Guide the Many-to-One Assignment: Open Information Extraction via IoU-aware Optimal Transport

Kaiwen Wei[1,2]*, Yiran Yang[1,2]*, Li Jin[1,2]†, Xian Sun[1,2], Zequn Zhang[1,2], Jingyuan Zhang[3], Xiaoyu Li[1,2], Linhao Zhang[1,2], Jintao Liu[1,2], Zhi Guo[1,2]

[1]Key Laboratory of Network Information System Technology, Aerospace Information Research Institute, Chinese Academy of Sciences

[2]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences

[3]DAMO Academy, Alibaba Group

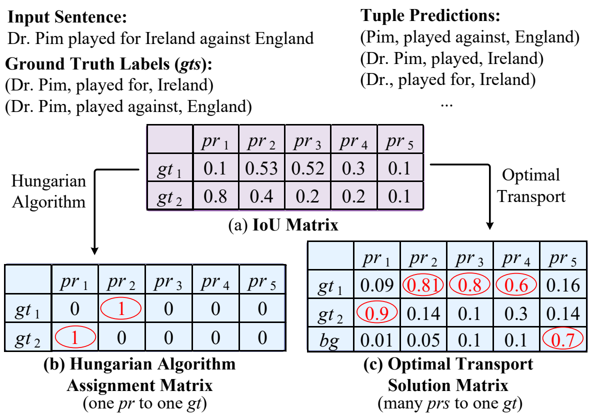
{weikaiwen19,yangyiran19}@mails.ucas.ac.cn, jinlimails@gmail.com

Abstract

Open Information Extraction (OIE) seeks to extract structured information from raw text without the limitations of close ontology. Recently, the detection-based OIE methods have received great attention from the community due to their parallelism. An essential step of those models is how to assign ground truth labels to the parallelly generated tuple proposals, which remains under-exploited. The commonly utilized Hungarian algorithm for this procedure is restricted to handling one-to-one assignment among the desired tuples and tuple proposals, which ignores the correlation between proposals and affects the recall of the models. To solve this problem, we propose a dynamic many-to-one label assignment strategy named **IOT**. Concretely, the label assignment process in OIE is formulated as an Optimal Transport (OT) problem. We leverage the intersection-over-union (IoU) as the assignment quality measurement, and convert the problem of finding the best assignment solution to the one of solving the optimal transport plan by maximizing the IoU values. To further utilize the knowledge from the assignment, we design an Assignment-guided Multi-granularity (AM) loss by simultaneously considering word-level and tuple-level information. Experiment results show the proposed method outperforms the state-of-the-art models on three benchmarks.

1 Introduction

Open Information Extraction (OIE) aims to extract structured information from the given text without the restriction of pre-defined ontology schema, and it is typically formed as a tuple ⟨subject, relation, object⟩ (Yates et al., 2007). For example, given the sentence "Dr. Pim played for Ireland against England", an OIE system needs to extract ⟨Dr. Pim, played for, Ireland⟩ and ⟨Dr. Pim, played against, England⟩, where there could be several overlapped elements. Due to the domain-independence and scalability (Mausam, 2016), OIE is widely used in various downstream tasks, such as word embedding generation (Stanovsky et al., 2015), knowledge graph completion (Han et al., 2020), multi-document question answering (Fan et al., 2019).



(a)	pr_1	pr_2	pr_3	pr_4	pr_5
gt_1	0.1	0.53	0.52	0.3	0.1
gt_2	0.8	0.4	0.2	0.2	0.1

(b)	pr_1	pr_2	pr_3	pr_4	pr_5
gt_1	0	1	0	0	0
gt_2	1	0	0	0	0

(c)	pr_1	pr_2	pr_3	pr_4	pr_5
gt_1	0.09	0.81	0.8	0.6	0.16
gt_2	0.9	0.14	0.1	0.3	0.14
bg	0.01	0.05	0.1	0.1	0.7

Figure 1: The toy example of the label assignment process from the detection-based OIE models with the Hungarian algorithm and the optimal transport strategies. The element in the matrices represents the IoU value (for (a)) or assignment probability (for (b) and (c)) between the implicit tuple proposal (pr) and ground truth tuple (gt). Elements in red circles mean these pr-gt pairs are selected for training. 'bg' is short for the background label. The Hungarian algorithm assigns only one pr for each gt, while in optimal transport many prs could be matched with the same gt.

Recently, since the sequence labelling-based OIE systems (Stanovsky et al., 2018; Zhan and Zhao, 2020) do not model the inherent dependencies among the extractions, those methods typically have relatively low accuracy. In contrast, the sequence generation-based OIE methods (Cui et al., 2018; Kolluru et al., 2020a) perform better, but the autoregressive strategies heavily reduce the in-

*Equal contribution.
†Corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, pages 4971–4984
July 9-14, 2023 ©2023 Association for Computational Linguistics

ference speed. Kolluru et al. (2020a) propose the OpenIE6 model to be the trade-off between the two kinds of methods, but the inference speed is still unsatisfactory. Different from previous paradigms, Vasilkovsky et al. (2022) first introduce a detection-based model named DetIE, which formulates OIE as a direct set prediction problem. It not only has a much faster inference speed than existing methods, but also obtains promising results. Concretely, during the training stage, DetIE generates a fixed number of tuple proposals (feature tensors) for each sentence in parallel, where each proposal represents a certain tuple. Since those proposals are implicit, DetIE adopts the Hungarian algorithm (Kuhn, 1955) to match ground truth labels (*gt*) with the proposals (*pr*), and utilizes the intersection-over-union (IoU) as the assignment quality measurement. A larger IoU value indicates the model could make a more accurate tuple prediction. During the testing stage, it utilizes the proposals to get the tuple predictions (more details are illustrated in section 3.1).

However, existing detection-based methods ignore the correlation between proposals during the label assignment process. As illustrated in Fig. 1 (a)(b), the Hungarian algorithm assigns only one *pr* for each *gt* (i.e., one-to-one), thus some informative tuple proposals (e.g., pr_3 that have similar IoU values with pr_2 for gt_1) do not participate in the training process and are forced to pull away from related proposals. Statistically, 78.3% of sentences in the widely-used CaRB (Bhardwaj et al., 2019) test set contain less than three tuples, which is far less than the number of proposals. For those situations where the number of *gt* is much smaller than the number of *pr*, most of the proposals would be discarded, which is not conducive to promoting the coverage and the learning efficiency of the model. Intuitively, taking more *prs* into account could bring better fault tolerance. To achieve the global optimal assigning result under the many *prs* to one *gt* situation, in this paper, we formulate the label assignment process in OIE as an IoU-aware Optimal Transport (OT) problem. Specifically, we define the transporting cost of each *gt*-*pr* pair as the negative IoU value, and propose a dynamic k strategy to determine how many *prs* should be assigned for each *gt* based on the IoU values. Under this formulation, finding the best assignment solution is converted to solving the optimal transport plan, which could be solved by the Sinkhorn-Knopp Iteration (Cuturi, 2013).

Moreover, in the many-to-one label assignment pattern, the proposals have different correlation strengths, so they should be treated differently during training. For example, in Fig. 1 (a)(c), $\{pr_2,pr_3,pr_4\}$ are all assigned to gt_1, but the IoU value of the pr_4-gt_1 pair is much smaller than others, indicating pr_4 makes an imperfect prediction, so the importance of pr_4 should be less than the other two proposals. To model the importance among different *prs*, an alternative way is to apply re-weight strategies. In the OIE task, because a sentence may be extracted with multiple overlapped tuples, an OIE system should treat each tuple as a whole, rather than just extracting at the word-level. As defined in the OT process, the IoU values reflect the correlations between the proposals and the golden labels, which could provide supervision from a higher perspective: the tuple-level. As a result, in this paper, we design an Assignment-guided Multi-granularity (AM) loss, where the proposal weights are dynamically determined by the combination of the predicted logits (word-level) and the IoU values (tuple-level), and those IoU values are selected by the OT assignment strategy.

We name the proposed **I**oU-aware **O**ptimal **T**ransport method with the Assignment-guided Multi-granularity loss for OIE as **IOT**. Experiments on three datasets illustrate that IOT surpasses state-of-the-art methods and achieves a fast inference speed. In summary, the contributions of this paper are as follows:

1) This is the first work to formulate the label assignment process in OIE as an IoU-aware optimal transport problem, which allows multiple tuple proposals to dynamically match with the same ground truth tuple for training.

2) We argue that IoU provides a higher level of supervision and introduce an Assignment-guided Multi-granularity (AM) loss, which explicitly considers word-level and tuple-level information to weigh different tuple proposals.

3) Experiment results show the proposed approach outperforms state-of-the-art models on three benchmarks with a fast inference speed.

2 Related Work

Open Information Extraction (OIE) aims at extracting the ⟨subject, relation, object⟩ tuples from unstructured text without limitations on the predefined relation type. RnnOIE (Stanovsky et al., 2018) and their improved versions (Roy et al., 2019;

Ro et al., 2020a; Zhan and Zhao, 2020) fulfil the OIE task by sequence labelling, but those methods are relatively less accurate. The sequence generation approaches (Cui et al., 2018; Kolluru et al., 2020b) produce tuples as sequences, but the utilized autoregressive strategies heavily restrict inference speed. Kolluru et al. (2020a) introduce the OpenIE6 model to balance the pros and cons of the two kinds of methods, but the inference speed is still unsatisfactory. Vasilkovsky et al. (2022) propose a detection-based OIE model named DetIE, which owns the advantages of fast speed and great performance. Nevertheless, they adopt the one-to-one Hungarian algorithm (Kuhn, 1955) during the label assignment process of OIE, ignoring the relations between different tuple proposals.

Label assignment aims at assigning proper predictions to each golden label. Sui et al. (2020) apply non-autoregressive models to information extraction and leveraged the Hungarian algorithm for bipartite matching. Tan et al. (2021) formulate named entity recognition (NER) as a sequence-to-set task. However, those methods utilize the one-to-one label assignment strategy. Shen et al. (2022) develop a one-to-many NER model named PIQN based on machine reading comprehension (MRC), where each entity can be assigned to multiple instance queries. However, there are several differences with the proposed IOT: (1) PIQN utilizes a linear transformation of the predicted probability as the cost function, but IOT utilizes the IoU matrix; (2) PIQN is an MRC-based model for extracting entities. In contrast, IOT is a detection-based model, which seeks to extract tuples. (3) PIQN assigns a fixed number of queries for each entity for training, but IOT dynamically determines the number of proposals for each ground truth tuple.

Re-weight Function aims to strengthen important samples while curbing insignificant ones. For example, weighted CE loss (Ronneberger et al., 2015) utilizes manually designed weights to control the importance of diverse classes. Focal loss (Lin et al., 2020) adds a modulated factor on the cross entropy loss to focus on hard samples. Dice Loss (Milletari et al., 2016) and DSC loss (Li et al., 2020) are inspired by the dice coefficient and Sørensen–Dice coefficient to get the weights of classes. However, those methods only consider the importance of each word. Li et al. (2019) also develop GHM loss in the computer version field, but it is mainly designed for foreground/background two classes.

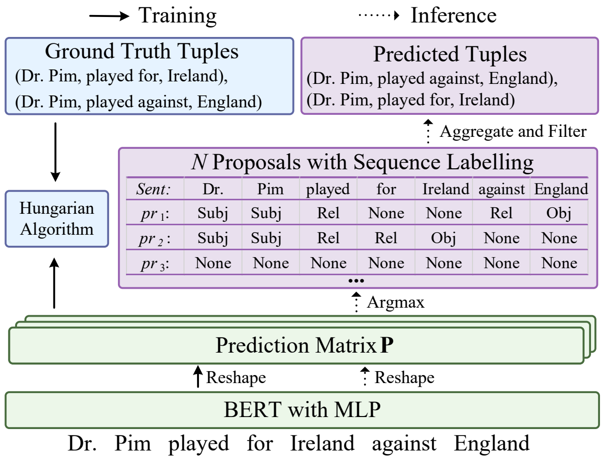


Figure 2: The architecture of the detection-based OIE model, which generates N proposals for every sentence and extracts tuples in parallel. The solid and dashed arrows indicate the training process and the inference process, respectively.

3 Method

In this section, we first introduce the process of the detection-based OIE model and then present how we formulate the label assignment process in OIE as an IoU-aware Optimal Transport (OT) problem.

3.1 Detection-based OIE Model

Inspired by one-stage anchor-based object detection methods in computer vision (Liu et al., 2016; Tan et al., 2020), Vasilkovsky et al. (2022) propose the DetIE model, which extracts tuples in parallel. The architecture is illustrated in Fig. 2. Formally, given an input sequence $\boldsymbol{W} = \{w_1, w_2, \dots w_T\}$ containing T words and this sequence has M golden tuples, DetIE leverages BERT (Devlin et al., 2019) as the backbone to get the encoded features. The output features are fed to a multilayer perceptron (MLP), and then transformed into a feature matrix $\mathbf{P} \in \mathbb{R}^{T \times N \times C}$, where N and C are the pre-defined number of possible extracted tuples and the number of classes, respectively. Every sentence will have N proposals (N is set as 20 in this experiment), where each proposal is a feature tensor implicitly representing a certain tuple.

During the training process, since the semantics of those proposals are implicit and it is hard to assign golden labels in advance, DetIE calculates the intersection-over-union (IoU) between the proposal (*pr*) and golden tuple (*gt*). The IoU matrix

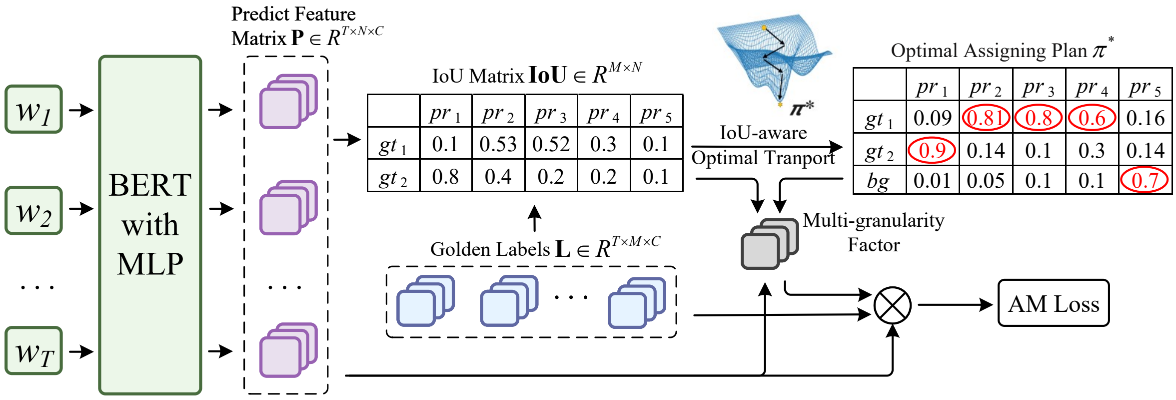


Figure 3: The workflow of IOT: (1) The sentence is encoded and reshaped into a feature matrix $\mathbf{P}$. (2) Get the IoU matrix from $\mathbf{P}$ and golden labels $\mathbf{L}$. (3) Utilize the IoU matrix as the input to solve the IoU-aware OT problem. (4) Leverage $\mathbf{P}$, $\mathbf{L}$, IoU matrix, and the optimal assigning plan $\boldsymbol{\pi}^*$ to calculate the AM loss.

$\mathbf{IoU} \in \mathbb{R}^{M \times N}$ is defined as:

$$\begin{aligned} IoU_{mn} &= \frac{I_{mn}}{U_{mn}} \\ I_{mn} &= \sum_{t,c} l_{tmc} \cdot p_{tnc} \\ U_{mn} &= \sum_{t,c} p_{tnc} + \sum_{t,c} l_{tmc} - I_{mn} \end{aligned} \tag{1}$$

where p_{tnc} is the element from the output feature tensor $\mathbf{P}$, and l_{tmc} is the element from ground truth matrix $\mathbf{L}$. Then, DetIE utilizes the Hungarian algorithm (Kuhn, 1955) that matches one proposal for each ground truth with the global minimum IoU.

During the inference stage, with the sequence labelling mechanism, the model leverages matrix $\mathbf{P}$ to classify each word of each proposal whether belonging to one of the pre-defined classes: *subject*, *relation*, *object*, or *None*, then assembles the predicted results into a set of tuples.

3.2 Background: Optimal Transport

The definition of OT is as follows: assuming there are M suppliers and N demanders in a certain area. The m-th supplier has s_m units of goods and the n-th demander needs d_n units of goods. The transporting cost for each unit of good from the m-th supplier to the n-th demander is c_{mn}. The goal of optimal transport is to find the best transportation plan $\boldsymbol{\pi}^* = \{\pi_{mn} \mid m = 1, 2, \ldots M, n = 1, 2, \ldots N\}$ that requires minimal transportation cost and all goods from suppliers can be transported to the demanders:

$$\begin{aligned} &\min_{\pi} \sum_{m=1}^{M} \sum_{n=1}^{N} c_{mn}\pi_{mn} \\ &\text{s.t.} \sum_{m=1}^{M} \pi_{mn} = d_n, \quad \sum_{n=1}^{N} \pi_{mn} = s_m, \\ &\sum_{m=1}^{M} s_m = \sum_{n=1}^{N} d_n, \quad \pi_{mn} \geq 0, \\ &m = 1, 2, \ldots M, n = 1, 2, \ldots N. \end{aligned} \tag{2}$$

OT is a linear program which can be solved in polynomial time. In this work, we leverage the Sinkhorn-Knopp Iteration (Cuturi, 2013) to solve this OT problem.

3.3 IoU-aware OT for Label Assignment

In the context of OIE, we view the ground truth tuples as suppliers and the model output proposals as demanders. Besides, we view the number of transporting goods as the number of labels. Supposing there are M golden tuples in a sentence and each one can provide $s_m = k$ units of labels. Likewise, let N be the pre-defined number of possible extracted tuples, and each one needs one unit of label (i.e., $d_n = 1$). In addition, the cost c_{mn} for transporting one unit of label could be defined as the negative IoU between the m-th *gt* and the n-th *pr*, which means $c_{mn} = -IoU_{mn}$.

In practice, some proposals are invalid to get tuples, so we introduce another supplier: *background*. The cost for transporting one unit from *background* to any *pr* is set as zero, i.e., $c^{bg} = \mathbf{0}$, and append to the last row of the cost matrix. Meanwhile, since the total supply should be equal to the total demand in standard OT problems, we make the background supplier owns $s_m = N - k * M$ units. The supplying function s could be formulated as:

$$s_m = \begin{cases} k, & \text{if} \quad m \leq M \\ N - M \times k, & \text{if} \quad m = M + 1 \end{cases} \tag{3}$$

Until now, we have obtained the cost matrix $\mathbf{C} \in \mathbb{R}^{(M+1)\times N}$, the supplying vector $\mathbf{s} \in \mathbb{R}^{M+1}$, and the demanding vector $\mathbf{d} \in \mathbb{R}^{N}$. Then we could find the assignment solution $\boldsymbol{\pi} \in \mathbb{R}^{(M+1)\times N}$ by applying the Sinkhorn-Knopp Iteration (Cuturi, 2013) to Eq. 2. According to the solution $\boldsymbol{\pi}$, we could get the optimal assignment plan $\boldsymbol{\pi}^*$ by assigning each *pr* with a *gt* (including *background*) that transports the largest amount of labels (i.e., $\boldsymbol{\pi}^* = \arg\max_{\mathrm{dim}=0}(\boldsymbol{\pi})$).

3.4 Dynamic *k* Strategy

A naive way for the OT problem is making each golden tuple *gt* has a fixed number of goods (i.e., setting k in Eq. 3 as a constant). However, those proposals with similar IoUs to the same golden label should be grouped together (e.g., $\{pr_2, pr_3, pr_4\}$ and $\{pr_1\}$ in Fig. 1 (c)). As a result, it is more reasonable to assign different golden tuples with different numbers of proposals.

To achieve this, we propose a simple but effective method, named dynamic *k* strategy, to roughly estimate how many proposals need to be assigned for each golden tuple. Specifically, given a sentence containing M golden tuples, for the m-th golden tuple, instead of using fixed k, we calculate the top q_m proposals according to IoU values. The top q_m number could be obtained by summing up the corresponding proposals' IoU values from the IoU matrix:

$$q_m = \lceil \sum_n IoU_{mn}) \rceil \tag{4}$$

where $\lceil . \rceil$ indicates rounding up to an integer. Such an estimation method is based on the following intuition: the value of q_m should be positively correlated with the number of proposals that are well-assigned by the IoU-aware OT.

3.5 AM Loss

Since there are numerous overlapping tuples in the OIE task, more than just utilizing word-level information for extraction, an OIE system should also regard tuples as wholes. As a result, we introduce an Assignment-guided Multi-granularity (AM) loss to weigh tuple proposals from the perspective of both word-level and tuple-level.

The intuition of AM is that the larger the predicted logits and the larger the IoU, the more accurately the tuple is predicted, thus the larger the training weights should be assigned. Specifically, according to the OT assignment solution $\boldsymbol{\pi}^*$, every proposal from each word would be assigned a golden label. We define a vector $\mathbf{o} \in \mathbb{R}^N$, where each element o_n represents the IoU value of every proposal and its matched golden label from the IoU matrix. For those proposals matched with the background, we set the IoU values as 0. The same IoU vector is leveraged for every word in the sentence, and the sentence assignment IoU matrix $\mathbf{O} \in \mathbb{R}^{T\times N}$ could be obtained by concatenating vector $\mathbf{o}$ over the words. Then, we define the multi-granularity factor ω_{tn} that considers both IoU values (tuple-level) and predicted logits (word-level):

$$\begin{aligned} a_{tn} &= o_{tn} \sum_c p_{tnc} * (1 - \log(p_{tnc})) \\ \omega_{tn} &= \| e^{\alpha \cdot a_{tn}} + \beta \|_2 \end{aligned} \tag{5}$$

where p_{tnc} is the element from the model logit feature $\mathbf{P}$, o_{tn} is the element from $\mathbf{O}$. α and β are hyper-parameters to control the degree of linearity. Besides, the multi-granularity factor ω_{tn} is normalized by the L2 normalization.

We utilize the multi-granularity factor ω_{tn} to dynamically control the whole classification process and finally get the AM loss as follows:

$$\mathrm{AM} = -\frac{1}{T} \sum_t^T \sum_n^N \sum_c^C \omega_{tn} l_{tnc} \log(p_{tnc}) \tag{6}$$

where l_{tnc} is the element from the ground truth label matrix $\mathbf{L}$.

3.6 Inference and Post-process

As illustrated in Fig. 2, during the inference stage, we need to aggregate the results from the proposals. Specifically, for each proposal of each word, the argmax operation is applied to get the predicted type. Moreover, we filter the aggregated results by restricting the output tuples that should contain subject, relation, and object simultaneously.

Additionally, since IOT is a many-to-one assignment strategy, one golden value could be assigned with multiple proposals during the training process. As a result, the model may produce multiple duplicate predictions. In order to solve this problem,

Algorithm 1 The procedure of IOT

Input: $\boldsymbol{W}$ is the input sentence
$\boldsymbol{L}$ is the ground truth label matrix
N is the pre-defined number of tuple proposals
M is the golden number of tuples;
Output: The predicted tuples set Ψ
// Training process
$\mathbf{P} \leftarrow$ Forward($\mathbf{W}$)
$\mathbf{IoU} \leftarrow (\mathbf{P}, \mathbf{L})$ based on Eq. 1
$s_m(m = 1, 2, \ldots, M) \leftarrow$ dynamic k strategy
$s_{m+1} = N - k * M \leftarrow$ *background* supplier
$\mathbf{C} = -\mathbf{IoU}$
Append c^{bg} to $\mathbf{C}$
$\boldsymbol{\pi} \leftarrow$ SinkhornKnopp($\mathbf{C}, \mathbf{s}, \mathbf{d}$) based on Eq. 2
$\boldsymbol{\pi}^* = \underset{\dim=0}{\arg\max}(\boldsymbol{\pi})$
$loss \leftarrow$ AM($\mathbf{P}, \mathbf{L}, \mathbf{IoU}, \boldsymbol{\pi}^*$) based on Eq. 6
// Inference process
$\Psi \leftarrow \arg\max(\mathbf{P})$, aggregate, filter, and post-process
return Ψ

we design a simple post-process method, which directly removes exactly the same prediction results at the inference time, obtaining the final set of tuples Ψ. The architecture of the proposed IOT is illustrated in Fig. 3. The detailed procedure of IOT is shown in Algorithm 1.

4 Experiment

4.1 Experiment Setup

Datasets and Evaluations. Following the experimental setup of Vasilkovsky et al. (2022), we conduct the experiments on two English benchmarks: IMoJIE (Kolluru et al., 2020b), LSOIE (Solawetz and Larson, 2021), and all the models are evaluated on LSOIE, CaRB (Bhardwaj et al., 2019) corpora. To validate the multi-lingual ability of the models, we also introduce a multilingual benchmark Synth (Vasilkovsky et al., 2022) for training, and evaluate the models on MultiOIE2016 (Ro et al., 2020b). The multilingual benchmarks contain data in English (EN), Spanish (ES), and Portuguese (PT). Table 1 gives the detailed data statistics. For evaluation, we adopt CaRB(1-1), CaRB, OIE16-C, Wire57-C as the evaluation metrics. Except for Wire57-C metric that only calculates F1 score, the others requires to report both F1 score and AUC. The AUC is measured by the area under receiver operating characteristic (ROC) curve, where the horizontal axis is the false positive rate (FPR) and the vertical axis is the true positive rate (TPR). As for those methods that does not provide confidence scores, the AUC values are approximated from a single TPR-FPR point.

Implementations. All the experiments are conducted with Pytorch Lightning[1] on one V100 GPU. We utilize *bert-base-multilingual-cased* weight from HuggingFace (Wolf et al., 2020). The batch size is set as 32, and the maximum number of training epochs is set to 100. We apply the early stop strategy to avoid over-fitting. We train the models with an Adam weight decay optimizer with an initial learning rate of 4e-5. The detection number N is set as 20. The Sinkhorn-Knopp iterates 50 times during label assignment. α and β for the AM loss are set as 1 and -0.5, respectively. The optimal hyper-parameters are obtained by grid search.

Split	Dataset	Sentences	Tuples
Train	IMoJIE	91,725	190,661
	LSOIE	34,780	100,862
	Synth	10,000	41,645
Test	LSOIE	7,900	17,459
	CaRB	641	2,715
	MultiOIE2016	595	1,508

Table 1: Dataset statistics, where MultiOIE2016 and Synth are multilingual datasets, and the numbers are given for each language.

4.2 Experimental Results

Baselines. We compare the proposed methods on the CaRB and LSOIE datasets with the state-of-the-art approaches, which include: (1) non-neural models: MinIE (Gashteovski et al., 2017); ClausIE (Corro and Gemulla, 2013); OIIIE (Mausam et al., 2012); ReVerb (Fader et al., 2011); OpenIE4 (Christensen et al., 2011); OpenIE5 (Saha et al., 2017; Saha and Mausam, 2018); (2) sequence labelling based methods: RnnOIE (Stanovsky et al., 2018); SenseOIE (Roy et al., 2019); SpanOIE (Zhan and Zhao, 2020); (3) generation based methods: NeuralOIE (Cui et al., 2018); IMoJIE (Kolluru et al., 2020b); OpenIE6 (Kolluru et al., 2020a); (4) detection-based model DetIE (Vasilkovsky et al., 2022), which could also be applied to the IGL-CA model in OpenIE6 with 'simplified' texts. Please note that we utilize the provided checkpoint from DetIE to reproduce the experiment, so the results may be different from the original paper.

Main Results. The experiment results on CaRB and LSOIE benchmarks are illustrated in Table 2 and Table 3, respectively. We can conclude that: (1) Compared to regular sequence labelling-based or sequence generation-based methods, detection-based models generally obtain better experiment

[1] https://github.com/Lightning-AI/lightning

Model	CaRB evaluation schemes							Speed(sent./sec)
	CaRB		CaRB(1-1)		OIE16-C		Wire57-C	
	F1	AUC	F1	AUC	F1	AUC	F1	
MinIE	41.0	-	38.4	-	52.3	-	28.5	8.9
ClausIE	45.0	22.0	40.2	17.7	61.0	38.0	33.2	4.0
OpenIE4	51.6	29.5	40.5	20.1	54.3	37.1	34.4	20.1
OpenIE5	48.0	25.0	42.7	20.6	59.9	39.9	35.4	3.1
SenseOIE	28.2	-	23.9	-	31.1	-	10.7	-
SpanOIE	48.5	-	37.9	-	54.0	-	31.9	19.4
RnnOIE	49.0	26.0	39.5	18.3	56.0	32.0	26.4	149.2
NeuralOIE	51.6	32.8	38.7	19.8	53.5	37.0	33.3	11.5
IMoJIE	53.5	33.3	41.4	22.2	56.8	39.6	36.0	2.6
IGL-OIE	52.4	33.7	41.1	22.9	55.0	36.0	34.9	142.0
CIGL-OIE	**54.0**	35.7	42.8	24.6	59.2	40.0	36.8	142.0
OpenIE6	52.7	33.7	**46.4**	26.8	65.6	48.4	**40.0**	31.7
DetIE(LSOIE)	42.2	26.4	31.0	16.7	48.7	31.8	29.3	**708.6**
DetIE(IMoJIE)	49.6	34.4	37.8	22.2	53.3	36.0	34.2	**708.6**
DetIE(LSOIE)+IGL-CA	36.6	27.0	34.0	22.7	62.4	46.8	30.4	112.2
DetIE(IMoJIE)+IGL-CA	44.2	33.3	40.5	27.8	66.0	51.0	35.5	112.2
IOT(LSOIE)	42.5	27.1	32.2	17.7	53.3	36.0	30.0	691.7
IOT(IMoJIE)	52.7	**37.0**	40.1	24.0	55.9	38.5	36.1	691.7
IOT(LSOIE)+IGL-CA	39.2	28.3	35.9	23.6	64.6	50.1	32.9	108.5
IOT(IMoJIE)+IGL-CA	48.3	35.8	43.7	**29.5**	**67.9**	**53.8**	38.4	108.5

Table 2: Experiment results on CaRB test set. Best results are shown in **bold**, and second bests are in underlined.

Model	Rec.	F1	AUC
OIIIIE	-	36.8	16.7
ReVerb	-	36.8	16.9
OpenIE4	-	54.6	32.3
OpenIE5	-	49.5	25.8
CIGL-OIE	-	59.7	48.0
OpenIE6	-	51.6	32.7
DetIE(IMoJIE)	47.0	46.3	34.2
DetIE(LSOIE)	59.5	62.7	50.4
DetIE(IMoJIE)+IGL-CA	47.5	35.1	30.4
DetIE(LSOIE)+IGL-CA	59.9	49.4	42.6
IOT(IMoJIE)	49.3	52.2	39.3
IOT(LSOIE)	**64.7**	**65.8**	**54.0**
IOT(IMoJIE)+IGL-CA	49.1	40.9	34.2
IOT(LSOIE)+IGL-CA	63.4	51.1	45.5

Table 3: Experiment results on LSOIE test set with the original CaRB evaluation scheme. 'Rec.' means recall.

results and more accelerated inference speed, illustrating the superiority of capturing features from multiple tuples in parallel. (2) Among the detection-based methods, IOT achieves the best or the second-best results while retaining the same amount of inference speed (the inference speed reduction is mainly due to the post-process). We attribute those improvements to that the many-to-one label assignment method from OT outperforms those in a one-to-one manner. Meanwhile, the tuple proposals could be effectively weighted by the AM loss. (3) For the experiments on the LSOIE test set, IOT surpasses strong baseline DetIE. Especially when trained with LSOIE training set, IOT obtains a gain of 3.1% and 3.6% on F1 and AUC, respectively. The improved ratio is larger than in CaRB dataset, indicating OT allows for better label assignment for datasets with the same annotation principles. (4) An interesting finding is that IOT model typically gets a higher recall than DetIE, which is consistent with our intuition that compared to those one-to-one label assignment methods, the many-to-one mechanism could find the results more comprehensively.

Experiment on MultiOIE2016 dataset. To validate the generality of the model on multilingual OIE situations, we also conduct experiments on the MultiOIE2016 dataset. There are strong baselines: (1) ArgOE (Gamallo and García, 2015); (2) PredPatt (White et al., 2016); (3) Multi2OIE (Ro et al., 2020b). From the experiment results in Table 4, we observe that IOT obtains the best F1 scores in all three languages, showing the generalizability of the model on the multilingual OIE task.

Ablation Study. To investigate the effect of each component, we conduct an ablation study by: (1) replacing the Hungarian algorithm (HA) with the optimal transport algorithm (OT); (2) replacing the cross entropy loss (CE) with the AM mechanism; (3) appending the dynamic k strategy (*dyk*); (4) appending the post-process (*post*). The results are listed in Table 5, and we observe: first, introducing different modules brought certain performance improvements to the model, proving the effectiveness of each module. Secondly, AM further promotes the model performance for HA and OT methods,

Lang.	Model	F1	Prec.	Rec.
En	ArgOE	43.4	56.6	35.2
	PredPatt	53.1	53.9	52.3
	Multi2OIE	69.3	66.9	71.7
	DetIE(IMoJIE)	76.6	90.3	66.5
	DetIE(IMoJIE+Synth)	77.7	89.2	68.8
	IOT(IMoJIE)	77.8	90.4	68.3
	IOT(IMoJIE+Synth)	**79.1**	**91.1**	**69.8**
PT	ArgOE	38.3	46.3	32.7
	PredPatt	42.9	43.6	42.3
	Multi2OIE	59.1	56.1	62.5
	DetIE(IMoJIE)	73.4	89.2	62.4
	DetIE(IMoJIE+Synth)	73.9	**89.4**	63.1
	IOT(IMoJIE)	73.0	86.9	62.9
	IOT(IMoJIE+Synth)	**74.3**	88.4	**64.1**
ES	ArgOE	39.4	48.0	33.4
	PredPatt	44.3	44.8	43.8
	Multi2OIE	60.2	59.1	61.2
	DetIE(IMoJIE)	74.3	**89.5**	63.6
	DetIE(IMoJIE+Synth)	74.5	88.3	64.5
	IOT(IMoJIE)	74.4	88.6	64.1
	IOT(IMoJIE+Synth)	**75.8**	88.8	**66.1**

Table 4: Binary extraction performance on Multi-OIE2016, which contains data in English (EN), Spanish (ES), and Portuguese (PT). 'Prec.', and 'Rec.' are short for precision and recall, respectively.

Model	F1	AUC
HA + CE	49.6	34.4
HA + AM	51.1	35.4
OT + CE	51.2	35.2
OT + CE + *dyk*	51.7	35.9
OT + AM + *dyk*	52.4	36.8
OT + AM + *dyk* + *post*	52.7	37.0

Table 5: Ablation study results on CaRB test set with the original CaRB evaluation scheme.

demonstrating the generality of AM.

Effectiveness of k. We conduct experiments to show the results under different settings of k, which controls how many labels each ground truth label supplies. As listed in Table 6, we could find that: when $k = 1$, the OT mechanism becomes the one-to-one assigning strategy, which is the same as in the Hungarian algorithm, and it does not perform well. When k rises from 1 to 3, the model performance improves. However, as k continues raising, the performance decreases instead. Especially when $k = 8$, the performance is even worse than $k = 1$. The reasons are as follows: since the total number of *prs* is constant (20 in our experiments), setting k too large will result in many *prs* matching the same *gt*, resulting in some *prs* are not assigned to the corresponding *gt*. However, as for the dynamic k strategy, the value of k is updated continuously with the change of IoU, thus obtaining a more promising result.

k	F1	AUC
k=1	51.7	35.7
k=2	52.1	36.5
k=3	52.3	36.6
k=4	52.0	36.1
k=8	50.2	33.8
dynamic k	**52.7**	**37.0**

Table 6: Analysis of various of k and dynamic k strategy on the CaRB test set and CaRB evaluation scheme.

Model	word	tuple	F1	AUC
OT + CE	✓		51.9	36.0
OT + DSC	✓		52.0	36.2
OT + weighted CE	✓		52.2	36.5
OT + focal	✓		52.3	36.4
OT + I-DSC	✓	✓	52.4	36.5
OT + I-focal	✓	✓	52.5	36.7
OT + AM	✓	✓	**52.7**	**37.0**

Table 7: Effectiveness of different re-weight functions on CaRB test set. All the experiments are conducted with dynamic k and post-process mechanisms. 'word' and 'tuple' indicate whether to utilize the word-level and tuple-level supervision, respectively.

Effectiveness of AM. We compare the proposed AM loss with several strong re-weight functions: (1) weighted CE loss (Ronneberger et al., 2015); (2) DSC loss (Li et al., 2020); (3) focal loss (Lin et al., 2020). Besides, we also apply the IoU fusion mechanism in AM to focal loss (I-focal) and DSC loss (I-DSC). The experiment results are listed in Table 7, and we observe that: among those strong re-weight baselines, because our AM mechanism considers IoU and is able to further calculate the weight of each sample from the tuple perspective, it achieves the best experimental result. Moreover, after integrating the IoU supervision on focal loss and DSC loss, the performance further improves. This finding illustrates the effectiveness and generality of the supervision from IoU at the tuple-level.

Error Analysis. Following the settings in Vasilkovsky et al. (2022) and Kolluru et al. (2020a), we conduct the error analyzing experiment by randomly sampling 100 sentences from the CaRB test set, and we identify four typical errors: (1) **Boundary identification**, which counts for 9%. This error means that the predicted results are partially mismatched with golden tuples due to punctuation, extra words (e.g., *and*, *to*, *of*), etc. (2) **Tuple missing**, which accounts for 6%. This error means some tuples are not extracted from the final predictions. (3) **Co-reference**, which accounts for 5%. This error means there are co-reference problems (e.g.,

Boundary identification	Sentence	32.7 % of all households were made up of individuals...
	Golden	(32.7 % of all households, were made, up of individuals)
	Prediction	(32.7 % of all households, were made up of, individuals)
Co-reference	Sentence	A Democrat , he became the youngest mayor...
	Golden	(he, became, the youngest mayor)
	Prediction	(A Democrat, became, the youngest mayor)
Tuple missing	Sentence	A cafeteria is also located on the sixth floor, a chapel on the 14th floor...
	Golden	(A cafeteria, is also located, on the sixth floor) (a chapel, is located, on the 14th floor)
	Prediction	(A cafeteria, is also located, on the sixth floor)
Misunderstand semantics	Sentence	After this point many of the republicans were arrested in Free State " round ups " when they had come out of hiding and returned home...
	Golden	(they, had come out of, hiding) (they, had returned, home)
	Prediction	(they, had come out of hiding and returned, home)

Table 8: Case study on CaRB test set, where there are four typical errors, including: boundary identification, co-reference, tuple missing, and misunderstand semantics.

he and *a democrat*). (4) **Misunderstand semantics**, which accounts for 3%. This error means the model does not fully understand the semantics of the text. Several typical error cases are shown in Table 8.

5 Conclusions

This paper introduces IOT, which formulates the OIE task as an IoU-aware optimal transport problem. Under this formulation, multiple implicit tuple proposals could be dynamically matched with the same golden label during training. To determine how many proposals should be assigned for each golden tuple, we design a dynamic k strategy based on the IoU values. Moreover, to weigh different tuple proposals, we introduce an Assignment-guided Multi-granularity (AM) loss, which takes both word-level and tuple-level information into account. Experiments show that our method surpasses strong state-of-the-art baselines on three benchmarks with a fast inference speed.

Limitations

Although the effectiveness of the IoU-aware optimal transport mechanism and the Assignment-guided Multi-granularity loss has been verified by empirical results on three datasets, the proposed IOT framework may still suffer from imprecise boundary identification and co-reference handling, as identified in our earlier discussion in error analysis. Moreover, IOT post-processing may bring additional computational cost during inference, although we did not observe significant loss in efficiency during our experiments on datasets in moderate scale. Considering IOT outperforms other detection-based methods in terms of performance and its inference speed is significantly better than other sequence labelling-based or sequence generation-based methods, we believe this cost is acceptable.

Acknowledgement

We thank the support of Natural Science Foundation of China under Grant 62206267. And we sincerely thank Nayu Liu, Jin-Ge Yao, and Yating Zhang for their constructive collaboration.

References

Sangnie Bhardwaj, Samarth Aggarwal, and Mausam. 2019. Carb: A crowdsourced benchmark for open IE. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 6261–6266. Association for Computational Linguistics.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2011. An analysis of open information extraction based on semantic role labeling. In *Proceedings of the 6th International Conference on Knowledge Capture (K-CAP 2011), June 26-29, 2011, Banff, Alberta, Canada*, pages 113–120. ACM.

Luciano Del Corro and Rainer Gemulla. 2013. Clausie: clause-based open information extraction. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 355–366. International World Wide Web Conferences Steering Committee / ACM.

Lei Cui, Furu Wei, and Ming Zhou. 2018. Neural open information extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 407–413. Association for Computational Linguistics.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1535–1545. ACL.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2019. Using local knowledge graph construction to scale seq2seq models to multi-document inputs. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 4184–4194. Association for Computational Linguistics.

Pablo Gamallo and Marcos García. 2015. Multilingual open information extraction. In *Progress in Artificial Intelligence - 17th Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Portugal, September 8-11, 2015. Proceedings*, volume 9273 of *Lecture Notes in Computer Science*, pages 711–722. Springer.

Kiril Gashteovski, Rainer Gemulla, and Luciano Del Corro. 2017. Minie: Minimizing facts in open information extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2630–2640. Association for Computational Linguistics.

Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Jie Zhou, and Maosong Sun. 2020. More data, more relations, more context and more openness: A review and outlook for relation extraction. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*, pages 745–758. Association for Computational Linguistics.

Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. 2020a. Openie6: Iterative grid labeling and coordination analysis for open information extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3748–3761. Association for Computational Linguistics.

Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Mausam, and Soumen Chakrabarti. 2020b. Imojie: Iterative memory-based joint open information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5871–5886. Association for Computational Linguistics.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Buyu Li, Yu Liu, and Xiaogang Wang. 2019. Gradient harmonized single-stage detector. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8577–8584. AAAI Press.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 465–476. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2020. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):318–327.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: single shot multibox detector. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, pages 21–37. Springer.

Mausam. 2016. Open information extraction systems and downstream applications. In *Proceedings of the*

Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016, pages 4074–4077. IJCAI/AAAI Press.

Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 523–534. ACL.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*, pages 565–571. IEEE Computer Society.

Youngbin Ro, Yukyung Lee, and Pilsung Kang. 2020a. Multi^{2}oie: Multilingual open information extraction based on multi-head attention with BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1107–1117. Association for Computational Linguistics.

Youngbin Ro, Yukyung Lee, and Pilsung Kang. 2020b. Multiˆ2oie: Multilingual open information extraction based on multi-head attention with BERT. *CoRR*, abs/2009.08128.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer.

Arpita Roy, Youngja Park, Taesung Lee, and Shimei Pan. 2019. Supervising unsupervised open information extraction models. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 728–737. Association for Computational Linguistics.

Swarnadeep Saha and Mausam. 2018. Open information extraction from conjunctive sentences. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2288–2299. Association for Computational Linguistics.

Swarnadeep Saha, Harinder Pal, and Mausam. 2017. Bootstrapping for numerical open IE. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 317–323. Association for Computational Linguistics.

Yongliang Shen, Xiaobin Wang, Zeqi Tan, Guangwei Xu, Pengjun Xie, Fei Huang, Weiming Lu, and Yueting Zhuang. 2022. Parallel instance query network for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 947–961. Association for Computational Linguistics.

Jacob Solawetz and Stefan Larson. 2021. LSOIE: A large-scale dataset for supervised open information extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2595–2600. Association for Computational Linguistics.

Gabriel Stanovsky, Ido Dagan, and Mausam. 2015. Open IE as an intermediate structure for semantic tasks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 303–308. The Association for Computer Linguistics.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 885–895. Association for Computational Linguistics.

Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, Xiangrong Zeng, and Shengping Liu. 2020. Joint entity and relation extraction with set prediction networks. *CoRR*, abs/2011.01675.

Mingxing Tan, Ruoming Pang, and Quoc V. Le. 2020. Efficientdet: Scalable and efficient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10778–10787. Computer Vision Foundation / IEEE.

Zeqi Tan, Yongliang Shen, Shuai Zhang, Weiming Lu, and Yueting Zhuang. 2021. A sequence-to-set network for nested named entity recognition. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3936–3942. ijcai.org.

Michael Vasilkovsky, Anton Alekseev, Valentin Malykh, Ilya Shenbin, Elena Tutubalina, Dmitriy Salikhov, Mikhail Stepnov, Andrey Chertok, and Sergey Nikolenko. 2022. Detie: Multilingual open information extraction inspired by object detection. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*.

Aaron Steven White, Dee Ann Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on universal dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1713–1723. The Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Alexander Yates, Michele Banko, Matthew Broadhead, Michael J. Cafarella, Oren Etzioni, and Stephen Soderland. 2007. Textrunner: Open information extraction on the web. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, April 22-27, 2007, Rochester, New York, USA*, pages 25–26. The Association for Computational Linguistics.

Junlang Zhan and Hai Zhao. 2020. Span model for open information extraction on accurate corpus. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9523–9530. AAAI Press.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In Limitaion Section

☑ A2. Did you discuss any potential risks of your work?
In Limitaion Section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
In Abstract and Introduction sections.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4.1

☑ B1. Did you cite the creators of artifacts you used?
Section 4.1

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. We use the public dataset.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4.1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4.1

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4.2

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4.1

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.