TRUE-UIE: Two Universal Relations Unify Information Extraction Tasks

Yucheng Wang*, Bowen Yu, Yilin Liu, Shudong Lu
Shanghai Artificial Intelligence Laboratory
Independent Researcher
University of Southern California
Beijing University of Posts and Telecommunications
{wychengsch,yubowen.ph}@gmail.com
yliu8258@usc.edu
lushudong@bupt.edu.cn

Abstract

Information extraction (IE) encounters challenges due to the variety of schemas and objectives that differ across tasks. Recent advancements hint at the potential for universal approaches to model such tasks, referred to as Universal Information Extraction (UIE). While handling diverse tasks in one model, their generalization is limited since they are actually learning task-specific knowledge. In this study, we introduce an innovative paradigm known as TRUE-UIE, wherein all IE tasks are aligned to learn the same goals: extracting mention spans and two universal relations named `NEXT` and `IS`. During the decoding process, the `NEXT` relation is utilized to group related elements, while the `IS` relation, in conjunction with structured language prompts, undertakes the role of type recognition. Additionally, we consider the sequential dependency of tokens during span extraction, an aspect often overlooked in prevalent models. Our empirical experiments indicate that TRUE-UIE achieves state-of-the-art performance on established benchmarks encompassing 16 datasets, spanning 7 diverse IE tasks. Further evaluations reveal that our approach effectively share knowledge between different IE tasks, showcasing significant transferability in zero-shot and few-shot scenarios.

1 Introduction

Information Extraction (IE) refers to the task of automatically extracting structured knowledge, including entities, relations, events, and sentiments, from unstructured textual data. The primary aim is to condense text into structured, machine-friendly formats, aiding downstream tasks such as question answering (Allam and Haggag, 2012) and sentiment analysis (Medhat et al., 2014).

In the era of Large Language Models (LLMs), structured knowledge enhances, validates, and grounds LLM outputs (Pan et al., 2023). Researchers are increasingly focusing on Universal Information Extraction (UIE), aiming to develop unified frameworks for various IE tasks. Two primary approaches have gained prominence: generative methods and linking-based methods. Generative methods generate a unified Structure Extraction Language to express various extraction targets (Lu et al., 2022; Cong et al., 2023). Linking-based methods, on the other hand, devise a set of directed token linking operations to break down information extraction tasks into multiple token pair labeling problems (Lou et al., 2023; Yan et al., 2023; Ping et al., 2023). Although both claim to be universal information extraction methods, We hold the belief that *a true UIE should maintain a uniform learning objective for all IE tasks, enabling comprehensive knowledge sharing.* Generative methods deviate from this criterion, generating specific structure languages for different IE tasks (Lu et al., 2022). For instance, structures generated for Named Entity Recognition tasks (NER)

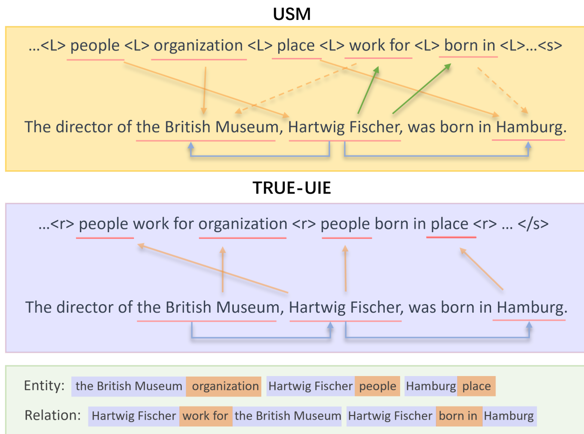

Figure 1: TRUE-UIE's superiority over USM: unifying its framework with (1) structure language prompts and (2) only two relations, `IS` (yellow) and `NEXT` (blue), circumventing the inconsistent learning objectives encountered by USM.

* Corresponding author

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1863–1876

lack the use of nesting"()", while those for relation and event extraction structures involve varying degrees of nesting "()". Existing linking-based methods also fail to meet this criterion. Take the prominent work USM (Lou et al., 2023) as an example (Figure 1): both the dashed and solid yellow arrows are defined identically but serve different purposes in NER and RE tasks. This leads to distinct learning objectives and limited knowledge sharing. Furthermore, the relations represented by green and blue arrows are only used in the RE task and receive no training in the NER task. Similar inconsistencies are evident in other linking-based methods (Yan et al., 2023; Ping et al., 2023). Additionally, all existing UIE methods face challenges in handling complex IE tasks, like discontinuous NER and open information extraction.

In this paper, we introduce **TRUE-UIE**, **T**wo **U**niversal **RE**lations **U**nify **I**nformation **E**xtraction Tasks, a novel approach distinguishes itself from prior work by modeling all information extraction tasks as a common task, with the aim of conducting two universal relation extractions. This achievement marks a paradigm shift towards the applicability of universal model outputs, moving away from outputs tailored to specific tasks. The success of TRUE-UIE hinges on two distinct designs: (1) **Structure Language Prompt**: The structured information of schemes is preserved, and placeholders for the `IS` relation are left for target mentions in the text. For instance, in the task of relation extraction, we organize prompts as *<subject type> <relation type> <object type>* as shown in Figure 1, in contrast to USM which separately enumerate entity and relation types. (2) **Only two relations are employed**: `IS` and `NEXT`. The `IS` relation aligns spans with corresponding placeholders in the prompt. As depicted in Figure 1, the entity "Hartwig Fischer" is linked to the entity type "people" in the triplet scheme *people work for organization*, indicating that "Hartwig Fischer" is involved in a relation of type *work for* and is categorized as "people". On the other hand, the `NEXT` relation establishes a connection between the current span and the subsequent span within the same structural knowledge instance. For instance, "Hartwig Fischer" is linked to "Hamburg" through the `NEXT` relation, indicating their membership in the same triplet. Using this approach, the `IS` relation is utilized to identify span types, while the `NEXT` relation groups these spans effectively. Additionally, this method tackles the challenge of a span appearing in several instances of the same knowledge type, a common challenge in overlapping relation extraction. (Wang et al., 2020). This is also why USM must employ the green relation in the RE task.

We conducted comprehensive experiments on 16 datasets covering 7 IE tasks, including flat NER, relation extraction, event extraction, sentiment extraction, nested NER, discontinuous NER, and open information extraction. These experiments demonstrate that TRUE-UIE surpasses both state-of-the-art task-specific and universal IE models across all datasets. Additionally, further zero-shot and few-shot experiments indicate that TRUE-UIE's universal relations enable more effective knowledge transfer across tasks.

2 Related Work

Information Extraction (IE) is the task of extracting relevant spans or tuples of spans from plain text. There are various specific IE tasks, including Flat/Nested/Discontinuous Named Entity Recognition (Nadeau and Sekine, 2007), Relation Extraction (Nasar et al., 2021), Event Extraction (Li et al., 2022b), Sentiment Extraction (Schouten and Frasincar, 2015), and Open Information Extraction (Zhou et al., 2022). For an extended period, researchers have focused on devising task-specific and independent methods to address these diverse IE tasks. However, in recent years, the emergence of pretraining techniques has sparked considerable interest in pretraining a versatile model capable of handling multiple IE tasks. Yan et al.2021b were the first to propose a universal approach to tackling different NER tasks. Yan et al.2021a unified various aspect-based sentiment analysis tasks. Lu et al.2022 introduced UIE, which employs a Structured Extraction Language to frame all IE tasks. Building upon UIE, Cong et al.2023 incorporated meta-pretraining to enhance the model's ability to extract complex structures. In contrast to UIE's use of a sequence-to-sequence structure to directly generate diverse target information structures, borrowing the idea from token pair linking (Wang et al., 2020, 2021b; Yu et al., 2022), USM (Lou et al., 2023) introduces three unified token linking operations to capture the skills of structuring and conceptualizing. Similarly, UTC-IE (Yan et al., 2023) decomposes several IE tasks into token pair classification tasks, utilizing the starting and ending tokens to locate spans, and using start-to-start and end-to-end token pairs to establish relations.

UniEX (Ping et al., 2023) also uniformly dissects all extraction objectives into joint span detection, classification, and association problems through a unified extractive framework. However, existing generative (Lu et al., 2022; Paolini et al., 2021; Cong et al., 2023; Wang et al., 2023) or token pair linking methods (Lou et al., 2023; Yan et al., 2023; Ping et al., 2023; Zhu et al., 2023) still struggle to unifying all Information Extraction (IE) tasks into a single learning objective, thus maximizing knowledge sharing and generalization. In contrast, our proposed True-UIE utilizes two universal relations to harmonize all tasks.

3 Methodology

Information extraction is the process of extracting knowledge from unstructured textual sources. The primary objective of UIE is to establish a single, universal model that can handle various information extraction tasks. The challenges of current SOTA method USM encompass two main dimensions: (1) Adapting the model to address the continually evolving complexities of information extraction, particularly in contexts where discontinuities and overlapping issues emerge; (2) Enhancing the model's generalization capabilities to ensure a broader degree of knowledge transferability and sharing across diverse tasks.

In this section, we begin by outlining the overall architecture and core principles of TRUE-UIE. Subsequently, we elucidate how TRUE-UIE addresses the aforementioned challenges. This entails two pivotal ideas: First, the introduction of a structural language prompt. By incorporating structured information from the schema into the prompt, we aim to enhance the model's comprehension of tasks and alleviate its learning burden. Second, utilizing two universal relational edges in conjunction with the structural prompt, we manage to unify seven IE tasks, transmuting them into a unified linking task with universal scheme. This strategy seeks to maximize the potential for knowledge to be shared seamlessly across tasks. Lastly, we introduce the main mathematical formulas and training objectives involved in the model.

3.1 The Overall Architecture

As illustrated in Figure 2, TRUE-UIE creates a structural prompt (enclosed in a purple dashed line) based on the extraction demands of the task, and concatenates it with the input text. The combined input is first passed through an encoder to obtain hidden states. These output hidden states are then processed by two fully connected layers, resulting in two distinct representations. Both representations are fed into the Semi-Matrix BiLSTM module and the Multiplicative Attention module. The operations of these two modules, shown on the right, produce presentations of spans and the corresponding relation scores. The span presentations are further used to compute the scores of spans through a fully connected layer.

3.2 Linking Scheme

Given an input text, TRUE-UIE combines the structure language prompt with the text to cater to varying extraction requirements. The adoption of this particular prompt arises from a notable distinction from previous work, where the structured information from the schema was not incorporated into the prompt. This forced the model to learn the intricate structure for each individual task. Regrettably, this knowledge could not be easily transferred across tasks, as each task possessed its unique structure.

The combined text is then input into the model, leading to the creation of a linking matrix that captures the relationships between tokens. In this framework, the `IS` relation aligns spans with their corresponding concept placeholders in the prompt, while the `NEXT` relation establishes a connection between a current span and the following span within the same instance of structural knowledge, such as within a triplet, an event, or an open fact. Next, we will provide a detailed presentation of the linking specifics for each IE task.

Relation Extraction: As illustrated in Figure 1, entity types and a relation type are amalgamated into a triplet prompt in the format of $<$*subject type*$>$ $<$*relation type*$>$ $<$*object type*$>$. Given that relation types often function as predicates, this design renders the prompt akin to a natural language expression, which facilitates semantic matching by the model. In cases of pure relation extraction where entity type annotations are absent, entity types default to "subject" or "object." When two utterance spans are connected by a `NEXT` relation and individually link to the subject type and object type surrounding the same relation type, a triplet is ascertained. Throughout this process, both entity types and the relation type are simultaneously determined. Even when a triplet involves multiple identified entity types, this decoding method does not introduce errors. Conversely, models with naive

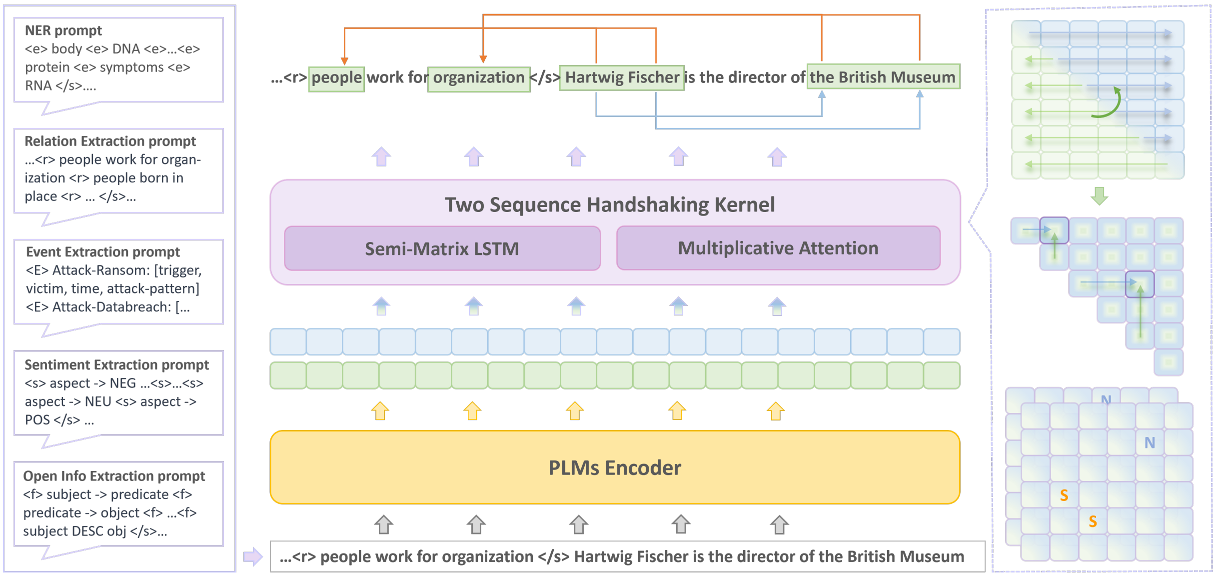

Figure 2: The overall architecture of TRUE-UIE.

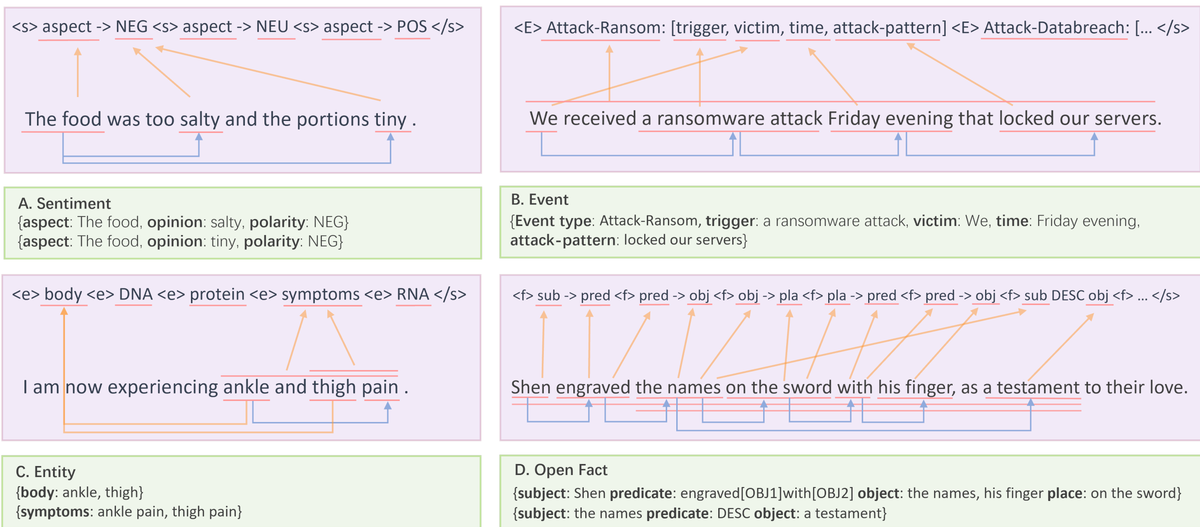

Figure 3: Unify different knowledge structures as two universal relations: `IS` (yellow lines) and `NEXT` (blue lines).

prompts struggle as they cannot discern which entity type(s) correspond to the recognized relation triplet, as they identify the entity type and relation type separately.

Sentiment Extraction: As illustrated in Figure 3.A, TRUE-UIE constructs a prompt for each sentiment type using the format *aspect* $\rightarrow$ $<$*polarity*$>$.This approach is analogous to relation extraction. When two spans are connected by a `NEXT` relation, and individually link to the "aspect" and the $<$polarity$>$ surrounding the same $\rightarrow$, a sentiment triplet is thereby determined.

Event Extraction: For representing an event, TRUE-UIE constructs a prompt using the format $<$*event type*$>$*: [argument role1, argument role2, ...]*, where the trigger is also considered as an argument, as depicted in Figure 3.B. During the decoding process, all spans that are linked to argument roles by the `IS` relation are grouped according to the preceding event type. Within the entire event span (indicated by the long red line above the text), only those paths that consist of argument spans sequentially linked by the `NEXT` relation and extending from one boundary to the other are outputted as individual event instances. Through this decoding logic, the model can effortlessly ascertain to which event type and trigger an argument span belongs, thereby smoothly resolving the event overlapping issue, where an argument may serve different roles within different instances of the same event type.

Conversely, models employing naive prompts grapple with this overlapping problem.

Nested and Discontinuous NER: For this task, TRUE-UIE employs a prompt similar to the naive one used in previous models. However, by utilizing the relation `NEXT`, TRUE-UIE gains the ability to handle discontinuous entities. Specifically, TRUE-UIE examines every span linked to an entity type to determine if there exists a continuous path within it, comprised of shorter spans, stretching from one boundary to the other. If such a path is found, it is output as a discontinuous entity, and the longer span is disregarded, as illustrated by *ankle pain* in Figure 3.C. If no path is found, the span is considered as a continuous entity. Additionally, if a short span is encompassed within a longer one without a connecting path, both are recognized as entities, reflecting a nested situation. An example of this is the term *thigh*, which appears within the spans *ankle and thigh pain* and *thigh pain*, but is not part of any path. As a result, *thigh* is identified as a *body* entity based on the `IS` relation, *thigh pain* is recognized as a *symptom* entity, and *ankle and thigh pain* is omitted, as previously described.

Open Information Extraction: This task involves identifying common role types such as subject, predicate, object, place, time, qualifier, etc., as demonstrated in Figure 3.D. This task faces challenges such as discontinuous arguments and role overlapping (e.g., "the names" serving as both object and subject). To tackle these complexities, TRUE-UIE uses the path decoding method with long spans and NEXT relations, as previously mentioned in discontinuous NER and event extraction. It avoids linking spans to a singular role through the `IS` relation, as this would not resolve the overlapping issue. Instead, TRUE-UIE recognizes roles in pairs like *<role1>* $\rightarrow$ *<role2>*, where two spans sequentially linked by `NEXT` and associated with role1 and role2 nearby the same $\rightarrow$ determine the roles. This ensures that every begin-to-end path within a long span is outputted as a fact instance. In situations where a predicate is missing, TRUE-UIE checks if subject and object spans are linked to predefined predicates, adding them to the fact instance if needed. An example of this can be found in the descriptive (DESC) fact in Figure 3.D.

3.3 Model Architecture

In previous linking-based UIE methods, span extraction often focuses only on the beginning and ending tokens of a span, neglecting the information embedded within the inner tokens. This can leave valuable sequential dependencies unexploited, particularly those crucial to the extraction of spans. In contrast, TRUE-UIE explicitly utilizes all tokens within a span. By employing semi-matrix LSTM operations to efficiently embeds this information into the span features. Given a sequence of n tokens $[t_1, \ldots, t_n]$, each token t_i is initially transformed into a low-dimensional contextual vector h_j utilizing a pretrained language model encoder such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019). Subsequently, two distinct representations, h_j^b and h_j^e, are computed to serve as features, specifically denoting the beginning and ending tokens of span boundaries:

$$h_j^b = W_b \cdot h_j + b_b, \tag{1}$$

$$h_j^e = W_e \cdot h_j + b_e. \tag{2}$$

Herein, W_* represents a parameter matrix, and b_* is a bias vector, both of which are subject to optimization during the training process.

For both h^b and h^e, TRUE-UIE constructs two matrices B and E by repeating each vector n times, each of dimensions $n \times n$, where n is the number of tokens. Next, TRUE-UIE employs a forward LSTM to encode the upper triangular region of E and a backward LSTM to encode the lower triangular region of B. The result is two new matrices B' and E', both of dimensions $n \times n$. In these matrices, the element $B'_{i,j}$ comprises the sequential information extending from token j to token i, while the corresponding element $E'_{i,j}$ embodies the sequential information extending from token i to token j. Subsequently, TRUE-UIE transposes B', and the sum of B' and E' yields a new matrix, denoted as S, where only the upper triangular region is saved, and the element $S_{i,j}$ encompasses the sequential information from token i to j as well as from j to i. This structured transformation facilitates TRUE-UIE's capacity to discern intricate dependencies between the tokens, thereby aligning with the overarching objective of span extraction. The mathematical formulations for scoring a span are provided as follows:

$$S_{i,j} = BiLSTM([h_i, \ldots, h_j]), \tag{3}$$

$$s_{i,j}^p = W_s \cdot S_{i,j} + b_s. \tag{4}$$

Herein, the $BiLSTM$ serves as a succinct expression for encoding the sequential information mentioned above. The score $s_{i,j}^p$ represents the output score for the span extending from token i to token j.

Additionally, when decoding the relations between two spans, a relation (IS or NEXT) is determined to exist only if both the beginning and ending tokens of the spans share this relation. TRUE-UIE adopts a multiplicative attention operation to fuse the features of these token pairs, feeding the integrated information to relation scorers:

$$s^*_{i,j} = h^*_i \cdot h^{*T}_j, \tag{5}$$

where h^* denotes the previously described features associated with the span boundaries, as expressed in Equations 1 and 2, the asterisk (*) symbolizes either b for beginning or e for ending of a span. The score $s^*_{i,j}$ signifies the relation score between the two boundary tokens i and j.

3.4 Learning Objective

The training process encounters a class imbalance issue, where the relation IS tends to occur more frequently than NEXT across all tasks. This disproportion is particularly pronounced in NER tasks, where discontinuous entities make up a small proportion, resulting in the relative sparsity of the NEXT relationship. To address this challenge, following USM (Lou et al., 2023), we implement optimization on class imbalance loss (Su et al., 2022):

$$L = \sum_{t\in T} \log\left(1 + \sum_{(i,j)\in t^+} e^{-s^*_{(i,j)}}\right) \tag{6}$$

$$+ \log\left(1 + \sum_{(i,j)\in t^-} e^{s^*_{(i,j)}}\right) \tag{7}$$

In this part, let T denote the set of label types, where t^+ corresponds to the target class, and t^- represents the non-target class. In this context, $s^*_{(i,j)}$ designates the scores as defined in Equations 4 and 5, with the asterisk (*) symbol taking on the values p for a span, b for the beginning pair, and e for the ending pair.

4 Experiment

In this section, comprehensive experiments are undertaken in both the supervised setting and few-shot/zero-shot scenarios. We also provide ablation study on each component of TRUE-UIE in Appendix.

4.1 Experimental Setup

In the supervised setting, we conduct experiments across 4 information extraction tasks commonly utilized in previous research (Yan et al., 2023; Lou et al., 2023; Ping et al., 2023), including namely, flat named entity recognition, relation extraction, event extraction, and sentiment extraction. Moreover, to further substantiate TRUE-UIE's scalability and effectiveness, we have added 3 additional tasks (nested, discontinuous named entity recognition, and open information extraction). Thus, this part of the experimentation covers seven information extraction tasks and utilizes 16 publicly available benchmark datasets only for research purposes, consistent with their intended use. The datasets employed include ACE04 (Mitchell et al., 2005), ACE05 (Walker et al., 2006); CoNLL03 (Sang and De Meulder, 2003), GENIA (Kim et al., 2003), Cadec (Karimi et al., 2015), CoNLL04 (Roth and Yih, 2004), SciERC (Luan et al., 2018), NYT (Riedel et al., 2010), CASIE (Satyapanich et al., 2020), SemEval-14/15/16 (Pontiki et al., 2014, 2015, 2016), and Saoke (Sun et al., 2018). The evaluation metrics align with those employed by Lu et al. (2022).

We primarily contrast TRUE-UIE with the previous SOTA model, USM (Lou et al., 2023), adhering to the same settings they employ for experiments. During the pretraining phase, we follow USM to use three corpus:

- D_{task} refers to Ontonotes (Pradhan et al., 2013), a widely used IE dataset. Each instance comes with a gold annotation, enabling the acquisition of in-task knowledge.
- D_{dist} represents the datasets obtained through distant supervision, wherein each instance aligns the text with Wikidata and Freebase (Cabot and Navigli, 2021; Riedel et al., 2013). Distant supervision is employed to gather large-scale training signals (Mintz et al., 2009), supplementing in-task supervised signals.
- D_{ind} denotes the indirect supervision dataset, comprising instances derived from sources outside the IE tasks. Following the USM setting, we leverage comprehension datasets from MRQA (Fisch et al., 2019) to offer a more enriched label semantic context for pretraining. Within this setting, questions are treated as labels.

Dataset	Tailored Model		UIE	UniEX	UTC-IE	USM*	USM†	USM^u	TRUE*	TRUE†	TRUE^u
ACE04	P-NER	88.72	86.89	87.12	87.54	87.79	87.62	87.34	88.92	89.34	**89.91**
ACE05-Ent	P-NER	88.26	85.78	87.02	87.75	86.98	87.14	-	88.31	**90.10**	-
CoNLL03	BS	93.65	92.99	92.65	93.45	92.76	93.16	92.97	92.88	93.51	**94.13**
Genia	PIQN	81.77	-	76.69$^-$	80.45	-	-	-	80.46	81.83	**82.56**
Cadec	W^{2}NER	73.21	-	-	-	-	-	-	72.06	73.25	**73.83**
Cadec$_D$	Mac	44.40	-	-	-	-	-	-	46.31	47.15	**47.51**
ACE05-Rel	PURE	69.40	66.06	66.06	67.79$^+$	66.54	67.88	-	67.93	**70.84**	-
CoNLL04	REBEL	75.40	75.00	73.40	-	75.40	75.86	78.84	73.05	77.84	**78.94**
NYT	UniRel	93.70	93.54	-	-	93.96	94.07	94.01	93.98	94.33	**94.83**
SciERC	PFN	38.40	36.53	38.00	38.77$^+$	37.05	37.36	37.42	37.40	38.06	**38.85**
ACE05-Evt$_T$	QE	73.60	73.36	74.08	73.44$^+$	71.68	72.41	72.31	72.51	74.63	**76.42**
ACE05-Evt$_A$	QE	55.10	54.79	53.92	57.68$^+$	55.37	55.83	55.57	55.21	56.41	**56.81**
CASIE$_T$	Txt2Evt	68.98	69.33	71.46	-	70.77	71.73	71.56	71.32	72.53	**73.02**
CASIE$_A$	Txt2Evt	60.37	61.30	62.91	-	63.05	63.26	63.00	62.78	63.66	**63.90**
14-res	GAS	72.16	74.52	74.77	-	76.35	77.26	77.29	77.11	77.82	**78.13**
14-lap	GAS	60.78	63.88	65.23	-	65.46	65.51	66.60	66.03	66.94	**67.07**
15-res	Sp-ASTE	63.27	67.15	68.58	-	68.80	69.86	-	69.92	**70.78**	-
16-res	Sp-ASTE	70.26	75.07	76.02	-	76.73	78.25	-	77.76	**78.83**	-
SAOKE	DragonIE	46.10	-	-	-	-	-	-	43.34	46.51	**47.11**

Table 1: The main results in the supervised setting. TRUE-UIE employs RoBERTa-large for English tasks and employs XLM-RoBERTa-large for SAOKE, as the latter needs to be trained on both Chinese and English datasets. The symbol $\star$ indicates that the model is initialized from the original pre-trained language model, $\dagger$ and ^{u} separately denote the models that were pre-trained on $D_{task,dist,ind}$ and fine-tuned on a single task and multi-task except for overlapped datasets: ACE05-Ent/Rel and 15/16-res. The symbol $+$ is used to represent results derived from models that are domain-specific or larger in size compared to RoBERTa-large. Cadec$_D$ refers to the subset of entities that are discontinuous.

Unseen/All	10/12	7/9	6/7	8/9	7/8	8/9	4/5	12/17	Avg	Improv
D_{task}	32.1/ **33.9**	2.5/ **4.3**	1.6/ **2.8**	10.7/ **12.2**	52.4/ **53.9**	45.9/ **47.4**	11.2/ **12.7**	14.1/ **15.4**	21.3/ **23.1**	+ 1.8
$D_{task,ind}$	39.8/ **41.9**	14.7/ **16.2**	20.6/ **22.5**	24.1/ **26.1**	56.2/ **57.9**	44.2/ **46.1**	32.9/ **34.5**	44.3/ **45.9**	34.6/ **36.3**	+ 1.7
$D_{task,dist}$	35.4/ **38.6**	21.1/ **24.2**	40.6/ **43.0**	27.6/ **30.3**	57.0/ **60.2**	49.3/ **52.1**	43.7/ **46.1**	44.1/ **47.3**	39.8/ **42.7**	+ 2.9
$D_{task,ind,dist}$	42.1/ **45.3**	26.0/ **29.1**	44.4/ **47.3**	34.9/ **38.1**	65.7/ **68.9**	60.1/ **63.1**	56.7/ **59.9**	55.3/ **58.5**	48.1/ **51.3**	+ 3.2
Δ	10.0/ **11.4**	23.5/ **24.8**	42.7/ **44.5**	24.2/ **25.9**	13.3/ **15.0**	14.1/ **15.7**	45.5/ **47.2**	41.1/ **43.1**	26.8/ **28.2**	-

Table 2: Comparison of zero-shot transfer performance on unseen entity label subset with different supervision signals between USM and TRUE-UIE, with two scores separated by "/". "Unseen" indicates label types that do not appear in the pre-training dataset. "Avg" represents average scores under pretraining; "Improv" indicates average improvement against USM; Δ signifies the enhancement difference from $D_{task,ind,dist}$ to D_{task}.

Placeholder	CoNLL04	Model Size
GPT-3	18.10	175B
DEEPSTRUCT	25.80	10B
USM	25.95	356M
TRUE-UIE	**27.13**	374M

Table 3: Zero-shot performance on relation extraction.

In addition to USM, we also make comparisons with two other linking-based UIE models (Yan et al., 2023; Ping et al., 2023) and a Generative UIE model (Lu et al., 2022). Towards providing a thorough evaluation of TRUE-UIE's performance relative to contemporary approaches, task-tailored models are also in comparison: PIQN (Shen et al., 2022), W^{2}NER (Li et al., 2022a), Mac (Wang et al., 2021b), Txt2Evt (Lu et al., 2021), PURE (Zhong and Chen, 2021), DragonIE (Yu et al., 2022), BS (Zhu and Li, 2022), P-NER (Shen et al., 2023), REBEL (Huguet Cabot and Navigli, 2021), UniRel (Tang et al., 2022), PFN (Yan et al., 2021c), QE (Wang et al., 2021a), GAS (Zhang et al., 2021), Sp-ASTE (Xu et al., 2021).

For additional details regarding the datasets, metrics, and training implementation, please consult Appendix A.

4.2 Experiments in the Supervised Setting

Table1 presents the performance of TRUE-UIE and strong baselines. Through the observation of experimental results, we identify several advantages of the TRUE-UIE framework, setting new state-of-the-art in the field of UIE.

1) TRUE-UIE offers a universal design that facilitates seamless sharing of learned knowledge across tasks. USM's decline in performance on several datasets after multi-task training (USM† vs. USM^{u}) suggests that its design may hinder proper knowledge sharing across tasks, potentially leading to conflicts among them. TRUE-UIE overcomes this by transforming multi-tasks into a unified common task, demonstrating more stable growth under the same experimental settings (TRUE† vs. TRUE^{u}). 2) TRUE-UIE is not merely a more universal framework but also exhibits a strong advantage in initial performance before pretraining. It surpasses other pretrained UIE methods even before pre-training. Particularly in NER tasks, where TRUE-UIE's prompt and linking style are almost identical to USM's design, it still significantly outperforms USM on various datasets. This improvement is attributed to the token sequential information embedded in the span features, which, apart from the prompt and linking style, is the main distinction from USM. 3) TRUE-UIE showcases the ability to tackle discontinuous and overlapping issues, a capability lacking in earlier linking-based UIE models. Although the initial performance of TRUE-UIE falls short of task-specific state-of-the-art models, after pre-training, it attains improvements of 3.11 on Cadec$_{D}$ and 1.01 on SAOKE, respectively. TRUE-UIE's universal design, prioritizing overall performance across all tasks, explains why it might not excel in specific tasks without prior pre-training. 4) It is noteworthy that after multi-task fine-tuning on English datasets, TRUE-UIE demonstrates a slight improvement on SAOKE (+0.6), a Chinese dataset. This reveals TRUE-UIE's promising ability to generalize knowledge across languages.

4.3 Experiments in the Zero-shot Setting

In zero-shot NER setting, aligned with USM, TRUE-UIE is trained using 4 different combinations of pretraining datasets and then evaluated across 8 diverse NER datasets (Liu et al., 2013; Strauss et al., 2016; Liu et al., 2021). As illustrated in Table 2, in four pre-training settings, TRUE-UIE consistently outperforms USM across all datasets, highlighting its strong zero-shot transferability across various domains. This shows a more robust generalization capability than USM. Moreover, comparative analysis reveals a notable expansion in the performance growth gap for TRUE-UIE under the $D_{task,dist}$ and $D_{task,ind,dist}$ configurations, with average improvements of 2.9 and 3.2 percentage points over USM, respectively. This indicates that TRUE-UIE can adeptly generalize knowledge learned from relation extraction tasks to NER tasks within pre-training settings involving D_{dist}, despite the absence of annotated entity types.

Regarding zero-shot relation extraction, following USM, TRUE-UIE is trained on all available pre-training datasets, and benchmarked against GPT-3 175B (Brown et al., 2020) and DEEPSTRUCT 10B (Wang et al., 2022) on the Conll04 dataset. As shown in Table 3, despite having a smaller model size, TRUE-UIE not only surpasses robust zero-shot baselines such as GPT-3 and DEEPSTRUCTURE, but also demonstrates competitive performance compared to USM, which is of a comparable size. These findings robustly affirm the efficacy of the TRUE-UIE framework. Compared to multi-task models like USM, common task models manifest a superior capacity for generalization.

4.4 Experiments in the Few-shot Setting

In our few-shot transfer experiments, we followed the data preprocessing and experimental settings from previous studies (Lu et al., 2022; Lou et al., 2023). Table 4 shows the performance of 7 IE tasks in few-shot scenarios, with the average results from 1/5/10-shot experiments labeled as "Avg." TRUE-UIE*, representing the initial model without IE pre-training, is used as the baseline for discontinuous NER and Open IE tasks where UIE and USM are not applicable. The results indicate that TRUE-UIE outperforms both baseline models, achieving an average improvement of 6.29 and 1.17 on the first five datasets. This suggests a superior generalization ability over the other two baseline models. Moreover, TRUE-UIE surpasses its preliminary model, TRUE-UIE*, by an average score of 14.46 for the final three tasks. This demonstrates that TRUE-UIE is not only capable of expanding to more complex IE tasks but also effectively generalizes the knowledge gained during pretraining to novel tasks. These remarkable results stem from its architecture, which models IE tasks as a shared task using two universal relation extraction processes, maximizing knowledge sharing and robust scalability for vari-

Title	Model	1-Shot	5-Shot	10-Shot	Avg.
CoNLL03	UIE	57.53	75.32	79.12	70.66
	USM	71.11	83.25	84.58	79.65
	TRUE-UIE	**73.56**	**84.78**	**85.66**	**81.33**
CoNLL04	UIE	34.88	51.64	58.98	48.50
	USM	36.17	53.20	60.99	50.12
	TRUE-UIE	**36.77**	**53.94**	**62.21**	**50.97**
ACE05-Evt (trigger)	UIE	42.37	53.07	54.35	49.93
	USM	40.86	55.61	58.79	51.75
	TRUE-UIE	**41.33**	**56.88**	**59.93**	**52.71**
ACE05-Evt (argument)	UIE	14.56	31.20	35.19	26.98
	USM	19.01	36.69	42.48	32.73
	TRUE-UIE	**19.64**	**37.10**	**43.55**	**33.43**
Sentiment (16res)	UIE	23.04	42.67	53.28	39.66
	USM	30.81	52.06	58.29	47.05
	TRUE-UIE	**32.03**	**54.02**	**60.12**	**48.72**
Genia	TRUE-UIE*	6.10	29.33	33.44	22.96
	TRUE-UIE	**37.34**	**55.54**	**57.97**	**50.28**
Cadec_D	TRUE-UIE*	2.01	9.63	15.81	9.15
	TRUE-UIE	**10.17**	**20.13**	**27.64**	**19.31**
SAOKE	TRUE-UIE*	2.32	5.74	7.61	5.22
	TRUE-UIE	**5.61**	**10.34**	**17.44**	**11.13**

Table 4: Comparison of few-shot perfromace across various tasks. TRUE-UIE* indicates that the model is initialized from the original pre-trained language model.

ous tasks. Contrastingly, UIE's need to learn varied schema structure languages leads to a large decoding search space and restricted knowledge sharing, presenting substantial learning challenges in low-resource settings. While USM reduces this search space via semantic matching, it fails to learn more universal relations, resulting in varied knowledge acquisition across tasks.

5 Conclusion

In this study, we've introduced an innovative approach called **TRUE-UIE**, which presents a unified framework for various information extraction (IE) tasks. By leveraging only two universal relations, namely `IS` and `NEXT`, we have established a consistent methodology across all IE tasks. This ensures that all components and definitions within the method remain uniform for different IE tasks, and can be applied to tasks such as discontinuous NER and open information extraction that are challenging for existing top-performing methods. The experimental results demonstrate that TRUE-UIE achieves state-of-the-art performance across 7 IE tasks and 16 datasets. It also showcases robust generalization capabilities in scenarios involving zero-shot and few-shot transfers. Notably, TRUE-UIE offers both adaptable task scalability and the ability to seamlessly transfer pre-trained knowledge to novel tasks. We hope that TRUE-UIE can drive further development in the field of UIE to better explore the relevant knowledge between tasks.

References

Ali Mohamed Nabil Allam and Mohamed Hassan Haggag. 2012. The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, 2(3).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381.

Xin Cong, Bowen Yu, Mengcheng Fang, Tingwen Liu, Haiyang Yu, Zhongkai Hu, Fei Huang, Yongbin Li, and Bin Wang. 2023. Universal information extraction with meta-pretrained self-retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4084–4100.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. *arXiv preprint arXiv:1910.09753*.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022a. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10965–10973.

Qian Li, Jianxin Li, Jiawei Sheng, Shiyao Cui, Jia Wu, Yiming Hei, Hao Peng, Shu Guo, Lihong Wang, Amin Beheshti, et al. 2022b. A survey on deep learning event extraction: Approaches and applications. *IEEE Transactions on Neural Networks and Learning Systems*.

Jingjing Liu, Panupong Pasupat, Scott Cyphers, and Jim Glass. 2013. Asgard: A portable architecture for multilingual dialogue systems. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8386–8390. IEEE.

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

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13452–13460.

Jie Lou, Yaojie Lu, Dai Dai, Wei Jia, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2023. Universal information extraction as unified semantic matching. *arXiv preprint arXiv:2301.03282*.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event 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 1: Long Papers)*, pages 2795–2806, Online. Association for Computational Linguistics.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv:1808.09602*.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Mike Mintz, Steven Bills, Rion Snow, and Dan 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.

Alexis Mitchell et al. 2005. Ace 2004 multilingual training corpus ldc2005t09. Web Download.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2021. Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys (CSUR)*, 54(1):1–39.

Jeff Z Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, et al. 2023. Large language models and knowledge graphs: Opportunities and challenges. *arXiv preprint arXiv:2308.06374*.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *arXiv preprint arXiv:2101.05779*.

Yang Ping, JunYu Lu, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Pingjian Zhang, and Jiaxing Zhang. 2023. UniEX: An effective and efficient framework for unified information extraction via a span-extractive perspective. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16424–16440, Toronto, Canada. Association for Computational Linguistics.

Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop* *on semantic evaluation (SemEval 2015)*, pages 486–495.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*, pages 19–30. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III 21*, pages 148–163. Springer.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 74–84.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the eighth conference on computational natural language learning (CoNLL-2004) at HLT-NAACL 2004*, pages 1–8.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Taneeya Satyapanich, Francis Ferraro, and Tim Finin. 2020. Casie: Extracting cybersecurity event information from text. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8749–8757.

Kim Schouten and Flavius Frasincar. 2015. Survey on aspect-level sentiment analysis. *IEEE transactions on knowledge and data engineering*, 28(3):813–830.

Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. 2023. Promptner: Prompt locating and typing for named entity recognition.

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)*, pages 947–961, Dublin, Ireland. Association for Computational Linguistics.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine De Marneffe, and Wei Xu. 2016. Results of the wnut16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144.

Jianlin Su, Mingren Zhu, Ahmed Murtadha, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2022. Zlpr: A novel loss for multi-label classification. *arXiv preprint arXiv:2208.02955*.

Mingming Sun, Xu Li, Xin Wang, Miao Fan, Yue Feng, and Ping Li. 2018. Logician: A unified end-to-end neural approach for open-domain information extraction. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 556–564.

Wei Tang, Benfeng Xu, Yuyue Zhao, Zhendong Mao, Yifeng Liu, Yong Liao, and Haiyong Xie. 2022. UniRel: Unified representation and interaction for joint relational triple extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7087–7099, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Christopher Walker et al. 2006. Ace 2005 multilingual training corpus ldc2006t06. Web Download.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022. Deepstruct: Pretraining of language models for structure prediction. *arXiv preprint arXiv:2205.10475*.

Sijia Wang, Mo Yu, Shiyu Chang, Lichao Sun, and Lifu Huang. 2021a. Query and extract: Refining event extraction as type-oriented binary decoding. *arXiv preprint arXiv:2110.07476*.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023. Instructuie: multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1572–1582.

Yucheng Wang, Bowen Yu, Hongsong Zhu, Tingwen Liu, Nan Yu, and Limin Sun. 2021b. Discontinuous named entity recognition as maximal clique discovery. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 764–774, Online. Association for Computational Linguistics.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. Learning span-level interactions for aspect sentiment triplet 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 1: Long Papers)*, pages 4755–4766, Online. Association for Computational Linguistics.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021a. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2416–2429.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021b. A unified generative framework for various ner subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5808–5822.

Hang Yan, Yu Sun, Xiaonan Li, Yunhua Zhou, Xuanjing Huang, and Xipeng Qiu. 2023. UTC-IE: A unified token-pair classification architecture for information extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4096–4122, Toronto, Canada. Association for Computational Linguistics.

Zhiheng Yan, Chong Zhang, Jinlan Fu, Qi Zhang, and Zhongyu Wei. 2021c. A partition filter network for joint entity and relation extraction. *arXiv preprint arXiv:2108.12202*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Bowen Yu, Zhenyu Zhang, Jingyang Li, Haiyang Yu, Tingwen Liu, Jian Sun, Yongbin Li, and Bin Wang. 2022. Towards generalized open information extraction. *arXiv preprint arXiv:2211.15987*.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. 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 504–510, Online. Association for Computational Linguistics.

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

Shaowen Zhou, Bowen Yu, Aixin Sun, Cheng Long, Jingyang Li, Haiyang Yu, Jian Sun, and Yongbin Li. 2022. A survey on neural open information extraction: Current status and future directions. *arXiv preprint arXiv:2205.11725*.

Enwei Zhu and Jinpeng Li. 2022. Boundary smoothing for named entity recognition. *arXiv preprint arXiv:2204.12031*.

Tong Zhu, Junfei Ren, Zijian Yu, Mengsong Wu, Guoliang Zhang, Xiaoye Qu, Wenliang Chen, Zhefeng Wang, Baoxing Huai, and Min Zhang. 2023. Mirror: A universal framework for various information extraction tasks. *arXiv preprint arXiv:2311.05419*.

A More Dataset Details

A.1 Datasets for Evaluation

We carry out evaluations on 7 information extraction tasks, spanning 16 distinct datasets. Comprehensive statistics for each of these datasets are presented in Table 5. We follow the pre-processing steps and data split of previous works (Lu et al., 2022; Lou et al., 2023).

A.2 Datasets for Pretraining

Details regarding the pretraining datasets are outlined as follows:

- For D_{task}, all 60K samples are utilized.
- D_{dist} consists of 356K samples. From this, the Rebel dataset is narrowed down to the 230 most frequently occurring relation types, and 300K instances are randomly selected for pretraining, in accordance with Lou et al. (2023).
- D_{ind} contains 195K samples, drawn from several datasets: HotpotQA (Yang et al., 2018), Natural Questions (Kwiatkowski et al., 2019), NewsQA (Trischler et al., 2016), SQuAD (Rajpurkar et al., 2016), and TriviaQA (Joshi et al., 2017). For each instance, the selection is restricted to a maximum of 5 questions, and

Datasets	Ent	Rel/Pol	Evt	#Train	#Val	#Test
ACE04	7	-	-	6,202	745	812
ACE05-Ent	7	-	-	7,299	971	1,060
CoNLL03	4	-	-	14,041	3,250	3,453
Genia	5	-	-	16,692	1,854	1,854
Cadec	1	-	-	5,340	1,097	1,160
ACE05-Rel	7	6	-	10,051	2,420	2,050
CoNLL04	4	5	-	922	231	288
NYT	3	24	-	56,196	5,000	5,000
SciERC	6	7	-	1,861	275	551
ACE05-Evt	7	-	33	19,216	901	676
CASIE	21	-	5	11,189	1,778	3,208
14res	2	3	-	1,266	310	492
14lap	2	3	-	906	219	328
15res	2	3	-	605	148	322
16res	2	3	-	857	210	326
SAOKE	6	7	-	37,544	4,693	4,693

Table 5: The statistics for evaluation datasets

any samples where the combined text length exceeds 500 tokens are excluded.

- For the Chinese open information extraction (IE) dataset, Saoke, we deviate from the above datasets for pretraining. Instead, we assemble a large-scale distant supervision dataset by aligning Wikidata with the Chinese version of Wikipedia.

B Implementation Details

In all our experiments, the optimization of our model is performed using the Adam algorithm (Kingma and Ba, 2014). During the pretraining phase, we set the learning rate at 2×10^{-5}, the global batch size at 96, and run the process for 5 epochs. For the fine-tuning phase, we explore a variety of hyper-parameters, adjusting the learning rate within the range $\{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}, 4 \times 10^{-5}, 5 \times 10^{-5}\}$ and the batch size from among $\{8, 12, 16, 32, 64, 96\}$. With 3 random seeds, we select the optimal hyper-parameter configuration based on the performance on the development set. For multi-task learning, we choose the best checkpoint based on the average performance across all datasets. All experiments are carried out on NVIDIA A100 (80G) GPUs and repeated 3 times to reported the averaged F1 scores.

We evaluate the model using span-based offset Micro-F1 as the primary metric, with different criteria for different aspects of the information extraction task:

- **Entity:** An entity mention is deemed correct if both its offsets and type correspond to a reference entity.
- **Relation (Strict Match):** A relation is considered correct if its type matches and both the offsets and entity types of the related entity mentions are correct.
- **Relation (Triplet Match):** A relation is considered correct if its type matches, and the offsets of the subject and object are correct.
- **Event Trigger:** An event trigger is considered correct if its offsets and event type align with a reference trigger.
- **Event Argument:** An event argument is marked as correct if its offsets, role type, and event type match a reference argument mention.
- **Sentiment Triplet:** To consider a sentiment triplet correct, the offsets boundaries of both the aspect and the opinion span must be correct, and the sentiment polarity must also be accurate.

These criteria ensure a comprehensive evaluation of the model's ability to correctly identify various elements of information extraction tasks.

C Ablation Study

In Table 6, we performed ablation studies on three components: token sequential dependency (Seq Dep) in span features, structure language prompt (SLP), and novel linking style for two universal relation extraction (TUR). We replaced span features with multiplicative attention and substituted SLP and TUR with USM's naive prompt and linking style, excluding discontinuous NER and Open IE from the experiments since the naive method can not extend to these two tasks. Our conclusions:

1) Token sequential dependency is vital for all four IE tasks. Its removal led to a substantial performance decline, confirming its effectiveness.

2) Ablating SLP & TUR didn't affect NER, as our prompt and linking style are similar to USM on the NER task. Other tasks showed declines, highlighting TRUE-UIE's prompt and linking style's effectiveness on IE tasks. The relatively noticeable performance decline in relation extraction and

event extraction demonstrates that this design effectively enhances the unification of learning objectives, allowing knowledge gained in NER to be shared across the relation extraction and event extraction tasks.

Task	Ent	Rel	Evt-Tri	Evt-Arg	Senti.
TRUE-UIE	**96.89**	**68.91**	**73.12**	**58.33**	**81.73**
w/o Seq Dep	95.26	67.52	72.79	57.34	80.91
w/o SLP & TUR	95.18	66.48	71.97	56.83	80.53

Table 6: Ablation study for TRUE-UIE on 4 tasks: entity recognition (CoNLL03), relation extraction (ACE-Rel), event extraction (ACE05-Evt), and sentiment analysis (16res).

D Limitations

The Structure Language Prompt might lead to performance decline in certain datasets where default entity types or coarse entity types are commonly used in many triplet schemes. This occurs as the same type of text, such as “people”, appears in different schemes, causing confusion. For instance, in Figure 1, “people” is used in both “work for” and “born in” relations, but an entity of the type “people” may not always be involved in both relations. If the model, post-training, represents “people” similarly across different schemes, it could lead to confusion, resulting in high recall but low precision. Our solution is to employ an attention mask strategy as following Figure 4, enabling the model to focus only on text within the scheme group. For example, the first “people” would only pay attention to “work for organization”, and the second “people” to “born in place”.

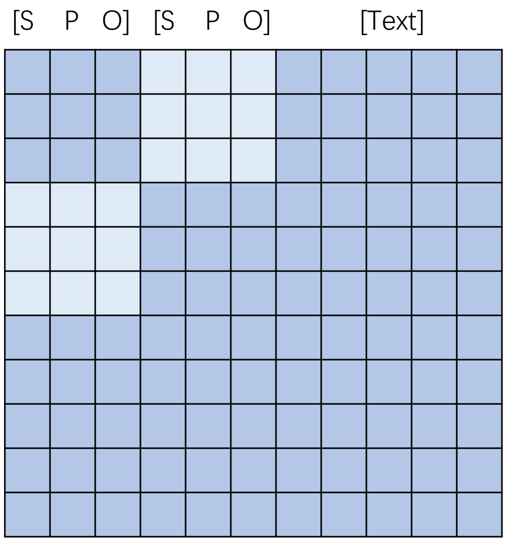

Figure 4: The figure illustrates TRUE-UIE’s attention mask approach for handling datasets with numerous duplicate entity/role types.

E Help from AI assistants

When necessary, we use ChatGPT or Copilot for guidance on how to write regular expressions, like the *tokenize_uni* function in *utils.py*.