

LREC-COLING 2024

A Streamlined Span-based Factorization Method for Few Shot Named Entity Recognition

Author: WenJie Xu JianQuan OuYang

Institution: Xiangtan University

Contact: xuwenjie444@gmail.com ojq@xtu.edu.cn

博学笃行 感德日新



湘潭大学

XIANGTAN UNIVERSITY

湘潭大學
XIANGTAN UNIVERSITY

目录

CONTENTS

1

Research Background

2

Model Introduction

3

Experimental Results

4

Conclusion

Part I



Research Background



Research Background



Named Entity Recognition (NER) is an important task in the field of Natural Language Processing. It involves automatically identifying entities with specific meanings mentioned in the text and categorizing these entities into predefined categories, such as names of **people, places, organizations**, etc.

Sequence labeling-based method: This approach uses the BIO/BLOU strategy to convert the task of entity recognition into a token label prediction task.

Drawback: It does not perform well with nested entities.

Span-based method: This approach involves recognizing entities across all spans. For example, by designing **enumeration and span models**, etc.

Drawback: The recognition of entity span boundaries is not accurate and the computational complexity is high.

Under the few-shot scenario, the performance is poor, and acquiring a large amount of high-quality annotated data is very time-consuming and costly, especially in certain specialized fields, such as medicine and law, where professional knowledge is required for accurate annotation.

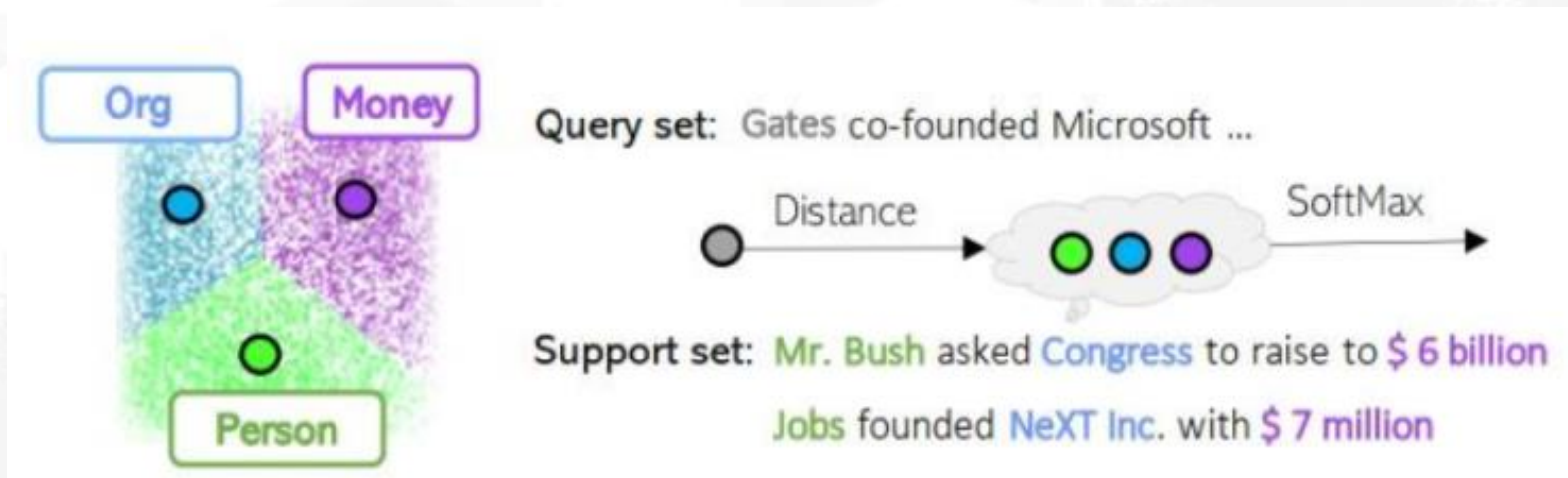
Research Background



Few-shot Named Entity Recognition aims to recognize and classify named entities in a given text when only a small number of annotated samples are available.

In the context of few-shot learning, the model needs to be able to quickly adapt to new entity categories, even if there are only a few annotated instances for these categories. This poses a challenge to the generalization ability of the model as it must be able to learn from limited information and make accurate predictions.

The core idea of prototypical networks based on meta-learning is to find a **vector representation for each category**, the "prototype" and then classify by comparing the **similarity** between the sample to be predicted and these prototypes.



Part II

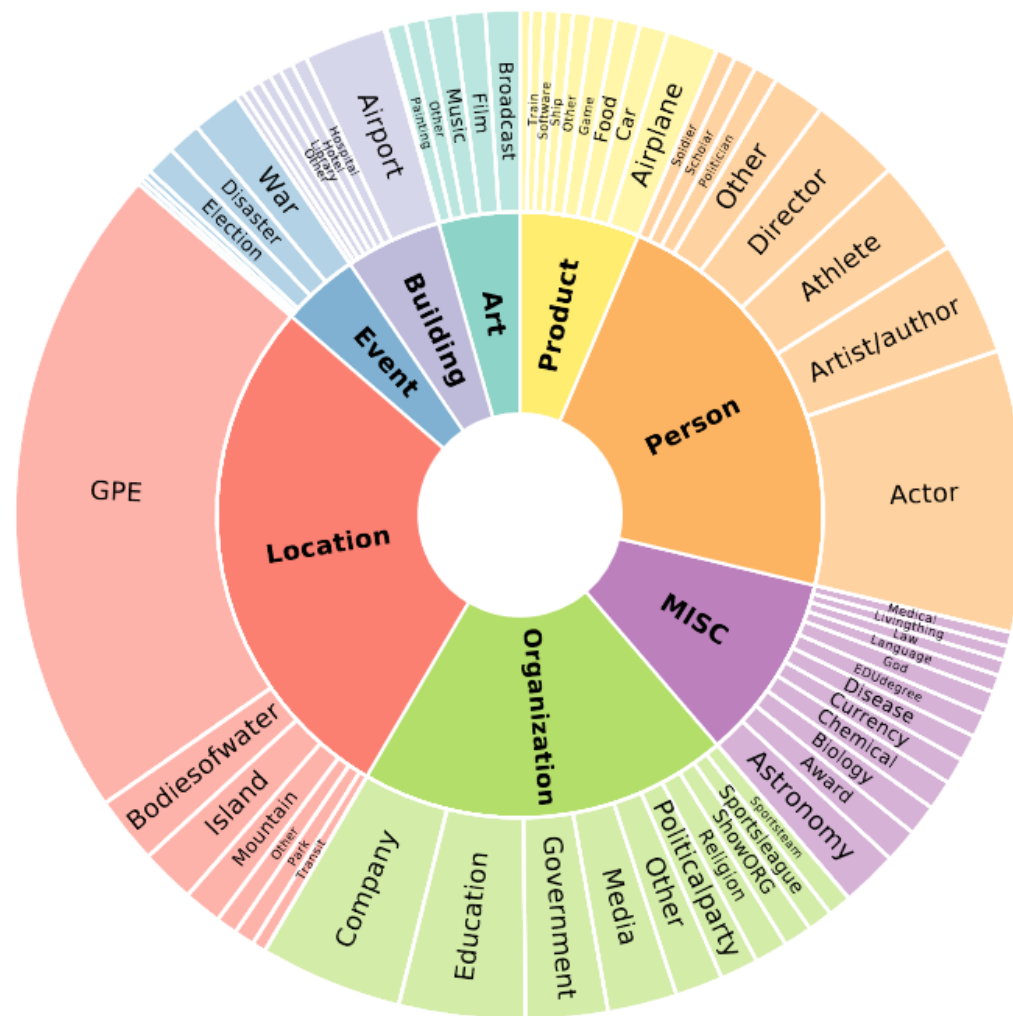
Model Introduction



Few-NERD数据集: A large-scale, manually annotated dataset for few-shot NER tasks. This dataset includes **8 coarse-grained** and **66 fine-grained** entity types, each entity label has a hierarchical structure of coarse-grained + fine-grained, and contains **188,238** sentences from Wikipedia and **4,601,160** words. It is divided into four settings:

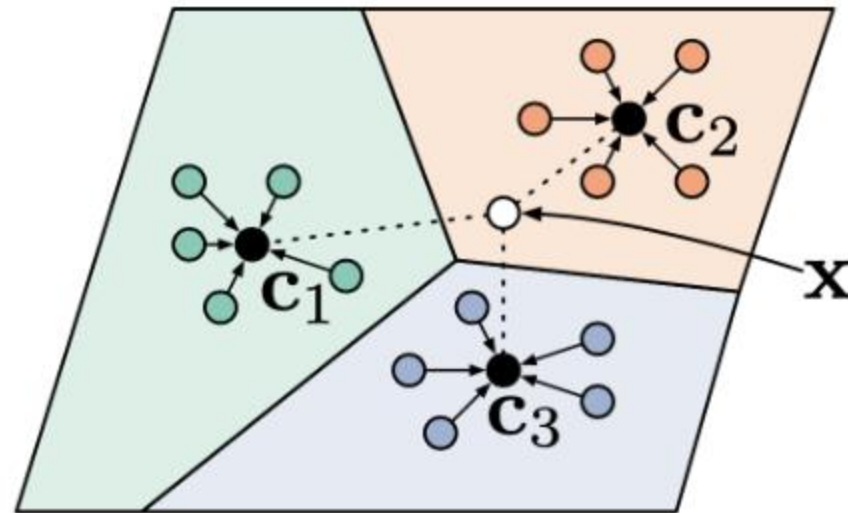
- 5-way 1~2 shot, 5-way 5~10 shot
- 10-way 1~2 shot, 10-way 5~10 shot

Target Types \mathcal{Y}	[person-actor], [art-film]
Support set \mathcal{S}	(1) <i>Brad Pitt</i> _[person-actor] is an accomplished and talented film actor. (2) <i>Titanic</i> _[art-film] is a classic and beloved romantic drama film.
Query Set \mathcal{Q}	Tom Cruise starred in Top Gun, a classic '80s action movie.
Expected output	<i>Tom Cruise</i> _[person-actor] starred in <i>Top Gun</i> _[art-film] , a classic '80s action movie.

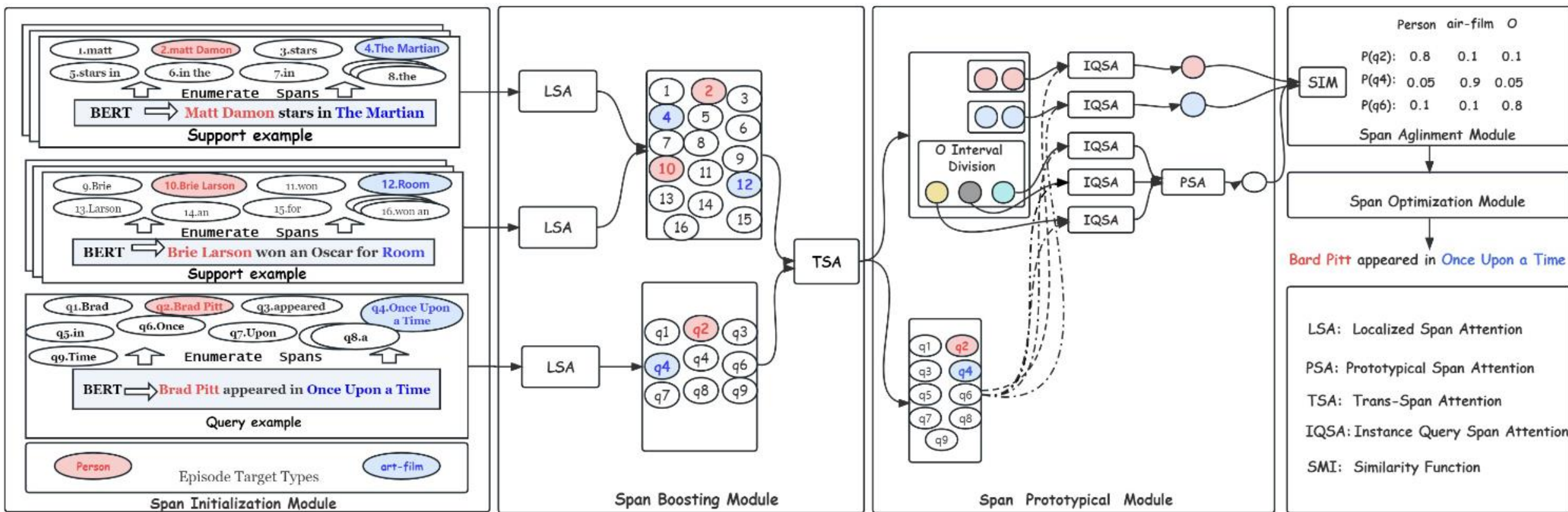


Prototypical Network

The central idea of a **prototypical network** is to learn a center point or "prototype" for each category in the feature space. These prototypes represent the feature vectors of their respective categories. When classifying new samples, **the prototypical network calculates the distance between the sample features and the prototypes of each category**, usually using Euclidean distance or cosine similarity, and then classifies the sample into the category represented by the nearest prototype.



Core: How to obtain better prototype representation vectors?



Innovations:

1. In response to the issue that the prototype network representation is not detailed enough, **Localized Span Attention** and **Trans-Span Attention** are proposed to enhance prototype representation.
2. To address the issue of potential conflicts in spans during the inference stage, a Beam search combined with SoftNMS algorithm is proposed to resolve conflicts.

Part III

Experiment results



Experiment results



Models	Intra					Inter				
	1 ~ 2 shot		5 ~ 10 shot		Avg.	1 ~ 2 shot		5 ~ 10 shot		Avg.
	5 way	10 way	5 way	10 way		5 way	10 way	5 way	10 way	
ProtoBERT	20.76 \pm 0.84	15.04 \pm 0.44	42.54 \pm 0.94	35.40 \pm 0.13	28.44	38.83 \pm 1.49	32.45 \pm 0.79	58.79 \pm 0.44	52.92 \pm 0.37	45.75
NNShot	25.78 \pm 0.91	18.27 \pm 0.41	36.18 \pm 0.79	27.38 \pm 0.53	26.90	47.24 \pm 1.00	38.87 \pm 0.21	55.64 \pm 0.63	49.57 \pm 2.73	47.83
StructShot	30.21 \pm 0.90	21.03 \pm 1.13	38.00 \pm 1.29	26.42 \pm 0.60	28.92	51.88 \pm 0.69	43.34 \pm 0.10	57.32 \pm 0.63	49.57 \pm 3.08	50.53
CONTaiNER	40.40	33.82	53.71	47.51	43.86	56.1	48.36	61.90	57.13	55.87
ESD	36.08 \pm 1.6	30.00 \pm 0.70	52.14 \pm 1.5	42.15 \pm 2.6	40.09	59.29 \pm 1.25	52.16 \pm 0.79	69.06 \pm 0.80	64.00 \pm 0.43	61.13
DecomposedMeta	49.48 \pm 0.85	42.84 \pm 0.46	62.92 \pm 0.57	57.31 \pm 0.25	53.14	64.75 \pm 0.35	58.65 \pm 0.43	71.49 \pm 0.47	68.11 \pm 0.05	65.75
SpanProto	54.49 \pm 0.39	45.39 \pm 0.72	65.89 \pm 0.82	59.37 \pm 0.47	56.29	73.36 \pm 0.18	66.26 \pm 0.33	75.19 \pm 0.77	70.39 \pm 0.63	71.30
MSDP	56.35\pm0.28	47.13\pm0.69	66.80\pm0.78	64.69\pm0.51	58.74	76.86\pm0.22	69.78\pm0.31	84.78\pm0.69	81.50\pm0.71	78.23
MeTNet	55.79 \pm 0.23	47.18 \pm 0.89	65.41 \pm 0.35	60.71 \pm 0.17	57.27	74.42 \pm 0.61	67.91 \pm 0.68	76.28 \pm 0.32	71.96 \pm 0.35	72.64
PromptNER	55.32 \pm 1.03	50.29 \pm 0.61	67.26 \pm 1.02	60.42 \pm 0.73	58.32	64.92 \pm 0.71	62.28 \pm 0.39	72.64 \pm 0.16	70.13 \pm 0.67	67.49
SSF (Ours)	60.80\pm0.75	50.31\pm0.45	74.09\pm0.55	61.69\pm0.55	61.72	83.21\pm0.80	75.87\pm0.50	91.24\pm0.30	85.95\pm0.50	84.06

1. Compared to the current state-of-the-art model MSDP, it achieves an average improvement of **3%** in the intra-scenario and **6%** in the inter-scenario.
2. Compared to other models, it achieves an average improvement of **10%**.

Experiment results



	Models	We	Mu	PI	Bo	Se	Re	Cr	Avg.
1-SHOT	TransferBERT	55.82 \pm 2.75	38.01 \pm 1.74	45.65 \pm 2.02	31.63 \pm 5.32	21.96 \pm 3.98	41.79 \pm 3.81	38.53 \pm 7.42	39.06 \pm 3.86
	MN+BERT	21.74 \pm 4.60	10.68 \pm 1.07	39.71 \pm 1.81	58.15 \pm 0.68	24.21 \pm 1.20	32.88 \pm 0.64	69.66 \pm 1.68	36.72 \pm 1.67
	ProtoBERT	46.72 \pm 1.03	40.07 \pm 0.48	50.78 \pm 2.09	68.73 \pm 1.87	60.81 \pm 1.70	55.58 \pm 3.56	67.67 \pm 1.16	55.77 \pm 1.70
	Ma2021	-	-	-	-	-	-	-	69.3 _(unk)
	L-TapNet+CDT	71.53 \pm 4.04	60.56 \pm 0.77	66.27 \pm 2.71	84.54 \pm 1.08	76.27 \pm 1.72	70.79 \pm 1.60	62.89 \pm 1.88	70.41 \pm 1.97
	ESD	78.25 \pm 1.50	54.74 \pm 1.02	71.15 \pm 1.55	71.45 \pm 1.38	67.85 \pm 0.75	71.52 \pm 0.98	78.14 \pm 1.46	70.44 \pm 0.47
	SSF (Ours)	85.59 \pm 1.50	69.25 \pm 0.75	83.48 \pm 0.65	74.48 \pm 2.29	84.40 \pm 0.45	79.44 \pm 0.66	94.64 \pm 1.31	81.64 \pm 0.47
5-SHOT	TransferBERT	59.41 \pm 0.30	42.00 \pm 2.83	46.07 \pm 4.32	20.74 \pm 3.36	28.20 \pm 0.29	67.75 \pm 1.28	58.61 \pm 3.67	46.11 \pm 2.29
	MN+BERT	36.67 \pm 3.64	33.67 \pm 6.12	52.60 \pm 2.84	69.09 \pm 2.36	38.42 \pm 4.06	33.28 \pm 2.99	72.10 \pm 1.48	47.98 \pm 3.36
	ProtoBERT	67.82 \pm 4.11	55.99 \pm 2.24	46.02 \pm 3.19	72.17 \pm 1.75	73.59 \pm 1.60	60.18 \pm 6.96	66.89 \pm 2.88	63.24 \pm 3.25
	Retriever	82.95 _(unk)	61.74 _(unk)	71.75 _(unk)	81.65 _(unk)	73.10 _(unk)	79.54 _(unk)	51.35 _(unk)	71.72 _(unk)
	ConVEx	71.5 _(unk)	77.6 _(unk)	79.0 _(unk)	84.5 _(unk)	84.0 _(unk)	73.8 _(unk)	67.4 _(unk)	76.8 _(unk)
	Ma2021	89.39 _(unk)	75.11 _(unk)	77.18 _(unk)	84.16 _(unk)	73.53 _(unk)	82.29 _(unk)	72.51 _(unk)	79.17 _(unk)
	L-TapNet+CDT	71.64 \pm 3.62	67.16 \pm 2.97	75.88 \pm 1.51	84.38 \pm 2.81	82.58 \pm 2.12	70.05 \pm 1.61	73.41 \pm 2.61	75.01 \pm 2.46
	ESD	84.50 \pm 1.06	66.61 \pm 2.00	79.69 \pm 1.35	82.57 \pm 1.37	82.22 \pm 0.81	80.44 \pm 0.80	81.13 \pm 1.84	79.59 \pm 0.39
SSF (Ours)	91.05 \pm 0.70	77.90 \pm 0.65	89.52 \pm 1.50	94.87 \pm 0.57	95.13 \pm 0.20	87.99 \pm 0.48	96.54 \pm 0.30	90.35 \pm 0.39	

Compared to other models, it shows an average increase of **10%**

Ablation study

Ablation Models	F1
SSF	91.24 \pm 0.30
Ours w/o Localized Span Attention	83.10 \pm 0.4
Ours w/o Trans-Span Attention	80.6 \pm 1.5
Ours w/o Instance Query Span Attention	84.2 \pm 0.6
Ours w/o O-type Division	81.7 \pm 1.3
Ours w/o ASBNMS	85.3 \pm 1.4

1. Trans-Span Attention is important than Localized Span Attention.

2. It is important to segment prototypes with no entities.

Part IV ▶

Conclusion



This paper proposes a span-based decomposition method specifically designed to solve the few-shot named entity recognition task. The method employs a span-based prototypical network and breaks down the task into four modules: **Span Boosting, Span Prototype, Span Alignment, and Span Optimization**, in order to improve the model's generalization ability and accuracy in few-shot scenarios. **Experimental results indicate that this method outperforms the current state-of-the-art MSDP model on general datasets.**

Thanks

