Think Rationally about What You See: Continuous Rationale Extraction for Relation Extraction

Xuming Hu
Tsinghua University
hxm19@mails.tsinghua.edu.cn

Zhaochen Hong
Tsinghua University
hongzc20@mails.tsinghua.edu.cn

Chenwei Zhang
Amazon
cwzhang910@gmail.com

Irwin King
The Chinese University of Hong Kong
king@cse.cuhk.edu.hk

Philip S. Yu
University of Illinois at Chicago
psyu@cs.uic.edu

ABSTRACT

Relation extraction (RE) aims to extract potential relations according to the context of two entities, thus, deriving rational contexts from sentences plays an important role. Previous works either focus on how to leverage the entity information (e.g., entity types, entity verbalization) to inference relations, but ignore context-focused content, or use counterfactual thinking to remove the model's bias of potential relations in entities, but the relation reasoning process will still be hindered by irrelevant content. Therefore, how to preserve relevant content and remove noisy segments from sentences is a crucial task. In addition, retained content needs to be fluent enough to maintain semantic coherence and interpretability. In this work, we propose a novel rationale extraction framework named RE^2, which leverages two continuity and sparsity factors to obtain relevant and coherent rationales from sentences. To solve the problem that the gold rationales are not labeled, RE^2 applies an optimizable binary mask to each token in the sentence, and adjust the rationales that need to be selected according to the relation label. Experiments on four datasets show that RE^2 surpasses baselines.

CCS CONCEPTS

• **Computing methodologies → Information Extraction**.

KEYWORDS

Continuous Rationale Extraction, Relation Extraction

ACM Reference Format:
Xuming Hu, Zhaochen Hong, Chenwei Zhang, Irwin King, and Philip S. Yu. 2023. Think Rationally about What You See: Continuous Rationale Extraction for Relation Extraction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3592072

1 INTRODUCTION

Relation extraction (RE) is a crucial part of many information retrieval (IR) systems, which could extract relations between entities from sentences. These structured triplets such as *(Ryan, Yaz, per/per/alumni)* (Figure 1) from heterogeneous sources could benefit multiple downstream applications like question answering [18, 19] and natural language understanding [14, 20, 21]. To obtain structured triples, we need to exploit relevant and noise-free sentences from the entity context, so that the correct relation can be extracted between the two available entities. *Entity Thinking* methods such as Hu et al. [15, 16, 17] inject reserved special tokens <e> and </e> before and after the entity, and focus on the contextualized features of the entity through these special tokens. However, the semantic information of entities cannot be specified in special tokens. Therefore, Zhou and Chen [36] and Lu et al. [23] respectively introduce entity type and entity verbalization to better reveal the contextual semantic representation of entities and infer the relations between entities. Although Entity Thinking methods can better capture the context semantics of the entity, but cannot automatically remove noisy and irrelevant contents. Such noisy content tends to destroy correct relational inference. Taking Figure 1 as an example, the Entity Thinking method has no idea in judging the relevance of the content such as "they knew each other from school" and "me and my friend" for predicting the relation between entities. Therefore, the model may be misled by the word "friend", and mispredicts the relation as `Per/per/peer`. To remove the potential impact of the content on the relation extraction between entities, *Counterfactual Thinking* methods [24, 32] remove the model's bias against different words and entities. However, these methods do not focus on explicitly removing noisy contextual content, thus, the model's prediction can still be misled as `Per/per/siblings`.

To remove irrelevant and noisy content in sentences, we first propose rational thinking methods which could extract relevant

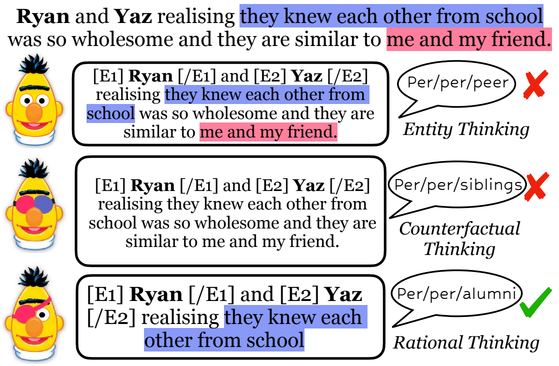
Ryan and **Yaz** realising they knew each other from school was so wholesome and they are similar to me and my friend.
[E1] **Ryan** [/E1] and [E2] **Yaz** [/E2] realising they knew each other from school was so wholesome and they are similar to me and my friend.
Per/per/peer ✗
Entity Thinking
[E1] **Ryan** [/E1] and [E2] **Yaz** [/E2] realising they knew each other from school was so wholesome and they are similar to me and my friend.
Per/per/siblings ✗
Counterfactual Thinking
[E1] **Ryan** [/E1] and [E2] **Yaz** [/E2] realising they knew each other from school
Per/per/alumni ✓
Rational Thinking

Figure 1: Different models "see" different content in sentences by thinking differently. Rational thinking predicts the correct relation label `per/per/alumni` between two entities `Ryan` and `Yaz` by seeing the relevant and correct content.

SIGIR '23July 23–27, 2023Taipei, Taiwan
2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3592072

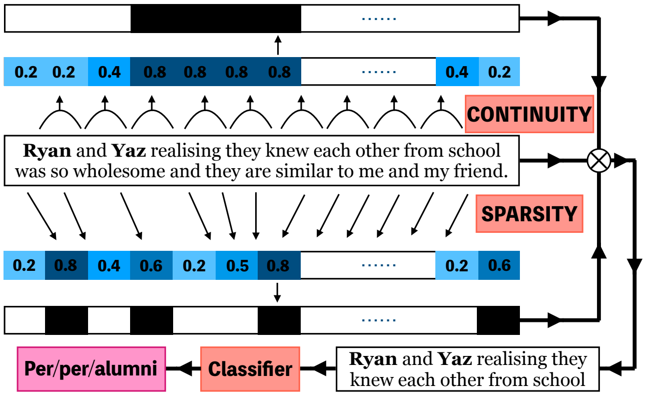
0.2 0.2 0.4 0.8 0.8 0.8 0.8 ...... 0.4 0.2
CONTINUITY
Ryan and Yaz realising they knew each other from school was so wholesome and they are similar to me and my friend.
SPARSITY
0.2 0.8 0.4 0.6 0.2 0.5 0.8 ...... 0.2 0.6
......
Per/per/alumni
Classifier
Ryan and Yaz realising they knew each other from school

Figure 2: Architecture: The rationale extractor obtains the rationales from the input using a binary mask consists of continuity and sparsity factors.

and noise-free rationales in RE task. Although the methods of rational thinking have been verified in the various downstream tasks of information retrieval such as question answering [33], we still face two crucial challenges to leverage the rational thinking methods for the RE task: (1) Gold rationales which are relevant to the relation label in the sentences are not available, therefore, we cannot train a rationale extractor in a supervised learning manner, (2) Extracted rationales are encouraged to be continuous, which can not only improve the interpretability of the rationales, but also express coherent semantics to predict the relation labels between entities.

In addressing the two main challenges, we present two new continuity and sparsity factors in this study to manage the coherence and quantity of chosen rationale tokens. Sparsity imposition aids in striking a balance between eliminating irrelevant material and preserving relevant content. Promoting continuity is advantageous for obtaining continuous rationales, leading to a more coherent semantic representation. Furthermore, we employ an adjustable binary mask for rationale selection and modify the rationale tokens necessary for the relation extraction task using relation labels. As a result, the unavailability of gold rationales can be addressed through end-to-end training. Our primary contributions include: (1) Introducing a novel end-to-end training system, RE^2, which treats rationale extraction as an adjustable binary mask for the relation extraction task and retains relevant, noise-free rationales via continuity and sparsity factors. (2) Experiments on four commonly used datasets demonstrate that RE^2 significantly improves best-reported baselines in both full data and low-resource settings.

2 PROPOSED MODEL

2.1 Continuous Rationale Extractor

In the continuous rationale extractor module of the model, we mask the tokens that are irrelevant to the relation extraction task, and keep the continuous tokens to improve extraction performance.

2.1.1 *Sentence Representation and Importance Matrix.*

We can obtain semantic embedding of each token through its contextualized sentence representation. In practice, we adopt the BERT [8] to encode the token representations as: $\boldsymbol{x_{sent}} \in \mathbb{R}^{D \times L}$, where D is the dimension of the embedding and L is the number of tokens in the sentence. To select the tokens most relevant to the entities in the sentence for relation extraction task, we first calculate the importance matrix: $\boldsymbol{s} = \boldsymbol{x_{sent}}^\top (\boldsymbol{x_{e_1}} + \boldsymbol{x_{e_2}})$ with the token embeddings of the two entities $\boldsymbol{x_{e_1}}$ and $\boldsymbol{x_{e_2}}$ extracted from $\boldsymbol{x_{sent}}$. We denote $\boldsymbol{s} = (s_1, ..., s_L)^\top$ and obtain the importance score s_i which represents the importance of the i^{th} token towards RE task.

2.1.2 *Factor Graph.*

We represent the token selection using a binary vector $\boldsymbol{m} = (m_1, m_2, ..., m_L)^\top$, where $m_i \in \{0, 1\}$. The value 0 or 1 is used to indicate whether the i^{th} token is selected. In this way, we could transform the structured prediction problem of rational sequence generation into the assignment of values to multiple variables. To maintain semantic coherence in token selection, it's important to consider the continuity in token choice. Additionally, to emphasize the importance and relevance of tokens to entities, we need to limit the number of tokens with sparsity. Therefore, we introduce the factor graph $\mathcal{F}$ and decompose these requirements into multiple local factors for optimal token selection. More specifically, we adopt the pairwise factor CONTINUITY and L-ary factor SPARSITY. In the following sections, we will formulate these two factors and provide their score functions.

CONTINUITY (CON): To improve the continuity of tokens selected for RE task, we adopt the CONTINUITY (CON) factor, which could examine whether each pair of consecutive tokens are both selected. We adopt the factor CON $(m_i, m_i{+}1; r_{i,i+1})$ to represent the constraint on the continuous selection of the i^{th} and $(i+1)^{th}$ tokens. As shown in Figure 2, if both tokens are selected, we encourage this continuous selection by adding the edge score $r_{i,i+1} \geq 0$ in the score function. Formally, the score function for CON factor can be denoted as:

$$\text{score}_{\text{CON}}(m_i, m_{i+1}; r_{i,i+1}) = m_i m_{i+1} r_{i,i+1}. \tag{1}$$

As illustrated in Section 2.1.1, we adopt the importance matrix $\boldsymbol{s}$ to measure the tokens that are relevant to the entities, which is a critical metrics to token selection. We add the scores of the selected i^{th} and $(i+1)^{th}$ tokens in the score function and finalize the score function as:

$$\text{score}_{\text{CON}}(m_i, m_{i+1}; r_{i,i+1}) = m_i m_{i+1} r_{i,i+1} + m_i s_i + m_{i+1} s_{i+1}. \tag{2}$$

We can impose continuity constraints on the token selection for the original sentence by leveraging the combination of the pairwise factors. Formally, the factor graph can be formulated as:

$$\mathcal{F} = \{\text{CON}(m_i, m_{i+1}; r_{i,i+1}) : 1 \leq i < L\}. \tag{3}$$

SPARSITY (SPA): To control the sparsity in token selection for RE task, we adopt the L-ary factor SPARSITY (SPA) by imposing a limit K on the maximum number of selected tokens as a restriction. In practice, K can also be the proportion of all tokens in a sentence. The SPA factor is a hard constraint by definition. We can formulate the SPA factor with the following score function:

$$\text{score}_{\text{SPA}}(m_1, m_2, \cdots, m_L, K) = \begin{cases} 0, & \sum m_i \leq K, \\ -\infty, & \sum m_i > K. \end{cases} \tag{4}$$

Overall, to consider continuity and sparsity together, we obtain the factor graph $\mathcal{F}$ by instantiating with the L binary variables and combining the CON and SPA factors:

$$\mathcal{F} = \{\text{SPA}(m_1, ..., m_L; K)\} \cup \{\text{CON}(m_i, m_{i+1}; r_{i,i+1}). \tag{5}$$

To utilize the both continuity and sparsity constraints and find an optimal solution for token selection, the score functions of factor graph $\mathcal{F}$ need to sum the local sub-problems $\{\mathtt{CON}(m_i, m_{i+1}; r_{i,i+1}) : 1 \leq i < L\}$ and $\{\mathtt{SPA}(m_1, ..., m_L; K)\}$ as:

$$\text{score}(m; s) = \begin{cases} \sum_{i=1}^{L} m_i s_i + \sum_{i=1}^{L-1} m_i m_{i+1} r_{i,i+1}, & \sum m_i \leq K, \\ -\infty, & \sum m_i > K. \end{cases} \quad (6)$$

The hard constraint of $\mathcal{F}$ is inherited from the SPA factor, which specify that the total number of selected tokens should not exceed K. The soft constraint is inherited from the CON factors, encouraging to select consecutive tokens that are relevant to the RE task. To find a solution that satisfies these constraints well, we approach the problem of solving the variables as a Maximum A Posteriori (MAP) inference problem that maximizes the score function: score $(\boldsymbol{m}; \boldsymbol{s})$. We can represent this problem as maximization of the score function under the constraint that $|\boldsymbol{m}|_1 \leq K$:

$$\hat{\boldsymbol{m}} = \arg \max_{|\boldsymbol{m}|_1 \leq K, \boldsymbol{m} \in \{0,1\}^L} \underbrace{\left(\boldsymbol{s}^\top \boldsymbol{m} + \sum_{i=1}^{L-1} m_i m_{i+1} r_{i,i+1}\right)}_{\text{score}(\boldsymbol{m};\boldsymbol{s})}. \quad (7)$$

In fact, maximizing the score function is essentially a complex structured problem involving sub-problems with interrelated global agreement constraints, making it difficult to find an accurate maximization algorithm [25]. To solve this problem, we consider the Marginal Inference with Lagrange Multiplier.

2.1.3 Marginal Inference with Lagrange Multiplier. We can solve the maximization problem with the Gibbs distribution and get an approximate solution. We construct a Gibbs distribution so that $p(\boldsymbol{m}; \boldsymbol{s}) \propto exp(\text{score}(\boldsymbol{m}; \boldsymbol{s}))$. In this way, we can sample from $\hat{\boldsymbol{m}} \sim p(\boldsymbol{m}; \boldsymbol{s})$ and obtain an approximate optimal solution. However, obtaining unbiased samples is challenging. To address this issue, we use Perturb-and-MAP [26], an approximate sampling strategy.

Another problem is that the score function in Eq. 6 is a piecewise function, making the Gibbs distribution $p(\boldsymbol{m}; \boldsymbol{s}) \propto exp(\text{score}(\boldsymbol{m}; \boldsymbol{s}))$ discontinuous. As marginal inference in discontinuous Markov Random Fields is hard to solve, we reformulate the hard constraint: SPA in Eq. 6 with Lagrange multiplier, which express hard constraints in the form of continuous functions. Specifically, we use a Lagrange Multiplier $\lambda > 0$, and add $\lambda(K - |\boldsymbol{m}|_1)$ to the objective function in Eq. 7. We finalize the Eq. 7 as:

$$\hat{\boldsymbol{m}} = \arg \max_{|\boldsymbol{m}|_1 \leq K, \boldsymbol{m} \in [0,1]^L} \left(\boldsymbol{s}^\top \boldsymbol{m} + \sum_{i=1}^{L-1} m_i m_{i+1} r_{i,i+1} + \lambda(K - |\boldsymbol{m}|_1)\right), \quad (8)$$

where the Gibbs distribution should be reformulated as $p(\boldsymbol{m}; \boldsymbol{s}) \propto exp(\text{score}(\boldsymbol{m}; \boldsymbol{s}) + \lambda(K - |\boldsymbol{m}|_1))$. Therefore, the reformulated Gibbs distribution becomes continuous, enabling us to calculate the optimal $\boldsymbol{m}$ that maximizes the score function, and obtain the rationales which are relevant to the relation extraction task.

2.2 Relation Classifier

Finally, the classifier makes relation predictions conditioned on the selected rationales and the entities $\boldsymbol{e}$: $\hat{\boldsymbol{y}} = \mathtt{pred}(\boldsymbol{m} \odot \boldsymbol{x} \parallel \boldsymbol{e})$ to obtain the relation label distributions. $\odot$ and $\parallel$ denote the element-wise product and concatenation, respectively. The relation classification loss could be calculated as: $\mathcal{L} = -\sum_{i=1}^{N} \boldsymbol{y}_i \log \hat{\boldsymbol{y}}_i$, where N is the number of training sentences in an epoch, and $\boldsymbol{y}_i$ is the ground-truth tag vector of the sentence x_i. The relation classification loss could jointly train the Continuous Rationale Extractor module and Relation Classifier module in an end-to-end manner.

3 EXPERIMENTS AND ANALYSES

3.1 Experimental Setup and Baselines

Setup: We evaluate the model on four widely-used RE datasets: SemEval [13], which contains 6,507/1,493/2,717 samples in train/dev/test sets and 19 relation types. TACRED [35] and TACRED-Revisit [1], which contain 68,124/22,631/15,509 samples and 42 relation types. Re-TACRED [31], which contains 58,465/19,584/13,418 samples and 42 relation types. Following prior effort [32], we adopt Micro F1 as the evaluation metric. Under the low-resource setting, we randomly sample 10%, 25%, and 50% of the training set as the small-scale training sets for evaluation, and evaluate our model on the test set. We use the BERT-Base default tokenizer with a max-length of 128 to preprocess data. We set K as 60%.

Baselines: We first introduce SOTA models as base model on the RE task, and then adopt various baselines. We adopt SURE [23] as the base model. We compare RE^2 with the following baselines: *Entity thinking* baselines: (1) MTB [30], (2) Entity Mask [35], (3) Typed Entity Marker [36]. *Counterfactual thinking* baselines adopt causal inference to remove bias in RE tasks: (4) CFIE [24], (5) CORSAIR [28] (6) CORE [32]. *Rationale thinking* baselines could predict sparse binary masks over input tokens for RE tasks: (7) HardKuma [2], (8) IB objective [27], (9) UNIREX [3]. Note that the entity thinking method is also used in the SURE, all baselines of entity thinking are used to replace the methods in SURE.

3.2 Results and Analysis

Overall Performance. Table 1 displays mean and standard deviation results for 5 training and testing runs on four datasets. Using SURE [23] for entity information verbalization yields an average 0.6% F1 improvement across datasets over other entity thinking methods, thus we adopt SURE as our base model. Both counterfactual and rationale thinkings contribute a 0.3% F1 boost across datasets. Our rational thinking method tackles (1) end-to-end training for rationale extraction and relation classification, and (2) continuous rationales extraction, leading to a significant 0.9% F1 improvement across all RE datasets, including low-resource ones, outperforming the previous SOTA: UNIREX by an extra 0.4%. Notably, RE^2 shows better improvements in low-resource settings (10% of training set), with 1.1% vs. 0.9%, indicating its robustness with limited training data. The low-noise, relevant rationales from RE^2 notably enhance the base model's F1 performance.

Ablation Study. We perform an ablation study to demonstrate the effectiveness of our model's various modules on the test set. RE^2 *without Continuity* and RE^2 *without Sparsity* eliminate the continuity and sparsity elements in the factor graphs in rationale extraction, respectively. RE^2 *without Adding Entities* removes the entities added in the relation classifier module, using only the rationales for relation classification. Table 1 generally concludes that all modules

Table 1: Average micro F1 results in four RE datasets. "re." means that we will replace the entity information verbalization in SURE with the corresponding entity thinking baselines. We mark the (standard deviation) of the results.

Methods/Datasets	SemEval				TACRED				TACRED-Revisit				Re-TACRED			
	10%	25%	50%	100%	10%	25%	50%	100%	10%	25%	50%	100%	10%	25%	50%	100%
SURE	77.2	81.5	83.9	86.3	67.9	70.4	71.9	73.3	72.3	75.1	77.4	79.2	78.5	82.6	84.7	88.2
re. MTB [30]	76.3	80.8	82.9	85.4	67.2	69.8	71.1	72.3	71.4	74.3	76.6	78.4	77.6	81.8	83.8	87.2
re. Entity Mask [35]	76.7	80.9	83.3	86.0	67.3	70.0	71.4	72.5	71.8	74.7	76.9	78.5	77.8	82.2	84.3	87.6
re. Typed Marker [36]	77.0	81.3	83.7	86.3	67.8	70.4	71.8	73.2	72.1	75.0	77.2	79.0	78.2	82.4	84.5	88.1
+CFIE [24]	77.3	81.7	84.1	86.6	68.2	70.5	72.2	73.5	72.5	75.2	77.6	79.4	78.6	82.8	85	88.4
+CORSAIR [28]	77.4	81.8	84.1	86.7	68.3	70.7	72.4	73.7	72.7	75.4	77.7	79.6	78.9	83.1	85.1	88.6
+CORE [32]	77.5	82.0	84.2	86.7	68.3	70.8	72.3	73.7	72.8	75.6	77.8	79.6	78.9	83.2	85.3	88.7
+HardKuma [2]	77.4	81.4	84.0	86.5	68.0	70.6	72.2	73.5	72.5	75.3	77.6	79.4	78.7	82.7	85.0	88.3
+IB Objective [27]	77.5	81.7	84.3	86.6	68.3	70.7	72.3	73.6	72.6	75.3	77.8	79.5	79.0	83.0	85.1	88.6
+UNIREX [3]	77.8	81.9	84.4	86.7	68.5	70.9	72.5	73.8	72.8	75.7	78.0	79.7	79.2	83.1	85.4	88.8
+**RE**^{2}	**78.2(0.2)**	**82.3(0.1)**	**84.8(0.1)**	**87.2(0.2)**	**68.9(0.3)**	**71.3(0.2)**	**73.0(0.3)**	**74.2(0.1)**	**73.3(0.2)**	**76.0(0.2)**	**78.4(0.1)**	**80.1(0.1)**	**79.6(0.3)**	**83.6(0.2)**	**85.7(0.1)**	**89.1(0.2)**
w/o Continuity	77.8(0.3)	81.9(0.4)	84.2(0.2)	86.8(0.4)	68.5(0.3)	70.7(0.2)	72.5(0.3)	73.8(0.2)	72.8(0.4)	75.5(0.3)	77.9(0.3)	79.7(0.2)	79.1(0.4)	83.2(0.3)	85.3(0.2)	88.8(0.3)
w/o Sparsity	77.6(0.2)	81.8(0.2)	84.3(0.1)	86.6(0.3)	68.4(0.1)	70.5(0.2)	72.4(0.3)	73.6(0.1)	72.7(0.2)	75.5(0.3)	77.8(0.3)	79.8(0.2)	79.0(0.2)	83.1(0.3)	85.1(0.1)	88.6(0.1)
w/o Adding Entities	78.0(0.2)	82.2(0.1)	84.6(0.2)	87.1(0.2)	68.8(0.3)	71.1(0.2)	72.8(0.2)	74.0(0.1)	73.1(0.2)	75.9(0.2)	78.2(0.1)	80.0(0.1)	79.3(0.3)	83.4(0.2)	85.6(0.2)	89.0(0.1)

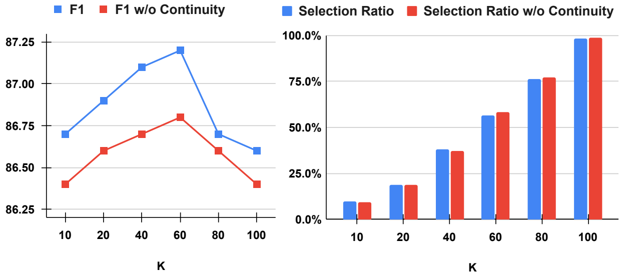

Figure 3: Effect of Two Factors (Continuity and Sparsity). K **is the hyper-parameter to control the sparsity of the token selection. Continuity is imposed to improve contiguity.**

positively impact performance. Specifically, the absence of continuity leads to discontinuous rationales, affecting the coherence of semantic representations and causing a 0.4% F1 performance drop. Removing sparsity selects noisier rationales, resulting in a 0.5% F1 performance reduction. Interestingly, removing added entities has minimal effect on F1 performance (0.1%). We find that 89% of rationales contain two entities and 97% contain at least one entity, indicating that adding entities provides little additional information.

Effect of Two Factors: As shown in Figure 3, we display F1 scores and token selection rates in relation to varying K values on SemEval. As K rises, more tokens within sentences are chosen as rationales. Nonetheless, the F1 score for RE^2 doesn't increase consistently with higher K, due to the incorporation of unrelated rationales. Optimal performance occurs at $K = 60$, meaning 60% of tokens are selected as rationales on average. Eliminating the Sparsity factor entirely causes the model's F1 score to decline from 87.2 to 86.6. Additionally, the Continuity constraint benefits the model, as RE^2 with Continuity constraints consistently produces improved outcomes.

Coherence Analysis of Rationales. RE^2 utilizes the continuity factor to control the generation of rationales that are more semantically coherent, which can express more fluent semantics. We analyze the coherence of the rationales through perplexity based on GPT-3 [29]. From Table 2, RE^2 could obtain the lowest average perplexity, approaching that of the original sentences.

Human Evaluation. We conduct human evaluations of rationales with a 15-member annotation team, involving 5 members in data validation. Annotators predict relation labels using original sentences and extracted rationales, then rate information sufficiency (on a 1-5 scale) for both. Higher scores signify greater sufficiency.

Table 2: Perplexity of the extracted rationales. Original means the original sentences. Lower perplexity is better.

Methods / Datasets	SemEval	TACRED	TACRED-Revisit	Re-TACRED
HardKuma [2]	11.37	13.35	12.21	11.93
IB Objective [27]	10.37	12.23	11.56	10.74
UNIREX [3]	13.42	12.64	14.22	13.87
RE^2	**5.13**	**5.68**	**6.02**	**5.75**
Original Sentences	3.75	4.02	3.83	3.68

Table 3: Human evaluation (Micro F1 / Information Sufficiency) of the original sentences and extracted rationales.

Datasets	SemEval	TACRED	TACRED-Revisit	Re-TACRED
Extracted Rationales	95.5 / 4.4	89.3 / 4.0	93.5 / 4.2	96.8 / 4.5
Original Sentences	94.3 / 4.5	87.6 / 4.3	92.3 / 4.2	96.0 / 4.6

To ensure consistency, we perform inter-annotator agreement and manual validation. Table 3 shows that annotators can provide more accurate relation labels even with lower information sufficiency in rationales than original sentences, suggesting that removing irrelevant details from sentences can decrease noise and enhance relational prediction accuracy.

4 CONCLUSION AND FUTURE WORK

In this work, we introduce a unique rationale extraction scheme, RE^2, incorporating continuity and sparsity for relevance and coherence in the RE task. We employ marginal inference and a Lagrange multiplier to optimize the two-factor score function. Consequently, we can concurrently train the rationale extraction and relation classification tasks end-to-end, without need for rationale gold annotations. RE^2's effectiveness is demonstrated in experiments on four datasets. In the future, we can extend the research on RE to the construction of knowledge graphs [6, 7, 9, 34], the matching of knowledge graphs [10–12, 22], and the acceleration of IR [4, 5].

5 ACKNOWLEDGMENTS

Chenwei Zhang is the corresponding author. Xuming Hu and Zhaochen Hong have made equal contributions. Source code: https://github.com/THU-BPM/RE2. The work described in this paper was partially supported by the National Key Research and Development Program of China (No. 2018AAA0100204) and by the Research Grants Council of the Hong Kong Special Administrative Region, China (RGC GRF 2151185; CUHK 14222922), NSF under grant III-1909323.

REFERENCES

[1] Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task. In *Proc. of ACL*. 1558–1569.

[2] Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable Neural Predictions with Differentiable Binary Variables. In *Proc. of ACL*. 2963–2977.

[3] Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan, Shaoliang Nie, Xiaochang Peng, Xiang Ren, and Hamed Firooz. 2022. Unirex: A unified learning framework for language model rationale extraction. In *Proc. of ICML*. PMLR, 2867–2889.

[4] Yankai Chen, Yixiang Fang, Yifei Zhang, and Irwin King. 2023. Bipartite Graph Convolutional Hashing for Effective and Efficient Top-N Search in Hamming Space. In *Proc. of WWW*. ACM.

[5] Yankai Chen, Huifeng Guo, Yingxue Zhang, Chen Ma, Ruiming Tang, Jingjie Li, and Irwin King. 2022. Learning binarized graph representations with multi-faceted quantization reinforcement for top-k recommendation. In *Proc. of SIGKDD*. 168–178.

[6] Yankai Chen, Menglin Yang, Yingxue Zhang, Mengchen Zhao, Ziqiao Meng, Jianye Hao, and Irwin King. 2022. Modeling scale-free graphs with hyperbolic geometry for knowledge-aware recommendation. In *Proc. of WSDM*. 94–102.

[7] Yankai Chen, Yaming Yang, Yujing Wang, Jing Bai, Xiangchen Song, and Irwin King. 2022. Attentive knowledge-aware graph convolutional networks with collaborative guidance for personalized recommendation. In *Proc. of ICDE*. IEEE, 299–311.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL-HLT*. 4171–4186.

[9] Yixiang Fang, Reynold Cheng, Yankai Chen, Siqiang Luo, and Jiafeng Hu. 2017. Effective and efficient attributed community search. *The VLDB Journal* 26 (2017), 803–828.

[10] Yunjun Gao, Xiaoze Liu, Junyang Wu, Tianyi Li, Pengfei Wang, and Lu Chen. 2022. ClusterEA: Scalable Entity Alignment with Stochastic Training and Normalized Mini-batch Similarities. In *KDD*. 421–431.

[11] Congcong Ge, Xiaoze Liu, Lu Chen, Baihua Zheng, and Yunjun Gao. 2021. Make It Easy: An Effective End-to-End Entity Alignment Framework. In *SIGIR*. 777–786.

[12] Congcong Ge, Xiaoze Liu, Lu Chen, Baihua Zheng, and Yunjun Gao. 2022. LargeEA: Aligning Entities for Large-scale Knowledge Graphs. *PVLDB* 15, 2 (2022), 237–245.

[13] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O 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 *Proc. of SemEval*. 33–38.

[14] Xuming Hu, Zhijiang Guo, Yu Fu, Lijie Wen, and Philip S. Yu. 2022. Scene Graph Modification as Incremental Structure Expanding. In *Proc. of COLING*. 5707–5720.

[15] Xuming Hu, Fukun Ma, Chenyao Liu, Chenwei Zhang, Lijie Wen, and Philip S Yu. 2021. Semi-supervised Relation Extraction via Incremental Meta Self-Training. In *Proc. of EMNLP: Findings*.

[16] Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and Philip Yu. 2020. SelfORE: Self-supervised Relational Feature Learning for Open Relation Extraction. In *Proc. of EMNLP*. Online, 3673–3682.

[17] Xuming Hu, Chenwei Zhang, Yawen Yang, Xiaohe Li, Li Lin, Lijie Wen, and Philip S. Yu. 2021. Gradient Imitation Reinforcement Learning for Low Resource Relation Extraction. In *Proc. of EMNLP*. 2737–2746.

[18] Aiwei Liu, Xuming Hu, Li Lin, and Lijie Wen. 2022. Semantic Enhanced Text-to-SQL Parsing via Iteratively Learning Schema Linking Graph. In *Proc. of KDD*. 1021–1030.

[19] Aiwei Liu, Xuming Hu, Lijie Wen, and Philip S Yu. 2023. A comprehensive evaluation of ChatGPT's zero-shot Text-to-SQL capability. *arXiv preprint arXiv:2303.13547* (2023).

[20] Aiwei Liu, Honghai Yu, Xuming Hu, Shu'ang Li, Li Lin, Fukun Ma, Yawen Yang, and Lijie Wen. 2022. Character-level White-Box Adversarial Attacks against Transformers via Attachable Subwords Substitution. In *Proc. of EMNLP*.

[21] Shuliang Liu, Xuming Hu, Chenwei Zhang, Shu'ang Li, Lijie Wen, and Philip S. Yu. 2022. HiURE: Hierarchical Exemplar Contrastive Learning for Unsupervised Relation Extraction. In *Proc. of NAACL-HLT*. 5970–5980.

[22] Xiaoze Liu, Junyang Wu, Tianyi Li, Lu Chen, and Yunjun Gao. 2023. Unsupervised Entity Alignment for Temporal Knowledge Graphs. In *WWW*.

[23] Keming Lu, I Hsu, Wenxuan Zhou, Mingyu Derek Ma, Muhao Chen, et al. 2022. Summarization as Indirect Supervision for Relation Extraction. In *Proc. of EMNLP*.

[24] Guoshun Nan, Jiaqi Zeng, Rui Qiao, Zhijiang Guo, and Wei Lu. 2021. Uncovering Main Causalities for Long-tailed Information Extraction. In *Proc. of EMNLP*. 9683–9695.

[25] Vlad Niculae and André F. T. Martins. 2020. LP-SparseMAP: Differentiable Relaxed Optimization for Sparse Structured Prediction. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 7348–7359.

[26] George Papandreou and Alan L Yuille. 2011. Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *Proc. of ICCV*. IEEE, 193–200.

[27] Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An Information Bottleneck Approach for Controlling Conciseness in Rationale Extraction. In *Proc. of EMNLP*. 1938–1952.

[28] Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. In *Proc. of ACL-IJCNLP*. 5434–5445.

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[30] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *Proc. of ACL*. 2895–2905.

[31] George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset. In *Proc. of AAAI*, Vol. 35. 13843–13850.

[32] Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022. Should We Rely on Entity Mentions for Relation Extraction? Debiasing Relation Extraction with Counterfactual Analysis. In *Proc. of NAACL-HLT*. 3071–3081.

[33] Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 4 (2018), e1253.

[34] Xinni Zhang, Yankai Chen, Cuiyun Gao, Qing Liao, Shenglin Zhao, and Irwin King. 2022. Knowledge-aware Neural Networks with Personalized Feature Referencing for Cold-start Recommendation. *arXiv preprint arXiv:2209.13973* (2022).

[35] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proc. of EMNLP*. 35–45.

[36] Wenxuan Zhou and Muhao Chen. 2022. An Improved Baseline for Sentence-level Relation Extraction. *Proc. of AACL-IJCNLP* (2022), 161.