3.3 幽默等级识别层

该层由全连接层及softmax层组成。首先将局部和全局语义信息进行融合，然后通过全连接层和softmax层，得到幽默等级的概率分布，计算公式如下：

$$T = [v_{s,ave}, v_{s,max}, u_s, v_{p,ave}, v_{p,max}, u_p] \tag{22}$$

$$humor_{cls} = softmax\left(T\right) = \frac{e^{ti}}{\sum_{i=1}^{12h} e^{ti}} \tag{23}$$

其中$T \in R^{10h}$，$humor_{cls} \in R^C$是概率分布，C是幽默等级数量。本文采用交叉熵作为损失函数，其形式化表示如下：

$$loss = -\sum_{i=1}^{Num}\sum_{j=1}^{C} y_i^j \, log\widehat{y}_i^j + \lambda \parallel \theta \parallel^2 \tag{24}$$

其中，Num是训练集样本数，i是样本序号，j是标签序号，y_i^j是样本的真实标签类别，$\widehat{y}_i^j$是样本的预测标签类别，λ是L_2正则化项的超参数，θ是模型参数的集合。

4 实验结果

本节首先介绍了实验数据、评价指标、实验设置和基线方法，然后对比了基线方法和本文提出的MSIN方法的幽默等级识别性能，最后通过实验分析了本文提出方法的有效性。

4.1 实验数据与评价指标

Reddit数据集：该数据集由Weller等(2019)构建。幽默语句来自Reddit中带有“humor”标签的文本，采用众包方式对幽默语句的“铺垫”和“笑点”进行了标注，且对幽默语句的强弱进行了人工标注。数据集规模详见下表。

	弱幽默	强幽默	总计
训练集	9719	9719	19438
验证集	304	304	608
测试集	304	304	608

Table 2: Reddit幽默数据集统计信息

评价指标：为了便于和基线方法进行比较，本文采用了被广泛接受并应用于文本分类任务中的精确率（Acc）、准确率（P）、查全率（R）和F1 Score （F1）作为评价指标。

4.2 实验设置

词嵌入：在训练过程中，词嵌入表示分别采用了Glove以及Word2Vce(Mikolov et al., 2013)，维度均为300，词嵌入在训练的过程中固定。对未登录词使用$(-0.01, 0.01)$上的平均分布随机初始化。

超参数：在实验中，设置L_2正则化项的超参数$\lambda = 10^{-5}$，Bi-LSTM的神经元个数为128，CNN三个卷积核的尺寸分别为2、3和5，优化方法为Adam(Kingma and Ba, 2014)，Batch大小为64，dropout为0.5。为了防止过度拟合，在训练过程中使用了学习率衰减和早停机制。为了便于和基线模型对比，采用了Weller等(2019)对数据的划分。

4.3 基线方法

本文使用下述基线方法进行对比实验：

- Human(Weller and Seppi, 2019)*：人工预测结果。
- CNN(Weller and Seppi, 2019)*：采用CNN自动提取幽默语句的潜在语义特征并进行幽默等级识别。

https://nlp.stanford.edu/projects/glove/
https://code.google.com/archive/p/word2vec/

Method			Precision	Recall	F1_Score	Accuracy
Human(Weller et al.)*			-	-	-	66.30
分类任务	分类模型	CNN(Weller et al.)*	-	-	-	68.80
		CNN(Kim et al.)	68.22	68.85	68.18	68.16
		LSTM(Hochreiter et al.)	69.46	68.09	68.77	69.08
		Bi-LSTM-Attention	68.70	73.62	71.02	69.98
		Transformer(Weller et al.)*	-	-	-	72.40
		BERT(Devlin et al.)	72.06	74.67	73.34	72.86
推理任务	表示模型	CNN(Kim et al.)	69.78	69.87	69.41	69.42
		LSTM(Hochreiter et al.)	70.66	71.61	70.89	70.71
		BiLSTM-Attention	69.93	73.88	71.70	70.97
		BERT(Devlin et al.)	73.27	73.03	73.15	73.19
	交互模型	ESIM(Chen et al.)	73.38	70.72	72.03	72.53
		MSIN	**74.10**	**74.34**	**74.22**	**74.18**

Table 3: Reddit数据集实验结果

- CNN(Kim, 2014)：本文复现的基于CNN的方法，使用3种不同尺寸卷积核的CNN提取幽默文本特征进行幽默等级识别。
- LSTM(Hochreiter and Schmidhuber, 1997)：使用LSTM提取幽默特征并进行幽默等级识别。
- Bi-LSTM-Attention：使用双向LSTM和注意力机制提取幽默文本特征，并对幽默等级进行识别。
- Transformer(Weller and Seppi, 2019)*：使用基于transformer结构(Vaswani et al., 2017)的预训练模型对幽默文本整体做特征提取，以进行幽默等级识别。
- BERT(Devlin et al., 2018)：本文复现的基于BERT方法的结果，在任务语料上做微调然后进行幽默等级识别。
- ESIM(Chen et al., 2016)：只基于局部语义交互信息进行幽默等级识别。
- MSIN：本文提出的多粒度语义交互理解网络，综合使用语义嵌入、局部语义交互和全局语义交互进行幽默等级识别。

4.4 实验结果分析

本文在Reddit数据集上的实验结果见表3。表格整体分为三部分，第一部分为人工进行幽默等级识别的结果；第二部分采用之前研究的通用方法，将幽默等级识别视作文本分类任务，把幽默文本整体编码后进行分类；第三部分基于本文观点，即可将幽默等级识别任务视作自然语言推理任务，把幽默文本划分为铺垫和笑点两个语义部分，以这两部分作为模型的输入，使用表示型模型或交互型模型预识别文本蕴含的幽默等级。

在第二部分，本文的使用的CNN与Weller等(2019)的CNN结果相近，且两者均取得了明显好于人工预测的结果，证明了神经网络在幽默等级识别上的有效性。然而CNN由于卷积核尺寸固定，难以捕获长距离的语义关系，这对需要充分理解上下文的幽默等级识别任务是不利的。相比CNN，LSTM使用隐态向量捕获句子在长距离上的语义关系，可对时间序列进行有效建模，在数据集上取得了好于CNN的结果。然而LSTM是有偏倚的模型，后送入模型的信息会比先送入模型的信息拥有更大的权重，因此文本又使用Bi-LSTM+Attention进行改进。一方面，BiLSTM可以编码句子从前到后和从后到前两个方向上的信息，获取的特征更丰富，另一方面，Attention将所有时间步上的隐态向量赋予权重，让模型关注在文本分类过程中起关键作用的部分，缓解了由于LSTM的偏倚性造成的信息损失，因此模型相比LSTM取得了更好的结果。最后，本文使用BERT识别文本的幽默等级，其结果与Weller等(2019)使用Transformer的结果相近，并且两者均明显优于之前的模型。

Method	Precision	Recall	F1_Score	Accuracy
MSIN+Glove	73.20	70.07	71.60	72.20
MSIN+Word2Vec	73.33	72.37	72.85	73.03
MSIN+Both	**74.10**	**74.34**	**74.22**	**74.18**

Table 4: 不同词向量使用方式结果比较

Method	Precision	Recall	F1_Score	Accuracy
Word Level	72.64	73.36	73.00	72.86
Sub-sentence Level	70.68	73.22	71.91	71.41
MSIN	**74.10**	**74.34**	**74.22**	**74.18**

Table 5: 不同粒度实验结果比较

在第三部分，本文分别使用表示型和交互型两类模型进行幽默等级识别。

表示模型分别将铺垫和笑点编码为向量，然后将两向量与他们之间作差及点乘的结果拼接以捕获两部分的关系，最后基于拼接后的向量进行分类。为方便与第一部分的结果作比较，本文仍采用CNN、LSTM、Bi-LSTM-Attention和BERT四个模型。首先做内部比较，可以发现四个模型的结果依次递增，与第一部分的趋势保持一致；其次将表示模型与第一部分比较，发现四个模型的结果均高于第一部分中对应的模型，证明将幽默文本拆分为铺垫和笑点两部分，并让模型学习两部分之间的关系信息有助于幽默等级的识别。

在交互模型部分，本文使用ESIM与本文提出的MSIN进行比较。ESIM通过计算两部分文本之间单词的相似度矩阵来构建局部语义交互表示，并以此来推断前后文本的关系，在没有大量预训练知识的情况下，取得了略低于BERT的结果。本文提出的MSIN综合考虑交互过程中局部和全局语义信息的影响，取得了好于ESIM的最优结果。因此可以证明，相比表示模型，交互模型可以更好地捕捉到铺垫和笑点之间的关系；本文提出的多粒度语义交互理解模型融合单词和子句两个级别的交互信息，在幽默等级识别任务上取得了提升。

同时，本文进行消融实验，证明了词向量融合及多粒度交互两个结构的有效性，实验结果分别见表4和表5。表4前两行分别为只使用Glove和只使用Word2Vec的结果，第三行是使用融合词向量的结果，可以发现，融合之后效果更佳。表5前两行分别为只使用单词和子句交互的结果，第三行为融合两个粒度进行交互的结果，可以发现，多粒度交互网络取得了最优结果。

5 结论

本文将幽默文本划分为铺垫和笑点两部分，提出对两者之间的关系进行建模可以显著提升模型识别幽默等级的性能。基于这个观点，首先，本文在融合多种嵌入表示的基础上，从局部和全局两个粒度来对幽默中的语义关系进行理解和建模。其次，本文对幽默中“铺垫”和“笑点”两部分的关联信息做交互建模，从而实现充分挖掘铺垫和笑点之间的关系。最后，本文在Reddit幽默数据集上进行实验，取得了最优结果，同时结合消融实验证实了模型设计的有效性。在以后的工作中，我们将在幽默文本自动切分及基于铺垫的笑点文本生成方面做更多的探索。

参考文献

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv: Computation and Language*.

Francesco Barbieri and Horacio Saggion. 2014. Automatic detection of irony and humour in twitter. *Process Biochemistry*, 40(8):2637–2642.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 6: Siamese lstm with attention for humorous text comparison. pages 390–395.

Dario Bertero and Pascale Fung. 2016a. Deep learning of audio and language features for humor prediction. page 496.

Dario Bertero and Pascale Fung. 2016b. A long short-term memory framework for predicting humor in dialogues. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Vladislav Blinov, Valeria Bolotovabaranova, and Pavel Braslavski. 2019. Large dataset and language model fun-tuning for humor recognition. pages 4027–4032.

Andrew Cattle and Xiaojuan Ma. 2016. Effects of semantic relatedness between setups and punchlines in twitter hashtag games. pages 70–79.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2016. Enhanced lstm for natural language inference.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Tomas Engelthaler and Thomas T. Hills. 2017. Humor norms for 4,997 english words. *Behavior Research Methods*, 50(1):1–9.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. "president vows to cut hair": Dataset and analysis of creative text editing for humorous headlines. *arXiv: Computation and Language*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. pages 1746–1751.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *Computer Science*.

Lizhen Liu, Donghai Zhang, and Wei Song. 2018a. Exploiting syntactic structures for humor recognition. pages 1875–1883.

Lizhen Liu, Donghai Zhang, and Wei Song. 2018b. Modeling sentiment association in discourse for humor recognition. 2:586–591.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Twenty-Sixth International Joint Conference on Artificial Intelligence*.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. pages 531–538.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Computer Science*.

Donald R. Morse. 2007. Use of humor to reduce stress and pain and enhance healing in the dental setting. *J N J Dent Assoc*, 78(4):32–36.

John Allen Paulos. 1980. *Mathematics and Humor*. University of Chicago Press.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Semeval-2017 task 6: hashtagwars: Learning a sense of humor. In *International Workshop on Semantic Evaluation*.

Victor Raskin. 1979. Semantic mechanisms of humor. *Synthese Language Library*, 5(4):409–415.

J. M. SULS. 1972. A two-stage model for the appreciation of jokes and cartoons : An information-processing analysis. *Psychology of Humor*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. pages 5998–6008.

Orion Weller and Kevin D Seppi. 2019. Humor detection: A transformer gets the last laugh. pages 3619–3623.

Chris Westbury and Geoff Hollis. 2018. Wriggly, squiffy, lummox, and boobs: What makes some words funny? *Journal of Experimental Psychology General*, 148(1).

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2018. Double embeddings and cnn-based sequence labeling for aspect extraction.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. pages 2367–2376.

Renxian Zhang and Naishi Liu. 2014. Recognizing humor on twitter. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 889–898.

Zhenjie Zhao, Andrew Cattle, Evangelos E Papalexakis, and Xiaojuan Ma. 2019. Embedding lexical features via tensor decomposition for small sample humor recognition. pages 6375–6380.

杨勇，杨亮，邹艳波，任鸽，樊小超. 2020. 基于音形义特征和层次注意力机制的幽默识别. 计算机工程, pages 1–12.