LREC-COLING 2024

Majority Rules Guided Aspect-Category based Sentiment Analysis via Label Prior Knowledge

Authors: Lin Li[1], Shaopeng Tang[1], Renwei Wu[1]

Wuhan University of Technology[1]

武漢理工大學

WUHAN UNIVERSITY OF TECHNOLOGY

Aspect-Category based Sentiment Analysis

- In some review text, it is important to know the customer's sentiment polarity towards specific aspect and category, which will help the merchant to further improve.

	L#P	L#G	L#D
It 's truly a great lap-top for the price.	0(neutral)	1(positive)	-
This laptop is a great price and has a sleek look.	1(positive)	-	1(positive)

Table 1: Two review examples (from LAP16). L#P is Laptop#Price, L#G is Laptop#General and L#D is Laptop#Design_features.

- Aspect-Category based Sentiment Analysis (ACSA) aims to distinguish aspect categories in text modality while simultaneously predicting the sentiment polarity associated with each of these categories.

In this research field, prompt learning for ACSA has achieving promising results. However, existing methods have not paid enough attention to annotation difference which commonly appear in sentiment analysis due to the ability differences of annotators in understandings of the labeling standards.

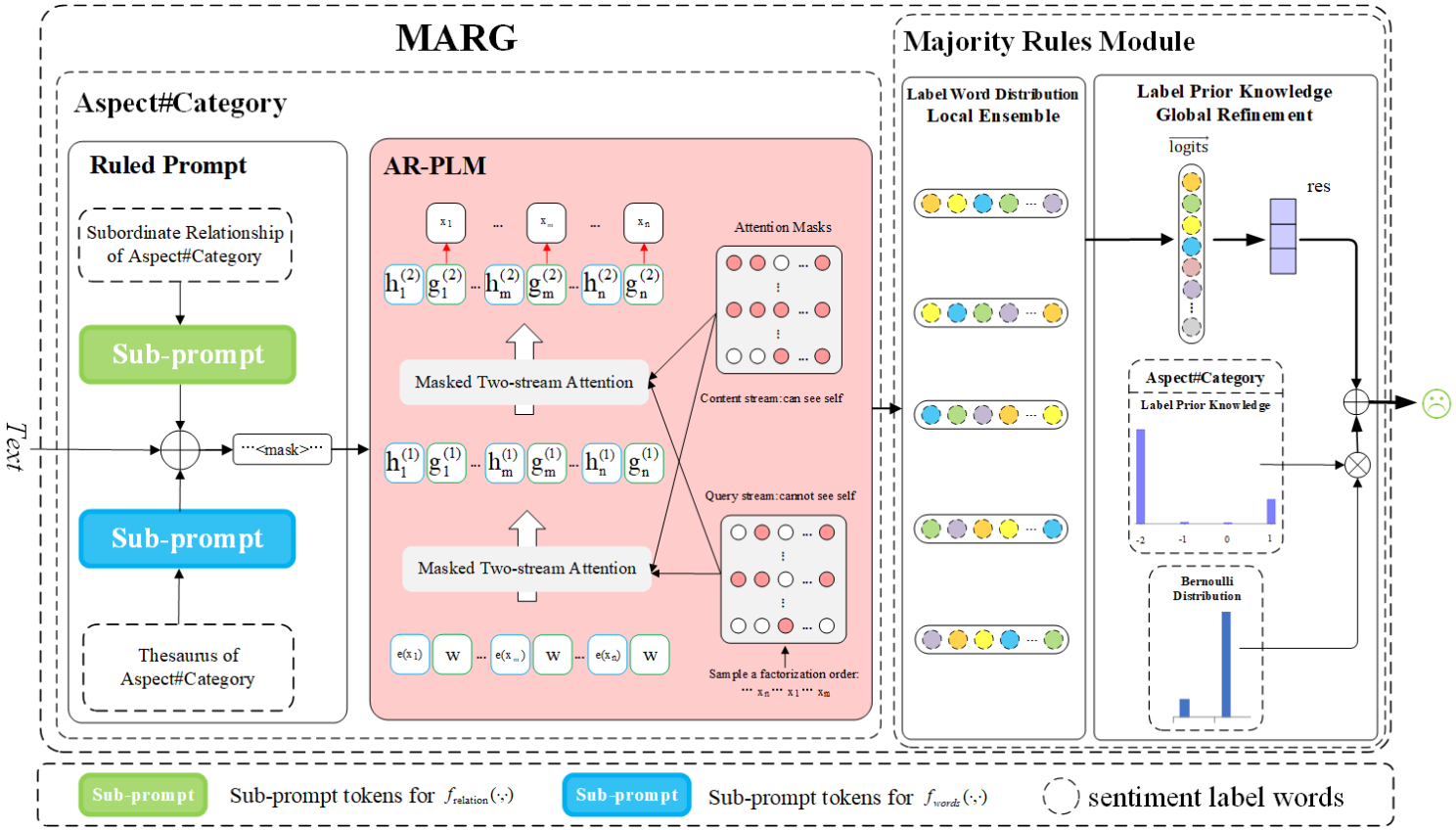
	L#P	L#G	L#D
It 's truly a great lap-top for the price.	0(neutral)	1(positive)	-
This laptop is a great price and has a sleek look.	1(positive)	-	1(positive)

Table 1: Two review examples (from LAP16). L#P is Laptop#Price, L#G is Laptop#General and L#D is Laptop#Design_features.

Take the two reviews in Table 1 as a case study. Although the descriptions of two reviews are basically the same on Aspect#Category: Laptop#Price, the sentiment polarity annotation is different. The highly semantically similar text has polysemous sentiments to different people, which leads to annotation difference.

Our Focus

To mitigate the interference of the annotation difference on the performance of ACSA, we design a majority rules module which contains two parts: 1) The local majority rule is the ensemble of label word distributions. 2) The global majority rule is the refinement based on the label prior knowledge of aspect categories.



Methodology

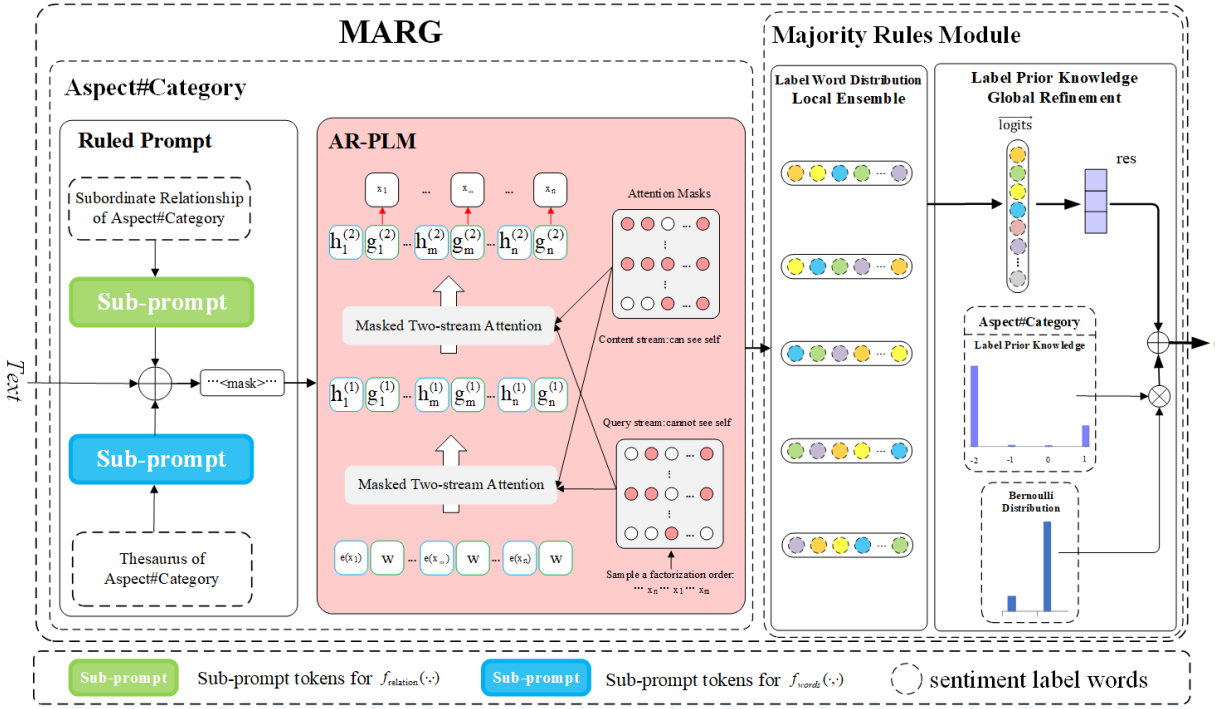
武汉理工大学
Wuhan University of Technology

In our paper, we propose a majority rules guided framework (MARG) for ACSA.

(1) Ruled Prompt: we design Ruled Prompt by building several sub-prompts which contain subordinate relationships and thesauri of aspect categories to generate prompts.

(2) AR-PLM: We train an AutoRegressive Pre-trained Language Model, to generate sentiment label word distributions and obtain token-wise semantic consistency.

(3) Majority Rules Module: we mitigate the impact to the performance of the model caused by the difference through two majority rules.



Ruled Prompt: Manually designed prompt might contain bias, and lack of coverage capability. To address this problem, Han et al. proposed a rule-based method to generate prompts. Each sub-prompt is generated according to the generation rules of different tasks. Inspired by this, we design our Ruled Prompt for ACSA.

$$f_{text}(\cdot) \wedge f_{relation}(\cdot,\cdot) \wedge f_{words}(\cdot,\cdot)$$

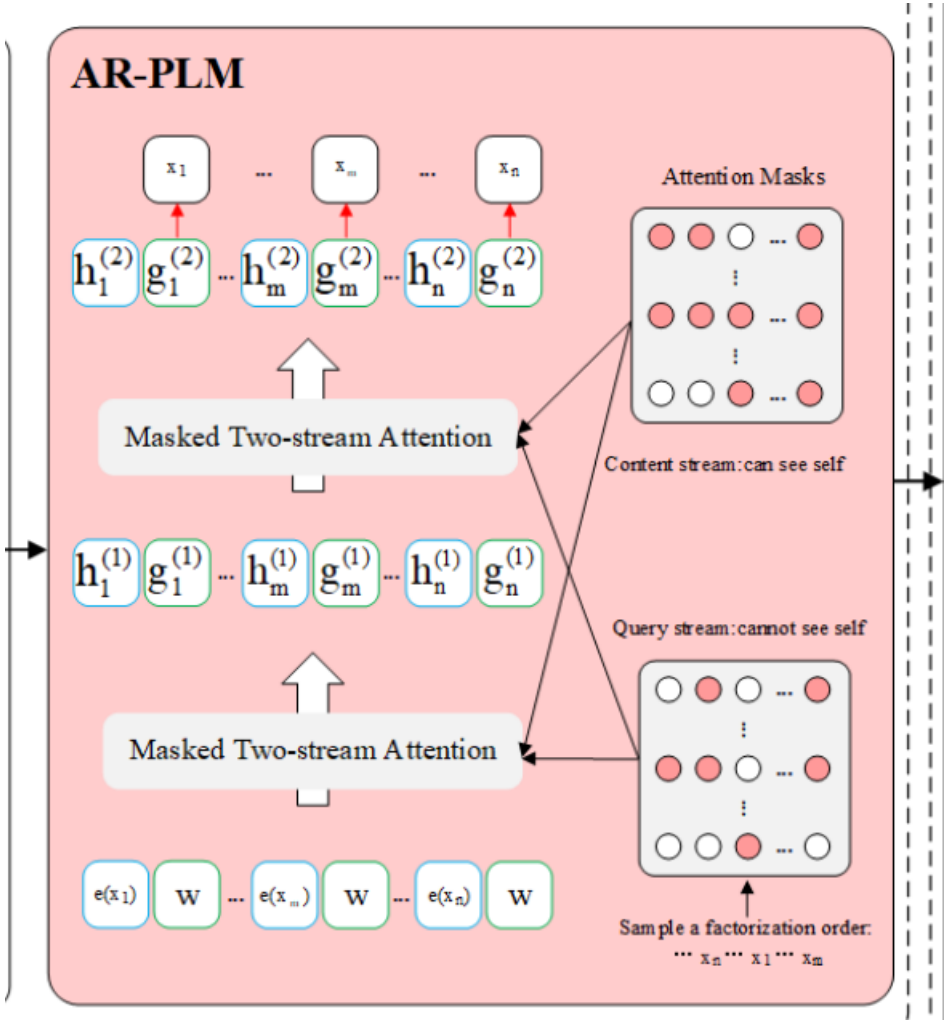
$f_{text}(\cdot)$ is a unary function to determine the sub-prompt generated from the original input text. $f_{relation}(\cdot,\cdot)$ is a binary function to determine the sub-prompt generated from aspect and category to indicate subordinate relationship of the both. $f_{words}(\cdot,\cdot)$ is a binary function to determine the sub-prompt generated from thesaurus of aspect and category containing the prior knowledge.

The generated prompts of Food#Quality are shown in Table 2. Generally, each generated prompt consists of three parts, i.e., $f_{text}(\cdot)$ (black font), $f_{relation}(\cdot,\cdot)$ (green font) and $f_{words}(\cdot,\cdot)$ (blue font), which respectively indicate the original semantic information, the subordinate relationship of aspect and category, and the prior knowledge of specific domain.

Input	The generated prompts
Text	**Prompt 1:** Text. The quality of the food is <mask>.
	Prompt 2: Text. So the quality of food is <mask> that refers to the taste, the freshness, the texture, the consistency, the temperature, the preparation, the authenticity, the cooking or general quality.
	Prompt 3: Text. Thus, considering the taste, freshness, texture, consistency, temperature, preparation, authenticity, cooking or general quality, I feel the quality of the food is <mask>.
	Prompt 4: The quality of food is <mask> that refers to the taste, the freshness, the texture, the consistency, the temperature, the preparation, the authenticity, the cooking or general quality. Text.
	Prompt 5: Considering the taste, freshness, texture, consistency, temperature, preparation, authenticity, cooking or general quality, I feel the quality of the food is <mask>. Text.

Table 2: The generated prompts by Ruled Prompt for Food#Quality

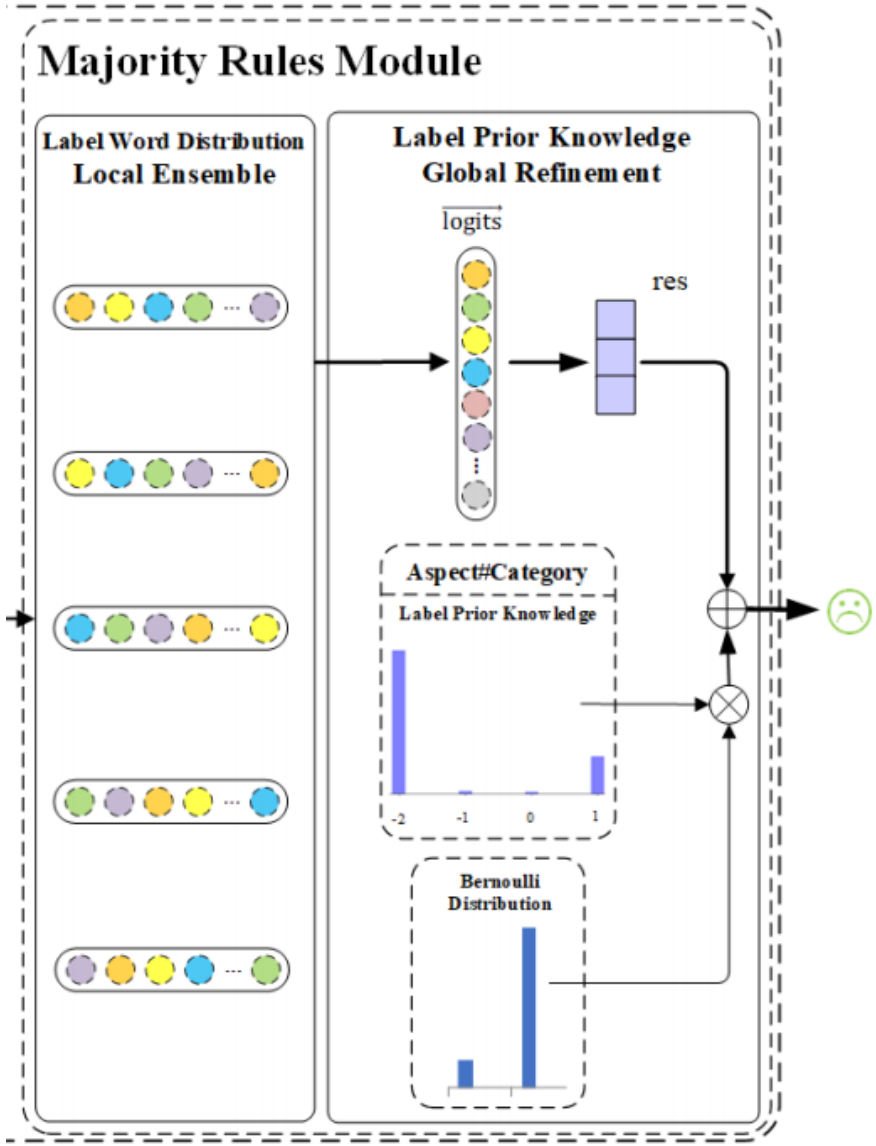
AR-PLM: we propose the Autoregressive Prompting to stimulate the potential of the pretrained language model and generate labels with token-wise semantic consistency. Pre-Trained XLNet is employed as the pre-trained language model for our prompt-tuning which combines the advantages of AutoRegressive LM and AutoEncoder LM.



Majority Rules Module: To mitigate the interference of the annotation difference on the performance of ACSA, we design a majority rules module which contains two parts:

1) The local majority rule is the ensemble of label word distributions.

2) The global majority rule is the refinement based on the label prior knowledge of aspect categories.



Datasets

Our experiments are performed on public SemEval 2015 and SemEval 2016.

Each of them contains both of restaurant and laptop domains.

Baselines

1. AddOneDim-LSTM (EMNLP, 2018)
2. CapsNet-BERT (EMNLP, 2019)
3. GIN-BERT (NLPCC, 2020)
4. MIMLLN-BERT (EMNLP, 2020)
5. Hier-GCN-BERT (COLING, 2020)
6. AAGCN-BERT (EMNLP, 2021)
7. Prompt_ACSA (ACM Comput. Surv., 2023)

Dataset	A#C	Train	Test
REST15	30	1102	572
LAP15	198	1397	644
REST15	30	1680	580
LAP16	198	2037	572

Overall Performance

As shown in Table 3, We divide baselines into two main types: non-LLM and LLM-based. Obviously, The non-LLM solutions are generally less effective than the LLM-based ones. Compared to AddOneDim-LSTM from scratch, MARG improves F1-score by 67.68% on REST16. Prompt_ACSA is based on pre-trained model and prompt-tuning, and MARG improves F1-score by 12.72% on LAP15 compared to it. Among other five baselines in LLM-based solutions, AAGCN-BERT achieved good performances, and our MARG further exceeds it with 4.07% improvement F1-score on REST15.

Solution Type	Model	REST15		LAP15		REST16		LAP16	
		ACC.	F1	ACC.	F1	ACC.	F1	ACC.	F1
non-LLM	AddOneDim-LSTM (2018, EMNLP)	-	37.32	-	-	-	50.50	-	-
	CapsNet (2019, EMNLP)	78.14	61.56	74.71	61.75	83.79	61.36	76.31	61.07
	GIN (2020, NLPCC)	81.17	62.38	75.93	63.18	87.05	65.03	78.92	62.93
	MIMLLN (2020, EMNLP)	78.27	60.59	75.30	61.39	85.76	63.52	78.57	62.63
	AAGCN (2021, EMNLP	82.79	67.43	80.02	65.87	88.32	72.55	81.76	65.96
LLM-based	CapsNet-BERT (2019, EMNLP)	81.89	61.85	82.19	59.75	86.50	62.12	80.53	61.03
	GIN-BERT (2020, NLPCC)	83.96	66.03	82.97	65.29	89.47	74.87	82.76	63.77
	MIMLLN-BERT (2020, EMNLP)	82.76	65.10	82.98	62.36	88.12	73.05	82.57	63.26
	Hier-GCN-BERT (2020, COLING)	-	64.23	-	62.13	-	74.55	-	54.15
	AAGCN-BERT (2021, EMNLP)	87.92	71.75	85.82	72.39	92.83	80.77	85.24	69.68
	Prompt_ACSA (2023b, ACM Comput. Surv.)	85.87	69.72	83.43	65.78	90.59	77.86	85.38	66.75
LLM-based	MARG (ours)	**88.94**	**74.67**	**87.05**	**74.15**	**95.34**	**84.68**	**87.89**	**71.65**
	w/o Global Refinement	87.34	73.98	85.90	72.67	93.25	82.12	87.33	71.40
	w/o Global Refinement & Local Ensemble	80.56	60.18	81.20	59.24	84.53	60.95	78.55	58.04

Table 3: Experimental results on SemEval 2015 and SemEval 2016 (%). The score marked as **bold** means the best performance among all models. The score marked with an underline means the best one among the baselines.

Ablation Study

To better analyze the effect of Refinement and Ensemble, we conduct two ablation experiments. The result indicates that the difference will clearly affect the performance of the model. MARG outperforms all in terms of Accuracy and F1-score, indicating that Majority Rules Module can mitigate the interference through two majority rules.

Solution Type	Model	REST15		LAP15		REST16		LAP16	
		ACC.	F1	ACC.	F1	ACC.	F1	ACC.	F1
non-LLM	AddOneDim-LSTM (2018, EMNLP)	-	37.32	-	-	-	50.50	-	-
	CapsNet (2019, EMNLP)	78.14	61.56	74.71	61.75	83.79	61.36	76.31	61.07
	GIN (2020, NLPCC)	81.17	62.38	75.93	63.18	87.05	65.03	78.92	62.93
	MIMLLN (2020, EMNLP)	78.27	60.59	75.30	61.39	85.76	63.52	78.57	62.63
	AAGCN (2021, EMNLP	82.79	67.43	80.02	65.87	88.32	72.55	81.76	65.96
LLM-based	CapsNet-BERT (2019, EMNLP)	81.89	61.85	82.19	59.75	86.50	62.12	80.53	61.03
	GIN-BERT (2020, NLPCC)	83.96	66.03	82.97	65.29	89.47	74.87	82.76	63.77
	MIMLLN-BERT (2020, EMNLP)	82.76	65.10	82.98	62.36	88.12	73.05	82.57	63.26
	Hier-GCN-BERT (2020, COLING)	-	64.23	-	62.13	-	74.55	-	54.15
	AAGCN-BERT (2021, EMNLP)	87.92	71.75	85.82	72.39	92.83	80.77	85.24	69.68
	Prompt_ACSA (2023b, ACM Comput. Surv.)	85.87	69.72	83.43	65.78	90.59	77.86	85.38	66.75
LLM-based	MARG (ours)	**88.94**	**74.67**	**87.05**	**74.15**	**95.34**	**84.68**	**87.89**	**71.65**
	w/o Global Refinement	87.34	73.98	85.90	72.67	93.25	82.12	87.33	71.40
	w/o Global Refinement & Local Ensemble	80.56	60.18	81.20	59.24	84.53	60.95	78.55	58.04

Table 3: Experimental results on SemEval 2015 and SemEval 2016 (%). The score marked as **bold** means the best performance among all models. The score marked with an underline means the best one among the baselines.

Additional Analysis

1. Considering the datasets are not large, we conduct a double-independent sample T-test on the corresponding prediction results (Acc. and F1-score) for MARG and some representative baselines. As shown in Table 4, the p values of MARG and CapsNet-BERT, MIMLLN-BERT, and Prompt_ACSA are 0.017, 0.028, and 0.041, respectively, which are less than 0.05. This shows that the effect of MARG has some statistical significance and reliability.

2. To further validate the power of Majority Rules Module on the large-scale dataset, we conduct extensive experiments on the Challenger 2018 (120K, Chinese) . As shown in Table 5, the addition of the module can bring about an increase in results, which indicates it can improve the model on large-scale datasets as well.

	MARG	
	t statistic	p value
CapsNet-BERT	-2.72	0.017
MIMLLN-BERT	-2.44	0.028
Prompt_ACSA	-2.25	0.041

Table 4: Significance test of MARG experimental results

Model	Precision	Recall	F1
MARG	**72.74**	70.15	**71.00**
w/o Refinement	72.58	**70.24**	70.77
w/o Refinement & Ensemble	69.06	66.15	68.32

Table 5: The results of extensive experiments on the Challenger 2018 (%)

Conclusions & Future Work

Conclusions

- Our work studies how to mitigate the impact of the annotation difference both locally and globally, and proposes a majority rules guided framework called MARG.
- Experiments show that it outperforms the state-of-the-art models and still works on the large-scale dataset, demonstrating its efficacy.

Future Work

- MARG is effective on the large-scale dataset in Chinese, and yet its effectiveness on the large dataset in English and in a real-world system is up for exploration.
- There is still a manual involvement in Ruled Prompt. PPT (Pre-trained Prompt Tuning) based on continuous prompts Gu et al. recently proposed is a promising approach.


תודה
Dankie
Gracias
Спасибо
شكراً
Merci
Takk
Köszönjük
Terima kasih
Grazie
Dziękujemy
Dėkojame
Ďakujeme
Vielen Dank
Paldies
Täname teid
Kiitos
谢谢
Thank You
Tak
感謝您
Obrigado
Teşekkür Ederiz
감사합니다
Σας Ευχαριστούμ
ขอบคุณ
Bedankt
Děkujeme vám
ありがとうございます
Tack