FSUIE: A Novel Fuzzy Span Mechanism for Universal Information Extraction

Tianshuo Peng[1,†], **Zuchao Li**[1,2,†,*], **Lefei Zhang**[1,2], **Bo Du**[1,2] **and Hai Zhao**[3]
[1]National Engineering Research Center for Multimedia Software,
School of Computer Science, Wuhan University, Wuhan, 430072, P. R. China
[2]Hubei Luojia Laboratory, Wuhan 430072, P. R. China
[3]Department of Computer Science and Engineering, Shanghai Jiao Tong University
{pengts,zcli-charlie,zhanglefei,dubo}@whu.edu.cn,
zhaohai@cs.sjtu.edu.cn

Abstract

Universal Information Extraction (UIE) has been introduced as a unified framework for various Information Extraction (IE) tasks and has achieved widespread success. Despite this, UIE models have limitations. For example, they rely heavily on span boundaries in the data during training, which does not reflect the reality of span annotation challenges. Slight adjustments to positions can also meet requirements. Additionally, UIE models lack attention to the limited span length feature in IE. To address these deficiencies, we propose the Fuzzy Span Universal Information Extraction (FSUIE) framework. Specifically, our contribution consists of two concepts: *fuzzy span loss* and *fuzzy span attention*. Our experimental results on a series of main IE tasks show significant improvement compared to the baseline, especially in terms of fast convergence and strong performance with small amounts of data and training epochs. These results demonstrate the effectiveness and generalization of FSUIE in different tasks, settings, and scenarios.

1 Introduction

Information Extraction (IE) is focused on extracting predefined types of information from unstructured text sources, such as Named Entity Recognition (NER), Relationship Extraction (RE), and Sentiment Extraction (SE). To uniformly model the various IE tasks under a unified framework, a generative Universal Information Extraction (UIE) was proposed in (Lu et al., 2022) and has achieved widespread success on various IE datasets and benchmarks. Due to the necessity of a powerful generative pre-training model for the generative UIE, the time overhead is extensive and the efficiency is not satisfactory. For this reason, this paper examines span-based UIE to unify various IE tasks, conceptualizing IE tasks as predictions of spans.

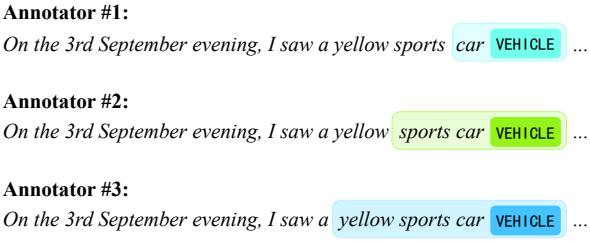

Figure 1: An example of annotation ambiguity for "vehicle" entity in sentence "*On the 3rd September evening, I saw a yellow sports car drive past my house*".

However, UIE models still have some limitations. First, as it is the process of training machine learning models to extract specific information from unstructured text sources, IE relies heavily on human annotation which involves labeling the data by identifying the specific information to be extracted and marking the corresponding span boundaries in the text manually. However, due to the complexity of natural language, determining the correct span boundaries can be challenging, leading to the phenomenon of annotation ambiguity. As shown in Figure 1, different annotated spans can be considered reasonable. In the span learning of UIE models, the method of teacher forcing is commonly used for loss calculation, making the model dependent on the precise span boundaries given in the training data. This can cause performance bottlenecks due to annotation ambiguity.

When the model structure in UIE places too much emphasis on the exact boundaries of IE tasks, it leads to insufficient utilization of supervision information. In order to predict span boundaries, positions closer to the ground-truth should be more accurate than those relatively farther away, as shown

*Corresponding author. † Equal contribution. This work was supported by the Fundamental Research Funds for the Central Universities (No. 2042023kf0133), the Special Fund of Hubei Luojia Laboratory under Grant 220100014 and the National Science Fund for Distinguished Young Scholars under Grant 62225113. Hai Zhao was funded by the Key Projects of National Natural Science Foundation of China (U1836222 and 61733011).

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 16318–16333
July 9-14, 2023 ©2023 Association for Computational Linguistics

in Figure 1. For example, words close to the target "car" are more likely to be correct than the word "evening" which is farther away from the target. Under the premise of positioning to the span where "car" is located, both the "yellow car" and the "yellow sports car" can be regarded as vehicle entities. This means that the span model learned should be fuzzy rather than precise.

In addition, the use of pre-trained Transformer (Vaswani et al., 2017) in UIE to extract the start and end position representations also poses a problem. The Transformer model is designed to focus on the global representation of the input text, while UIE requires focusing on specific parts of the text to determine the span boundaries. This mismatch between the Transformer's focus on global representation and UIE's focus on specific parts of the text can negatively impact the performance of the model.

When there is a mismatch between the Transformer architecture and the span representation learning, the model may not make good use of prior knowledge in IE. Specifically, given the start boundary (end boundary) of the label span, the corresponding end boundary (start boundary) is more likely to be found within a certain range before and after, rather than throughout the entire sequence. This is a prior hypotheses that span has limited length, which is ignored in the vanilla UIE model. To address this, a fuzzy span attention mechanism, rather than fixed attention, should be applied.

In this paper, we propose the Fuzzy Span Universal Information Extraction (FSUIE) framework that addresses the limitations of UIE models by applying the fuzzy span feature, reducing over-reliance on label span boundaries and adaptively adjusting attention span length. Specifically, to solve the issue of fuzzy boundaries, we design the *fuzzy span loss* that quantitatively represents the correctness information distributed on fuzzy span. At the same time, we introduce *fuzzy span attention* that sets the scope of attention to a fuzzy range and adaptively adjusts the length of span according to the encoding. We conduct experiments on various main IE tasks (NER, RE, and ASTE). The results show that our FSUIE has a significant improvement compared to the strong UIE baseline in different settings. Additionally, it achieves new state-of-the-art performance on some NER, RE, and ASTE benchmarks with only bert-base architecture, outperforming models with stronger pre-trained language models and complex neural designs. Furthermore, our model shows extremely fast convergence, and good generalization on low-resource settings. These experiments demonstrate the effectiveness and generalization of FSUIE in different tasks, settings, and scenarios.

2 FSUIE

In FSUIE, incorporating fuzzy span into base UIE model involves two aspects. Firstly, for the spans carrying specific semantic types in the training data, the boundary targets should be learned as fuzzy boundaries to reduce over-reliance on span boundaries. To achieve this, we propose a novel *fuzzy span loss*. Secondly, during the span representation learning, the attention applied in span should be dynamic and of limited length, rather than covering the entire sequence. To achieve this, we propose a novel *fuzzy span attention*.

2.1 Fuzzy Span Loss (FSL)

The introduction of FSL is a supplement to traditional teacher forcing loss (usually implemented as Cross Entrophy), to guide the model in learning fuzzy boundaries. The challenge for FSL is how to quantify the distribution of correctness information within the fuzzy boundary. Specifically, for a given label span S, conventional target distributions (one-hot) indicate the correct starting and ending boundaries. This form actually follows the Dirac delta distribution that only focuses on the ground-truth positions, but cannot model the ambiguity phenomenon in boundaries.

To address the challenge discussed above, we propose a fuzzy span distribution generator (FSDG). In our method, we use a probability distribution of span boundaries to represent the ground-truth, which is more comprehensive in describing the uncertainty of boundary localization. It consists of two main steps: 1) determining the probability density distribution function f; 2) mapping from the continuous distribution to a discrete probability distribution based on f.

Specifically, let $q \in S$ be a boundary of the label span, then the total probability value of its corresponding fuzzy boundary $\hat{q}$ can be represented as follows:

$$\hat{q} = \int_{R_{min}}^{R_{max}} xQ(x)dx, \quad q \in S \tag{1}$$

where x represents the coordinate of boundaries within the fuzzy range $[R_{min}, R_{max}]$, R_{min} and

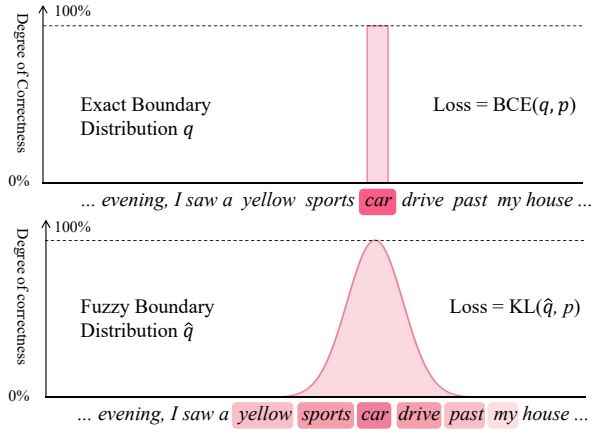

Figure 2: Illustration of exact boundary and fuzzy boundary.

R_{max} are the start and end positions of the fuzzy range, q^{gt} represents the ground-truth position for boundary q, and $Q(x)$ represents the corresponding coordinate probability.

The traditional Dirac delta distribution can be viewed as a special case of Eq. (1), where $Q(x) = 1$ when $x = q^{gt}$, and $Q(x) = 0$ otherwise. Through a mapping function F, we can quantifying continuous fuzzy boundaries into a unified discrete variable $\hat{\mathbf{q}} = [F(q_1), F(q_2), \cdots, F(q_n)]$ with n subintervals. $[q_1, q_2, \cdots, q_n]$ represent continuous coordinates in fuzzy range where $q_1 = R_{min}$ and $q_n = R_{max}$, the probability distribution of each given boundary of the label span can be represented within the range via the softmax function.

Since the Dirac delta distribution only assigns non-zero probability to a single point, it is not suitable for modeling uncertainty or ambiguity in real-world data. Thus in FSUIE, we choose the Gaussian distribution $N(\mu, \sigma^2)$ as the probability density function f. Compared with other probability distributions, the Gaussian distribution assigns non-zero probability to an entire range of values has the following advantages: (1) it is continuous, symmetrical, and can well represent the distribution of correctness information within the fuzzy boundary including the gold position; (2) it is a stable distribution with fewer peaks and offsets, and can ensure that the correctness information is more concentrated on the gold position while distributed on the fuzzy boundary; (3) the integral of the Gaussian distribution is 1, which can ensure that the accuracy distribution after softmax is more gentle.

To get the discrete variable $\hat{\mathbf{q}}$, Four parameters are involved here: variance σ, mean μ, sampling step s, and sampling threshold θ. These parameters are used to control the range, peak position, and density of the fuzzy boundary, respectively. Specifically, the parameter μ is set to q^{gt} and the Gaussian distribution is determined using a pre-determined σ. Assuming $q_g \in [q_1, q_2, \cdots, q_n] = q^{gt}$, F can represented as:

$$F(q_i) = \begin{cases} \varepsilon, & \varepsilon \geq \theta \\ 0, & \varepsilon < \theta \end{cases}, \quad \varepsilon = f(\mu + (i - g)s). \tag{2}$$

Given that values in the marginal regions of Gaussian distribution are quite small, the sampling threshold θ here acts as a filter to eliminate information from unimportant locations. The specific choice of parameters is discussed in the following experimental section. We use $\hat{\mathbf{q}}$ as the distribution of correctness information on the fuzzy boundaries. The beginning and end fuzzy boundaries together make up the fuzzy span. Then, we calculate the KL divergence between the model's predicted logits and the gold fuzzy span distribution as the *fuzzy span loss*. The exact boundary and fuzzy boundary distribution is shown in Figure 2. This *fuzzy span loss* is then incorporated into the original teacher-forcing loss function with a coefficient:

$$\mathcal{L}_{FS} = D_{KL}(\hat{\mathbf{q}} \| p) = \sum_{i=1}^{N} \hat{\mathbf{q}}(x_i) \left(\log \frac{\hat{\mathbf{q}}(x_i)}{p(x_i)} \right),$$
$$\mathcal{L} = \mathcal{L}_{ori} + \lambda \mathcal{L}_{FS}$$

where p represents the predicted distribution of the model and $\hat{\mathbf{q}}$ represents the generated fuzzy span distribution from FSDG according to the annotation in training data. $\mathcal{L}_{ori}$ is the original Binary Cross Entropy (BCE) loss of the model in UIE, and λ is the coefficient of the *fuzzy span loss*.

2.2 Fuzzy Span Attention (FSA)

We construct a FSA based on a multi-head self-attention mechanism with relative positional encoding (RPE), since RPE is more suitable for span representation learning with fuzzy bounds. In conventional multi-head attention with RPE, for a token at position t in the sequence, each head computes the similarity matrix of this token and the tokens in the sequence. The similarity between token t and token r can be represented as:

$$s_{tr} = y_t^\top W_q^\top \left(W_k y_r + p_{t-r} \right) \tag{3}$$

where W_k and W_q are the weight matrices for "key" and "query" representations, y_t and y_r are the representations of token t and r, and p_{t-r} is the relative position embedding, the corresponding attention weight can be obtained through a softmax function

$$a_{tr} = \frac{\exp(s_{tr})}{\sum_{q=0}^{t-1} \exp(s_{tq})}. \tag{4}$$

Conventional self-attention focus on global representations, mismatching the requirement of fuzzy spans. To address this issue, we present a novel attention mechanism, called *Fuzzy Span Attention* (FSA), to control attention scores of each token, aiming to learn a span-aware representation. The fuzzy span mechanism of FSA consists of two aspects: (1) the length of the range applying full attention is dynamically adjusted; and (2) the attention weights on the boundary of the full attention span are attenuating rather than truncated. Specifically, inspired by (Sukhbaatar et al., 2019), we design a mask function g_m to control the attention score calculation. Assuming the maximum length of the possible attention span is L_{span}, the new attention scores can be represented as:

$$a_{tr} = \frac{g_m(t-r)\exp(s_{tr})}{\sum_{q=t-L_{span}}^{t-1} g_m(t-r)\exp(s_{tq})}. \tag{5}$$

The following process is divided into two stages: (1) determining the attention changing function g_a on the fuzzy span, and (2) constructing the mask function g_m based on g_a for span-aware representation learning. According to the characteristics of fuzzy span, we set g_a as a monotonically decreasing linear function. To adjust the attention span length, we define a learnable parameter $\delta \in [0, 1]$. The $g_a(x)$ and corresponding $g_m(x)$ can be represented as follows:

$$\begin{aligned} g_a(z) &= \frac{-z + l + d}{d}, \\ l &= \delta L_{span}. \end{aligned} \tag{6}$$

$$g_m(z) = \begin{cases} 1, & g_a(z) > 1 \\ 0, & g_a(z) < 0 \\ g_a(z), & otherwise \end{cases}. \tag{7}$$

where l controls the length of the full attention range and d is a hyper-parameter that governs the length of the attenuated attention range.

In Figure 3, an illustration of the g_m function is depicted. The dashed lines represent alternative

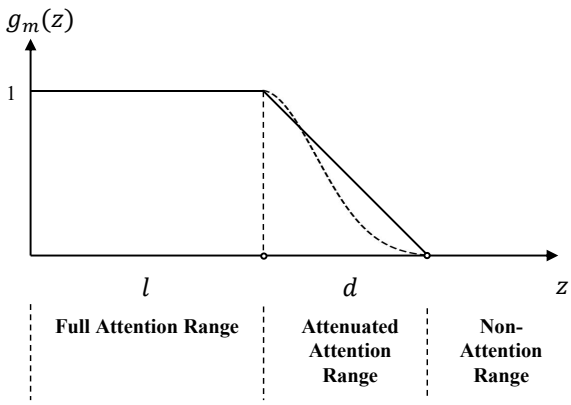

Figure 3: Illustration of attention mask function g_m.

choices of g_a functions, such as

$$g_a'(z) = \begin{cases} 1, & z \le l \\ 0, & z > l \end{cases},$$

$$g_a''(z) = \begin{cases} 1, & z \le l \\ \frac{1}{\sqrt{2\pi}\cdot\frac{d}{3}} \exp\left(-\frac{(z-l)^2}{2(\frac{d}{3})^2}\right), & z > l \end{cases}.$$

Through experimentation, we found that the linear attenuated function performs best (refer to comparison in Appendix A). Iterative optimization of δ allows the model to learn the optimal attention span lengths for a specific task. It is important to note that different heads learn the attention span length independently and thus obtain different optimal fuzzy spans. In our implementation, instead of using multiple layers of fuzzy span attention layers, we construct the span-aware representation with a single fuzzy span attention layer on top of Transformer encoder, and it does not participate in the encoding process. Therefore, although the maximum range of fuzzy span attention is limited by L_{span}, it only affects span decisions and does not have any impact on the representation of tokens in the sequence.

3 Experiments

3.1 Setup

Tasks We conducted experiments on 4 datasets for 3 common information extraction tasks: NER, RE, and ASTE. The datasets we used include ACE2004, ACE2005, ADE (Gurulingappa et al., 2012) for NER and RE tasks, and ASTE-Data-V2 (Xu et al., 2020) for ASTE task. We evaluate our model using different metrics for the three IE tasks. For NER, we use the Entity F_1 score, in

Models	PLM	ACE04			ACE05			ADE		
		P	R	F_1	P	R	F_1	P	R	F_1
BENSC (Tan et al., 2020)	BERT-base	85.80	84.80	85.30	83.80	83.90	83.90	-	-	-
(Ma et al., 2020)	BERT-large	-	-	-	-	-	88.60	-	-	-
SpERT (Eberts and Ulges, 2019)	BERT-base	-	-	-	-	-	-	88.69	89.20	88.95
SpERT.PL (Santosh et al., 2021)	Bio-BERT	-	-	-	-	-	-	90.05	91.69	90.86
BoningKnife (Jiang et al., 2021)	BERT-base	85.98	86.86	86.41	84.77	86.16	85.46	-	-	-
JCBIE (He et al., 2022)	Bio-BERT	-	-	-	-	-	-	-	-	87.80
GLOBAL POINTER (Su et al., 2022)	BERT-base	-	-	-	-	-	-	-	-	90.10
BS (Zhu and Li, 2022)	RoBERTa-base	88.43	87.53	87.98	86.25	88.07	87.15	-	-	-
Triaffine (Yuan et al., 2022)	BERT-large	87.13	87.68	87.40	86.70	86.94	86.82	-	-	-
Triaffine (Yuan et al., 2022)	ALBERT-xxlarge	88.88	88.24	**88.56**	87.39	90.31	**88.83**	-	-	-
(Yan et al., 2021)	BERT-large	87.27	86.41	86.84	83.16	86.38	84.74	-	-	-
Generative UIE (SEL only) (Lu et al., 2022)	T5-v1.1-large	-	-	86.52	-	-	85.52	-	-	-
Generative UIE (Lu et al., 2022)	T5-v1.1-large	-	-	86.89	-	-	85.78	-	-	-
UIE-base	BERT-base	88.25	80.31	84.09	86.19	83.12	84.63	87.85	91.56	89.67
FSUIE-base	BERT-base	85.67	84.82	85.24	87.05	85.40	86.22	91.17	92.17	92.49
FSUIE-large	BERT-large	86.15	86.17	86.16	88.06	85.79	86.91	93.82	92.21	**93.08**

Table 1: NER experimental results on ACE04, ACE05, and ADE datasets.

which an entity prediction is correct if its span and type match a reference entity. For RE, we use the Relation Strict F_1 score, where a relation is considered correct only if its relation type and related entity spans are all correct. For ASTE, we use the Sentiment Triplet F_1 score, where a triplet is considered correct if the aspect, opinion, and sentiment polarity are all correctly identified.

Training Details We trained two variations of FSUIE, FSUIE-base and FSUIE-large, which are based on the BERT-base and BERT-large model architecture and pre-training parameters respectively. In addition, we also trained a UIE-base based on BERT-base as a baseline without using FSL and FSA layers. In FSUIE, we added the FSA layer and the span boundary prediction layer to both models. Specifically, FSUIE-base has 12 layers of 12-head Transformer layers, with a hidden size of 768, while FSUIE-large has 24 layers of 16-head Transformer layers, with a hidden size of 1024. During training, we set the parameters of the Gaussian distribution in FSL as $\sigma = 0.5$, the distribution value truncation threshold θ to 0.3, sampling step s to 0.3, and the loss coefficient λ to 0.01. And the parameter μ is set to the coordinate of annotation boundary. The hyper-parameters L_{span} and d involved are determined based on the statistics of the target length on the UIE training data. During training, we set L_{span} to 30 and d to 32, and experimentation results have shown that the model's performance is not significantly sensitive to the choice of these hyper-parameters (refer to comparison in Appendix C).

We trained both models for 50 epochs with a learning rate of 1e-5 on the datasets of each task, and selected the final model based on the performance on the development set. The code is available at https://github.com/pengts/FSUIE.

3.2 Results on NER tasks

We report the results of NER task in Table 1. By comparing the results of our baseline UIE-base with other methods, it can be seen that UIE-base has achieved comparable results compared to other methods that use the same BERT-base architecture. It serves as a strong baseline to visually demonstrate enhancements made by FSL and FSA. By introducing FSL and FSA, our FSUIE-base achieves significant performance improvements over the UIE-base that does not have fuzzy span mechanism (+1.15, +1.59, +1.99 F1 scores). Our proposed FSUIE model shows the most significant improvement on the ADE dataset. This is primarily due to the smaller scale of training datasets in the ADE dataset, which allows the model to easily learn generalized fuzzy span-aware representations. This demonstrates the superiority of the FSUIE model.

FSL and FSA enable the model to reduce over-dependence on label span boundaries and learn span-aware representations. When compared to existing NER models, FSUIE achieves new state-of-the-art performance on the ADE dataset even with the BERT-base backbone. FSUIE-large even achieve significant improvement (+1.42) on FSUIE-base. FSUIE-large also achieves comparable results on the ACE04 and ACE05 datasets, even when compared to models using stronger pre-trained language models such as ALBERT-xxlarge. Furthermore, our FSUIE demonstrates an advantage in

Models	PLM	ACE04			ACE05			ADE		
		P	R	F_1	P	R	F_1	P	R	F_1
SpERT (Eberts and Ulges, 2019)	BERT-base	-	-	-	-	-	-	77.77	79.96	78.84
SpERT.PL (Santosh et al., 2021)	Bio-BERT	-	-	-	-	-	-	80.11	84.18	82.03
JCBIE (He et al., 2022)	Bio-BERT	-	-	-	-	-	-	-	-	74.18
(Ma et al., 2020)	BERT-base	-	-	-	-	-	66.10	-	-	-
(Ma et al., 2020)	BERT-large	-	-	-	-	-	68.10	-	-	-
Tabel-Sequence Encoder (Wang and Lu, 2020)	ALBERT-xxlarge	-	-	63.30	-	-	67.60	-	-	80.10
PL-Marker (Ye et al., 2022)	BERT-base	-	-	66.70	-	-	69.00	-	-	-
PL-Marker (Ye et al., 2022)	ALBERT-xxlarge	-	-	69.70	-	-	73.00	-	-	-
Generative UIE (SEL only) (Lu et al., 2022)	T5-v1.1-large	-	-	-	-	-	64.68	-	-	-
Generative UIE (Lu et al., 2022)	T5-v1.1-large	-	-	-	-	-	66.06	-	-	-
UIE-base	BERT-base	88.73	53.46	66.72	95.38	52.04	67.34	67.61	56.31	61.45
FSUIE-base	BERT-base	91.78	58.99	71.82	96.79	57.69	72.29	91.10	75.87	82.79
FSUIE-large	BERT-large	89.01	61.13	**72.48**	98.84	59.34	**74.16**	92.02	78.23	**84.57**

Table 2: RE experimental results on ACE04, ACE05, and ADE datasets.

terms of its structure prediction compared to the generative UIE model. As it does not require the generation of complex IE linearized sequences, our FSUIE-base, which only uses BERT-base as its backbone, outperforms the generative UIE model that uses T5-v1.1-large on the ACE05 dataset.

3.3 Results on RE tasks

In Table 2, we present the results of the RE tasks. Compared to the baseline, UIE-base, which does not incorporate fuzzy span mechanism, our proposed FSUIE-base, which incorporates FSL and FSA, also achieves a significant improvement on the RE task using same backtone. Furthermore, when compared to the Table-Sequence Encoder approach (Wang and Lu, 2020), our method learns label span boundary distribution and span-aware representations, resulting in optimal or competitive results on the RE task even with FSUIE-base, despite using a simpler structure and smaller PLM backbones.

Compared to span-based IE models, our method outperforms the traditional joint extraction model by performing a two stage span extraction and introducing the fuzzy span mechanism. Specifically, on the ADE dataset, our method performs better than joint extraction methods using Bio-BERT, a domain-specific pre-trained language model on biomedical corpus, even using BERT-base as the pre-trained language model. This demonstrates that the fuzzy span mechanism we introduced can extract general information from the data, giving the model stronger information extraction capabilities, rather than simply fitting the data.

Compare to generative UIE models, our span-based FSUIE reflects the reality of the structure of IE task and does not require additional sequence generation structures, achieving higher results with less parameters even with FSUIE-base. Compared to models that perform relation extraction using a pipeline approach, like PL-Marker, our FSUIE improves performance in both stages of the pipeline by introducing FSL and FSA. As a result, it results in an overall improvement in relation extraction. Additionally, our model achieves new state-of-the-art results on ACE04 and ADE datasets,even using only BERT-base as the backbone, and on ACE05 dataset with FSUIE-large, compared to other models that use more complex structures. This demonstrates the model's ability to effectively extract information through our proposed method.

3.4 Results on ASTE tasks

In Table 3, we present the results of our experiments on the ASTE task. Due to the small scale of the ASTE-Data-V2, FSUIE-large is not needed to achieve better results, and this section only uses FSUIE-base for comparison. It can be seen that by introducing the fuzzy span mechanism, our FSUIE model significantly improves ASTE performance compared to the baseline UIE-base. This also demonstrates the effectiveness and generalization ability of FSUIE in IE tasks. Additionally, our FSUIE-base model achieves state-of-the-art results on three datasets (14lap, 15res, 16res) and demonstrates competitive performance on the 14res dataset. This indicates that the fuzzy span mechanism is effective in improving the model's ability to exploit and extract information, as well as its performance on specific tasks without increasing model parameters.

Furthermore, our FSUIE model has a relatively simple architecture, compared to other models, which shows that FSUIE is able to improve per-

Models	PLM	14lap			14res			15res			16res		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
JET (Xu et al., 2020)	BERT-base	55.39	47.33	51.04	70.56	55.94	62.40	64.45	51.96	57.53	70.42	58.37	63.83
Dual-MRC (Mao et al., 2021)	BERT-base	57.39	53.88	55.58	71.55	69.14	70.32	63.78	51.87	57.21	68.60	66.24	67.40
ASTE-RL (Mao et al., 2021)	BERT-base	64.80	54.99	59.50	70.60	68.65	69.61	65.45	60.29	62.72	67.21	69.69	68.41
GAS (Zhang et al., 2021)	BERT-base	-	-	60.78	-	-	72.16	-	-	62.10	-	-	70.10
Span-ASTE (Xu et al., 2021)	BERT-base	63.44	55.84	59.38	72.89	70.89	71.85	62.18	64.45	63.27	69.45	71.17	70.26
SBN (Chen et al., 2022)	BERT-base	65.68	59.88	62.65	76.36	72.43	74.34	69.93	60.41	64.82	71.59	72.57	72.08
Generative UIE (SEL only) (Lu et al., 2022)	T5-v1.1-large	-	-	63.15	-	-	73.78	-	-	66.10	-	-	73.87
Generative UIE (Lu et al., 2022)	T5-v1.1-large	-	-	63.88	-	-	**74.52**	-	-	67.15	-	-	75.07
UIE-base	BERT-base	65.21	57.64	61.19	75.32	71.53	73.38	71.11	65.98	68.45	74.84	70.04	72.36
FSUIE-base	BERT-base	69.49	62.06	**65.56**	76.22	72.23	74.17	72.71	68.66	**70.63**	77.98	73.74	**75.80**

Table 3: ASTE experimental results on ASTE-DATA-V2 datasets (14lap, 14res, 15res, and 16res).

formance without the need for complex structures. The gap in performance between UIE models and other models can be attributed, in part, to the advantage of UIE pre-training, which is further enhanced by our proposed fuzzy span mechanism. Compared to models that decompose the ASTE task into two subtasks of opinion recognition and sentiment classification, and use separate models to handle each, our FSUIE model achieved better performance using a unified model architecture.

For ASTE, span-based UIE models, as opposed to generative UIE models, can leverage the complete semantic information of the predicted aspect span to assist in extracting opinions and sentiments. The fuzzy span mechanism enhances the model's ability to exploit the semantic information within the fuzzy span, where possible opinions and sentiments reside, while ensuring span-aware representation learning, resulting in significant improvements.Furthermore, FSUIE is a reaction to the real structure of IEtask, avoiding the extra parameters that sequence generation structures bring, and therefore outperforms generative UIE models with fewer parameters.

We notice that FSUIE improves relatively less on the RE task compared to the ASTE task. In the RE task, the model has to learn different entities, different types of relationships, and binary matching skills. In contrast, in ASTE tasks, the model only needs to learn different entities, two relationships that differ significantly in semantics (opinion and sentiment), and ternary pairing tips. From this perspective, RE tasks are more challenging than ASTE tasks.

3.5 Results on Low-resource Settings

To demonstrate the robustness of our proposed FSUIE method in low-resource scenarios, we conducted experiments using a reduced amount of training data on ACE04 for NER and RE tasks, and 14res for ASTE task. Specifically, we created three subsets of the original training data at 1%, 5%, and 25% of the original size. In each low-resource experiment, we trained the model for 200 epochs instead of 50 epochs. The results of these experiments were compared between FSUIE-base and UIE-base and are presented in Table 4.

Entity F1	**1%**	**5%**	**25%**	**100%**
UIE-base	63.47	72.98	83.08	84.63
FSUIE-base	70.09	77.20	83.49	85.22
Relation Strict F1	**1%**	**5%**	**25%**	**100%**
UIE-base	6.68	45.55	64.43	66.72
FSUIE-base	9.73	53.44	66.08	71.82
Sentiment Triplet F1	**1%**	**5%**	**25%**	**100%**
UIE-base	45.66	63.12	73.73	73.38
FSUIE-base	46.87	63.79	74.27	74.17

Table 4: Experimental results on low-resource settings.

The results of the low-resource experiments further confirm the superior performance of FSUIE over UIE in handling low-resource scenarios. With only a small fraction of the original training data, FSUIE is still able to achieve competitive or even better performance than UIE. This demonstrates the robustness and generalization ability of FSUIE in dealing with limited data. Overall, the results of the low-resource experiments validate the ability of FSUIE to effectively handle low-resource scenarios and extract rich information through limited data.

We also found that the model both performed better on NER and ASTE taks than on RE task under low-resource settings. This is because NER and ASTE tasks are simpler than RE, so less data can bring better learning performance. Additionally, we noticed a small performance decrease in the ASTE task for the 100% set compared to the

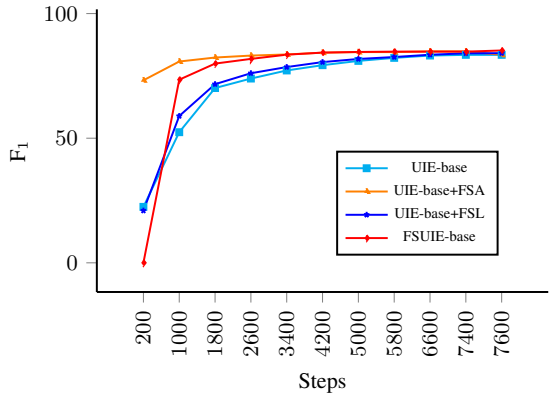

Figure 4: NER performance of different models on ACE04 test set.

25% set. This change may be due to the fact that the training data is unbalanced, and reducing the training size can alleviate this phenomenon.

3.6 Ablation Study

Since FSUIE has been verified to make more effective use of the information in the training set, in order to verify this, we verify it from the perspective of the model training process. Specifically, we recorded the effects of baseline UIE-base, UIE-base+FSL, UIE-base+FSA and full model FSUIE-base on different training steps on the NER ACE04 test set, and the results are shown in Figure 4.

We noticed that the models with FSA have a significantly faster convergence speed, indicating that by learning span-aware representations, which are closer to the span prediction goal, the span learning process becomes more easy and efficient. With FSA, the model can focus its attention on the necessary positions and capture the possible span within a given sequence. While for FSL, it have a similar convergence trend with the baseline, thus may not improve the convergence speed.

Models	P	R	F1
UIE-base	87.85	91.56	89.67
UIE-base + FSL	89.61	90.58	90.09
UIE-base + FSA	89.21	90.09	89.65
FSUIE-base	91.17	92.17	92.49

Table 5: Ablation study of FSL and FSA on NER task using ADE dataset.

To further investigate the contribution of FSL and FSA to the improvement of model performance, we conduct ablation experiments on the NER task using the ADE dataset. The specific experimental results are shown in Table 5. It can be seen that the introduction of FSL alone can improve model performance individually. When using FSA alone, the performance of the model drops slightly. However, when both FSL and FSA are used together, the model is significantly enhanced.

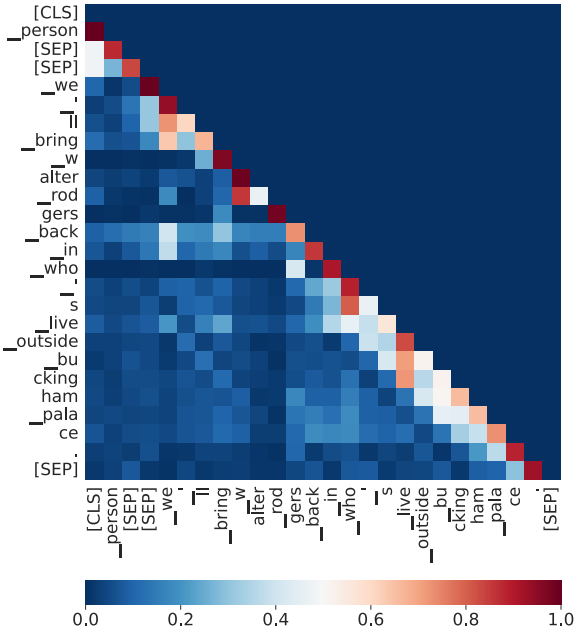

Figure 5: Illustration of attention scores distribution in FSA layer. The extracting target are "walter rodgers" and "who".

From our perspectives, the separate introduction of FSA makes the model focus on specific parts of the sequence rather than global representation, resulting in a loss of information from text outside the span. This may explain the slight drop in performance when using UIE+FSA. However, this also demonstrates that in the IE task, sequence information outside a specific span has a very limited impact on the results. The introduction of FSL alleviates the model's over-dependence on label span boundaries, allowing the model to extract more information, resulting in an improvement in both settings. When FSA and FSL operate simultaneously, the model extracts more information from the text and FSA guides the model to filter the more critical information from the richer information, resulting in the most substantial improvement.

3.7 Visualization of FSA

To further examine the effectiveness of the fuzzy span mechanism, we visualized the attention distribution of the FSA layer in FSUIE-large as shown in Figure 5. It should be noted that FSA is only placed at the top layer for constructing span-aware

representation and does not participate in the encoding process, thus only affects span decisions rather than the representation of tokens in the sequence.

The attention distribution indicates that, for a given input text, each token in the final encoding sequence tends to focus on semantic information within a limited range of preceding tokens rather than on the global representation of the input text. This aligns with our expectation for the design of the fuzzy span mechanism and confirms that fuzzy span mechanism does indeed guide the model for appropriate attention distribution for IE tasks.

4 Related Work

Universal Model Building universal model structures for a wide range of NLP tasks has been a hot research area in recent years. The focus is on building model structures that can be adapted to different sources of data, different types of labels, different languages, and different tasks. Several universal models have been proposed, such as models learning deep contextualized word representations (Peters et al., 2018; Devlin et al., 2019), event extraction models that can predict different labels universally (Lu et al., 2021), models that can handle multiple languages (Arivazhagan et al., 2019; Aharoni et al., 2019; Conneau et al., 2020), a universal fine-tuned approach to transfer learning (Howard and Ruder, 2018), models that learning syntactic dependency structure over many typologically different languages (Li et al., 2018; Sun et al., 2020) and models that can universally model various IE tasks in a unified text-to-structure framework (Lu et al., 2022). This paper builds upon the UIE by incorporating the fuzzy span mechanism to improve IE performance.

Information Extraction IE is the task of extracting structured information from unstructured text data. This includes NER, RE, ASTE, Event Extraction (EE), Aspect-Based Sentiment Analysis (ABSA), etc. Research has proposed numerous approaches for IE, such as rule-based (Appelt and Onyshkevych, 1998), machine learning (Téllez-Valero et al., 2005; Kolya et al., 2010), deep learning (Qin et al., 2018), active learning (Radmard et al., 2021), and logic fusion (Wang and Pan, 2020). There are still many task-specific models being proposed, based on previous approaches and structures, e.g., NER (Li et al., 2022; Zhu and Li, 2022; Yang et al., 2022; Zhang et al., 2019); RE (Nan et al., 2020; Lai et al., 2021), ABSA (Jing et al., 2021); and ASTE (Xu et al., 2021, 2020).

More related work about sparse attention please refer to Appendix B.

5 Conclusion

In this paper, we proposed the Fuzzy Span Universal Information Extraction (FSUIE) framework, an improvement for Universal Information Extraction. To make use of boundary information in the training data and learn a decision-closer span-aware representation, we proposed a fuzzy span loss and fuzzy span attention. Extensive experiments on several main IE tasks show that our FSUIE has a significant improvement compared to the UIE baseline, and achieves state-of-the-art results on ADE NER datasets, ACE04 RE, ACE05 RE and ADE RE datasets and four ASTE datasets. The experiments also reveal FSUIE's fast convergence and good generality in low-resource settings. All the results demonstrate the effectiveness and generalizability of our FSUIE in information extraction.

6 Limitations

This paper are based on the assumption that Universal Information Extraction (UIE) models have limitations, particularly with regards to over-reliance on label span boundaries and inflexible attention span length. Therefore, the proposed framework may be computationally and spatially expensive as it requires a more complex attention mechanism and additional computing power for training. Nevertheless, this limitation of the span-based UIE model can be overlooked in comparison to that of the generative UIE model, which uses a stronger language model. Additionally, the probability density functions explored in FSL are limited; thus, further research is needed to develop a more targeted strategy for adjusting the correct information distribution.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Douglas E. Appelt and Boyan Onyshkevych. 1998. The common pattern specification language. In TIPSTER

TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998, pages 23–30, Baltimore, Maryland, USA. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. CoRR, abs/1907.05019.

Yuqi Chen, Keming Chen, Xian Sun, and Zequn Zhang. 2022. A span-level bidirectional network for aspect sentiment triplet extraction. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. CoRR, abs/1904.10509.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

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

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training. CoRR, abs/1909.07755.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. J. Biomed. Informatics, 45(5):885–892.

Kai He, Rui Mao, Tieliang Gong, Erik Cambria, and Chen Li. 2022. Jcbie: a joint continual learning neural network for biomedical information extraction. BMC bioinformatics, 23(1):1–20.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Huiqiang Jiang, Guoxin Wang, Weile Chen, Chengxi Zhang, and Börje F. Karlsson. 2021. Boningknife: Joint entity mention detection and typing for nested NER via prior boundary knowledge. CoRR, abs/2107.09429.

Hongjiang Jing, Zuchao Li, Hai Zhao, and Shu Jiang. 2021. Seeking common but distinguishing difference, a joint aspect-based sentiment analysis model. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3910–3922, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anup Kumar Kolya, Asif Ekbal, and Sivaji Bandyopadhyay. 2010. A supervised machine learning approach for event-event relation identification. In Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, pages 447–454, Tohoku University, Sendai, Japan. Institute of Digital Enhancement of Cognitive Processing, Waseda University.

Tuan Lai, Heng Ji, ChengXiang Zhai, and Quan Hung Tran. 2021. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6248–6260, Online. Association for Computational Linguistics.

Zhuoran Li, Chunming Hu, Xiaohui Guo, Junfan Chen, Wenyi Qin, and Richong Zhang. 2022. An unsupervised multiple-task and multiple-teacher model for cross-lingual named entity recognition. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 170–179, Dublin, Ireland. Association for Computational Linguistics.

Zuchao Li, Shexia He, Zhuosheng Zhang, and Hai Zhao. 2018. Joint learning of POS and dependencies for multilingual Universal Dependency parsing. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 65–73, Brussels, Belgium. Association for Computational Linguistics.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2795–2806, Online. Association for Computational Linguistics.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In Proceedings of the 60th Annual Meeting

of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

Youmi Ma, Tatsuya Hiraoka, and Naoaki Okazaki. 2020. Named entity recognition and relation extraction using enhanced table filling by contextualized representations. CoRR, abs/2010.07522.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. A joint training dual-mrc framework for aspect based sentiment analysis. 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, pages 13543–13551. AAAI Press.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1546–1557, Online. Association for Computational Linguistics.

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, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust distant supervision relation extraction via deep reinforcement learning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2137–2147, Melbourne, Australia. Association for Computational Linguistics.

Puria Radmard, Yassir Fathullah, and Aldo Lipani. 2021. Subsequence based deep active learning for named entity recognition. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4310–4321, Online. Association for Computational Linguistics.

TYSS Santosh, Prantika Chakraborty, Sudakshina Dutta, Debarshi Kumar Sanyal, and Partha Pratim Das. 2021. Joint entity and relation extraction from scientific documents: Role of linguistic information and entity types.

Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. 2022. Global pointer: Novel efficient span-based approach for named entity recognition. CoRR, abs/2208.03054.

Sainbayar Sukhbaatar, Edouard Grave, Piotr Bojanowski, and Armand Joulin. 2019. Adaptive attention span in transformers. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 331–335, Florence, Italy. Association for Computational Linguistics.

Kailai Sun, Zuchao Li, and Hai Zhao. 2020. Cross-lingual universal dependency parsing only from one monolingual treebank. arXiv preprint arXiv:2012.13163.

Chuanqi Tan, Wei Qiu, Mosha Chen, Rui Wang, and Fei Huang. 2020. Boundary enhanced neural span classification for nested named entity recognition. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 9016–9023. AAAI Press.

Alberto Téllez-Valero, Manuel Montes-y-Gómez, and Luis Villaseñor Pineda. 2005. A machine learning approach to information extraction. In Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing 2005, Mexico City, Mexico, February 13-19, 2005, Proceedings, volume 3406 of Lecture Notes in Computer Science, pages 539–547. Springer.

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

Jue Wang and Wei Lu. 2020. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1706–1721, Online. Association for Computational Linguistics.

Wenya Wang and Sinno Jialin Pan. 2020. Integrating deep learning with logic fusion for information extraction. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 9225–9232. AAAI Press.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. Learning span-level interactions for aspect sentiment triplet extraction. In Proceedings of the 59th Annual Meeting of the Association for Computational

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4755–4766, Online. Association for Computational Linguistics.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2339–2349, Online. Association for Computational Linguistics.

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

Yifei Yang, Zuchao Li, and Hai Zhao. 2022. Nested named entity recognition as corpus aware holistic structure parsing. In Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022, pages 2472–2482. International Committee on Computational Linguistics.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4904–4917, Dublin, Ireland. Association for Computational Linguistics.

Zheng Yuan, Chuanqi Tan, Songfang Huang, and Fei Huang. 2022. Fusing heterogeneous factors with triaffine mechanism for nested named entity recognition. In Findings of the Association for Computational Linguistics: ACL 2022, pages 3174–3186, Dublin, Ireland. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 504–510, Online. Association for Computational Linguistics.

Zhuosheng Zhang, Bingjie Tang, Zuchao Li, and Hai Zhao. 2019. Modeling named entity embedding distribution into hypersphere. CoRR, abs/1909.01065.

Enwei Zhu and Jinpeng Li. 2022. Boundary smoothing for named entity recognition. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7096–7108, Dublin, Ireland. Association for Computational Linguistics.

Yimeng Zhuang, Jing Zhang, and Mei Tu. 2022. Long-range sequence modeling with predictable sparse attention. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 234–243, Dublin, Ireland. Association for Computational Linguistics.

7 Appendix

7.1 g_a in FSA

	P	R	F1
UIE-base	87.85	91.56	89.67
FSUIE-base (g_a')	89.84	90.69	90.26
FSUIE-base (g_a'')	90.93	90.41	90.67
FSUIE-base (g_a^l)	91.17	92.17	92.49

Table 6: Performance of models using different g_a in FSA

In Table 6, we present the performance of models using various g_a functions in the FSUIE technique on the ADE NER test set, where g_a^l denotes the linear attenuated function employed in FSUIE. Compared to the UIE-base, which does not integrate the fuzzy span mechanism, all FSUIE-based models employing different g_a functions obtain better results, thus illustrating the superiority of FSUIE. Regarding the different g_a strategies, FSUIE-base (g_a') shows minimal enhancement. This is likely because the fuzzy span of attention attenuation adequately reflects the real reading context and enables the model to take advantage of more abundant information within the boundary of the attention span. The best performance is achieved by FSUIE-base (g_a^l), which indicates that the attention should not decay too quickly at the boundary of the attention span, as evidenced by the results of g_a''.

7.2 Related Work on Sparse Attention

The high time and space complexity of Transformer ($O(n^2)$) is due to the fact that it needs to calculate the attention information between each step and all previous contexts. This makes it difficult for Transformer to scale in terms of sequence length. To address this issue, sparse attention was proposed (Child et al., 2019). This refers to attention mechanisms that focus on a small subset of the input elements, rather than processing the entire input sequence. This method allows attention to be more focused on the most contributing value factors, thus reducing memory and computing capacity requirements.

Based on the idea of sparse attention, various approaches have been proposed, such as an adaptive width-based attention learning mechanism and a dynamic attention mechanism that allows different heads to learn only the region of attention (Sukhbaatar et al., 2019). Zaheer et al. (2020) proposed an $O(N)$ complexity model with three different sparse attentions. Zhuang et al. (2022) sought to make the sparse attention matrix predictable. This paper, however, based on adaptive span attention (Sukhbaatar et al., 2019) to establish a fuzzy span attention, which aims at learning a span-aware representation with the actual needs of information extraction tasks. Our approach differs from previous work in that we aim to obtain a fuzzy span of attention in the process of locating the target, rather than reducing computational and memory overhead.

7.3 L_{span} and d in FSA

d	P	R	F1
16	91.45	93.45	92.44
32	92.58	92.40	92.49
48	92.32	92.50	92.41

Table 7: Performance of FSUIE-base models with different d

L_{span}	P	R	F1
16	91.65	93.83	92.73
30	92.58	92.40	92.49
48	92.25	92.69	92.47

Table 8: Performance of FSUIE-base models with different L_{span}

In Table 7, we present the performance of FSUIE-base models using various hyper-parameter d on the ADE NER test set. In Table 8, we present the performance of FSUIE-base models using various hyper-parameter L_{span} on the ADE NER test set. The results demonstrate that the model's performance is not significantly affected by the choice of these hyper-parameters.

7.4 Single-Side and Both-Side Ambiguity in FSL

FSL strategies	P	R	F1
both-side	92.58	92.40	92.49
single-side	91.99	92.69	92.34

Table 9: Performance of FSUIE-base models with different FSL strategies

Actually, there may be cases of single-side ambiguity in the labeling of entity boundaries in the text. Therefore, we demonstrate FSUIE-base models' performance with different FSL strategies in

Table 9, where "single-side" means applying FSL only on start boundary and "both-side" means applying FSL on both start and end boundary. The results suggests that the influence of single-sided and both-sided fuzziness on the model's performance is limited, because not all head words are at the end or start, and FSL only performs limited left/right extrapolation on precise boundaries, without affecting the important information provided by the original boundary. For generalization purposes, we utilized both-sides fuzzy span in FSUIE.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section #Limitation

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section #1

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

B ☒ Did you use or create scientific artifacts?

Left blank.

☑ B1. Did you cite the creators of artifacts you used?
Left blank.

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

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

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

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

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

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. Left blank.

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

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

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

Left blank.

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

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

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

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

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