (2019) | 36.7 | 45.4 | 38.5 | 41.7 |
| Our reproduction | 37.0 | 43.3 | 42.0 | 42.6 |
| w. PubMedBERT | 34.9 | 47.7 | 41.6 | 44.5 |
| MULTISCALE | | | | |
| Jia et al. (2019) | 39.7 | 48.1 | 38.9 | 43.0 |
| Our reproduction | 39.6 | 48.7 | 37.8 | 42.6 |
| w. PubMedBERT | 39.5 | 46.8 | 43.1 | 44.9 |

Table 7: Comparison on the original CKB test set: reported results in Jia et al. (2019), results from our reproduction, and results after replacing LSTM with PubMedBERT as the encoder.

used in Lee et al. (2018); Joshi et al. (2020):

$$s(x, y) = s_m(x) + s_m(y) + s_c(x, y)$$
$$s_m(x) = \text{FFNN}_2(x)$$
$$s_c(x, y) = \text{FFNN}_3([x, y, x \circ y])$$

where $\circ$ denotes element-wise multiplication. $\text{FFNN}_2$ and $\text{FFNN}_3$ are two-layer feed-forward networks using ReLU as activation functions.

## B  Multiscale with PubMedBERT

Replacing LSTM with PubMedBERT (Gu et al., 2021) as the encoder generally leads to comparable or better performance by the Multiscale system (Jia et al., 2019). Table 7 shows the results on the original CKB test set as used in Jia et al. (2019). Note that these results shouldn't be compared with the main results in Table 4, as the latter are obtained on the latest CKB with considerably more cross-paragraph relations.