Function-words Adaptively Enhanced Attention Networks for Few-Shot Inverse Relation Classification

Chunliu Dou[1*], **Shaojuan Wu**[1*], **Xiaowang Zhang**[1,2†], **Zhiyong Feng**[1] and **Kewen Wang**[3]

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China, 300350
[2]Tianjin University-Aishu Data Intelligence Joint Laboratory, Tianjin, China
[3]School of Information and Communication Technology, Griffith University, Brisbane, Australia

Abstract

The relation classification is to identify semantic relations between two entities in a given text. While existing models perform well for classifying inverse relations with large datasets, their performance is significantly reduced for few-shot learning. In this paper, we propose a function words adaptively enhanced attention framework (FAEA) for few-shot inverse relation classification, in which a hybrid attention model is designed to attend class-related function words based on meta-learning. As the involvement of function words brings in significant intra-class redundancy, an adaptive message passing mechanism is introduced to capture and transfer inter-class differences. We mathematically analyze the negative impact of function words from dot-product measurement, which explains why the message passing mechanism effectively reduces the impact. Our experimental results show that FAEA outperforms strong baselines, especially the inverse relation accuracy is improved by 14.33% under 1-shot setting in FewRel1.0.

1 Introduction

Relation Classification (RC) aims to classify the relation between two given entities based on their related context. Specifically, given a sentence in a natural language, a set of relation names, and two entities, we want to determine the correct relation between two entities. RC is widely used in natural language processing, such as knowledge base completion [Dong *et al.*, 2020], dialog system[Lee, 2021]. Most existing approaches to RC are based on supervised learning, and the datasets used in training depend on manually labeled data, which limits the classification performance with only a few instances. Therefore, it is necessary to make the RC models able to handle relations with few instances. Due to the success of few-shot learning in computer vision domain [Wu *et al.*, 2021; Yang *et al.*, 2022], Han *et al.* [2018] first investigated the problem of few-shot relations classification (FSRC) and proposed the dataset FewRel1.0 for evaluating the performance of FSRC models. Since then, several other FSRC models have been reported in the literature and they demonstrate remarkable performance on FewRel1.0.[Gao *et al.*, 2019a; Qu *et al.*, 2020; Yang *et al.*, 2021].

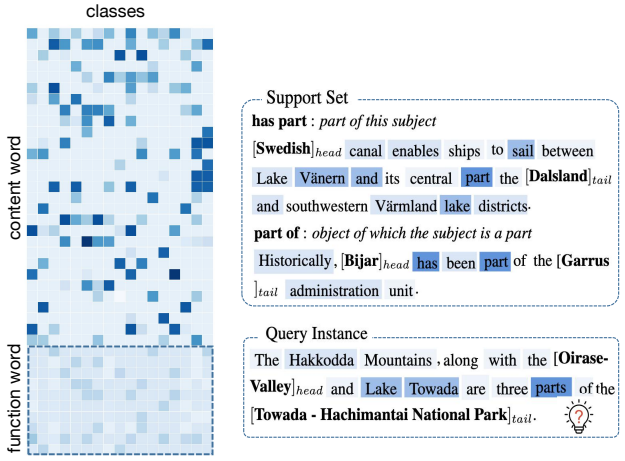

Figure 1: The left figure shows the words attention visualization of TD-PROTO, where a darker unit indicates a higher value. We observe content words region is deeper than function words. The right is a 2-way-1-shot task of FSIRC, involving two relations and each relation only with one support instance.

However, our experiments show that the performance of existing models for FSRC is significantly reduced when one relation is the inverse of another relation in the given set of relations. Figure 1 shows a few-shot inverse relations classification (FSIRC) task where the relations set contains two relations '*has part*' and '*part of*'. From the two support instances, we can see that the relation '*has part*' is the inverse relation of '*part of*'. We note that '*has part*' and '*part of*' have the same content word '*part*' but different function words '*its*' and '*of*'. Existing models[Sun *et al.*, 2019; Bao *et al.*, 2020] for FSRC focus on characterizing the differences between content words but ignore the differences of function words. As a result, these models do not perform well when one relation is the inverse of another relation.

In practical applications, it is often the case that one relation is the inverse of another relation. For example, we found that 21.25% of relations in FewRel1.0 dataset are inverse.

[*]These authors contributed equally to this work and should be considered co-first authors.
[†]Corresponding authors.

However, it is useful but challenging to classify relations with the presence of inverse relations in few-shot scenarios, as the lack of sufficient samples makes it hard to determine the importance distribution of class-related function words.

In order to address this issue, we propose a new approach to the problem of FSIRC, called *Function-words Adaptively Enhanced Attention Networks (FAEA)*. As shown in Figure 2, in the instance encoder, besides considering the importance of keywords, we use a hybrid attention to capture function words. Specifically, the class-general attention mechanism learns general function words importance distribution. As function words appearing in the same phrase with keywords are more likely to be informative [Zhang *et al.*, 2020], we design a class-specific attention by strengthening function words importance adjacent to keywords. From our experience, in some cases, function words far from keywords are also important. For this reason, we introduce semantic-related attention for computing the direct semantic relevance between function words and keywords. However, the introduction of function words may increase the intra-class differences. So we present a message passing mechanism to capture and transfer inter-class differences and intra-class commonalities between instances. But when inter-class differences are large, they will bring in noises and thus useful relation semantics can be lost. To avoid this issue, we adaptively control the proportion of transferred inter-class message. Our experiments show that FAEA significantly outperforms major baseline models for FSIRC. Our code is available at https://github.com/DOU123321/FAEA-FSRC.

In a nutshell, our contributions are listed as follows:

- We present FAEA that uses a hybrid attention to capture class-related function words and an adaptive message passing mechanism to reduce intra-class redundancy caused by function words.
- We mathematically show that the involvement of function words will increase intra-class differences from dot-product measurement and the designed message passing mechanism effectively reduces the redundancy.
- We conduct experiments with two datasets and the results show that our model significantly outperforms the baselines, especially for FSIRC task. Ablation experiments demonstrate the effectiveness of the proposed modules.

2 Related Work

Few-shot relation classification predicts novel relations by exploring a small number of labeled instances. Existing methods can be mainly divided into two categories: Gradient-based and metric-based models. A gradient-based method [Finn *et al.*, 2017; Abiola and Andreas, 2019; Qu *et al.*, 2020] can quickly adapt the model to a specific task through a few update steps. MAML[Finn *et al.*, 2017] is a representative model, learning appropriate initialization parameters of the model from base classes and transferring these parameters to novel classes. And metric-based models[Snell *et al.*, 2017; Gao *et al.*, 2019a; Ye and Ling, 2019; Wen *et al.*, 2021] learn the distance distributions among classes, and the same class instances are adjacent in the distance space. As a representative model, Prototypical Network (PN) [Snell *et al.*, 2017] calculates the prototype for each class and classifies query instances by calculating their Euclidean distances. Some models[Ye and Ling, 2019; Wen *et al.*, 2021] add attention mechanisms to enhance PN for highlighting crucial instances and features, but they ignore the intra-instance differences. Some models [Bao *et al.*, 2020; Yang *et al.*, 2021] capture local content words to obtain fine-grained information and ignore function words. However, inverse relations of FSRC has not been effectively handled. This work focuses on inverse relations and proposes a hybrid function words attention to model subtle variations across inverse relations.

3 Our Method

3.1 Problem Statement

FSRC is defined as a task to predict the relation y between the entity pair (h, t) mentioned in a query instance x^q, given a support set $\mathcal{S}$ and a relation set $\mathcal{R}$, $\mathcal{S} = \left\{\left(x_k^i, h_k^i, t_k^i, r^i, y^i\right), i = 1, \ldots, N; k = 1, \ldots, K\right\}$ and $\mathcal{R} = \left\{y^1, y^2, \ldots, y^N\right\}$, where $\left(x_k^i, h_k^i, t_k^i, r^i, y^i\right)$ means there is a relation y^i between the entity pair $\left(h_k^i, t_k^i\right)$ in the instance x_k^i, and r^i is corresponding relation description. N is the number of relations, and each relation with quite small K labeled instances. For a FSIRC task, the relation set $\mathcal{R}$ includes some pairs of inverse relations. For example, '*participant*' and '*participant of*' are inverse relations, and their relation descriptions are "*person that actively takes part in the event*" and "*event a person was a participant in*", respectively.

3.2 Overall Framework

As shown in Figure 2, our model consists of three parts:

- **Instance Encoder**. Given an instance and entity pair, we employ instance-level global encoder and phrase-level local encoder to encode the instance into an embedding.
- **Function-words Enhanced Attention**. Phrase-level local encoder utilizes function-words enhanced attention to capture important function words in instances.
- **Adaptive Message Passing**. After computing embeddings, we transfer commonalities between same class instances and differences between different class instances.

3.3 Instance Encoder

Given an instance $x = \{w_1, \ldots, w_l\}$ mentioning two entities with l words, we use BERT [Devlin *et al.*, 2019] as the encoder to get corresponding embeddings $\mathbf{X} = \{\mathbf{w}_1, \ldots, \mathbf{w}_l\}$, where each word embedding $\mathbf{w}_i \in \mathbb{R}^d$ and d is embedding dimension. For i-th relation r^i, we encode the name and description to get relation word embeddings $\mathbf{R}^i \in \mathbb{R}^{l \times d}$, and use hidden states of [CLS] token to obtain features of relations $\mathbf{r}^i \in \mathbb{R}^{2d}$.

For instance x_k^i in $\mathcal{S}$ and query instance x^q, our model generates global instance embeddings and local phrase embeddings to form hybrid instance embeddings $\mathbf{x}_k^i$ and $\mathbf{x}^q$. The following takes $\mathbf{x}_k^i$ as an example to explain.

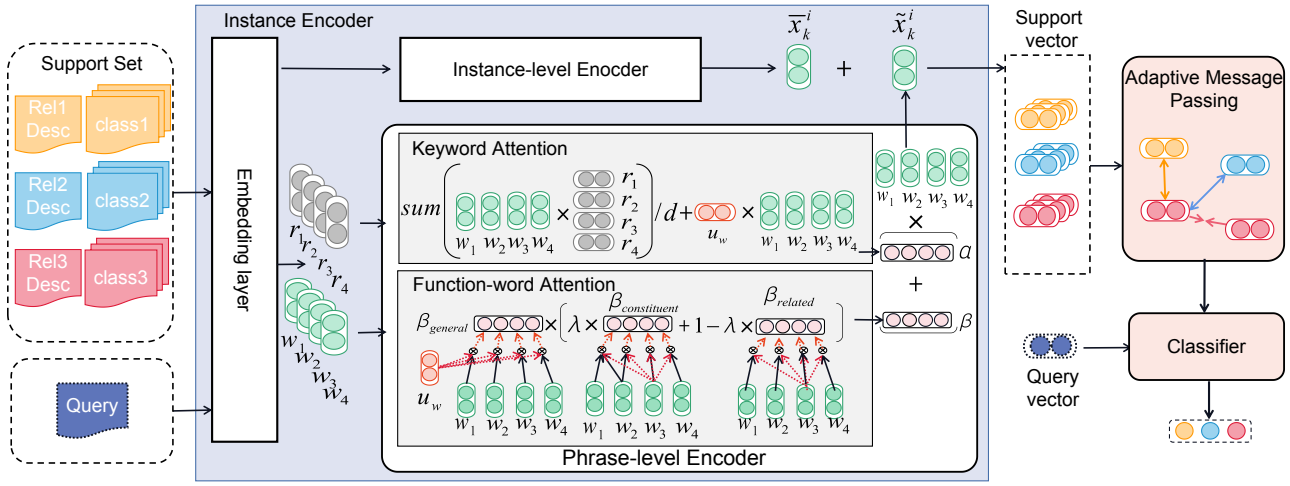

Figure 2: The overall framework of FAEA. The input of Instance Encoder is an instance with corresponding relation description.

Instance-level Global Encoder. The global features $\{\overline{\mathbf{x}}_k^i \in \mathbb{R}^{2d}, i = 1, \ldots, N; k = 1, \ldots, K\}$ are obtained by concatenating corresponding hidden states of two entity mentions according to [Soares *et al.*, 2019].

Phrase-level Local Encoder. The main process consists of learning a keyword attention $\boldsymbol{\alpha}_k^i \in \mathbb{R}^l$ and a function-words attention $\boldsymbol{\beta}^s \in \mathbb{R}^l$. $\boldsymbol{\alpha}_k^i$ can be computed as follows:

$$\boldsymbol{\alpha}_k^i = \text{softmax}\left(\mathbf{X}_k^i \mathbf{u}_w + \text{sum}\left(\frac{\mathbf{X}_k^i (\mathbf{R}^i)^T}{\sqrt{d}}\right)/d\right) \tag{1}$$

where the memory unit $\mathbf{u}_w \in \mathbb{R}^d$ is a trainable parameter. It can help us to select general keywords from instances. Then we obtain $\boldsymbol{\beta}^s$ according to 3.4 and form phrase-level local features $\tilde{\mathbf{x}}_k^i$, computed as follows:

$$\tilde{\mathbf{x}}_k^i = \left(\boldsymbol{\beta}^s + \boldsymbol{\alpha}_k^i\right) \mathbf{X}_k^i \tag{2}$$

In short, we additionally attend local phrase-level information to learn the subtle differences of inverse relations.

3.4 Function-words Enhanced Attention

We utilize class-general attention to learn general function words importance distribution and leverage class-specific attention consisting of constituent attention and semantic-related attention to estimate class-specific importance.

Class-general Attention. By downweighing the importance of words related to $\mathbf{u}_w$ and upweighing the words importance irrelated to $\mathbf{u}_w$, we get general function-words importance $\boldsymbol{\beta}_{\text{general}} \in \mathbb{R}^l$, where $\mathbf{E} \in \mathbb{R}^l$ is an all-one vector:

$$\boldsymbol{\beta}_{\text{general}} = \text{softmax}\left(\mathbf{E} - \mathbf{X}_k^i \mathbf{u}_w\right) \tag{3}$$

Class-specific Attention. Considering function words importance varying by class, we learn a constituent prior matrix $\mathbf{C} \in \mathbb{R}^{l \times l}$ and a semantic-related matrix $\mathbf{S} \in \mathbb{R}^{l \times l}$ to strengthen the attention of function words adjacent to keywords.

The element $C_{i,j}$ means the probability that w_i and w_j in instance x_k^i belong to the same phrase, obtained as follows, where $[\cdot]_n$ is the n-th row of a matrix.

$$C_{i,j} = e^{\sum_{n=i}^{j-1} \log(a_n)} \tag{4}$$

$$s_{n,n+1} = \left([\mathbf{X}_k^i]_n \times [\mathbf{X}_k^i]_{n+1}^T\right)/\sqrt{d} \tag{5}$$

$$p_{n,n+1}, p_{n,n-1} = \text{softmax}\left(s_{n,n+1}, s_{n,n-1}\right) \tag{6}$$

$$a_n = \sqrt{p_{n,n+1} \times p_{n+1,n}} \tag{7}$$

We compute the score $s_{n,n+1}$ representing the tendency that w_n links to right neighbor w_{n+1}. Then, constrain w_n to either link to its right neighbor or left neighbor. This constraint is implemented by applying a softmax function to two attention links of w_n. As $p_{n,n+1}$ and $p_{n+1,n}$ may have different values, we average its two attention links.

We also use self-attention mechanism to obtain matrix $\mathbf{S}$ to attend necessary function words far from keywords:

$$\mathbf{S} = \frac{\mathbf{X}_k^i \times (\mathbf{X}_k^i)^T}{\sqrt{d}} \tag{8}$$

Next, we find the keywords index $\mathbf{I}_k^i$ in x_k^i, computed as:

$$\mathbf{I}_k^i = \max(\boldsymbol{\alpha}_k^i)_r \tag{9}$$

where $\max(\cdot)_r$ is used to get indexes of the top-r largest attention keywords, and r is the number of keywords. And then, we strengthen related function words according to $\mathbf{C}$ and $\mathbf{S}$:

$$\boldsymbol{\beta}_{\text{constituent}} = [\mathbf{C}]_{\mathbf{I}_k^i} \text{ and } \boldsymbol{\beta}_{\text{related}} = [\mathbf{S}]_{\mathbf{I}_k^i} \tag{10}$$

Finally, the model uses $\boldsymbol{\beta}_{\text{constituent}}$, $\boldsymbol{\beta}_{\text{related}}$ and $\boldsymbol{\beta}_{\text{general}}$ to form hybrid function-words attention vector $\boldsymbol{\beta}$, formalized as:

$$\begin{aligned} \boldsymbol{\beta}' &= \lambda \boldsymbol{\beta}_{\text{constituent}} + (1-\lambda)\boldsymbol{\beta}_{\text{related}} \\ \boldsymbol{\beta} &= \frac{1}{r}\sum_{i=1}^{r} [\boldsymbol{\beta}']_i + \boldsymbol{\beta}_{\text{general}} \end{aligned} \tag{11}$$

where λ is hyper-parameter.

All in all, inspired by MAML[Finn *et al.*, 2017] learning general model parameters and fine-tunning them to adapt to the specific task, we design class-general and class-specific attention to learn function-words variance in few-shot setting.

3.5 Adaptive Message Passing

Adaptive Message Passing reduces intra-class redundancy caused by function words and adaptively controls the proportion of transferred inter-class message.

Firstly, we construct a directed graph $G = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V} = [\mathbf{x}_1^1; \ldots; \mathbf{x}_N^K]$ is a set of instances features with $|V| = N \times K$ and E is the adjacency matrix. $[;]$ denotes the row-wise concatenation and $\mathbf{v}^i \in \mathbb{R}^d$ denotes the i-th row of matrix $\mathbf{V}$.

$$E_{ij} = \begin{cases} \frac{(\mathbf{v}^i)^T \mathbf{v}^j}{\|\mathbf{v}^i\|_2 \|\mathbf{v}^j\|_2} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \tag{12}$$

We design a new node updating way that captures and transfers inter-class differences and intra-class commonalities between instance nodes, according to the work of [Bo *et al.*, 2021]

$$\bar{e}_i = \frac{1}{2(N \times K - 1)} \Big(\sum\nolimits_{j \in \mathcal{N}_i^0} E_{ij} + \sum\nolimits_{j \in \mathcal{N}_i^1} E_{ij} \Big) \tag{13}$$

$$e_i^j = \frac{\max(E_{ij} - \bar{e}_i, 0)}{\sqrt{d_i d_j}} \tag{14}$$

$$\tilde{\mathbf{v}}^i = \mathbf{v}^i - \sum\nolimits_{j \in \mathcal{N}_i^0} e_i^j \mathbf{v}^j + \sum\nolimits_{j \in \mathcal{N}_i^1} e_i^j \mathbf{v}^j \tag{15}$$

where $\mathcal{N}_i^0$ and $\mathcal{N}_i^1$ denote the different and same class neighbor set as $\mathbf{v}^i$, respectively. d_i is the degree of node i.

For i-th relation, we average K supporting features to form prototype representation $\mathbf{p}^i$ following [Snell *et al.*, 2017].

$$\mathbf{p}^i = \frac{1}{K} \sum_{j=(i-1)\times K}^{i \times K} \mathbf{v}^j \tag{16}$$

With N prototype representations, we calculates the probability that query instance x^q belongs to the relation i:

$$z\left(y = i \mid \mathbf{x}^q\right) = \frac{\exp\left(\mathbf{x}^q \cdot \mathbf{p}^i\right)}{\sum_{n=1}^{N} \exp\left(\mathbf{x}^q \cdot \mathbf{p}^n\right)} \tag{17}$$

The final objective function is formally written as:

$$\mathcal{L}_{CE} = -\log\left(z_y\right) \tag{18}$$

where y is relation label, and z_y is estimated probability for the relation y.

In short, we design a new node updating method to capture inter-class differences in few-shot setting.

3.6 Theoretical Analysis

In this section, we theoretically prove that the involvement of function words increases intra-class differences (Theorem1) and the designed message passing mechanism makes different class nodes become discriminative and same class nodes similar (Theorem2).

Given any two instances x_i and x_j, let the corresponding keywords representations be $\mathbf{x}_i^c = \{a_{i1}, a_{i2}, \ldots, a_{id}\} \in \mathbb{R}^d$ and $\mathbf{x}_j^c = \{b_{j1}, b_{j2}, \ldots, b_{jd}\} \in \mathbb{R}^d$, the function words representations be $\mathbf{x}_i^f = \{a_i^1, a_i^2, \ldots, a_i^d\} \in \mathbb{R}^d$ and $\mathbf{x}_j^f = \{b_j^1, b_j^2, \ldots, b_j^d\} \in \mathrm{R}^d$, the instance representations considering function words be $\mathbf{x}_i = \mathbf{x}_i^c + \mathbf{x}_i^f$ and $\mathbf{x}_j = \mathbf{x}_j^c + \mathbf{x}_j^f$.

Theorem 1. *If* $\mathrm{norm}(x_i^f \odot x_j^f) \leq \mathrm{norm}(x_i^c \odot x_j^c)$*, then*

$$\mathrm{norm}(x_i \odot x_j) \leq \mathrm{norm}(x_i^c \odot x_j^c), \tag{19}$$

where $\mathrm{norm}(x_i \odot x_j)$ *is computed as follows:*

$$\mathrm{norm}(x_i \odot x_j) = \frac{|x_i \odot x_j|}{\|x_i \odot x_j\|_2} \tag{20}$$

and $\odot$ *indicates the inner product of vectors.*

Theorem 2. *Given any two instances considering function words* $\mathbf{x}_i$ *and* $\mathbf{x}_j$*, define the similarity measure between* $\mathbf{x}_i$ *and* $\mathbf{x}_j$ *as*

$$D\left(\mathbf{x}_i, \mathbf{x}_j\right) = \mathbf{x}_i \odot \mathbf{x}_j \tag{21}$$

The message passings between same class instances and different classes are respectively defined as follows:

$$\mathbf{x}_i' = \mathbf{x}_i + \mathrm{norm}\left(\mathbf{x}_i \odot \mathbf{x}_j\right) \mathbf{x}_j \tag{22}$$

and

$$\mathbf{x}_i' = \mathbf{x}_i - \mathrm{norm}\left(\mathbf{x}_i \odot \mathbf{x}_j\right) \mathbf{x}_j \tag{23}$$

- *If* $\mathbf{x}_i$ *and* $\mathbf{x}_i$ *belong to the same class, then,*

$$D\left(\mathbf{x}_i, \mathbf{x}_j\right) \leq D\left(\mathbf{x}_i', \mathbf{x}_j'\right) \tag{24}$$

- *If* $\mathbf{x}_i$ *and* $\mathbf{x}_i$ *belong to different classes, then,*

$$D\left(\mathbf{x}_i, \mathbf{x}_j\right) \geq D\left(\mathbf{x}_i', \mathbf{x}_j'\right) \tag{25}$$

4 Experiments

4.1 Baselines

- To demonstrate the usefulness of local-level features, compare with global-level models: metric-based models **Proto** [Snell *et al.*, 2017], **Proto-HATT** [Gao *et al.*, 2019a], **MLMAN** [Ye and Ling, 2019], **BERT-PAIR** [Gao *et al.*, 2019b] and **TPN** [Wen *et al.*, 2021]. And gradient-based models **MAML** [Finn *et al.*, 2017] and **GNN** [Satorras and Estrach, 2018].
- To prove function words importance, compare with word-level models: **TD-Proto** [Sun *et al.*, 2019], using memory network to learn content words importance. **ConceptFERE** [Yang *et al.*, 2021], designing attention mechanism to measure class-specific words importance.
- Models introducing external information: **REGRAB** [Qu *et al.*, 2020] utilizes relation graph knowledge. **CTEG** [Wang *et al.*, 2020] uses dependency trees.
- Pretrained RC models: **MTB** [Soares *et al.*, 2019] pretrained with matching the blank task. **CP** [Peng *et al.*, 2020], a framework pretrained with entity masked.

4.2 Datasets and Settings

We evaluate our model on **FewRel1.0** [Han *et al.*, 2018] and **FewRel2.0** [Gao *et al.*, 2019b] in terms of the accuracy under multiple N-way-K-shot tasks. And we select N to be 5 and 10, K to be 1 and 5 to form 4 test scenarios according to [Gao *et al.*, 2019a]. In addition, we take base-uncased BERT as the encoder of 768 dimensions for a fair comparison. The input max length is set to 128. Besides, the AdamW optimizer is applied with the learning rate as 2×10^{-5} and weight decay as 1×10^{-2}. Furthermore, hyper-parameter λ is set to 0.6 and $\mathbf{u}_w$ is randomly initialized following [Sun *et al.*, 2019].

Encoder	Model	5-way-1-shot	5-way-5-shot	10-way-1-shot	10-way-5-shot
CNN	Proto-CNN [Snell *et al.*, 2017] ▷	72.65 / 74.52	86.15 / 88.40	60.13 / 62.38	76.20 / 80.45
	Proto-HATT [Gao *et al.*, 2019a]	75.01 / – –	87.09 / 90.12	62.48 / – –	77.50 / 83.05
	MLMAN [Ye and Ling, 2019] ◇	78.85 / 82.98	88.32 / 92.66	67.54 / 73.59	79.44 / 87.29
Bert	Proto-Bert [Snell *et al.*, 2017] ○	82.92 / 80.68	91.32 / 89.60	73.24 / 71.48	83.68 / 82.89
	MAML [Finn *et al.*, 2017] ○	82.93 / 89.70	86.21 / 93.55	73.20 / 83.17	76.06 / 88.51
	GNN [Satorras and Estrach, 2018]	74.21 / 75.66	86.16 / 89.06	67.98 / 70.08	73.65 / 76.93
	BERT-PAIR [Gao *et al.*, 2019b] ▷	85.66 / 88.32	89.48 / 93.22	76.84 / 80.63	81.76 / 87.02
	REGRAB [Qu *et al.*, 2020] ○	87.93 / 90.30	92.58 / 94.25	80.52 / 84.09	87.02 / 89.93
	TD-Proto [Sun *et al.*, 2019]	83.43 / 84.53	90.26 / 92.38	72.45 / 74.32	82.10 / 85.19
	ConceptFERE [Yang *et al.*, 2021]	87.21 / 89.21	90.53 / 93.98	73.56 / 75.72	83.29 / 86.21
	TPN [Wen *et al.*, 2021]	– – / 80.14	– – / 93.60	– – / 72.67	– – / 89.83
	CTEG [Wang *et al.*, 2020] ◇	84.72 / 88.11	92.52 / 95.25	76.01 / 81.29	84.89 / 91.33
	FAEA(ours)	**90.81 / 95.10**	**94.24 / 96.48**	**84.22 / 90.12**	**88.74 / 92.72**
	MTB [Soares *et al.*, 2019] ▷	– – / 93.86	– – / 97.06	– – /89.20	– – / 94.27
	CP [Peng *et al.*, 2020]	– – / 95.10	– – / 97.10	– – / 91.20	– – / 94.70
	FAEA(ours)+CP	**94.11 / 96.36**	**89.55 / 97.85**	**86.59 / 93.82**	**93.64 / 96.29**

Table 1: Accuracy (%) of FSRC task on FewRel1.0 validation / test set. ▷ are from FewRel public leaderboard, ○ are reported by [Qu *et al.*, 2020], and ◇ are reported by [Wang *et al.*, 2020] .

Model	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
Proto-CNN	35.09	49.37	22.98	35.22
Proto-BERT	40.12	51.50	26.45	36.93
Proto-ADV	42.21	58.71	28.91	44.35
Bert-Pair	67.41	78.57	54.89	66.85
Our	**73.58**	**90.10**	**62.98**	**80.51**

Table 2: Accuracy (%) of few shot classification on the FewRel2.0 domain adaptation test set.

Model	2-way-1-shot		4-way-1-shot		5-way-1-shot		5-way-3-shot		5-way-5-shot	
	R	I	R	I	R	I	R	I	R	I
Proto-HATT	83.26	53.62	78.61	49.72	75.01	62.13	80.53	68.15	87.09	73.02
Bert-Pair	91.21	56.20	87.44	54.87	85.66	67.53	88.42	69.92	89.48	71.21
TD-Proto	89.69	53.81	85.36	52.31	83.25	63.21	84.21	65.32	85.21	70.19
ConceptFERE	92.57	62.21	88.89	59.76	87.21	69.47	88.79	71.15	90.53	76.24
Our	**97.65**	**78.96**	**92.21**	**75.45**	**90.81**	**80.02**	**91.96**	**82.26**	**94.24**	**85.63**

Table 3: Accuracy (%) of different few-shot settings on FewRel1.0. 'R' stands for 'Random' and 'I' stands for 'Inverse'.

Model	Id	5-way 1-shot	10-way 1-shot
Our	1	**90.81**	**84.22**
– phrase-level encoding	2	84.98	77.02
– function word attn	3	87.52	80.92
– general attn	4	88.81	82.08
– constituent attn	5	88.93	82.69
– related attn	6	89.43	83.51
– message passing	7	90.01	83.62
– mean	8	90.32	83.86

Table 4: Ablation study on FewRel1.0 validation set showing accuracy (%).

4.3 Results

Performance on FSRC. As shown in Table 1, our method is significantly better than the strong baselines, especially under 1-shot settings. Specifically, our model improves 5-way-1-shot and 10-way-1-shot tasks by 4.80 points and 6.06 points in terms of accuracy, demonstrating superior ability. In addition, our method achieves good performance on FewRel2.0, as shown in Table 2.

- Proto and GNN, as widely-used baselines for few-shot learning, perform not well on FSRC. Unlike low-level patterns can be shared across tasks in computer vision, words that are informative for one task may not be relevant for other tasks. But these models ignore such local words importance variations in learning. But FAEA leverages phrase-level attention to attend local features.
- TD-PROTO and ConceptFERE also use semantic-level attention to explore content words, but neglect function words maintaining syntactic structure differences. Since FAEA captures function words to form fine-grained features, it obtains better performance.
- When computing relation prototypes, Proto-HATT and TPN utilize intra-class commonalities, not considering inter-class differences. FAEA captures and leverages differences to get more discriminative representations.

Performance on FSIRC. To further illustrate the model effectiveness for FSIRC, we evaluate models on FewRel1.0 validation set with different scenarios, shown in Table 3. Random means general evaluation setting, that randomly samples 10,000 FSRC tasks from validation set. Inverse represents each evaluated task including inverse relations. The baselines perform well under random scenarios but drop significantly under inverse settings, around 26.98 points in 1-shot scenarios, which illustrates that FSIRC is extremely challenging. FAEA achieves the best accuracy, especially under inverse settings, proving that it can effectively capture function words and handle FSIRC tasks.

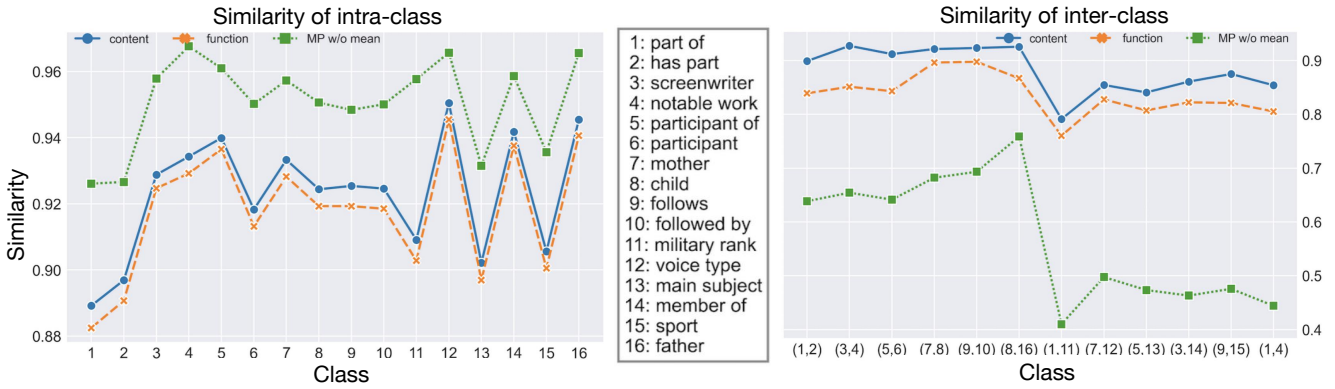

Figure 3: The similarity of intra-class and inter-class between some classes computed by dot-product.

5 Analysis

5.1 Analysis of Function Words Attention

This section discusses the effect of function-words attention. As shown in Table 4, removing phrase-level (Model 2) and function-words attention (Model 3) severely decreases the performance, indicating function words are also essential to represent relations. Furthermore, as shown in Figure 4, with the help of function-words attention, we highlight '*are*' and '*of*' to form '*are part of*', which appears in query and support instance of class '*part of*', then this support instance gets a higher score, and our model correctly classifies the query.

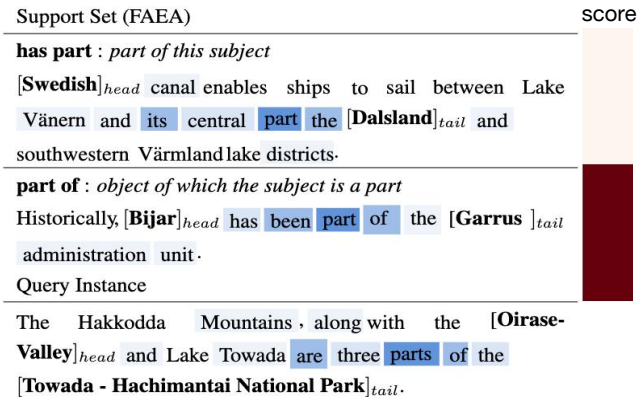

Figure 4: A 2-way-1-shot FSIRC task and visualizes attention scores of words by FAEA.

To demonstrate the effectiveness of three components of function-words attention, from model 4,5,6 of Table 4, we can see a performance decline if three components are removed separately. As shown in Figure 5, TD-Proto mainly attends content words such as '*parts*', '*Towada*' and '*Lake*'. FAEA without general attention enhances not only function words importance but also content words irrelated to keywords, such as '*mountains*' and '*Hakkodda*'. FAEA without constituent attention enhances some keywords-irrelated function words such as '*along*', '*and*'. FAEA without related attention tends to decrease some related function words importance far away from the keywords, such as '*are*'. FAEA further captures correct function words to form the phrase '*are three parts of*' , demonstrating that three components all contribute to enhancing function words importance.

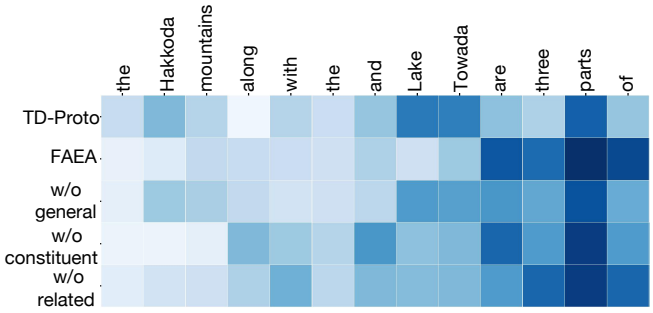

Figure 5: Attention scores of words under different models.

5.2 Analysis of Adaptive Message Passing

As shown in Table 4, we compare models without message passing (Model 7) and message passing without mean (Model 8). We observe that considering message passing achieves higher accuracy, and adding the mean to control the proportion of transferred message further improves the performance.

To further demonstrate the effectiveness of message passing, we choose some classes and visualize the similarity shown in Figure 3. We can see that only considering content words, the inter-class information of inverse relations has a high similarity score. And the introduction of function words effectively reduces it. But from the left part, function words reduce the similarity of intra-class information. The designed message passing without mean effectively increases intra-class commonalities and keeps inter-class differences. But from the right part, when the inter-class differences are large enough, with the message passing mechanism, the similarity of the inter-class sharp decline, and it will destroy the original relation semantic.

6 Conclusion

In this paper, we have presented FAEA, a framework that can effectively handle few-shot inverse relations by enhancing related function words importance. Experiments demonstrate that FAEA achieves new SOTA results on two NLP tasks on FewRel dataset. In future work, we will try to design a more effective and general function-words enhanced backbone network for various NLP tasks.

Acknowledgments

This work was supported by National Key R&D Program of China (2019YFB2102404), National Natural Science Foundation of China (NSFC) (61972455), National Natural Science Foundation of China (NSFC) (61976153), the Joint Project of AISHU.com and Bayescom. Xiaowang Zhang is supported by the program of Peiyang Young Scholars in Tianjin University (2019XRX-0032).

References

[Abiola and Andreas, 2019] Obamuyide Abiola and Vlachos Andreas. Model-agnostic meta-learning for relation classification with limited supervision. In *Proc. of ACL*, pages 5873–5879, 2019.

[Bao *et al.*, 2020] Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. Few-shot text classification with distributional signatures. In *Proc. of ICLR*, 2020.

[Bo *et al.*, 2021] Deyu Bo, Xiao Wang, Chuan Shi, and Huawei Shen. Beyond low-frequency information in graph convolutional networks. In *Proc. of AAAI*, pages 3950–3957, 2021.

[Devlin *et al.*, 2019] Jacob Devlin, Mingwei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*, pages 4171–4186, 2019.

[Dong *et al.*, 2020] Hang Dong, Wei Wang, Frans Coenen, and Kaizhu Huang. Knowledge base enrichment by relation learning from social tagging data. *Inf. Sci.*, 526(4):203–220, 2020.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. of ICML*, pages 1126–1135, 2017.

[Gao *et al.*, 2019a] Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *Proc. of AAAI*, pages 6407–6414, 2019.

[Gao *et al.*, 2019b] Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Fewrel 2.0: Towards more challenging few-shot relation classification. In *Proc. of EMNLP-IJCNLP*, pages 6249–6254, 2019.

[Han *et al.*, 2018] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proc. of EMNLP*, pages 4803–4809, 2018.

[Lee, 2021] Yohan Lee. Improving end-to-end task-oriented dialog system with a simple auxiliary task. In *Proc. of EMNLP*, pages 1296–1303, 2021.

[Peng *et al.*, 2020] Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Learning from context or names? An empirical study on neural relation extraction. In *Proc. of EMNLP*, pages 3661–3672, 2020.

[Qu *et al.*, 2020] Meng Qu, Tianyu Gao, Louis Pascal A. C. Xhonneux, and Jian Tang. Few-shot relation extraction via bayesian meta-learning on relation graphs. In *Proc. of ICML*, pages 7867–7876, 2020.

[Satorras and Estrach, 2018] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *Proc. of ICLR*, 2018.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Proc. of NeurIPS*, pages 4077–4087, 2017.

[Soares *et al.*, 2019] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proc. of ACL*, pages 2895–2905, 2019.

[Sun *et al.*, 2019] Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao Lv. Hierarchical attention prototypical networks for few-shot text classification. In *Proc. of EMNLP-IJCNLP*, pages 476–485, 2019.

[Wang *et al.*, 2020] Yuxia Wang, Karin Verspoor, and Timothy Baldwin. Learning from unlabelled data for clinical semantic textual similarity. In *Proc. of ClinicalNLP*, pages 227–233, 2020.

[Wen *et al.*, 2021] Wen Wen, Yongbin Liu, Chunping Ouyang, Qiang Lin, and Tonglee Chung. Enhanced prototypical network for few-shot relation extraction. *Inf. Process. Manag.*, 58(4):102596, 2021.

[Wu *et al.*, 2021] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Universal-prototype enhancing for few-shot object detection. In *Proc. of ICCV*, pages 9567–9576, 2021.

[Yang *et al.*, 2021] Shan Yang, Yongfei Zhang, Guanglin Niu, Qinghua Zhao, and Shiliang Pu. Entity concept-enhanced few-shot relation extraction. In *Proc. of ACL*, pages 987–991, 2021.

[Yang *et al.*, 2022] Fengyuan Yang, Ruiping Wang, and Xilin Chen. Sega: Semantic guided attention on visual prototype for few-shot learning. In *Proc. of WACV*, pages 1586–1596, 2022.

[Ye and Ling, 2019] Zhixiu Ye and Zhenhua Ling. Multi-level matching and aggregation network for few-shot relation classification. In *Proc. of ACL*, pages 2872–2881, 2019.

[Zhang *et al.*, 2020] Ji Zhang, Chengyao Chen, Pengfei Liu, Chao He, and Cane Wing-Ki Leung. Target-guided structured attention network for target-dependent sentiment analysis. *Trans. Assoc. Comput. Linguistics*, 8(3):172–182, 2020.