Article:
Sun, K., Zhang, R., Mensah, S. orcid.org/0000-0003-0779-5574 et al. (2 more authors) (2022) Learning implicit and explicit multi-task interactions for information extraction. ACM Transactions on Information Systems. ISSN 1046-8188

https://doi.org/10.1145/3533020

Learning Implicit and Explicit Multi-task Interactions for Information Extraction*

KAI SUN, SKLSDE, School of Computer Science and Engineering, Beihang University, China
RICHONG ZHANG†, SKLSDE, School of Computer Science and Engineering, Beihang University, China
SAMUEL MENSAH, Department of Computer Science, University of Sheffield, UK
YONGYI MAO, School of Electrical Engineering and Computer Science, University of Ottawa, Canada
XUDONG LIU, SKLSDE, School of Computer Science and Engineering, Beihang University, China

Information extraction aims at extracting entities, relations, etc., in text to support information retrieval systems. To extract information, researchers have considered multitask learning (ML) approaches. The conventional ML approach learns shared features across tasks, with the assumption that these features capture sufficient task interactions to learn expressive shared representations for task classification. However, such an assumption is flawed in different perspectives. First, the shared representation may contain noise introduced by another task; tasks coupled for multitask learning may have different complexities but this approach treats all tasks equally; the conventional approach has a flat structure which hinders the learning of explicit interactions. This approach however learns implicit interactions across tasks and often has a generalization ability which has benefited the learning of multitasks. In this paper, we take advantage of implicit interactions learned by conventional approaches while alleviating the issues mentioned above by developing a Recurrent Interaction Network with an effective Early Prediction Integration (RIN-EPI) for multitask learning. Specifically, RIN-EPI learns implicit and explicit interactions across two different but related tasks. To effectively learn explicit interactions across tasks, we consider the correlations among the outputs of related tasks. It is however obvious that task outputs are unobservable during training, so we leverage the predictions at intermediate layers (referred to as early predictions) as proxies as well as shared features across tasks to learn explicit interactions through attention mechanisms and sequence learning models. By recurrently learning explicit interactions, we gradually improve predictions for the individual tasks in the multitask learning. We demonstrate the effectiveness of RIN-EPI on the learning of two mainstream multitasks for information extraction: (1) entity recognition and relation classification, (2) aspect and opinion term co-extraction. Extensive experiments demonstrate the effectiveness of the RIN-EPI architecture, where we achieve state-of-the-art results on several benchmark datasets.

CCS Concepts: • **Information systems → Information extraction**.

Additional Key Words and Phrases: multitask learning, information extraction

*This article is an extension of the conference paper 'Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao and Xudong Liu: Recurrent Interaction Network for Jointly Extracting Entities and Classifying Relations. EMNLP2020. [52]'

†Corresponding Author

Authors' addresses: Kai Sun, SKLSDE, School of Computer Science and Engineering, Beihang University, Beijing, China, sunkai@act.buaa.edu.cn; Richong Zhang, SKLSDE, School of Computer Science and Engineering, Beihang University, Beijing, China, zhangrc@act.buaa.edu.cn; Samuel Mensah, Department of Computer Science, University of Sheffield, Sheffield, UK, s.mensah@sheffield.ac.uk; Yongyi Mao, School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada, yymao@site.uottawa.ca; Xudong Liu, SKLSDE, School of Computer Science and Engineering, Beihang University, Beijing, China, liuxd@act.buaa.edu.cn.

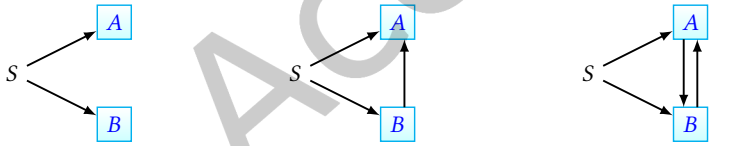
1046-8188/2022/6-ART $15.00
https://doi.org/10.1145/3533020

1 INTRODUCTION

Research in information retrieval has gained much traction due to the widespread access to information retrieval systems through the internet. The work in this field has branched into diverse areas such as information extraction, whose final aim is to facilitate information access. Specifically, the extraction of entities, relations, events, or any structured information from text can be considered as an instance of information extraction. This particular task has supported the construction of knowledge or ontology databases and has shown to be beneficial in information retrieval systems such as Google Search.

A diverse range of deep learning techniques have been developed and applied to a broad range of problem types under information extraction. Deep learning has enabled computational models to learn multiple abstraction levels of the underlying distribution of data. It has demonstrated great success in different learning paradigms including supervised and unsupervised learning. However, it may be noted that the benefits of deep learning we see today particularly comes from supervised learning. Supervised learning is defined by its use of labelled data to train models to make predictions on unlabelled data. It performs well on certain tasks as compared to unsupervised methods given the availability of sufficient labelled data. However, in most applications collecting labelled data is expensive. For this reason, one must work with limited available labelled data.

With the lack of sufficient labelled training data, researchers have explored transfer learning approaches. Here, the idea is to transfer knowledge from a related task or domain to improve learning on a target task. Among the different methodologies proposed for transfer learning, multi-task learning (MTL) has been used successfully across a wide range of machine learning applications including, speech recognition [8], computer vision [40, 75], natural language processing [4], etc. In brief, multi-task learning is a subfield of machine learning in which multiple tasks are solved simultaneously. While single task learning may achieve acceptable performance by training a single model, empirical evidence has shown that MTL benefits from the training signals from a different but related task.



(a) Flat Structure (b) Hierarchical Structure (c) Graph Structure

Fig. 1. Topological structures for multi-task learning. Here, A and B are two different tasks, and S is the shared information for the two tasks. The directed edges define the information flow.

In the context of deep learning, multitask learning is backed by parameter or information sharing across tasks. The difference in these sharing techniques lies in the type and flow of information across tasks, which in effect presents an MTL approach. Figure 1 shows existing MTL approaches which can be distinguished by their topological structure. In the figure, A and B are two different but related tasks and S is a shared information across the two tasks. The interest in illustrating two tasks is motivated by a large number of recent observations in information extraction that simultaneously learn two tasks in a multi-task setting [11, 23, 36, 43]. Although multi-task learning can be extended to more than two tasks, our work is tailored to learning two tasks in an effort to demonstrate the effectiveness of our approach on different information extraction applications.

An optimal choice when applying any of these approaches depends on a number of factors including the target task to be addressed, the distribution of the data, and the relatedness of the tasks coupled in the multi-task learning network, just to name a few. Early works [11, 14, 17, 23, 29, 36, 43, 62, 74] leveraging multitask learning typically follow the flat structure. The flat structure consists of a shared encoder and one decoder per task. Specifically,

the shared encoder takes an input and learns a shared representation by sharing parameters across all tasks. The shared representation is then fed to task-specific decoders to output task-specific predictions. The flat structure is effective to an extent because it helps to improve generalization performance on all tasks. Nonetheless, a number of studies [18, 32] have noted its weakness from different perspectives. Firstly, the shared representation used by the target task module may contain noise brought by another task which can lead to a negative influence on the prediction of the target task. Secondly, this approach has the implicit assumption that the shared encoder is sufficient to capture the correlations that exist among the tasks which may not always be true. Thirdly, this approach ignores the different levels of complexity for the individual tasks coupled in the network [18]. The hierarchical structure [18, 51] partially solves this problem by ordering the direction of information flow between tasks by considering the complexity of tasks. The graph structure [32] on the other hand makes no strong assumptions on the task complexity or the nature of interactions between tasks. This is based on the fact that the degree of relatedness between tasks is subject to change as it relies on the samples seen during training. Hence, the graph structure offers a natural analogy of the interactions existing between tasks.

MTL architectures that follow the graph structure mainly focus on learning the explicit interaction between tasks. Modelling interactions using the graph structure has gained increasing attention in sequence learning [32] as well as the biomedical field for the reconstruction of gene regulatory networks [39]. Current work can mainly be divided into two lines. The first line models the interactions by exploiting intermediate task-specific representations while the second line additionally consider the early task-specific predictions (i.e., predictions produced in the intermediate layers of the neural network). So far, a variety of works in the first line have been proposed [7, 10, 14, 27, 30, 32, 33, 37, 63, 67]. Among these works, gate and attention mechanism are commonly used to capture the relationships between tasks and control the information flow between task-specific representations. Although this line has shown some impressive results, for some multi-task learning problems (e.g., aspect and opinion terms co-extraction), there exist an inter-dependency between the task outputs. Therefore, only modeling the interactions between the intermediate task-specific representations may not be sufficient to capture the inter-dependency between tasks [19].

Recently, modeling explicit interactions between tasks by considering information relating to the task outputs have been proposed [19, 77]. This group of works fall under the second line. He et al. [19] propose an interactive multi-task learning framework that recurrently updates the shared representation produced by an MTL network. More specifically, the method linearly transforms a concatenation of early task predictions and a shared representation to improve the shared representation for the joint aspect term extraction and sentiment classification task. Zhao et al. [77] propose to model the explicit interaction in the joint medical named entity recognition and normalization task where both subtasks are formulated as sequence labelling problems. In this work, early task-specific predictions are linearly transformed into distributed vectors. These vectors are then added to shared representations to construct task-specific representations for sequence labelling.

Unlike these works that consider a shallow network structure, we consider a deep neural network comprising of multiple interaction layers to learn complex relationships between tasks. Besides, rather than employing simple linear transformations to fuse early task-specific predictions and representations, we employ sequence learning models and attention mechanisms to learn the complex dynamic interactions effectively. To this end, we propose a Recurrent Interaction Network with an effective Early Prediction Integration (RIN-EPI), a multi-task learning network based on the graph structure. The RIN-EPI architecture aims to capture interactions implicitly through a shared encoder and explicitly by sharing early task-specific predictions and representations across the intermediate layers of RIN-EPI to improve predictions on individual tasks. We demonstrate the effectiveness of RIN-EPI on the the following joint tasks: (1) entity recognition and relation classification, and (2) aspect and opinion term co-extraction. In the following subsections we will provide a background on the joint tasks.

1.1 Entity recognition and relation classification

The extraction of entities and relations from textual data comprises of two sub-tasks: entity recognition (ER) [44] and relation classification (RC) [20]. The ER task aims to extract all entities in a given text. The RC task aims to classify the relation between any pair of entities in the text. For instance, consider the sentence *"Adriel was born in London which is the capital of England"*. The goal of the entity recognition and relation extraction task is to identify all the factual relational triples (or relational facts) *(London, birth_place_of, Adriel)*, *(England, birth_place_of, Adriel)* and *(London, capital_of, England)*. Solving these tasks has contributed significantly in extracting structured knowledge from unstructured text for several applications, including knowledge base construction [72] and information extraction [38].

A straightforward approach to solve the ER and RC tasks is to first extract all entities in the sentence and then classify the relation between entity pairs. This is considered to be a pipeline-based approach which has been utilized by several traditional works [3, 68, 80]. However, the shortcomings of pipeline-based approaches cannot be ignored. Specifically, this approach disregards the correlation between ER and RC tasks. Hence, if a model proposed for ER produces biased predictions or make wrong predictions for specific samples, the ER prediction errors are propagated to the RC task, leading to wrong RC predictions. The major downside is that the misclassification by the RC task is irreversible.

A strategy to address this limitation is to learn the two tasks jointly so as to allow the ER and RC tasks communicate with each other to make informed decisions and produce reliable task-specific predictions. Proposed works can be categorized according to the type of neural framework. The first class of works address the task using a sequence labelling approach [5, 54, 61, 66, 79]. Among them, Zheng et al. [79] was the first to develop a tagging strategy to address the problem. However, the method fails to identify overlapping relations [71]. As a solution, Dai et al. [5] proposed a position-attentive tagging scheme that simultaneously tags entities and relation labels according to a query word position, producing a set of different position-aware sentence representations in which overlapping relations can easily be extracted. Other methods [54, 66] decompose the task into two sequence labelling sub-tasks, where one attempts to detect the subject of the relational fact and the other detects its object with respect to the relation type.

Another line of works [70, 71] treat the problem using a sequence-to-sequence (seq2seq) approach. Among them, Zeng et al. [71] designed a seq2seq model known as CopyRE that successively decodes the first entity, second entity and relation. However, since this method extracts a predefined number of relational triples, it risks missing relational facts when there are several in the text. Zeng et al. [70] showed the importance of the order in which relational facts are extracted. Specifically, the work shows that the information related to a relational fact in text can help extract another. Based on this observation, they proposed a seq2seq model based on reinforcement learning that learns the order of relational triple extraction. Notably, entities may span over multiple tokens. However, seq2seq models decode a single word for an entity which makes it undesirable for the joint extraction task.

The third class of works consider a multi-task learning approach to capture the correlation between the ER and RC tasks [1, 11, 41, 42, 69]. Among them, Miwa and Bansal [41] adopt a bi-directional LSTM and a bidirectional tree-structured LSTM [53] to capture sequential and dependency information to extract entities and relations. Adel and Schütze [1] employ convolutional neural networks to encode different parts of the text simultaneously and a linear-chain conditional random field output layer to predict entities and relations. Zeng et al. [69] developed CopyMTL, an MTL model that improves the entity recognition in CopyRE. Specifically, CopyMTL is a multi-task architecture that includes an encoder comprising of a CopyRE encoder and a sequence labelling module to help extract multi-token entities while a decoder (i.e., a seqseq model) is used to extract relational facts. Fu et al. [11] proposed GraphRel, a method that employs a graph convolutional network to extract relational facts by learning the linear and structural interactions between entities and relations in text. Although proposed MTL-based methods have shown promising success in addressing the joint extraction task, the architecture of these methods follow

the flat structure. However, identifying the relational facts in sentences is a difficult problem due to overlapping relations among other factors. Therefore the design of these models restricts the model to effectively learn the correlations between tasks.

Gupta et al. [16] considered the drawback of the flat structure and proposed a multi-task learning model that reduces the joint task into a table filling problem, where the inter-dependencies between the two tasks are modelled explicitly. However, Gupta et al. [16] models this interaction in a successive way, i.e., the output of one task is used to benefit the prediction of the other, indicating a hierarchical MTL topology. In effect, a wrong prediction by a task initiates an error cascade to the other task. Nonetheless, their work shows that the output of the ER task can be used to promote the prediction of the RC task, and vice versa, indicating a correlation between the outputs of two tasks. For example, consider the NYT dataset [50] which contains articles with annotated relational triples. It turns out that for the relation "nationality", 97.42% of head entities associated with this relation have type "person" while 98.59% of tail entities have type "country". Similarly, for the relation "company", 98.33% of head entities have type "person" while 79.93% of tail entities have type "organization". These statistics show a strong inter-dependency between the entity type and relation, implying the intermediate early predictions of the ER and RC tasks perhaps can be useful to learn interactions across tasks. Without modeling such explicit interactions, existing MTL-based methods [11, 41, 69] cannot effectively capture the correlation between the ER and RC tasks.

1.2 Aspect and Opinion Term Co-extraction

The task of aspect and opinion term co-extraction aims to jointly extract aspects describing features of an entity and opinion words describing the sentiments toward an aspect. Thus, the joint extraction task comprises of an aspect term extraction (ATE) task and an opinion word extraction (OTE) task. For example, given the sentence *"We ordered the special branzino, that was so infused with bone, it was unpalatable."* The goal is to extract the aspect term *"special brazino"* and opinion term *"unpalatable"*. This particular task has found strong applications in opinion retrieval [13] and aspect-based sentiment analysis [34]. Most mainstream methods fall into one of two categories: rule-based methods and neural network based methods. Rule-based methods commonly use manually designed rules to extract aspect and opinion terms [35, 47, 56, 81]. Among these works, [81] generates keywords to obtain a list of aspect and opinion terms. Then, rules based on the dependency grammar graph are applied to mine the target aspect and corresponding opinion term. [47] identified eight types of rules about grammar dependency relations to extract aspect and opinion terms. However, their method only considers noun aspect terms and adjective opinion terms. Meanwhile, [35] automatically selects a relevant subset of rules about grammar dependency relations to address the aspect and opinion term extraction problem. [56] evaluates on SE15-R (from SemEval 2015). Although their rule-based method extracted features that relatively addressed the extraction task as at that time, the rule designing process is labor intensive and requires elegant feature engineering which is usually affected by human ingenuity.

More recent works have focused on multitask learning based on neural networks. These architectures are adept in automatically extracting features for the co-extraction task [24, 58, 65]. Proposed works typically tag each word in the sentence using a **B**eginning, **I**nside, **O**utside (BIO) tagging scheme [48] to extract aspect or opinion terms. That is, the problem is casted into a sequence labelling problem for each task. Among proposed approaches, [24] first obtain features for each word in a sentence, which is then fed to a Conditional Random Field (CRF) [26] to obtain the sequence labelling tags. Other works including [58, 65] introduce dependency tree into the feature learning process to improve the prediction. Particularly, [65] employs an unsupervised approach to learn both word and dependency path features. The learned features are then fed to a CRF model for prediction. [58] on the other hand applies a recursive neural network on a dependency tree to learn features for a CRF model. Some works have explored different text encoders including autoencoders [73] and convolutional neural networks [64] for feature learning for this co-extraction task. The recent work [6] considered mining rules automatically from

data and introduces these rules as weak supervision on a neural network model during training to improve model performance. Some multi-task learning architectures exploit an attention mechanism [2] to model the relationship between aspect and opinion terms [28, 59]. Specifically, [59] proposed a multi-layer attention network where each layer consists of a couple of attention mechanisms that interact to propagate information across words to extract aspect and opinion terms. [28] proposed a Truncated History-Attention and Selective Transformation Network that exploits opinion summary and aspect detection history to extract aspect terms.

However, the multitask learning architecture used by these works are flat in structure, or the way they learn explicit interactions through an attention mechanism is executed in a rather shallow way. That is, interactions between task-specific representations is performed in one single step. This approach may not be sufficient to encode the relevant interaction information in task-specific representations because a counter-intuitive assumption is that if a model extracts some relevant interaction information in one step of interaction, then by stacking multiple of such steps, the model can gradually accumulate useful signals for supervision and finally capture the semantic relationship between tasks in a more comprehensive way. Therefore in our approach, we consider to learn deeper explicit interactions by recurrently utilizing previous predictions of task-specific networks as well as shared features to gradually improve the task-specific representations for classification.

2 MODEL

We study the problem of solving multiple tasks in parallel through a multitask learning network which we refer to as Recurrent Interaction Network with an effective Early Prediction Integration (RIN-EPI). The RIN-EPI architecture considers two tasks. Let $\mathcal{X}$ be the input space, $\mathcal{Y}$ and $\mathcal{Z}$ be the output spaces of the two tasks. Let $X \in \mathcal{X}$ be an input, $Y \in \mathcal{Y}$ and $Z \in \mathcal{Z}$ be outputs. We refer to the tasks that takes X and predicts Y and Z as the $\mathcal{Y}$-task and $\mathcal{Z}$-task respectively. It is assumed that the $\mathcal{Y}$-task and $\mathcal{Z}$-task are related and they exploit commonalities and differences across tasks through interaction layers to improve predictions on individual tasks. We also assume that $\mathcal{X}$ is the input space of texts. Although we allow the two tasks to share the same input space in this paper, in other settings the two different tasks may receive inputs from different spaces depending on the tasks to be solved. Figure 3 shows a high-level overview of RIN-EPI.

2.1 Conventional Multi-Task Learning Model

To provide a motivation to the RIN-EPI architecture, we first present a conventional multi-task learning model. Most MTL architectures assume a flat structure, where multiple tasks are grounded on a shared representation. This is what we refer to as a conventional multi-task learning model. Figure 2 shows the architecture of a conventional multi-task learning model.

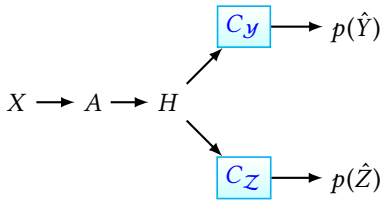


Fig. 2. Architecture of the conventional multi-task learning model.

In this type of architecture, the two tasks $\mathcal{Y}$-task and $\mathcal{Z}$-task share a common module A that takes the input $X \in \mathcal{X}$ and maps it to a representation $H \in \mathcal{H}$, i.e, $A : \mathcal{X} \rightarrow \mathcal{H}$. The representation H is then fed independently to the $\mathcal{Y}$-task and $\mathcal{Z}$-task for prediction. For example, suppose $X = \{x_1, x_2, \ldots, x_n\}$ is a word sequence of length

n, the module A maps X to the representation $H = \{h_1, h_2, \ldots, h_n\}$. There are a variety of text encoders that can be used to model A and extract word representations automatically. Some of the common text encoders include bidirectional LSTM [21] and BERT [9], which stands for Bidirectional Encoder Representations from Transformers. We will provide a brief description of text encoders we experiment with in Section 2.5.

Now suppose $C_{\mathcal{Y}}$ and $C_{\mathcal{Z}}$ are classifiers corresponding respectively to the $\mathcal{Y}$-task and $\mathcal{Z}$-task, $C_{\mathcal{Y}}$ and $C_{\mathcal{Z}}$ take H as input and independently output the predictions $p(\hat{Y})$ and $p(\hat{Z})$ over the respective output spaces $\mathcal{Y}$ and $\mathcal{Z}$. The mapping function is formulated as follows:

$$\begin{aligned} C_{\mathcal{Y}} &: \mathcal{H} \rightarrow p(\mathcal{Y}) \\ C_{\mathcal{Z}} &: \mathcal{H} \rightarrow p(\mathcal{Z}) \end{aligned} \tag{1}$$

As can be seen in the model architecture (see Fig 2), interactions are learned implicitly across tasks through module A, impeding dynamic learning of intrinsic correlations that may exist between the two tasks.

2.2 Recurrent Interaction Network with an effective Early Prediction Integration

Although the conventional multi-task learning model captures interaction across tasks, the interaction is captured implicitly through the shared module A. Thus, we have little control on the interaction information we will like to capture.

Indeed, there are correlations across related tasks that can be exploited explicitly. Take for instance an entity recognition (ER) and relation classification (RC) task. The ER task aims to identify entities in text while the RC task aims to extract relational factual triples in text by classifying the relations between any pair of entities. Thus, there exist an inter-dependency between the outputs of the two tasks. For instance in the NYT dataset, the relation /business/person/company contains about 98% of left-end entities with type person while other relations such as /people/person/nationality have almost 99% of its right-end entities with type country, suggesting strong correlations between relations (i.e., the output of the RC task) and entities (i.e., the output of the ER task).

However, the outputs of tasks are unobserved but one can exploit the correlations among their early predictions (i.e., predictions from the intermediate layers of the task-specific networks) to progressively improve the prediction performance on the individual tasks. Hence beyond capturing implicit interactions, RIN-EPI dynamically learns explicit interactions among tasks by exploiting early predictions as well as the shared word features of both tasks (or simply shared representation).

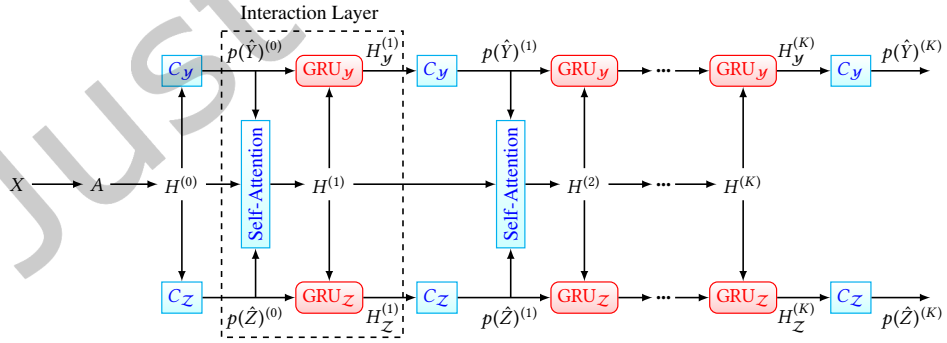


Fig. 3. A high-level overview of the RIN-EPI architecture.

As a solution, we present a Recurrent Interaction Network with an effective Early Prediction Integration (RIN-EPI). A high-level over of the model architecture is shown in Fig. 3. The RIN-EPI model is comprised of multiple

interaction layers. At each layer we learn two types of explicit interactions from different angles. The first is referred to as *global interactions*. As the name suggest, by modelling global interactions, we aim to capture the correlations across the two different tasks by learning the strength of relationship among the early predictions and the shared representation. The learned global interaction information is then compressed to construct a new shared representation. It is expected that this new shared representation provides a good summary of the relationship among the early predictions from the two different tasks and the text itself.

It should be noted that the early predictions for any task-specific network comes from applying a classifier on the current task-specific features. To progressively refine these predictions, it is imperative to refine the task-specific features. This leads us to consider the modelling of interactions at each task-specific network as a means to refine task-specific features. Thus, the second type of interactions we aim to model is referred to as *local interactions* since we consider the interaction at a specific task network. To achieve this, we model the dependencies among the early predictions and the new shared representation to guide the model to extract salient task-specific features for classification.

In the following subsections, we provide a formal description on how we model both global interactions and local interactions among the two networks.

2.2.1 Modelling Global Interactions. When there are correlations among tasks, particularly on the outputs of related tasks (e.g. entity recognition and relation classification tasks), it is advantageous to capture these correlations to help individual tasks perform better. We employ a self-attention [55] layer to capture such correlations across the two task-specific networks. That is, the problem is formulated as modelling the global interactions since it considers how the two task-specific networks interact.

The self-attention layer is primarily composed of a self-attention mechanism that takes as input the early predictions from both tasks as well as shared features. The self-attention mechanism allows the input to interact with each other to allow the model to pay attention to the salient features. Its output is an aggregate of these interactions which forms the new shared representation. Intuitively, the new shared representation can be broadly be interpreted as a vector that indicates the strength of relationship between early predictions across tasks and the shared features.

More formally, we first combine early predictions $p(\hat{Y})$ and $p(\hat{Z})$ and shared features H using a concatenation operator and apply a linear transformation to project the concatenated representation into the same space as the shared representation. That is, we compute a representation H_f that contains prediction information from the $\mathcal{Y}$-task and $\mathcal{Z}$-task. Next, we calculate the query, key, value of the self-attention by multiplying H_f with three weight matrices Q, K and V which are targets to be trained. We then compute the self-attention output H_s by taking the query QH_f, key KH_f and value VH_f as inputs in the self-attention mechanism.

Formally, let H_s^k be the self-attention output at layer k of RIN-EPI, the equations that govern the computation of H_s^k is given by:

$$
\begin{aligned}
H_f^{(k)} &= \sigma\left(\left[H^{(k)}; p(\hat{Y})^{(k)}; p(\hat{Z})^{(k)}\right] W_f\right) \\
H_s^{(k)} &= \text{softmax}\left(\frac{QH_f^k (KH_f^k)^T}{\sqrt{d}}\right) VH_f^k
\end{aligned}
\tag{2}
$$

where σ is a ReLu activation function, $H^{(k)}$, $Q \in \mathbb{R}^{d\times d}$, $K \in \mathbb{R}^{d\times d}$, $V \in \mathbb{R}^{d\times d}$ and W_f are learnable parameters. Finally, we compute the new shared representation $H^{(k+1)}$ in the next layer as follows:

$$
H^{(k+1)} = H^{(k)} + H_s^{(k)} \tag{3}
$$

Our motivation for constructing $H^{(k+1)}$ by adding $H^{(k)}$ to $H_s^{(k)}$ is to retain the contextual information contained in the previous shared representation.

We recall that the new shared representation $H^{(k+1)}$ captures the strength of relationship among the early task-specific predictions as well as the previous shared representation. With this new information, we can guide the model to refine the predictions at each task-specific network. Accordingly, it is important to learn the correlations among the new shared representation and the previous predictions at a specific task-specific network. That is, modelling the local interactions since it considers interactions at a task-specific network. We will discuss how this works in the next section.

2.2.2 Modeling Local Interactions. We cast local interaction modeling as a sequence learning problem on shared features and early predictions. More precisely, we employ two separate gated recurrent units (GRUs) to separately model local interactions on the $\mathcal{Y}$-task and $\mathcal{Z}$-task. Our choice of the GRU network is based on the fact that we need a model component that 1) takes two kinds of input vectors (i.e, early predictions and shared features) and generates one output vector (i.e., task-specific features), and 2) is sufficiently expressive and capable of modelling complex dynamics.

Suppose we consider the shared features $H^{(k)}$ as well as the early predictions $p(\hat{Y})^{(k-1)}$ and $p(\hat{Z})^{(k-1)}$ that respectively correspond to the $\mathcal{Y}$-task and $\mathcal{Z}$-task. The GRU networks take in $H^{(k)}, p(\hat{Y})^{(k-1)}$ and $p(\hat{Z})^{(k-1)}$ and extract task-specific features $H_{\mathcal{Y}}^{(k)}$ and $H_{\mathcal{Z}}^{(k)}$ at layer k. By this operation, we encode task-specific information (i.e., via task-specific predictions) and interaction information (i.e., via shared features) into task-specific representations which in turn facilitates the retaining and modelling of global interactions across tasks in the learning process.

Formally, let $\text{GRU}_{\mathcal{Y}}$ and $\text{GRU}_{\mathcal{Z}}$ represent the GRU networks for the $\mathcal{Y}$-task and $\mathcal{Z}$-task in an interaction layer. The task-specific features $H_{\mathcal{Y}}^{(k)}$ and $H_{\mathcal{Z}}^{(k)}$ at the k-th interaction layer is computed as follows:

$$\begin{aligned} H_{\mathcal{Y}}^{(k)} &= \text{GRU}_{\mathcal{Y}}\left(p(\hat{Y})^{(k-1)}, H^{(k-1)} | \theta_{\text{GRU}_{\mathcal{Y}}}\right) \\ H_{\mathcal{Z}}^{(k)} &= \text{GRU}_{\mathcal{Z}}\left(p(\hat{Z})^{(k-1)}, H^{(k-1)} | \theta_{\text{GRU}_{\mathcal{Z}}}\right) \end{aligned} \tag{4}$$

where $\theta_{\text{GRU}_{\mathcal{Y}}}$ and $\theta_{\text{GRU}_{\mathcal{Z}}}$ are trainable parameters for the respective $\text{GRU}_{\mathcal{Y}}$ and $\text{GRU}_{\mathcal{Z}}$ networks.

$H_{\mathcal{Y}}^{(k)}$ and $H_{\mathcal{Z}}^{(k)}$ are then fed to the respective classifiers $C_{\mathcal{Y}}$ and $C_{\mathcal{Z}}$ to make predictions $p(\hat{Y})^{(k)}$ and $p(\hat{Z})^{(k)}$ for the two tasks.

2.3 Baseline Model

We apply RIN-EPI on the joint entity recognition (ER) and relation classification (RC) task as a baseline. We formally describe the problem of the individual tasks and then the application of RIN-EPI to solve the tasks.

2.3.1 Problem Statement. Given a text $X = \{x_1, x_2, \cdots, x_n\}$ with length n and let $T = \{t_1, \cdots, t_l\}$ be a set of pre-defined relation types of length l. The problem to be solved is to extract relational factual triples given X. In this paper, a candidate relational triple is of the form $\langle x_i, t, x_j \rangle$, where x_i(or x_j) is an entity word, non-entity word or form part of a multi-token entity and $t \in T$. This requires two sub-tasks to be solved: the ER and RC tasks. The ER task is typically treated as sequence labelling problem using a BIOES labelling scheme [11], where each word in X is tagged as: beginning of a chunk (B), inside a chunk (I), outside a chunk (O), end of chunk (E), or single element in a chunk (S). For a given word pair (w_i, w_j), the RC task aims to predict the probability that $\langle x_i, t, x_j \rangle$ is factual for $t \in T$. Note, the ER task can identify the head and tail of multi-token entities which is essential in completing candidate entities in the extracted relational triple.

2.3.2 Model Description. Let the ER task correspond to the $\mathcal{Y}$-task and the RC task correspond to the $\mathcal{Z}$-task.

Given X, an initial shared representation $H^{(0)}$ is computed by employing a text encoder on X. We then feed $H^{(0)}$ to the classifiers $C_{\mathcal{Y}}$ and $C_{\mathcal{Z}}$ (which will be described shortly) to output the respective early predictions $p(\hat{Y})^{(0)}$ and $p(\hat{Z})^{(0)}$. At any given layer, we feed $H^{(k)}, p(\hat{Y})^{(k)}, p(\hat{Z})^{(k)}$ in a self-attention layer to model global interactions which in turn outputs the updated shared representation $H^{(k+1)}$. We also model local interactions at the $\mathcal{Y}$-task by feeding $\text{GRU}_{\mathcal{Y}}$ with $p(\hat{Y})^{(k-1)}$ and $H^{(k)}$ to output the task-specific representation $H_{\mathcal{Y}}^{(k)}$.

In detail, given $x \in X$ with corresponding shared word representation $h^{(k)} \in H^{(k)}$ and early prediction $p(\hat{y})^{(k-1)} \in p(\hat{Y})^{(k-1)}$, $\text{GRU}_{\mathcal{Y}}$ computes the corresponding task-specific word representation $h_{\mathcal{Y}}^{(k)} \in H_{\mathcal{Y}}^{(k)}$ as follows:

$$\begin{aligned} z &= \sigma\left(W_{\text{z}}(h^{(k)} \oplus p(\hat{y})^{(k-1)})\right) \\ u &= \sigma\left(W_{\text{u}}(h^{(k)} \oplus p(\hat{y})^{(k-1)})\right) \\ \check{h} &= \tanh\left(W_{\text{o}}((u * h^{(k)}) \oplus p(\hat{y})^{(k-1)})\right) \\ h_{\mathcal{Y}}^{(k)} &= (1-z) * h^{(k)} + z * \check{h} \end{aligned} \tag{5}$$

where $\oplus$ is a concatenation operator, W_z, W_u, W_o are trainable parameters of the $\text{GRU}_{\mathcal{Y}}$. We then feed $h_{\mathcal{Y}}^{(k)}$ into the classifier $C_{\mathcal{Y}}$ to output the probability distribution $p(\hat{y})^{(k)}$ over BIOES labels. The equation that governs the classifier $C_{\mathcal{Y}}$ is given by,

$$p(\hat{y})^{(k)} = \text{softmax}(W_y h_{\mathcal{Y}} + b_y), \tag{6}$$

where W_y, b_y are learnable model parameters.

In the RC task, note that we make predictions over the relation types for each word pair (x_i, x_j) in the text X. To delineate these predictions, we use $p(\hat{z}_{ij}) \in p(\hat{Z})$ to denote the probability distribution for the word pair (x_i, x_j) over the l relation types. To employ $\text{GRU}_{\mathcal{Z}}$ to model the task-specific word representation for the RC task, we first associate each word x_i to the set of early predictions $Z(x_i) =: \{p(\hat{z}_{ij}) \in p(\hat{Z}) | x_j \in X\}$ and extract a vector z_i using a maxpool function.

$$z_i = \text{maxpool}\left(Z(x_i)\right), \tag{7}$$

We interpret z_i as a vector of relation predictions that is associated with the word x_i. Now $\text{GRU}_{\mathcal{Z}}$ follows the same computation steps as $\text{GRU}_{\mathcal{Y}}$ as shown in (5). However, given $x_i \in X$, it considers the corresponding shared word features $h_i^{(k)}$ and relation predictions $z_i^{(k-1)}$ to compute the task-specific word representation $h_{i\mathcal{Z}}^{(k)}$ for the RC task.

Now, for a given pair of task-specific word representations $h_{i\mathcal{Z}}, h_{j\mathcal{Z}} \in H_{\mathcal{Z}}$, the remaining step is to compute the new probability distributions $p(\hat{z}_{ij})$ over the relation types by feeding the task-specific word representations into $C_{\mathcal{Z}}$ that performs the following steps:

$$\begin{aligned} m &= \phi\left(W_{\text{m}}(h_{i\mathcal{Z}} \oplus h_{j\mathcal{Z}})\right) \\ p(\hat{z}_{ij}) &= \sigma\left(W_{\text{r}} m + b_r\right) \end{aligned} \tag{8}$$

where $\oplus$ is a concatenation operation, $\phi(\cdot)$ is the ReLU activation function, $\sigma(\cdot)$ is the sigmoid activation function. $W_{\text{m}}, W_{\text{r}}, b_r$ are learnable model parameters. Previous work [11] employ the softmax for the classification task. However, we find that the sigmoid function offers a natural way of identifying multiple relations that may exist between word pairs.

Although we have shown how RIN-EPI can be applied to the joint and entity and relation extraction task, it is worth noting that RIN-EPI can also be adapted to other multi-task learning settings, e.g., aspect and opinion terms co-extraction task [6].

2.4 Training Objective

The RIN-EPI model ultimately outputs task-specific representations, which are fed into their corresponding classifiers for predictions. Let D_{XY} and D_{XZ} be the training samples for the $\mathcal{Y}$- and $\mathcal{Z}$-tasks respectively, and let $L_{\mathcal{YZ}}$ be the total loss over all training samples for the two tasks. Formally, $L_{\mathcal{YZ}}$ is given by

$$L_{\mathcal{YZ}} = \sum_{(X,Y)\in D_{XY}} \text{CE}\left(p(\hat{Y})^K, Y\right) + \sum_{(X,Z)\in D_{XZ}} \text{CE}\left(p(\hat{Z})^K, Z\right) \tag{9}$$

where Y and Z are the respective ground-truths of the $\mathcal{Y}$-task and $\mathcal{Z}$-task, and $p(\hat{Y})^K$ and $p(\hat{Z})^K$ are the final predictions at the last layer K. CE denotes the cross-entropy loss function.

2.5 Text Encoders

Suppose the sentence $X = \{x_1, x_2, \ldots x_n\}$ is given. We experiment with two types of text encoders when modeling A and apply either one of the text encoders on X to construct a shared contextualized representation H. The text encoders considered are the bi-directional LSTM (BiLSTM) and the BERT encoder which are commonly used in natural language processing.

2.5.1 BiLSTM Encoder. All word tokens in X are mapped to word embeddings $E = \{e_1, ..., e_n\}$ by looking up their corresponding pre-trained pre-trained word embeddings [45]. We then contextualize word embeddings by applying a forward LSTM on E to learn hidden state representations $\overrightarrow{H} = \{\overrightarrow{h_1}, \overrightarrow{h_2}, \ldots \overrightarrow{h_n}\}$ and a backward LSTM to learn hidden state representations $\overleftarrow{H} = \{\overleftarrow{h_1}, \overleftarrow{h_2}, \ldots \overleftarrow{h_n}\}$. We then concatenate the corresponding parallel representations of $\overrightarrow{H}$ and $\overleftarrow{H}$ to construct the final representations $H = \{h_1, ..., h_n\}$.

2.5.2 BERT Encoder. Alternatively, we can utilize a Bidirectional Encoder Representations from Transformers (BERT) [9] to obtain the input representations H. BERT is a multi-layer bidirectional Transformer based language model. It is designed to capture deep representations by jointly modeling both the left and right context of words. It is comprised of a stack of M Transformer blocks. BERT used as a sentence encoder works as follows,

$$H^{(0)} = SW_s + W_p \tag{10}$$

$$H^{(m)} = \text{Transformer}(H^{(m-1)}), \quad m \in [1, M] \tag{11}$$

where S is the matrix of one-hot vectors of sub-words indexes in the input sentence, W_s is the sub-words embedding matrix, W_p is the positional embedding matrix, $H^{(m)}$ is the output of the m-th layer Transformer. The output of the last layer Transformer is usually extracted to form the shared representation $H = H^{(M)}$.

3 EXPERIMENT

The RIN-EPI is a generalized model that can be adapted to other related tasks. On that note, we consider to evaluate our model on two important information extraction tasks, namely joint entity and relation extraction, and aspect and opinion terms co-extraction. We also study ablated models of RIN-EPI to validate our design decision. Ablated models include,

(1) RIN-EPI$_{\text{w/o interaction}}$: excludes the interaction network used in RIN-EPI. That is, the initial shared features $H^{(0)}$ modeled by the BiLSTM network is directly passed into the classifiers $C_{\mathcal{Y}}$ and $C_{\mathcal{Z}}$ for task-specific predictions.
(2) RIN-EPI$_{\text{w/o global}}$: does not consider the global interaction (i.e., the self-attention component) in the RIN-EPI architecture.

3.1 Experimental results on the joint extraction of entities and relations

3.1.1 Datasets. On the joint entity and relation extraction task, we conduct experiments on four public benchmark datasets, namely NYT [50], WebNLG [12], NYT10 [50] and NYT11 [22]. These datasets are originally created by distant supervised methods, and adapted by [71] for a relational triple extraction task. We directly use the preprocessed NYT and WebNLG datasets released by [71].[1] It is worth to note that the heads of entities in the preprocessed datasets are unmarked in NYT and WebNLG, we therefore take a further step to distinguish entities by tagging them using the conventional BIOES tagging scheme [11]. We also use the preprocessed NYT10 and NYT11 datasets released by [54].

To probe the effectiveness of our model in dealing with different types of overlapping relations, we follow the categorization framework of the previous study [60] and divide dataset samples into three categories according to their overlapping patterns, namely, *Normal*, *EntityPairOverlap* (EPO) and *SingleEntityOverlap* (SEO). Specifically, if two relational triples share a common entity pair in a sample, then the sample belongs to *EntityPairOverlap*. For instance, a sample containing the relational triples (Brown, born_in, England) and (Brown, country_from, England) belong to the category *EntityPairOverlap* because both relational triples share the same head and tail entities but differ in the relation.[2] On the other hand, if two relational triples in a sample share a single entity only, then the sample belongs to *SingleEntityOverlap*. An example of such a sample is one that contains the relational triples (Jack, live_in, Washington) and (Washington, capital_of, America). Note, due to the complex nature of language in general, it is worth noting that some samples may belong to both *EntityPairOverlap* and *SingleEntityOverlap* categories. We give an example as an illustration. Consider the sample "Brown was born in London which is the captial city of England." Based on the existence of the relational triples (Brown, born_in, London) and (London, capital_of, England) in the sample, the sample is categorised as SEO, but considering that the relational triple (England contains, London) also exists, we additionally categorize the sample as EPO.

The statistics of these datasets are summarized in Table 1.

	NYT		WebNLG		NYT10		NYT11	
Category	Train	Test	Train	Test	Train	Test	Train	Test
Normal	37013	3266	1596	246	59396	2963	53395	368
EPO	9782	978	227	26	5376	715	2100	0
SEO	14735	1297	3406	457	8772	742	7365	1
ALL	56195	5000	5019	703	70339	4006	62648	369

Table 1. Statistics of the relation extraction datasets.

	NYT		WebNLG	
Dataset	Train	Test	Train	Test
Multi-token entities	39.1%	38.9%	64.2%	63.8%
Single-token entities	60.9%	61.1%	35.9%	36.2%
Relations	24	24	170	170

Table 2. Percentages of multi-token entities and single-token entities, and the number of relations on NYT and WebNLG.

[1] https://github.com/xiangrongzeng/copy_re

[2] The direction of the relation is not considered as like previous studies [60]

	NYT		WebNLG		NYT10		NYT11	
Dataset	Train	Test	Train	Test	Train	Test	Train	Test
Complex	39.8%	41.1%	69.5%	66.3%	19.2%	33.0%	15.1%	0.0%

Table 3. The percentage of complex samples (i.e. samples with overlapping relations).

3.1.2 Evaluation Protocol. On the entities and relation extraction task, we evaluate our models using the *Partial Match* and *Exact Match* retrieval tasks. The Partial Match task requires the relation and the heads of both subject and object entities of the extracted relational triple to be correct. The *Exact Match* on the other hand is more strict. It requires that the heads and tails of both subject and object entities as well as the relation are all correct. In other words, the extracted relational triple completely matches the gold relational triple.

Recent works [5, 11, 15, 22, 31, 41, 49, 54, 61, 69–71, 78, 79] extracting entities and relations evaluate on either the NYT and WebNLG dataset pair, or NYT10 and NYT11 dataset pair. For a fair comparison, we separately compare the performance of our method with recent methods that evaluate on either one of the dataset pairs. Specifically, we report the Precision (Prec), Recall (Rec) and micro-F1 (F1) scores of our method on the datasets for Partial Match and Exact Match and compare with other recent methods according to the dataset pairs they conduct experiments on. All reported results of our model the mean results over five runs using different random seeds.

3.1.3 Implementation Details. We build our models by employing the pretrained Glove vectors [45] to represent word embeddings. Due to the success of BERT in several NLP applications, we also consider a variant of our models where we represent word embeddings using the pretrained BERT architecture [9].

On the Glove-based models, we represent each sentence with Glove word vectors and pass these as input embeddings to a BiLSTM network which goes on to output the shared representation H. We improve learning by using dropout regularization on the input embeddings. The models are trained using an Adam optimizer [25] with a learning rate of e^{-3} and a batch size of 32. On the BERT-based models, we employ the pretrained BERT architecture as an encoder which takes in the sentence and output the shared representations H. The models are trained using an Adam optimizer [25] with a learning rate e^{-5} and a batch size of 6.

On the joint entity and relation extraction task, we threshold the probabilities of the prediction and return only the relations with probability values ≥ 0.5. Noting that the datasets NYT10 and NYT11 have no access to an official development set (see Table 1), we randomly select 10% samples from the training set and use as the development set. The hyper-parameters including the dropout rate, word embedding dimension, BiLSTM embedding dimension, interaction layers are set empirically and manually tuned on the development set to select the best model. We implement our model using PyTorch on a Linux machine with a GPU device NVIDIA V100 NVLINK 32GB.

3.1.4 Performance Comparison. On the NYT and WebNLG, we compare our method RIN-EPI and ablated models RIN-EPI$_{\text{w/o interaction}}$ and RIN-EPI$_{\text{w/o global}}$ with several competitive methods including NovelTagging [79], OneDecoder [71], MultiDecoder [71], OrderRL [70], CopyMLT [69], GraphRel [11] and CASREL [61]. Among these methods, NovelTagging and CASREL consider a sequence labelling approach, OneDecoder, MultiDecoder and OrderRL consider a sequence-to-sequence (seq2seq) approach while CopyMLT and GraphRel consider a MTL-based approach to address the problem. On the NYT10 and NYT11 datasets, we compare with popular methods including MultiR [22], FCM [15], SPTree [41], CoType [49], NovelTagging [79], LSTM-CRF [78],MultiDecoder [71], PA-LSTM-CRF [5] and HRL [54].

Tables 4 and 5 present our results. Table 4 shows the prediction performance on NYT10 and NYT11 while Table 5 shows the performance on NYT and WebNLG. The upper portion of each table presents evaluation on the Partial Match task while the lower portion corresponds to the Exact Match Task. The separation of the results into two tables (i.e., Table 4 and 5) comes from the fact that proposed methods conduct experiments on either the

Evaluation	Model	NYT10			NYT11		
		Prec	Rec	F1	Prec	Rec	F1
Partial Match	MultiDecoder	56.9	45.2	50.4	34.7	53.4	42.1
	$\text{CASREL}_{\text{BERT}}$	77.7	68.8	73.0	50.1	58.4	53.9
	$\text{RIN-EPI}_{\text{w/o interaction}}$	74.8	65.1	69.6	50.6	54.4	52.4
	$\text{RIN-EPI}_{\text{with RP}}$	77.8	66.4	71.6	52.3	56.0	54.1
	$\text{RIN-EPI}_{\text{w/o global}}$	78.7	67.1	72.4	54.1	56.9	55.9
	RIN-EPI	80.3	67.3	73.2	**56.5**	57.3	56.8
	$\text{RIN-EPI}_{\text{BERT}}$	**79.4**	**70.5**	**74.7**	56.4	**59.2**	**57.8**
Exact Match	NovelTagging	59.3	38.1	46.4	46.9	48.9	47.9
	MultiDecoder	56.9	45.2	50.4	34.7	53.4	42.1
	ReHession	-	-	-	41.2	57.3	48.0
	LSTM-CRF	-	-	-	69.3	31.0	42.8
	SPTree	49.2	55.7	52.2	52.2	54.1	53.1
	PA-LSTM-CRF	-	-	-	49.4	59.1	53.8
	HRL	71.4	58.6	64.4	53.8	53.8	53.8
	$\text{RIN-EPI}_{\text{w/o interaction}}$	69.7	60.2	64.6	47.9	53.4	50.5
	$\text{RIN-EPI}_{\text{with RP}}$	75.3	63.4	68.9	52.0	54.5	53.3
	$\text{RIN-EPI}_{\text{w/o global}}$	76.9	64.9	70.4	53.9	56.3	55.0
	RIN-EPI	77.6	66.1	71.4	54.9	**57.3**	56.1
	$\text{RIN-EPI}_{\text{BERT}}$	**78.9**	**70.2**	**74.3**	**57.8**	56.2	**57.0**

Table 4. Precision, Recall and F1 performance of different models on NYT10 and NYT11 datasets. The results of compared models are retrieved from [61].

NYT10 and NYT11 dataset pair or the NYT and WebNLG dataset pair. To show the effectiveness of our approach we present the prediction performance on all datasets.

We first compare ablated models with previous methods. We observe that $\text{RIN-EPI}_{\text{w/o interaction}}$ and $\text{RIN-EPI}_{\text{w/o global}}$ significantly outperforms seq2seq models including OrderRL, CopyMTL-one and CopyMTL-Mul on the NYT and WebNLG datasets for the partial match task. In a recent related study [61], authors found out that seq2seq models find it difficult to deal with the overlapping relation problem for the task. This may explain why these models perform poorly. Our empirical results goes on to support the study. We also observe that CASREL shows a competitive performance with $\text{RIN-EPI}_{\text{w/o interaction}}$ on NYT and WebNLG. However, by additionally modelling local interactions for each individual task through multiple interaction layers, we show that our ablated mode $\text{RIN-EPI}_{\text{w/o global}}$ can achieve better results as compared to CASREL on the datasets. We also experimented with a variant of our model, namely, $\text{RIN-EPI}_{\text{with RP}}$; an RIN-EPI that uses the raw predictions instead of representing predictions as vectors. The idea is to investigate if utilizing predictions as vectors contribute to the performance since raw predictions in a similar setting has shown to improve performance elsewhere [77]. We found that using raw predictions can lead to suboptimal performance, specifically when comparing $\text{RIN-EPI}_{\text{with RP}}$ and RIN-EPI. It could be that the vector representations may capture latent interactions between the text input and the prediction itself, and this might explain why we achieve such good results. Notably, our main model RIN-EPI improves over the ablated model architectures suggesting the importance our local and global interactions, and more importantly RIN-EPI outperforms previous methods.

We perform a further introspection of our results and note that on the WebNLG dataset the F1 performance of RIN-EPI significantly drops from the partial match to the exact match task. As 60% of entities in WebNLG are multi-token entities (as shown in Table 2), the exact match task on this dataset will be exceptionally difficult since the task requires the model to extract both the head and tail of the entity. We assessed the sensitivity of RIN-EPI on

Evaluation	Model	NYT			WebNLG		
		Prec	Rec	F1	Prec	Rec	F1
Partial Match	OneDecoder	59.4	53.1	56.0	32.2	28.9	30.5
	MultiDecoder	61.0	56.6	58.7	37.7	36.4	37.1
	OrderRL	77.9	67.2	72.1	63.3	59.9	61.6
	CASREL	84.2	83.0	83.6	86.9	80.6	83.7
	$\text{CASREL}_{\text{BERT}}$	89.7	89.5	89.6	**93.4**	90.1	**91.8**
	$\text{RIN-EPI}_{\text{w/o interaction}}$	83.9	83.1	83.5	84.9	86.3	85.6
	$\text{RIN-EPI}_{\text{with RP}}$	85.8	86.2	86.0	85.9	87.4	86.7
	$\text{RIN-EPI}_{\text{w/o global}}$	87.2	86.6	86.9	87.6	87.0	87.3
	RIN-EPI	87.6	87.8	87.7	87.9	88.0	88.0
	$\text{RIN-EPI}_{\text{BERT}}$	**90.2**	**90.4**	**90.3**	91.2	**91.9**	91.6
Exact Match	NovelTagging	62.4	31.7	42.0	52.5	19.3	28.3
	GraphRel_{1p}	62.9	57.3	60.0	42.3	39.2	40.7
	GraphRel_{2p}	63.9	60.0	61.9	44.7	41.1	42.9
	CopyMTL-One	72.7	69.2	70.9	57.8	60.1	58.9
	CopyMTL-Mul	75.7	68.7	72.0	58.0	54.9	56.4
	$\text{RIN-EPI}_{\text{w/o interaction}}$	77.4	76.4	76.9	75.0	73.3	74.2
	$\text{RIN-EPI}_{\text{with RP}}$	81.7	83.0	82.4	75.9	76.7	76.2
	$\text{RIN-EPI}_{\text{w/o global}}$	82.7	84.3	83.5	77.3	76.8	77.0
	RIN-EPI	84.5	84.5	84.5	77.0	78.4	77.7
	$\text{RIN-EPI}_{\text{BERT}}$	**87.6**	**88.6**	**88.1**	**81.3**	**83.4**	**82.3**

Table 5. Precision, Recall and F1 performance of different models on NYT and WebNLG datasets. The results of CopyMTL are retrieved from its original paper, and results of other models are retrieved from [61]

the entity recognition task and noticed that the F1 performance for single-token entity extraction is 93.6% while that of multi-token entities is 84.8%. We infer from these results that RIN-EPI will therefore perform better on the partial match as compared to the exact match task on WebNLG.

As BERT-based models are currently ranked high on the leaderboards for several NLP tasks, we employ pre-trained BERT embeddings to improve the prediction performance of RIN-EPI. Results for our bert-based model $\text{RIN-EPI}_{\text{BERT}}$ shows significant improvement over RIN-EPI on both the partial match and exact match tasks. The results indicate the importance of incorporating prior knowledge induced by BERT for the joint extraction task. Notably, $\text{RIN-EPI}_{\text{BERT}}$ outperforms $\text{CASREL}_{\text{BERT}}$ on NYT10 and NYT11, shows competitive performance with $\text{CASREL}_{\text{BERT}}$ on WebNLG and slightly outperforms $\text{CASREL}_{\text{BERT}}$ on NYT. The slight or competitive performance of $\text{RIN-EPI}_{\text{BERT}}$ over $\text{CASREL}_{\text{BERT}}$ on NYT and WebNLG is a concern. In our analysis we count the number of samples with overlapping relations (or "complex" samples). Table 3 presents the results. We see that NYT and WebNLG have relatively large proportions of samples with overlapping relations as compared to NYT10 and NYT11. Hence extracting relational facts from NYT and WebNLG is more challenging for RIN-EPI, explaining the slight or competitive performance with $\text{CASREL}_{\text{BERT}}$. To additionally validate our claim, we conduct a detailed experiment in a later section to compare RIN-EPI and CASREL on samples with different overlapping relation types.

3.1.5 Impact of Interaction Layer K. We investigate the impact of the depth of interaction layers of $\text{RIN-EPI}_{\text{w/o global}}$ and RIN-EPI on the NYT10 and WebNLG datasets. Recall, the hyper-parameter K is the number of interaction layers of our model. Meaning, K measures the degree of freedom required to model explicit interactions between the ER and RC tasks. As complex interactions are modelled with increasing depth, we expect the prediction performance to increase accordingly on both tasks.

Figure 4 shows the F1 curves of RIN-EPI$_{\text{w/o global}}$ and RIN-EPI for increasing values of K on NYT10 and WebBLG. At $K = 0$ both RIN-EPI and RIN-EPI$_{\text{w/o global}}$ are reduced to RIN-EPI$_{\text{w/o interaction}}$, explaining why the performance is equal for both models on each dataset for the partial and exact match tasks. Besides, the performance for both models at $K = 0$ is generally lower when compared with the performance at $K > 0$, indicating the importance of modelling explicit interactions across tasks. We also observe that the performance curve is similar, with the rise and fall in model performance with increasing values of K. More specifically, we observe a sharp rise in performance from $K = 0$ to $K = 1$, which indicates that modelling interactions implicitly may not sufficiently capture the complex interactions between the two tasks. Also at deeper layers (around $K > 4$) of RIN-EPI/ RIN-EPI$_{\text{w/o global}}$ we observe a high performance degradation which may be a result of overfitting. However, RIN-EPI generally has a higher performance as compared to RIN-EPI$_{\text{w/o global}}$ with increasing K, suggesting the importance of modelling global interactions irrespective of the depth of the interaction mechanism across the ER and RC tasks.

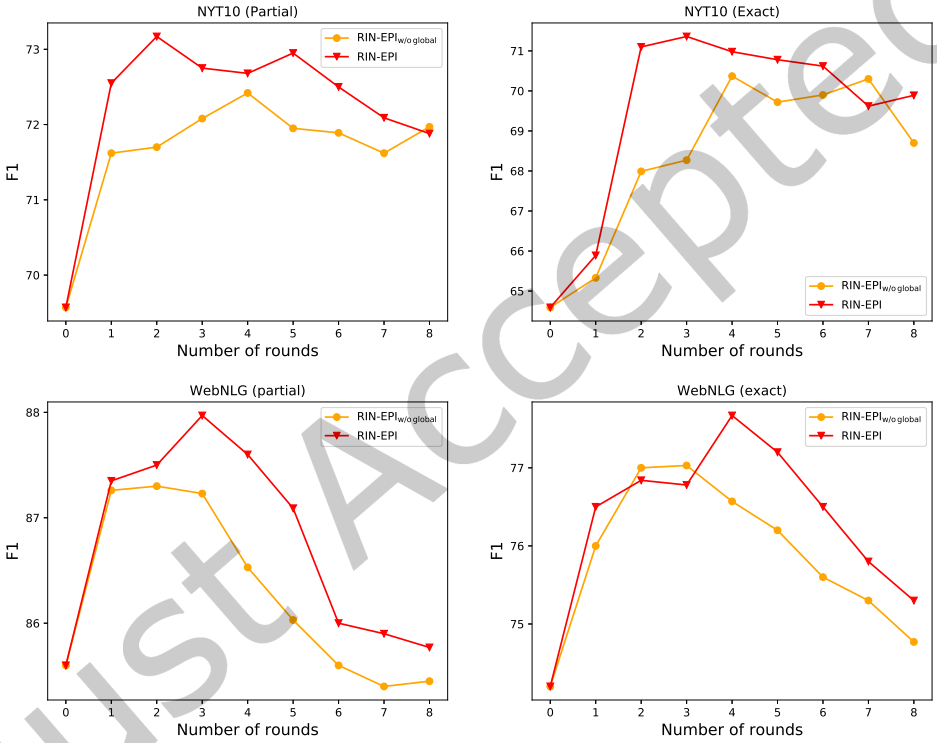


Fig. 4. Performance of different models on the NYT10 and WebNLG with different interaction rounds K.

3.1.6 Performance on Individual Relation types. In this section we conduct experiments on the NYT dataset to show the breakdown performance of models for different relation types and how they relate in their involvement with different entity types. We only focus on the top five relations, which account for 75% of the total number of relations in the train set. Table 6 shows the proportion of the top 5 relations. Table 7 also shows the largest proportion of entity types associated with the top 5 relations.

Table 8 shows the performance breakdown for different relation types in the NYT test set. Needless to say, the relation type with the largest proportion in the train set, in this case "llc", is expected to have the highest

Relation	#Percentage
/location/location/contains (llc)	48.6%
/people/person/nationality (ppn)	7.48%
/people/person/place_lived (ppp)	7.00%
/location/administrative_division/country (lac)	5.79%
/business/person/company (bpc)	5.29%

Table 6. The proportion of top 5 relations in the NYT training set.

Relation	#E1	#E2
/location/location/contains (llc)	country (34.98%)	city (47.59%)
/people/person/nationality (ppn)	person (97.42%)	country (98.59%)
/people/person/place_lived (ppp)	person (97.35%)	state_or_province (49.45%)
/location/administrative_division/country (lac)	city (61.48%)	country (98.67%)
/business/person/company (bpc)	person (98.33%)	organization (79.93%)

Table 7. The largest proportion of entity types associated with the the top 5 relations in the NYT training set. E1 and E2 denote the head and tail entity types respectively.

Model	llc	ppn	ppp	lac	bpc
RIN-EPI$_{\text{w/o interaction}}$	85.1	78.5	75.0	89.2	78.6
RIN-EPI$_{\text{w/o global}}$	88.5	84.1	76.7	94.2	83.2
RIN-EPI	89.5	85.0	78.1	94.6	86.1
RIN-EPI$_{\text{BERT}}$	91.8	87.4	82.3	95.0	88.2

Table 8. F1 Performance breakdown for different relation types in the NYT test set.

performance. However, we obtain some interesting results where we find "llc" being ranked the first runner up. It happens that the performance is also based on other factors - such as the proportion of entity types associated with the relation type. For instance the proportions of the relation types "ppn" and "ppp" are about 7% while "lac" and "bpc" are about 5%. But it can be noticed that the performance on "ppn" outperforms "ppp" while the performance on "lac" outperforms that of "bpc". By taking note of the proportion of entity types for the head (E1) and tail (E2) for a specific relation type, we believe RIN-EPI / RIN-EPI$_{\text{BERT}}$ also achieves performance when the proportion of tail entities with a specific entity type is relatively larger than that of head entities with a specific entity type.

3.1.7 Performance on Samples with Different Types of Overlapping Patterns. In this section we validate the effectiveness of our model in extracting overlapping relational triples from samples with different types of overlapping patterns (i.e., *Normal*, *EPO* and *SEO*) on NYT, WebNLG and NYT10. Almost all samples in the NYT11 test belong to *Normal* (with the exception of a single sample that belongs to *SEO*). This makes it impractical to evaluate the effectiveness of extracting relational triples of different overlapping patterns. As such, we ignore NYT11 in this particular study. On the NYT and WebNLG datasets we conduct experiments on RIN-EPI, the BERT-based models RIN-EPI$_{\text{BERT}}$ and CASREL$_{\text{BERT}}$ and the recent popular methods GraphRel and OrderRL. Meanwhile on NYT10 we particularly focus on the performance of RIN-EPI, RIN-EPI$_{\text{BERT}}$ and CASREL$_{\text{BERT}}$ for the different overlapping patterns.

Figure 5 shows the results of our experiments. With the increase in difficulty in extracting relational triples, starting from *Normal*, to *EPO*, then to *SEO*, capturing complex interactions between entities and relations becomes

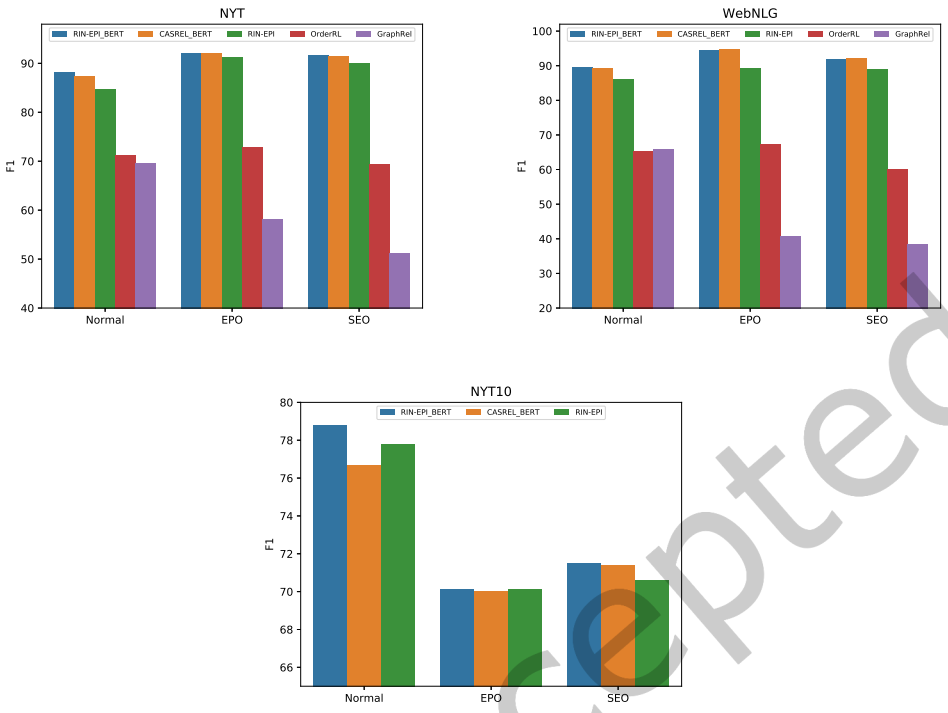


Fig. 5. F1 performance of models for different categories of overlapping samples on the partial match task.

the utmost importance for model performance. We find that GraphRel and OrderRel exhibit a decreasing trend in performance with the increase in difficulty when extracting samples with different types of overlapping patterns on NYT and WebNLG, insinuating are not adept in dealing with overlapping patterns. On the other hand, we find that RIN-EPI, the BERT-based models RIN-EPI$_{\text{BERT}}$ and CASREL$_{\text{BERT}}$ achieve comparable performance for the different types of overlapping patterns on NYT and WebNLG, suggesting their ability to deal with overlapping relations. However, on the NYT10 our models as well as CASREL$_{\text{BERT}}$ find it difficult to difficult to deal with overlapping relations. But the performance of RIN-EPI$_{\text{BERT}}$ as against CASREL$_{\text{BERT}}$ is quite notable, particularly for *Normal*, suggesting a well designed model architecture.

3.1.8 Ablation Study. In this section we conduct ablation experiments to provide valuable insights about the contribution of different components of the RIN-EPI architecture. To this end, we design the following ablated models:

(1) RIN-EPI$_{\text{w/o ER}}$: A RIN-EPI architecture where the self-attention mechanism of each layer takes only the previous shared representation and the prediction of the relation classification module as input, ignoring the entity prediction from the entity recognition module. That is, the entity prediction is not considered in intermediate layers is not considered when modelling explicit interactions across tasks.
(2) RIN-EPI$_{\text{w/o RC}}$: Similarly, this model follows the RIN-EPI architecture where the self-attention mechanism of each layer takes only the previous shared representation and the entity prediction, ignoring relation predictions

Model	NYT	WebNLG	NYT10	NYT11
RIN-EPI	**84.5**	**77.7**	**71.4**	**56.1**
RIN-EPI$_{\text{w/o ER}}$	82.2	77.0	69.8	54.7
RIN-EPI$_{\text{w/o RC}}$	78.9	76.1	67.4	52.3
RIN-EPI$_{\text{w/o GRU}}$	83.2	76.8	70.3	55.4

Table 9. F1 Performance of ablated model architectures on the exact match task.

from the relation classification module. Also here, it means the relation predictions in intermediate layers are not considered when modelling explicit interactions.

(3) RIN-EPI$_{\text{w/o GRU}}$: This is a RIN-EPI model that replaces the GRU with a multi-layer perceptron (MLP). The goal of this model is to examine how GRU can be compared with other neural networks.

We apply the ablated models on our datasets and report the results in Table 9. Results show that RIN-EPI deteriorates as we remove critical model components. Specifically, RIN-EPI$_{\text{w/o ER}}$ and RIN-EPI$_{\text{w/o RC}}$ both show a drop in performance, indicating the importance of utilising the intermediate predictions of both tasks for the modelling of explicit interactions across tasks. However, it seems the relation prediction is relatively more important than the entity prediction since the drop in performance of RIN-EPI$_{\text{w/o RC}}$ is quite substantial. Lastly, by replacing the GRU with the MLP we observe that the performance of RIN-EPI drops as shown by RIN-EPI$_{\text{w/o GRU}}$ on all datasets. This result suggest that the choice of neural network used to model local interactions is important, and GRU is adept for the task as compared to MLP.

3.2 Experimental results on aspect and opinion term extraction

In this section we demonstrate that the RIN-EPI architecture can be adapted for the aspect and opinion term co-extraction task. Particularly, both the aspect and opinion extraction tasks are treated as sequence labelling problems. That is, for the RIN-EPI architecture the $\mathcal{Y}-$ and $\mathcal{Z}-$ tasks may correspond to the aspect and opinion extraction tasks respectively.

3.2.1 Datasets. We conduct experiments and evaluate our model on two public benchmark datasets, namely SemEval-2014 Restaurants[3] and SemEval-2015 Restaurants[4]. These datasets are widely used for evaluation in recent works [6, 28, 57–59, 64, 65, 73]. Noting that the original datasets do not have annotations for opinion words, [58] and [59] manually annotated opinion words on the datasets. Later works make use of these annotations and preprocess the datasets. We directly use the preprocessed datasets provided by [6]. The dataset statistics are shown in Table 10.

Dataset	#Sentence	#AT	#OT
Rest14 (Train)	3044	3699	3528
Rest14 (Test)	800	1134	1021
Rest15 (Train)	1315	1279	1216
Rest15 (Test)	685	597	517

Table 10. Statistics of the aspect and opinion terms co-extraction datasets. #Sentence donates the number of sentences; #AT donates the number of aspect terms; #OT donates the number of opinion terms.

[3]https://alt.qcri.org/semeval2014
[4]https://alt.qcri.org/semeval2015

3.2.2 Evaluation Protocol. We refer to the BIO tags to extract aspect and opinion terms in the aspect extraction subtask and opinion term subtask. An extracted span in text is regarded as a correct aspect term (or opinion term) if the span exactly matches the golden span. Following previous works, we report the micro-F1 (F1) scores of our model on the two subtasks. All reported results of our model the mean results over five runs using different random seeds.

Note, the hyper-parameter settings of RIN-EPI in this application is similar to that used for the joint entity and relation classification task.

3.2.3 Performance Comparison. To further assess the effectiveness of RIN-EPI, we conduct experiments on the aspect and opinion terms co-extraction task. We compare with previous works including rule-based models WDEmb [65] and Elixa [57], dependency-based models RNCRF [58], attention-based models CMLA [59] and HAST [28], some recent SOTA baselines including NCRF-AE [73], DE-CNN [64], RINANTE-Double-Pre [6] and SpanMlt [76]. Table 11 presents the results.

	Rest14		Rest15	
Model	Aspect	Opinion	Aspect	Opinion
WDEmb	84.97	-	69.73	-
RNCRF	82.23	83.93	65.39	63.75
CMLA	82.46	84.67	68.22	70.50
NCRF-AE	83.28	85.23	65.33	70.16
Elixa	-	-	70.04	-
HAST	85.61	-	69.77	-
DE-CNN	85.20	-	68.28	-
RINANTE-Double-Pre	86.45	85.67	69.90	72.09
SpanMlt	85.24	85.79	71.07	75.02
RIN-EPI$_{\text{w/o interaction}}$	84.96	85.87	69.56	70.86
RIN-EPI$_{\text{w/o global}}$	85.69	86.53	70.93	72.97
RIN-EPI	85.97	87.24	71.98	74.99
RIN-EPI$_{\text{BERT}}$	**86.94**	**87.71**	**72.44**	**78.50**

Table 11. F1 performance of different models on Rest14 and Rest15 for the aspect and opinion terms co-extraction task. The results of the compared models are retrieved from [6].

First glance over the results show that RIN-EPI$_{\text{BERT}}$ achieves the best performance on Rest14 and Rest15, outperforming our own glove-based model RIN-EPI. The results suggest that BERT can easily adapt to other datasets on a different task and still achieve performance. Introspecting the results further, we notice that RIN-EPI outperforms the current state-of-the-art method SpanMLT. Note, RIN-EPI employs glove embeddings while SpanMLT employs contextualized embeddings, particularly ELMo emebddings [46]. However, SpanMLT cannot make the best of the prior knowledge of ELMo embeddings to achieve performance. Besides, authors of [76] earlier on explored BERT embeddings for SpanMLT but failed to achieve performance. Given that our models RIN-EPI and RIN-EPI$_{\text{BERT}}$ outperform SpanMLT, we can note that our model architecture is effective in utilizing pretrained knowledge captured in the word embeddings for performance. Interestingly, we also observe competitive performance between RINANTE-Double-Pre and RIN-EPI$_{\text{BERT}}$ on the aspect extraction task. We believe RINANTE-Double-Pre takes advantage of manually extracted rules specific to the domain to achieve such performance.

3.2.4 Case Study. In this section we present a comparative case study to demonstrate the importance of different model components operating on the aspect and opinion word co-extraction task. Table 12 presents the results of the study.

Among other observations, we find that the extraction of aspect and opinion terms are relatively difficult when the distance between these terms increases. Particularly, in Case1 where the opinion "good" and aspect "meal" lie side by side in the text, we observe that all models are able to effectively extract the aspect and opinion. However, the extraction gets difficult with increasing distance as seen in Case2 and Case3. Among the compared models, only RIN-EPI effectively extracts all aspect and opinion terms in Case2 and Case3. We attribute the performance of RIN-EPI to the modelling of local and global interactions that capture the dependencies between information relating to the aspect and opinion terms in the text. Without modelling such interactions, the model is prone to fail as seen in the result produced by RIN-EPI$_{\text{w/o interaction}}$ on Case2 and Case3.

Another observation is the impact on model performance when there are multiple aspects or opinion words in the text. In Case4, there are two opinion words "small" and "perfect", and one aspect "boths". We find that "both" is relatively close to "small" as compared to "perfect". As a result, it is no surprise that "boths" and "small" are identified by all models. But to identify "perfect" may require a higher-order modelling of complex interactions. Results of RIN-EPI on Case4 shows that this can be achieved by modelling global interactions.

Case1: *I paid just about $60 for a good meal, though.*	RIN-EPI$_{\text{w/o interaction}}$: good, meal RIN-EPI$_{\text{w/o global}}$: good, meal RIN-EPI: good, meal
Case2: *There was no tap beer that evening, which was a disappointment.*	RIN-EPI$_{\text{w/o interaction}}$: disappointment RIN-EPI$_{\text{w/o global}}$: disappointment, beer RIN-EPI: disappointment, beer
Case3: *The place is a bit hidden away, but once you get there, it's all worth it.*	RIN-EPI$_{\text{w/o interaction}}$: place RIN-EPI$_{\text{w/o global}}$: place RIN-EPI: worth, place
Case4: *The boths are not as small as some of the reviews make them out to look they're perfect for 2 people.*	RIN-EPI$_{\text{w/o interaction}}$: small, boths RIN-EPI$_{\text{w/o global}}$: small, boths RIN-EPI: small, perfect, boths

Table 12. Results on case examples for aspect and opinion word extraction. Opinion and aspect terms are marked in blue and orange texts respectively.

3.2.5 Attention Visualization for Global Interaction. Recall that our RIN-EPI model architecture as described in Section 2.2 employs a self-attention mechanism module in each layer to model global interactions across tasks (i.e., modelling the relationship across tasks). Thus, in the aspect and opinion terms co-extraction task, this module takes the previous shared representation H, aspect term predictions $p(Y)$ and opinion term predictions $p(Z)$ (subscript removed for conciseness), to extract a shared representation that capture the global interaction of the two tasks. To show that the self-attention mechanism is well designed for the co-extraction task we consider a variant that excludes the injection of the task-specific predictions $p(Y)$ and $p(Z)$. We take two case examples from the previous section, Case2 and Case3 (see Table 12), to visualize the attention placed by the self-attention module on different parts of the text. Figure 6 and 7 show results on self-attention heatmaps for Case2 and Case3 respectively. On the left is the self-attention weights without injecting task-specific predictions, while the right is with task-specific predictions injected.

The heatmaps provide an intuitive explanation on how different parts of the text interact with each other to pay attention to important words (i.e., aspect and opinion words). We observe that in each case example, the left heatmap is severely sparser than the right heatmap. Introspecting the right heatmaps, we find that the self-attention module can discriminate important and unimportant words by assigning high weights to important words and low weights to unimportant words. On Case2 the goal is to extract the aspect "beer" and opinion "disappointment", while on

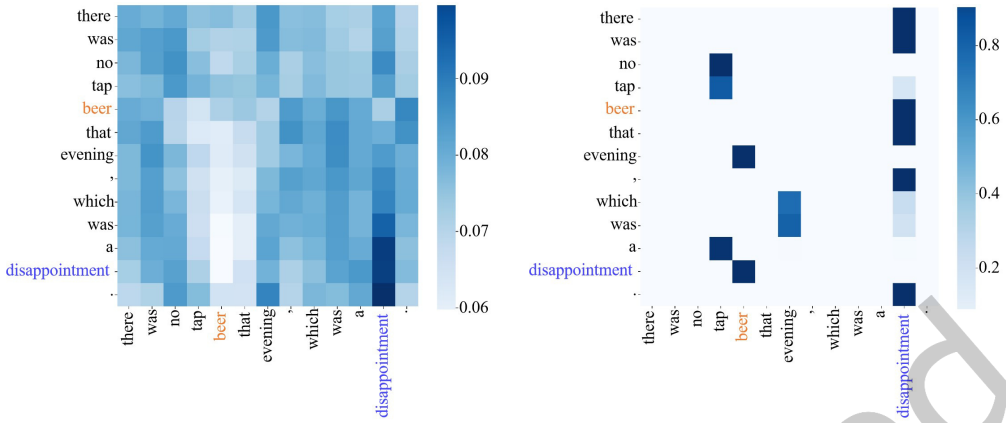


Fig. 6. Heatmaps for self-attention weights on **Case2** (from Table 12). On the left are the self-attention weights derived from the self-attention module without introducing previous task-specific predictions, while the right introduces previous task-specific predictions in the self-attention module. The aspect and opinion terms are marked in orange and blue respectively.

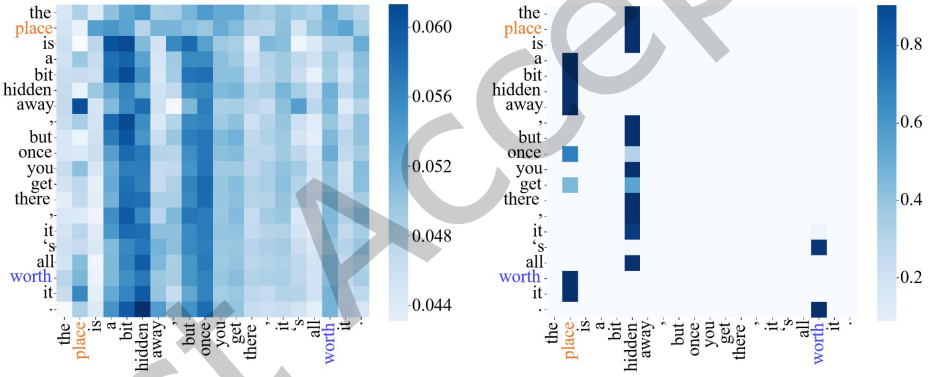


Fig. 7. Heatmaps for self-attention weights on **Case3** (from Table 12). On the left are the self-attention weights derived from the self-attention module without introducing previous task-specific predictions, while the right introduces previous task-specific predictions in the self-attention module. The aspect and opinion terms are marked in orange and blue respectively.

Case4 the goal is to extract the aspect “place” and opinion “worth”. By introducing the previous task-specific predictions to guide the extraction process, our proposed self-attention module effectively interacts and assign high attention weights to all important words for different positions of the text. More interestingly, is the attention placed on “tap” in Case2. In fact, “tap beer” in the text can be considered as a compound word. It is therefore not a surprise that the proposed self-attention module detects “tap” as an important word, although “beer” is the gold aspect in the text. We observe a similar behaviour on the right heatmap of Case3, where “hidden” is also identified as important. As a matter of fact, “hidden” conveys a negative opinion on the aspect “place”, and therefore can be extracted as an opinion word. This shows the importance of exploiting previous task-specific predictions.

4 CONCLUSION AND FUTURE WORK

In this work we introduced a Recurrent Interaction Network with an effective Early Prediction Integration (RIN-EPI) that builds on recent advances in multi-task learning based on deep neural networks. In particular, we first noted the correlations of early predictions generated in intermediate layers of individual task-specific networks. We exploit these correlations by developing a multi-task learning architecture that learns implicit interactions across tasks through a shared encoder and explicit interactions across tasks by effectively utilizing early predictions and shared features. Our experimental analysis shows strong improvements for the joint task of entity recognition and relation classification as well as aspect and opinion term co-extraction.

Our model offers new insights in utilizing early predictions for deep neural networks. Nonetheless, there is a possibility of randomness when it comes to early predictions since these predictions are dependent on the quality of the intermediate classifiers. Our model however, does not account for such randomness as deterministic maps (e.g. gated recurrent units) are used to construct the task-specific representations from both early predictions and shared features. Moreover, these early predictions cannot be naively considered as the ground-truth. As such, it is important to selectively choose only the relevant part of these early predictions that actually contribute to the task-specific representation modelling. This observation opens room for future research, which we intend to focus on.

5 ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China under Grant 2021ZD0110700, in part by the Fundamental Research Funds for the Central Universities, in part by the State Key Laboratory of Software Development Environment, and in part by a Leverhulme Trust Research Project Grant.

REFERENCES

[1] Heike Adel and Hinrich Schütze. 2017. Global Normalization of Convolutional Neural Networks for Joint Entity and Relation Classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. 1723–1729. https://doi.org/10.18653/v1/d17-1181

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1409.0473

[3] Yee Seng Chan and Dan Roth. 2011. Exploiting Syntactico-Semantic Structures for Relation Extraction. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*. 551–560. https://www.aclweb.org/anthology/P11-1056/

[4] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*. 160–167. https://doi.org/10.1145/1390156.1390177

[5] Dai Dai, Xinyan Xiao, Yajuan Lyu, Shan Dou, Qiaoqiao She, and Haifeng Wang. 2019. Joint Extraction of Entities and Overlapping Relations Using Position-Attentive Sequence Labeling. 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*. 6300–6308. https://doi.org/10.1609/aaai.v33i01.33016300

[6] Hongliang Dai and Yangqiu Song. 2019. Neural Aspect and Opinion Term Extraction with Mined Rules as Weak Supervision. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 5268–5277. https://doi.org/10.18653/v1/p19-1520

[7] Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. 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*. 2218–2229. https://doi.org/10.18653/v1/D19-1227

[8] Li Deng, Geoffrey Hinton, and Brian Kingsbury. 2013. New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 8599–8603.

[9] 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)*. 4171–4186. https://www.aclweb.org/anthology/N19-1423/

[10] Haihong E, Peiqing Niu, Zhongfu Chen, and Meina Song. 2019. A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 5467–5471. https://doi.org/10.18653/v1/p19-1544

[11] Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. GraphRel: Modeling Text as Relational Graphs for Joint Entity and Relation Extraction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. 1409–1418. https://www.aclweb.org/anthology/P19-1136/

[12] Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating Training Corpora for NLG Micro-Planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. 179–188. https://doi.org/10.18653/v1/P17-1017

[13] Shima Gerani, Mark Carman, and Fabio Crestani. 2012. Aggregation methods for proximity-based opinion retrieval. *ACM Transactions on Information Systems (TOIS)* 30, 4 (2012), 1–36.

[14] Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-Gated Modeling for Joint Slot Filling and Intent Prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*. 753–757. https://doi.org/10.18653/v1/n18-2118

[15] Matthew R. Gormley, Mo Yu, and Mark Dredze. 2015. Improved Relation Extraction with Feature-Rich Compositional Embedding Models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. 1774–1784. https://doi.org/10.18653/v1/d15-1205

[16] Pankaj Gupta, Hinrich Schütze, and Bernt Andrassy. 2016. Table Filling Multi-Task Recurrent Neural Network for Joint Entity and Relation Extraction. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*. 2537–2547. https://www.aclweb.org/anthology/C16-1239/

[17] Rujun Han, Qiang Ning, and Nanyun Peng. 2019. Joint Event and Temporal Relation Extraction with Shared Representations and Structured Prediction. 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*. 434–444. https://doi.org/10.18653/v1/D19-1041

[18] Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, 1923–1933. https://doi.org/10.18653/v1/d17-1206

[19] Ruidan He, Wee Sun Lee, and Hwee Tou Ng and‘ Daniel Dahlmeier. 2019. An Interactive Multi-Task Learning Network for End-to-End Aspect-Based Sentiment Analysis. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. 504–515. https://www.aclweb.org/anthology/P19-1048/

[20] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. 33–38.

[21] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[22] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke S. Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*. 541–550. https://www.aclweb.org/anthology/P11-1055/

[23] Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-Domain Targeted Sentiment Analysis via Span-Based Extraction and Classification. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. 537–546. https://doi.org/10.18653/v1/p19-1051

[24] Niklas Jakob and Iryna Gurevych. 2010. Extracting Opinion Targets in a Single and Cross-Domain Setting with Conditional Random Fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, 1035–1045. https://www.aclweb.org/anthology/D10-1101/

[25] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[26] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, Carla E. Brodley and Andrea Pohoreckyj Danyluk (Eds.). Morgan Kaufmann, 282–289.

[27] Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. 2017. Multi-task Attention-based Neural Networks for Implicit Discourse Relationship Representation and Identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. 1299–1308. https://doi.org/10.18653/v1/d17-1134

[28] Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect Term Extraction with History Attention and Selective Transformation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, Jérôme Lang (Ed.). ijcai.org, 4194–4200. https://doi.org/10.24963/ijcai.2018/583

[29] Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for End-to-End Aspect-based Sentiment Analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text, W-NUT@EMNLP 2019, Hong Kong, China, November 4, 2019*. 34–41. https://doi.org/10.18653/v1/D19-5505

[30] Xiangju Li, Kaisong Song, Shi Feng, Daling Wang, and Yifei Zhang. 2018. A Co-attention Neural Network Model for Emotion Cause Analysis with Emotional Context Awareness. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 4752–4757. https://doi.org/10.18653/v1/d18-1506

[31] Liyuan Liu, Xiang Ren, Qi Zhu, Shi Zhi, Huan Gui, Heng Ji, and Jiawei Han. 2017. Heterogeneous Supervision for Relation Extraction: A Representation Learning Approach. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. 46–56. https://doi.org/10.18653/v1/d17-1005

[32] Pengfei Liu, Jie Fu, Yue Dong, Xipeng Qiu, and Jackie Chi Kit Cheung. 2019. Learning Multi-Task Communication with Message Passing for Sequence Learning. 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*. 4360–4367. https://doi.org/10.1609/aaai.v33i01.33014360

[33] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Deep Multi-Task Learning with Shared Memory for Text Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, Jian Su, Xavier Carreras, and Kevin Duh (Eds.). The Association for Computational Linguistics, 118–127. https://doi.org/10.18653/v1/d16-1012

[34] Peng Liu, Lemei Zhang, and Jon Atle Gulla. 2021. Multilingual Review-aware Deep Recommender System via Aspect-based Sentiment Analysis. *ACM Transactions on Information Systems (TOIS)* 39, 2 (2021), 1–33.

[35] Qian Liu, Zhiqiang Gao, Bing Liu, and Yuanlin Zhang. 2015. Automated Rule Selection for Aspect Extraction in Opinion Mining. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, Qiang Yang and Michael J. Wooldridge (Eds.). AAAI Press, 1291–1297. http://ijcai.org/Abstract/15/186

[36] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-Task Deep Neural Networks for Natural Language Understanding. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. 4487–4496. https://doi.org/10.18653/v1/p19-1441

[37] Rui Mao and Xiao Li. 2021. Bridging Towers of Multi-task Learning with a Gating Mechanism for Aspect-based Sentiment Analysis and Sequential Metaphor Identification. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 13534–13542. https://ojs.aaai.org/index.php/AAAI/article/view/17596

[38] Mstislav Maslennikov and Tat-Seng Chua. 2010. Combining relations for information extraction from free text. *ACM Transactions on Information Systems (TOIS)* 28, 3 (2010), 1–35.

[39] Paolo Mignone, Gianvito Pio, Sašo Džeroski, and Michelangelo Ceci. 2020. Multi-task learning for the simultaneous reconstruction of the human and mouse gene regulatory networks. *Scientific Reports* 10, 1 (2020), 1–15.

[40] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. 2016. Cross-Stitch Networks for Multi-task Learning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 3994–4003. https://doi.org/10.1109/CVPR.2016.433

[41] Makoto Miwa and Mohit Bansal. 2016. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. https://www.aclweb.org/anthology/P16-1105/

[42] Makoto Miwa and Yutaka Sasaki. 2014. Modeling Joint Entity and Relation Extraction with Table Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. 1858–1869. https://www.aclweb.org/anthology/D14-1200/

[43] Toru Nishino, Shotaro Misawa, Ryuji Kano, Tomoki Taniguchi, Yasuhide Miura, and Tomoko Ohkuma. 2019. Keeping Consistency of Sentence Generation and Document Classification with Multi-Task Learning. 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. 3193–3203. https://doi.org/10.18653/v1/D19-1315

[44] Sinno Jialin Pan, Zhiqiang Toh, and Jian Su. 2013. Transfer joint embedding for cross-domain named entity recognition. *ACM Transactions on Information Systems (TOIS)* 31, 2 (2013), 1–27.

[45] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[46] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. 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)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 2227–2237. https://doi.org/10.18653/v1/n18-1202

[47] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion Word Expansion and Target Extraction through Double Propagation. *Comput. Linguistics* 37, 1 (2011), 9–27. https://doi.org/10.1162/coli_a_00034

[48] Lev-Arie Ratinov and Dan Roth. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning, CoNLL 2009, Boulder, Colorado, USA, June 4-5, 2009*, Suzanne Stevenson and Xavier Carreras (Eds.). ACL, 147–155. https://www.aclweb.org/anthology/W09-1119/

[49] Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, Tarek F. Abdelzaher, and Jiawei Han. 2017. CoType: Joint Extraction of Typed Entities and Relations with Knowledge Bases. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*. 1015–1024. https://doi.org/10.1145/3038912.3052708

[50] 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*. 148–163. https://doi.org/10.1007/978-3-642-15939-8_10

[51] Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 231–235.

[52] Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2020. Recurrent Interaction Network for Jointly Extracting Entities and Classifying Relations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 3722–3732. https://doi.org/10.18653/v1/2020.emnlp-main.304

[53] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. 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 1: Long Papers*. The Association for Computer Linguistics, 1556–1566. https://doi.org/10.3115/v1/p15-1150

[54] Ryuichi Takanobu, Tianyang Zhang, Jiexi Liu, and Minlie Huang. 2019. A Hierarchical Framework for Relation Extraction with Reinforcement Learning. 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*. 7072–7079. https://doi.org/10.1609/aaai.v33i01.33017072

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[56] Iñaki San Vicente, Xabier Saralegi, and Rodrigo Agerri. 2015. EliXa: A Modular and Flexible ABSA Platform. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, Daniel M. Cer, David Jurgens, Preslav Nakov, and Torsten Zesch (Eds.). The Association for Computer Linguistics, 748–752. https://doi.org/10.18653/v1/s15-2127

[57] Iñaki San Vicente, Xabier Saralegi, and Rodrigo Agerri. 2017. EliXa: A Modular and Flexible ABSA Platform. *CoRR* abs/1702.01944 (2017). arXiv:1702.01944 http://arxiv.org/abs/1702.01944

[58] Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive Neural Conditional Random Fields for Aspect-based Sentiment Analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, Jian Su, Xavier Carreras, and Kevin Duh (Eds.). The Association for Computational Linguistics, 616–626. https://doi.org/10.18653/v1/d16-1059

[59] Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled Multi-Layer Attentions for Co-Extraction of Aspect and Opinion Terms. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, Satinder P. Singh and Shaul Markovitch (Eds.). AAAI Press, 3316–3322. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14441

[60] Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2019. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction. arXiv:1909.03227 [cs.CL]

[61] Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. 1476–1488. https://www.aclweb.org/anthology/2020.acl-main.136/

[62] Changxing Wu, Xiaodong Shi, Yidong Chen, Yanzhou Huang, and Jinsong Su. 2016. Bilingually-constrained Synthetic Data for Implicit Discourse Relation Recognition. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. 2306–2312. https://doi.org/10.18653/v1/d16-1253

[63] Susmitha Wunnava, Xiao Qin, Tabassum Kakar, Xiangnan Kong, and Elke A. Rundensteiner. 2020. A Dual-Attention Network for Joint Named Entity Recognition and Sentence Classification of Adverse Drug Events. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020 (Findings of ACL, Vol. EMNLP 2020)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 3414–3423. https://doi.org/10.18653/v1/2020.findings-emnlp.306

[64] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double Embeddings and CNN-based Sequence Labeling for Aspect 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*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, 592–598. https://doi.org/10.18653/v1/P18-2094

[65] Yichun Yin, Furu Wei, Li Dong, Kaimeng Xu, Ming Zhang, and Ming Zhou. 2016. Unsupervised Word and Dependency Path Embeddings for Aspect Term Extraction. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, Subbarao Kambhampati (Ed.). IJCAI/AAAI Press, 2979–2985. http://www.ijcai.org/Abstract/16/423

[66] Bowen Yu, Zhenyu Zhang, Jianlin Su, Yubin Wang, Tingwen Liu, Bin Wang, and Sujian Li. 2019. Joint Extraction of Entities and Relations Based on a Novel Decomposition Strategy. *CoRR* abs/1909.04273 (2019). arXiv:1909.04273 http://arxiv.org/abs/1909.04273

[67] Jianfei Yu, Luís Marujo, Jing Jiang, Pradeep Karuturi, and William Brendel. 2018. Improving Multi-label Emotion Classification via Sentiment Classification with Dual Attention Transfer Network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 1097–1102. https://doi.org/10.18653/v1/d18-1137

[68] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel Methods for Relation Extraction. *J. Mach. Learn. Res.* 3 (2003), 1083–1106. http://jmlr.org/papers/v3/zelenko03a.html

[69] Daojian Zeng, Haoran Zhang, and Qianying Liu. 2019. CopyMTL: Copy Mechanism for Joint Extraction of Entities and Relations with Multi-Task Learning. arXiv:1911.10438 [cs.CL]

[70] Xiangrong Zeng, Shizhu He, Daojian Zeng, Kang Liu, Shengping Liu, and Jun Zhao. 2019. Learning the Extraction Order of Multiple Relational Facts in a Sentence with Reinforcement Learning. 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)*. Association for Computational Linguistics, Hong Kong, China, 367–377. https://doi.org/10.18653/v1/D19-1035

[71] Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Extracting Relational Facts by an End-to-End Neural Model with Copy Mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. 506–514. https://doi.org/10.18653/v1/P18-1047

[72] Richong Zhang, Samuel Mensah, Fanshuang Kong, Zhiyuan Hu, Yongyi Mao, and Xudong Liu. 2020. Pairwise link prediction model for out of vocabulary knowledge base entities. *ACM Transactions on Information Systems (TOIS)* 38, 4 (2020), 1–28.

[73] Xiao Zhang, Yong Jiang, Hao Peng, Kewei Tu, and Dan Goldwasser. 2017. Semi-supervised Structured Prediction with Neural CRF Autoencoder. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, 1701–1711. https://doi.org/10.18653/v1/d17-1179

[74] Xiaodong Zhang and Houfeng Wang. 2016. A Joint Model of Intent Determination and Slot Filling for Spoken Language Understanding. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*. 2993–2999. http://www.ijcai.org/Abstract/16/425

[75] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*. Springer, 94–108.

[76] He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and Hui Xue. 2020. SpanMlt: A Span-based Multi-Task Learning Framework for Pair-wise Aspect and Opinion Terms Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 3239–3248. https://doi.org/10.18653/v1/2020.acl-main.296

[77] Sendong Zhao, Ting Liu, Sicheng Zhao, and Fei Wang. 2019. A Neural Multi-Task Learning Framework to Jointly Model Medical Named Entity Recognition and Normalization. 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. 817–824. https://doi.org/10.1609/aaai.v33i01.3301817

[78] Suncong Zheng, Yuexing Hao, Dongyuan Lu, Hongyun Bao, Jiaming Xu, Hongwei Hao, and Bo Xu. 2017. Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing* 257 (2017), 59–66. https://doi.org/10.1016/j.neucom.2016.12.075

[79] Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. 1227–1236. https://doi.org/10.18653/v1/P17-1113

[80] Guodong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring Various Knowledge in Relation Extraction. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*. 427–434. https://www.aclweb.org/anthology/P05-1053/

[81] Li Zhuang, Feng Jing, and Xiaoyan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management, Arlington, Virginia, USA, November 6-11, 2006*, Philip S. Yu, Vassilis J. Tsotras, Edward A. Fox, and Bing Liu (Eds.). ACM, 43–50. https://doi.org/10.1145/1183614.1183625