文章编号：1003-0077(2021)01-0081-07

基于词性特征的明喻识别及要素抽取方法

赵琳玲[1]，王素格[1,2]，陈 鑫[1]，王 典[1]，张兆滨[1]

(1. 山西大学 计算机与信息技术学院，山西 太原 030006；
2. 山西大学 计算智能与中文信息处理教育部重点实验室，山西 太原 030006)

摘 要：比喻是一种利用事物之间的相似点建立关系的修辞方式。明喻是比喻中最常见的形式，具有明显的喻词，例如“像”，用于关联本体和喻体。近年来高考语文散文类鉴赏题中多有考查明喻句的试题，为了解答此类鉴赏题，需要识别比喻句中的本体和喻体要素。该文提出了基于词性特征的明喻识别及要素抽取方法。首先将句子中词向量化表示与词性特征向量化表示进行融合，将融合后的向量输入到 BiLSTM 中进行训练，然后利用 CRF 解码出全局最优标注序列；最后得到明喻识别和要素抽取的结果。公开数据集上的实验结果表明，该方法优于已有的单任务方法；同时也将该文方法应用于北京高考语文鉴赏题中比喻句的识别与要素抽取，验证了方法的可行性。

关键词：比喻；本体；喻体；BiLSTM；CRF

中图分类号：TP391 **文献标识码**：A

Part-of-Speech Based Simile Recognition and Component Extraction

ZHAO Linling[1], WANG Suge[1,2], CHEN Xin[1], WANG Dian[1], ZHANG Zhaobin[1]

(1. School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China;
2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan Shanxi 030006, China)

Abstract: Simile is the most common form in the metaphor, including obvious comparators, such as "like", used to relate tenor and vehicle. To better resolve the Chinese prose reading comprehension of the College Entrance Examination, this paper designs a method for the simile recognition and component extraction based on part-of-speech features. Firstly, the vector representation of the words in the sentence is merged with the representation of the part-of-speech. Then, the fused vector is input into BiLSTM model and the global optimal annotation sequence is decoded by CRF. Finally, the smile recognition and component extraction are generated according to annotated sequence. The experiment results show that the proposed method is better than the existing single task method on the open dataset.

Keywords: metaphor; tenor; vehicle; BiLSTM; CRF

0 引言

比喻是用于与其相似的事物、道理打比方的修辞格，使用比喻辞格的句子称为比喻句。利用比喻对事物的特征进行描述，可以使事物形象、生动、具体，给人留下深刻印象。在文学类作品中，利用比喻可增强语言的表现力，在近年的高考语文散文类鉴赏题中，多有涉及明喻句的考查。以 2016 年北京市高考语文第 24 题为例：

原文：我在这腔调里沉迷且陷入遐想，这是发自雄浑的关中大地深处的声响，抑或是渭水波浪的涛声，也像是骤雨拍击无边秋禾的啸响，亦不知无时节的好雨润泽秦川初春返青麦苗的细近于无的柔声，甚至让我想到柴烟弥漫的村巷里牛哞马叫的声音……

问题：文章第四段运用了多种手法，表达了作者对老腔的感受。请结合具体语句加以赏析。

部分参考答案：比喻，将老腔的腔调比喻为骤雨拍击秋禾的啸响、雨润麦苗的柔声等，既写出

收稿日期：2019-12-12 **定稿日期**：2019-12-31

基金项目：国家自然科学基金(62076158)；国家重点研发计划(2018YFB1005103)；山西省重点研发计划(201803D421024)

了……,又写出了……。

如果能自动识别明喻句并抽取句子中的本体和喻体,不仅可以解答鉴赏类问题,还可以进一步了解作者所表达的思想感情。

比喻句中相关要素定义[1]如下:

本体:被描写和说明的事物,即被比方的事物。

喻词/连接词:连接本体和喻体的词语,在明喻中称为“喻词”,例如,“像、如、好像”等,暗喻中为“连接词”,总称为“触发词”。

喻体:与本体相对,即用来打比方的事物。

喻解/喻底:使本体和喻体构成比喻关系的两者的相似点。

现代修辞学将比喻分为明喻、暗喻、借喻,将暗喻和借喻视为隐喻[2]。明喻就是直接打比方,有明显的喻词指引本体和喻体之间的关系[3]。通过研究明喻[4],分析和阐释明喻现象,了解明喻建立的意义和推理机制,探究明喻背后的认知过程,可以了解人的认知手段与过程。虽然与隐喻相比明喻研究相对容易,但由于数据的缺乏和相关研究较少,给研究也带来了挑战。

现代汉语明喻句的典型句式为“A像B”,此句中本体是A、喻体为B、喻词是像。针对此类明喻句,本文主要研究基于词性特征的明喻识别及要素抽取方法。由于双向的长短期记忆(BiLSTM)[5]能够充分利用上下文信息,而条件随机场(CRF)模型可以用来输出标签之间的前后依赖关系,因此,将词性特征融合到BiLSTM与CRF连接对序列化数据进行建模(BiLSTM-CRF)。在Chinese-Simile-Recognition[6]数据集上进行验证,实验结果表明,本文方法优于Liu等人[6]提出的单任务明喻识别和要素抽取方法。

1 相关工作

目前针对比喻句的已有研究主要是利用句法结构和深度学习的方法。

基于句法模式的分析方法是利用句子的句法结构(主谓宾结构)和词汇间的依存关系(并列,从属等)进行建模的方法。Niculae等人[7]提出了一种使用句法模式进行比较识别的方法,用于比喻句中的本体和喻体的抽取。该方法在处理明喻句中的短句时表现比较好,对于复杂或长句有时会导致本体和喻体抽取的不完整。Niculae等人[8]提出了在比较句中比喻的计算研究,探究了明喻的语言模式,发现领域知识是识别明喻的主要因素。

基于深度学习的比喻句识别,穆婉青[1]采用词和词性作为特征,提出了基于CNN_C的比喻句识别,正确率已达到94.7%,然而,并没有对要素进行抽取。对于要素抽取,研究者们利用多任务学习方法,通过在相关任务间共享表示信息,提升模型在原始任务上的泛化性能。Liu等人[6]提出了神经网络框架联合优化的三个任务。将明喻要素抽取看成序列标记问题,使用不同的前缀标签区分本体和喻体要素。CRF[9]能有效学习输出标签之间的前后依赖关系,近些年在自然语言处理领域中得到广泛应用。Huang等人[10]提出了一系列基于长短期记忆(LSTM)的序列标注模型,首次将BiLSTM-CRF模型应用于NLP基准序列标记数据集,并证明BiLSTM模型可以有效地利用过去和未来的输入特征。对于CRF层,它还可以使用句子级的标记信息,使方法具有较强的鲁棒性,而且对嵌入词的依赖性也较小。

本文将明喻识别和要素抽取作为序列标注任务。嵌入层将词性特征向量化得到的向量与词向量进行融合,采用BiLSTM学习文本中前向和后向距离特征来得到全局特征,在输出层添加CRF层得到文本的最优标注序列。

2 数据特征分析

为了对比喻句中本体和喻体准确的识别,本文选取两个数据集进行考察。

Chinese-Simile-Recognition(CSR)[6]:该数据集是由首都师范大学、科大讯飞等提供①,训练集共有7 262条句子,其中比喻句(明喻句)有3 315条,非比喻句有3 947条。

Simile-Recognition-SXU(SRS)[1]:该数据集是由山西大学研究团队构建,数据来源于高中语文课文、查字典网②、散文吧网站③和BCC网站④。该数据集共有3 207条,其中训练集有1 925条,开发集有641条,测试集有641条。人工标注明喻句中的本体和喻体是最简短的,且不带修饰语。

2.1 喻词分析

对于CSR,明喻句中喻词均为“像”,而明喻句

① https://github.com/cnunlp/Chinese-Simile-Recognition-Dataset

② https://www.Chazidian.com

③ https://www.sanwen8.cn

④ http://bcc.blcu.edu.cn

中的喻词不仅只有“像”，还有“如，好似，仿佛，若，似乎”等。对 SRS 中不同喻词的句子进行统计，结果如表 1 所示。人工校对部分分词，标注的本体和喻体是不带修饰语的名词短语。

表 1　SRS 中不同喻词统计

喻词	数量/条	百分比/%
像	2 396	74.7
如	571	17.8
好似	38	1.2
仿佛	81	2.5
若	23	0.7
似乎	12	0.4
其他	86	2.7

2.2　词性特征分析

通过对 CSR 和 SRS 两个数据集的统计，发现 CSR 中比喻句标出的本体与喻体包含多词的仅占 0.38%，而 SRS 本体与喻体中包含多词的仅占 1.07%，因此，本文只对 CSR 和 SRS 中本体和喻体为单个词的开展研究。再对 CSR 和 SRS 的本体与喻体按照词性进行统计，发现 CSR 和 SRS 中名词分别占 80.3%和 85.9%。而动词在句子中扮演着重要角色，它表征概念实体间的相互关系，是句子中名词实体的概念依存体。因此，词性特征对于识别明喻句中的本体和喻体可以提供更准确的信息。

3　明喻识别及要素抽取方法

通过第 2 节对比喻句特征的分析可以发现，本体和喻体的词性对比喻句的识别具有重要的作用。因此，将词性特征融合到词的表示中。由于 BiLSTM-CRF 模型在 BiLSTM 输出后增加了 CRF 层，所以它能够加强文本间信息的相关性，并同时考虑过去的与未来的特征。因此，本文将明喻识别及要素抽取问题看作序列标注问题。利用每个句子中的词表示和词性表示的联合特征，学习特征到标注结果的映射，得到特征到任意标签的概率，通过这些概率，得到最优序列结果，根据最后序列结果对明喻识别及要素抽取，整体框架如图 1 所示。

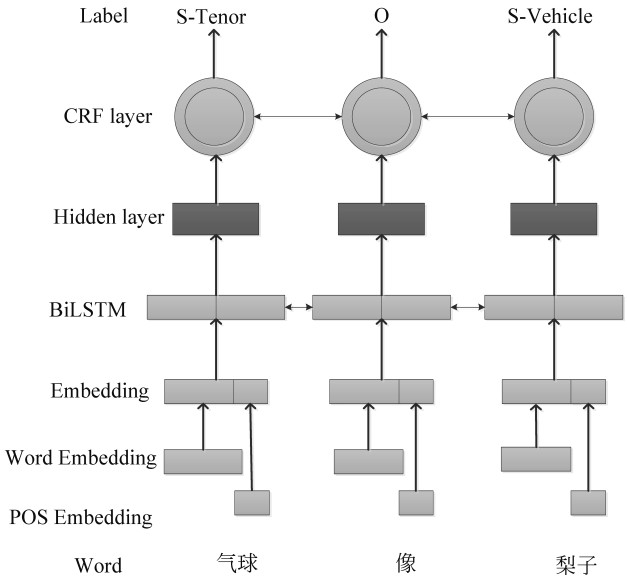

图 1　基于词性特征的明喻识别及要素抽取框架

图 1 中标注的实体有本体和喻体，分别用 Tenor 和 Vehicle 表示，将已标注的数据集转变为 IOBES 模式（O：既不是本体，也不是喻体，S：单独构成实体，B：实体的开始，I：实体的中间，E：实体的结尾），共得到九种标签，具体的标签类型 Typeset = {O，S-Tenor，B-Tenor，I-Tenor，E-Tenor，S-Vehicle，B-Vehicle，I-Vehicle，E-Vehicle}。例如，S-Tenor 表示单独一个词构成本体。对于明喻句的识别问题，当一个句子中所有词的标签都为“O”时，则判定此句为非明喻句，否则此句为明喻句。

3.1 词嵌入式表示

由于明喻句中的语言表达比较含蓄、委婉，直接从字面上对其识别比较困难。例如，图1中的例句，“气球像梨子”，之所以能这样说，主要原因是气球和梨子在形状上有某些相似之处，人们在使用时将“气球”比喻成“梨子”，也就是将有类似特征的词语放到一起使用。虽然它们在字面上属于不同领域的事物，但在上下文中又有一定的语义一致性，另外，它们所具有的词性都为名词。因此，可以建立词语和词性的深层语义表示。

3.1.1 词语的初始化表示[6]

为了刻画句子中词语深层语义表示，Word2Vec可以作为其初始化表示工具，其原因是Word2Vec是在大规模的语料库上进行训练所得，能使词语表达的深层语义更加丰富。

设给定一个句子 Sentence$=\{w_1,w_2,\cdots,w_n\}$，w_i 为句子 Sentence 中的第 i 个词语，利用Word2Vec工具获得 w_i 的初始化嵌入表示为 $\boldsymbol{c}_i$，得到句子的嵌入表示为 Sentence$=\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_n\}$，其中，$\boldsymbol{c}_i\in R^d(i=1,2,\cdots,n)$。

对于比喻句中的词性特征，直接利用结巴工具进行获取，其中名词用“1”表示，代词用“2”表示，动词用“3”表示，其他词性用“0”表示。词性特征也可使用Word2Vec工具获得，$\boldsymbol{p}_i$ 代表词性特征向量，其中 $\boldsymbol{p}_i\in R^d(i=1,2,\cdots,n)$。将词语嵌入和词性特征向量进行拼接，如式(1)所示。

$$\boldsymbol{x}_i=[\boldsymbol{c}_i;\boldsymbol{p}_i] \tag{1}$$

其中，“;”表示拼接操作，$\boldsymbol{x}_i\in R^{2d}$，代表拼接后的向量表示。

利用式(1)可得到句子表示为 Sentence$=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)$。

3.1.2 基于BiLSTM的词语上下文表示

由于BiLSTM[11]能较好地解决文本长距离依赖问题，同时可以在两个方向上进行文本语义表达，因而选用其作为词语的上下文表示。在包含 n 个词的句子 Sentence$=(x_1,x_2,\cdots,x_n)$ 中，每个词由一个 $2d$ 维的向量表示。BiLSTM的正向和反向的上下文信息分别为 $\overrightarrow{h_t}$ 和 $\overleftarrow{h_t}$，如式(2)、式(3)所示。

$$\overrightarrow{h_t}=\overrightarrow{\text{LSTM}}(x_t,\overrightarrow{h_{t-1}}) \tag{2}$$

$$\overleftarrow{h_t}=\overleftarrow{\text{LSTM}}(x_t,\overleftarrow{h_{t+1}}) \tag{3}$$

其中，$\overrightarrow{\text{LSTM}}$ 代表前向LSTM的神经单元，$\overleftarrow{\text{LSTM}}$ 代表后向LSTM的神经单元。

根据式(2)和式(3)分别得到正向和反向的上下文信息，将其进行组合[12]，可以有效地包含文本两个方向的上下文特征，获得句子的嵌入式表示为 Sentence$=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n)$，其中，$\boldsymbol{h}_i=[\overrightarrow{h_i};\overleftarrow{h_i}]$，$\boldsymbol{h}_i\in R^{4d}(i=1,2,\cdots,n)$。

3.2 基于BiLSTM-CRF的明喻要素预测

由于BiLSTM-CRF模型是在BiLSTM输出后增加CRF[13]层，可以增强词语间上下文信息相关性的特征，同时考虑当前每个词的隐层状态的特征。因此，为了对句子进行序列标注，将第3.1节得到的嵌入式表示，经过一个线性变换获得隐藏层每个词的新表示，这个新表示一方面作为CRF[13]的输入，另一方面作为该词在序列标注时标签的得分。

对于嵌入层表示后的句子 Sentence$=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n)$，再使用一个线性变换层，得到句子中每个词 w_i 隐藏状态的嵌入表示 $\boldsymbol{p}_i$，将句子中的词语从 $4d$ 维映射到 q 维空间，q 为标注序列中标签的个数，如式(4)所示。

$$\boldsymbol{p}_i=\boldsymbol{W}\cdot\boldsymbol{h}_i \tag{4}$$

其中，$\boldsymbol{W}\in R^{q\times 4d}$，$\boldsymbol{p}_i\in R^q$。

由式(4)获得句子 Sentence 输入CRF层的嵌入式表示为 $\boldsymbol{P}=(\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_n)$，其中，$\boldsymbol{P}\in R^{n\times q}$，$\boldsymbol{p}_i$ 中的每一个元素 $p_{i,j}$ 表示句子中第 i 个词语 x_i 得到第 j 个标签的得分。

在序列标注任务中，需要利用词的标签与周围词标签存在的依赖关系，然后解码出全局最优的标签序列，CRF正是针对这项工作的。因此，在BiLSTM网络输出层后加入CRF。

对于给定句子：Sentence$=(w_1,w_2,\cdots,w_n)$，预测标签序列为：$\hat{Y}=(\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_n)$。$\boldsymbol{P}=(\mathrm{p}_{i,j})_{n\times q}$ 为BiLSTM的输出标签得分矩阵，$\boldsymbol{A}=(a_{i,j})_{q\times q}$ 为标签间的转移得分矩阵，$a_{i,j}$ 表示从标签 i 到标签 j 的转移得分，y_0 和 y_{n+1} 是新增句子的开始标签和结束标签，因此，$\boldsymbol{A}$ 转化为 $\boldsymbol{B}_{(q+2)\times(q+2)}$ 矩阵。Sentence→Y的序列得分[14]，如式(5)所示。

$$s(\text{Sentence},\boldsymbol{Y})=\sum_{i=0}^{n}B_{y_i,y_{i+1}}+\sum_{i=1}^{n}p_{i,y_i} \tag{5}$$

对于预测句子 Sentence 的标签序列，由softmax计算获得，如式(6)所示。

$$p(Y\mid\text{Sentence})=\frac{e^{s(\text{Sentence},Y)}}{\sum_{Y}e^{s(\text{Sentence},Y)}} \tag{6}$$

在训练过程中，最大化正确标签序列的对数概

率[15]，如式(7)所示。

$$\log(p(Y \mid \text{Sentence})) = s(\text{Sentence}, Y) - \log(\sum_{Y} e^{s(\text{Sentence}, Y)}) \tag{7}$$

从式(7)中能够获得模型生成概率中最大的标签序列。在解码阶段，最后预测的输出序列是根据最高得分的标签序列所获得的，如式(8)所示。

$$\begin{aligned}\hat{Y} &= \arg\max_{Y}\{\log(p(Y \mid \text{Sentence}))\} \\ &= \arg\max_{Y}\{s(\text{Sentence}, Y)\}\end{aligned} \tag{8}$$

通过式(8)，可以获得句子 Sentence 中词的每个标签，其标签类型为第 3 节介绍的九种标签之一。

4　实验结果及分析

4.1　参数设置及评价指标

本文实验中词向量维度是 50，字向量维度是 100，均采用 Word2Vec 训练得到的向量。特征向量的维度设置为 50，LSTM 隐藏层的维度设置为 100，dropout 设置为 0.6。梯度下降优化算法采用 Adam[16]，学习率设置为 0.001。

本文采用第 2 节介绍的 CSR 和 SRS 作为实验数据集。对于一个明喻句，只有本体和喻体的边界和标签都标记正确时，才判定此明喻句要素抽取正确。因此，实验结果采用成对的评价指标①，精确率 P(precision)、召回率 R(recall)和 F_1 值。

4.2　对比方法介绍

为验证本文方法的有效性，设置如下方法对比实验。

CRF：直接利用分词特征，设计 CRF 的特征模板，窗口大小为 5。

RNN：以字向量作为输入的循环神经网络。

CNN：以字向量作为输入的卷积神经网络。

下面的方法仅说明其输入向量的方式，在此基础上采用 BiLSTM-CRF。

C：Embedding 层为每个字的字向量。

C+J：Embedding 层为每个字向量和位置信息的拼接，位置信息的表示是通过结巴分词得到分词信息特征。1 表示词的开始；2 表示词的中间；3 表示词的结尾；0 表示单个词。

Singletask(CE)：由 Liu 等人[6]提出的 Embedding 层为每个词的词向量。

W+T：Embedding 层为每个词的词向量和主题信息的拼接，主题信息是利用 LDA 聚类方法得到的。

W+F：Embedding 层为每个词的词向量和词性特征的拼接。

W+F+T：Embedding 层为每个词的词向量、词性特征和主题信息拼接。

4.3　实验结果及分析

实验 1　七种方法的明喻要素抽取比较

为了验证本文提出方法的有效性，在 CSR 和 SRS 上设置了如下对比实验，实验结果分别如表 2、表 3 所示。

表 2　七种抽取方法在 CSR 上明喻要素抽取的实验结果比较(%)

方法	P	R	F_1	Tenor			Vehicle		
				P	R	F_1	P	R	F_1
CRF	24.05	27.32	25.58	53.78	44.16	48.62	60.13	47.88	53.37
C	54.04	60.06	56.89	67.64	69.25	68.44	72.37	82.83	77.25
C+J	52.08	63.02	57.03	72.7	66.77	69.61	77.96	79.22	78.59
Singletask(CE)	54.04	66.67	59.69	74.92	67.76	71.16	76.46	86.75	81.28
W+T	53.56	69.03	60.32	67.03	73.63	70.18	76.61	85.84	80.97
W+F	**56.15**	69.82	**62.24**	70.33	73.13	**71.71**	78.16	85.14	**81.50**
W+F+T	51.33	72.09	59.97	74.58	66.57	70.35	77.62	85.64	81.43

① https://github.com/cnunlp/Chinese-Simile-Recognition-Dataset

表 3 七种抽取方法在 SRS 上明喻要素抽取的实验结果比较(%)

方法	P	R	F_1	Tenor			Vehicle		
				P	R	F_1	P	R	F_1
CRF	45.59	49.51	47.47	68.32	60.93	64.89	67.90	67.24	67.59
C	69.34	72.04	70.67	83.74	79.27	81.44	83.77	84.71	84.24
C+J	72.38	72.68	72.53	87.75	79.73	83.55	89.46	88.08	88.76
Singletask(CE)	72.03	71.43	71.73	86.06	80.95	83.42	91.64	86.12	88.79
W+T	72.24	73.75	72.99	84.60	81.25	82.89	89.10	87.10	88.09
W+F	**73.67**	**76.84**	**75.22**	87.13	81.55	**84.25**	89.07	90.32	**89.69**
W+F+T	71.41	74.48	72.91	86.36	82.01	84.13	90.45	87.66	89.03

由表 2 和表 3 可以看出:

(1) 以 CRF 方法作为基准方法,可以解决序列标注问题,但是与其他深度学习的方法相比,需要自定义特征模板,并没有学习到文本深层次的特征,因此抽取效果不及其他深度学习方法。

(2) 词向量表示优于字向量表示,主要原因是明喻句中的本体和喻体多数是一个词而不是一个字,而且词比字包含更多的语义信息。

(3) 喻体的抽取效果均比本体的抽取效果好。为了展示其原因,图 2 和图 3 给出了本体—喻词的相对距离以及喻体—喻词的相对距离,可以看出,相对于本体来说,喻体大多数分布在喻词的后面,比较集中,而本体的分布相对比较分散。

(4) 在七种方法的比较中,W+F 在明喻要素抽取中整体效果最好。

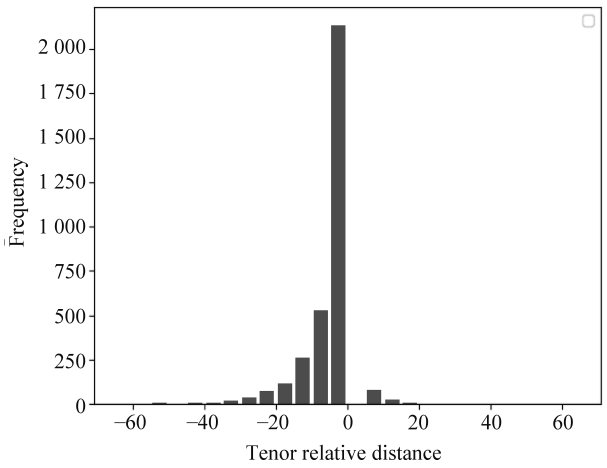

图 2 在 CSR 中本体—喻词的相对距离

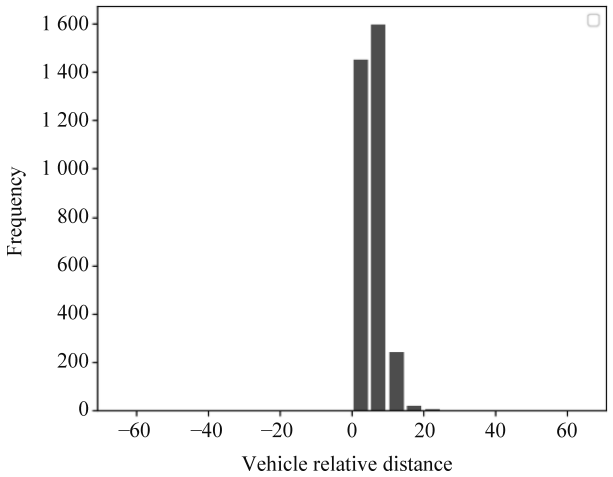

图 3 在 CSR 中喻体—喻词的相对距离

实验 2 九种方法在 CSR 上的明喻识别

为了与文献[6]中的单任务方法进行对比,仅在 CSR 上进行明喻识别对比实验,实验结果如表 4 所示。

表 4 各类识别方法在 CSR 上的实验结果比较(%)

方法	P	R	F_1
CRF	71.19	78.11	74.59
RNN	68.03	75.68	72.00
CNN	80.36	81.26	81.11
C	77.57	90.07	83.37
C+J	80.22	87.94	84.03
Singletask[6]	76.41	90.58	82.94
W+T	77.47	92.30	84.22
W+F	76.68	92.60	83.89
W+F+T	75.12	93.31	83.45

由表 4 可以看出:

(1) CRF、RNN、CNN 三类方法作为基准方法,识别效果均不及 BiLSTM-CRF 方法。这是因为

CRF、RNN、CNN 都无法充分考虑文本上下文信息，因此难以得到深层次的语义信息，不能较好地对文本语义进行特征建模。

（2）对于 Embedding 层分别仅用字向量和词向量表示，在 Recall 指标下，Singletask ＞ C，而在 F_1 指标下，C ＞ Singletask，这是因为判断是否为明喻句只是通过判断句中是否有本体或喻体的标注，而不能判断明喻要素抽取是否正确。

（3）对于 Embedding 层都包含词向量的，在 F_1 指标下，识别效果 Singletask ＜ W＋F＋T ＜ W＋F ＜ W＋T，W＋T 的识别效果最好，而 W＋F＋T 不如 W＋T，这是因为 W＋F＋T 中 Embedding 层融合了比较多的信息，而本文的数据集较小。

实验 3　W＋F 方法应用实验

为了验证 W＋F 方法的应用能力，对引言中高考鉴赏题进行实验，其结果如表 5 所示。

表 5　本文方法解答高考题示例

参考答案	比喻 本体：老腔的腔调（关中大地深处的声响和渭水波浪的涛声）喻体：骤雨拍击秋禾的啸响、雨润麦苗的柔声等
抽取答案	本体：声响、涛声 喻体：雨、柔声
识别答案	比喻
完整答案	比喻，关中大地深处的声响、渭水波浪的涛声比作骤雨、柔声

由表 5 可以看出，本文提出的方法可以抽取到本体和喻体最简洁的名词短语，并能识别出明喻句。抽取到的本体和喻体与参考答案相比不够完整，因此，需要根据抽取的本体和喻体，再与原文中抽取相关的修饰语相结合，形成最终的完整答案。完整答案与参考答案相比，可知本文方法能为解答散文类鉴赏题提供支持，以提升答题的准确率。

5　结论和展望

针对明喻句的识别及明喻句中本体和喻体的抽取问题，本文采用明喻句中本体和喻体的词性特征，设计了基于词性特征的明喻识别及要素抽取方法，并与现有的方法进行了对比，证明了本文所提方法的有效性。针对高考散文类鉴赏题，将本文所提的方法应用到答题中，可以获取部分答案信息。

由于在实际数据中句子比较复杂，而本文实验的数据集中大多数是句式简单的句子，并且其数据集中标注的本体和喻体均为不带修饰语的名词短语，训练集的数据集也比较小，所以深度学习学习到相应特征有限。在未来的工作里，将考虑加入注意力机制来识别带有修饰语的本体和喻体，并且从更深层次挖掘比喻句的特征来对隐喻进行研究。此外，创建更丰富的语料库也是我们下一步重点的工作方向。

参考文献

[1] 穆婉青. 基于修辞格识别的鉴赏类问题解答方法研究[D]. 太原：山西大学硕士学位论文，2018.

[2] 汪洋. 语文修辞[M]. 上海：上海交通大学出版社，2013：53-62.

[3] 严阵. 明喻综述[J]. 湖北教育学院学报，2005，22(3)：22-25.

[4] 徐亚芳. 面向计算的现代汉语明喻句的考察[D]. 南京：南京师范大学硕士学位论文，2015.

[5] Ma X，Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNS-CRF[J]. arXiv preprint arXiv：1603. 01354v5，2016.

[6] Liu L，Hu X，Song W，et al. Neural multitask learning for simile recognition[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing，2018：1543-1553.

[7] Niculae V，Yaneva V. Computational considerations of comparisons and similes[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics，2013：89-95.

[8] Niculae V，Danescu-Niculescu-Mizil C. Brighter than gold：Figurative language in user generated comparisons[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing，2014：2008-2018.

[9] Lafferty J，McCallum A，Pereira F. Conditional random fields：Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning，2001：282-289.

[10] Huang Z，Xu W，Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv：1508. 01991，2015.

[11] Graves A，Schmidhuber J. Framewise phoneme classification with bidirectional LSTM networks[C]//Proceedings of IEEE International Joint Conference on Neural Networks，2005：2047-2052.

[12] Rei M. Semi-supervised multitask learning for sequence labeling[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics，2017：2121-2130.

（下转第 95 页）

李训宇(1994—)，硕士研究生，主要研究领域为自然语言处理、机器翻译。
E-mail：1242041057@qq.com

余正涛(1970—)，博士，教授，博士生导师，主要研究领域为机器翻译、自然语言处理、信息检索。
E-mail：ztyu@hotmail.com

毛存礼(1977—)，博士，通信作者，副教授，主要研究领域为自然语言处理、信息检索、机器翻译。
E-mail：maocunli@163.com

(上接第 87 页)

[13] 沈龙骧，邹博伟，叶静，等. 基于双向 LSTM 与 CRF 融合模型的否定聚焦点识别[J]. 中文信息学报，2019，33(1)：25-34.

[14] 陈伟，吴友政，陈文亮，等. 基于 BiLSTM-CRF 的关键词自动抽取[J]. 计算机科学，2018，45(6)：91-113.

[15] Rei M，Crichton G K O，Pyysalo S. Attending to characters in neural sequence labeling models[J]. arXiv preprint arXiv：1611. 04361v1，2016.

[16] Kingma D，Ba J. Adam：A method for stochastic optimization[J]. arXiv preprint arXiv：1412. 6980，2014.

[17] Lample G，Ballesteros M，Subramanian S，et al. Neural architectures for named entity recognition[J]. arXiv preprint arXiv：1603. 01360v3，2016.

赵琳玲(1995—)，硕士研究生，主要研究领域为自然语言处理。
E-mail：2280950619@qq.com

陈鑫(1992—)，博士研究生，主要研究领域为情感分析。
E-mail：1315614497@qq.com

王素格(1964—)，通信作者，博士，教授，主要研究领域为自然语言处理。
E-mail：wsg@sxu.edu.cn