Multi-Granularity Fusion Text Semantic Matching Based on WoBERT

Number 779: Hongchun Yu Wei Pan Xing Fan Hanqi Li

1.China West Normal University

2.Nanchong Key Laboratory of IoT Awareness and Distributed Computing

COTENTS

Introduce

Text semantic matching is a crucial task in the field of natural language processing, involving the determination of semantic relationships between two or more texts.

ID	Sentence1	Sentence2	Matching
1	怎样学习英语才能又快又好？ How to be fast and efficient when learning English?	英语怎么学习才可以很快进步 How to improve quickly on learning English?	True
2	我喜欢上海的小吃 I like Shanghai snacks.	我不喜欢上海的小吃 I don’t like Shanghai snacks	False

Introduce

Typically, text semantic matching task is applied in various applications :

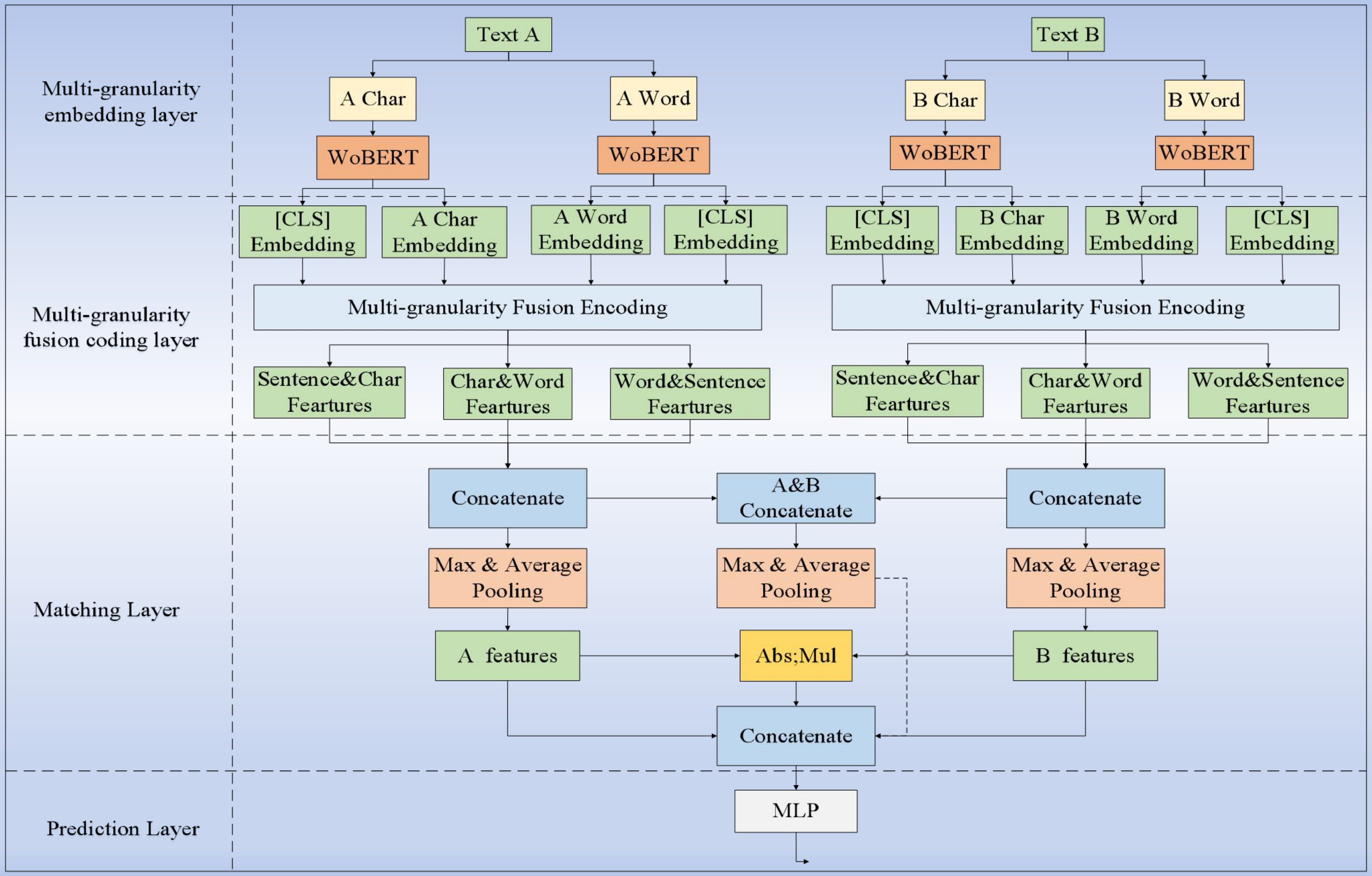
- text similarity computation
- question-answering systems
- information retrieval
- dialogue systems

Introduce

Difficulties

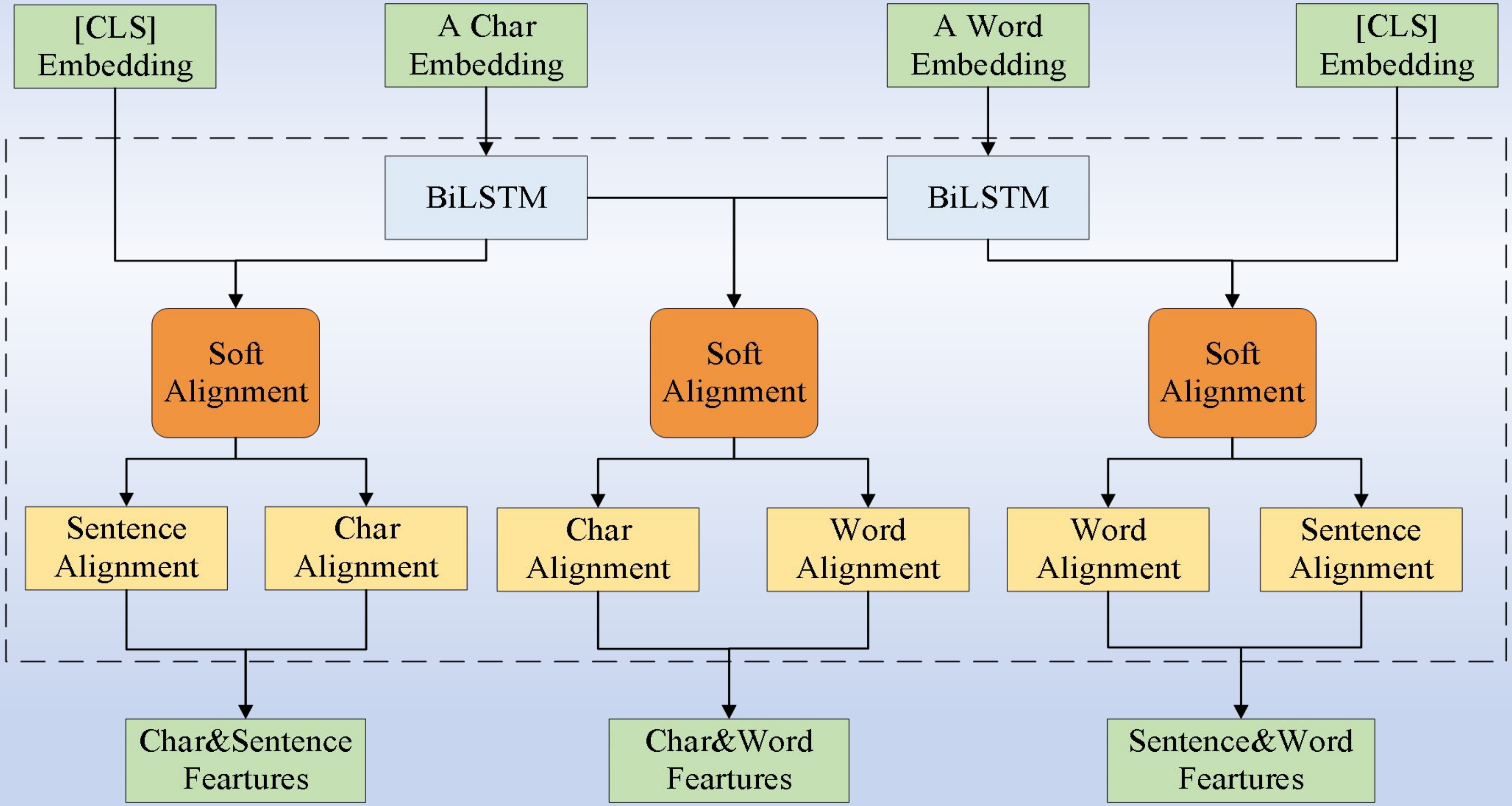
- Short texts have the characteristics of less information and lack of contextual background.
- Methods such as TF-IDF and Word2Vec cannot make correct judgment on sentences with literal similarity and opposite semantics such as "I like Shanghai snacks" and "I don't like Shanghai snacks", resulting in low matching accuracy.
- Feature extraction from a single perspective can not get more comprehensive semantic information

Methodology



Methodology

Multi-granularity fusion coding layer



Datasets

Distribution of BQ and LCQMC datasets

Dataset	Train	Dev	Test
BQ	100,000	10,000	10,000
LCQMC	238,766	8,802	12,500

Data example

Text A	Text B	Label
微信消费算吗 Does wechat consumption count	还有多少钱没还 How much money is still outstanding	0
第一次使用，额度多少？ First time use, how much is the limit?	我的额度多少钱 How much is my limit?	1

Results

Comparative Experimental Results

	BQ		LCQMC	
Model	ACC	F1	ACC	F1
Text-CNN	68.52	69.17	72.80	75.70
BiLSTM	73.51	72.68	73.50	77.5
BIMPM	81.85	81.73	83.30	84.90
ESIM	81.93	81.87	82.58	84.49
MGF	82.86	81.21	85.83	86.72
LET	83.22	83.03	84.81	86.08
MSEM	83.47	83.62	84.33	85.68
GMN	84.21	84.11	84.62	86.00
MFTM(ours)	**85.09**	**85.11**	**89.05**	**89.04**

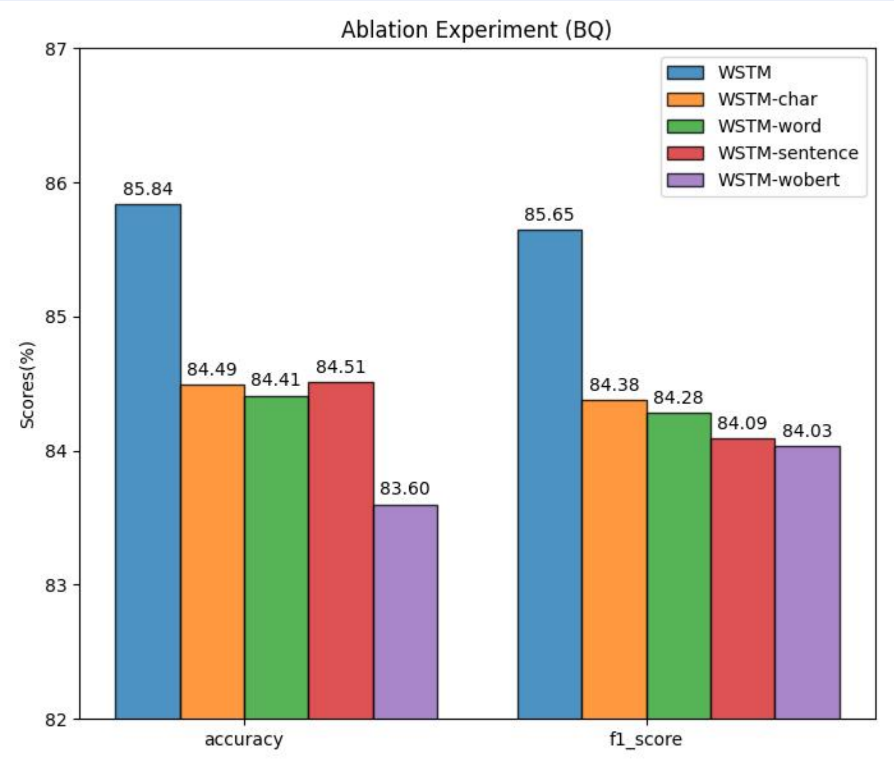
Results

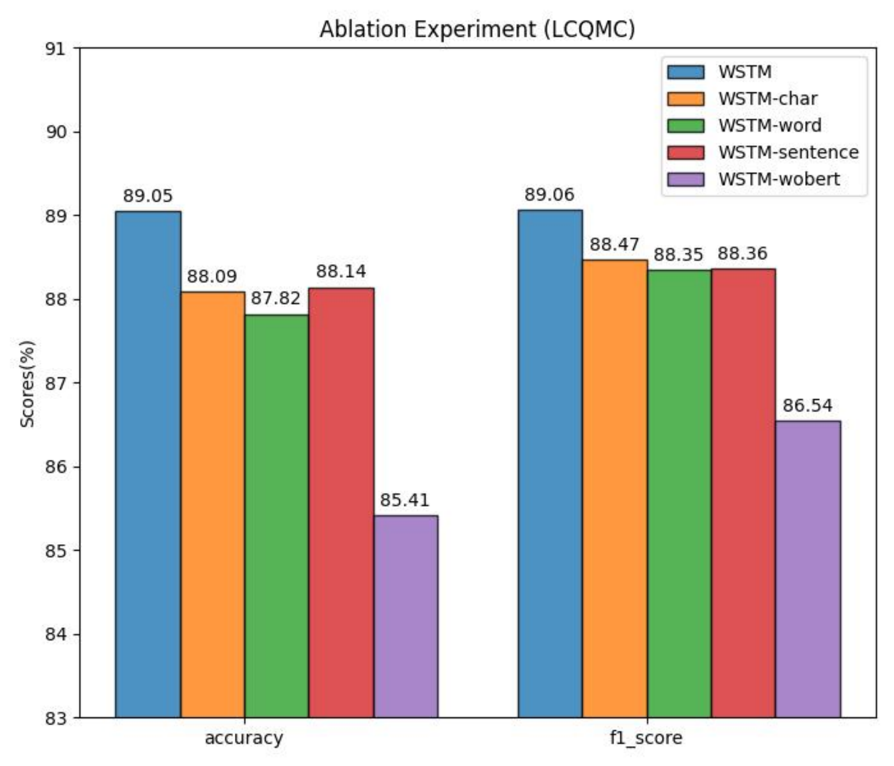
Comparative Experimental Results

	BQ		LCQMC	
Model	ACC	F1	ACC	F1
BERT	84.50	84.00	85.73	86.86
BERT-wwm	84.89	84.29	86.80	87.78
ERNIE	84.67	84.20	87.04	88.06
SBERT	-	-	87.28	-
WoBERT	-	-	88.47	88.21
MFTM(ours)	85.09	85.11	89.05	89.04

Results

Ablation Experiments Results





Conclusion

Advantages

- replacing the Word2Vec model with the large-scale pre-trained language model, WoBERT, enhanced the model's ability to comprehend semantic relationships within text.
- The use of multi-granularity text features makes up for the deficiency of single-granularity features in the comprehensive understanding of text semantics
- Experiments show that our method captures more comprehensive semantic features and enhances the performance of text matching tasks.

Conclusion

limitations

- The pre-trained model WoBERT was directly used to obtain the feature embedding, and the effect of the model was greatly affected by the pre-trained model.

- The way of feature fusion is also relatively simple, and the pertinence of different levels of features is not enough.

- There are also deficiencies in the processing of polysemy and domain-specific data.

- In the future, we will explore more efficient feature fusion methods to reduce computational costs and enhance matching performance.

Thank you for your listening !