文章编号：1003-0077（2019）00-0000-00

# 基于非对称孪生网络的新闻与案件相关性分析

赵承鼎[1,2] 郭军军[1,2] 余正涛[1,2] 黄于欣[1,2] 刘权[1,2] 宋燃[1,2]
（1. 昆明理工大学 信息工程与自动化学院，云南 昆明 650500；
2. 昆明理工大学 云南省人工智能重点实验室，云南 昆明 650500）

**摘要：**新闻与案件的相关性分析是法律领域新闻舆情分析的重要环节，可转化为新闻文本与案件文本的相似度计算任务。借助孪生网络计算文本相似度是一种有效途径，其对平衡样本具有良好的学习能力，但在新闻与案件的相关性计算中面临文本不平衡和新闻文本冗余的问题，因此，提出了基于非对称孪生网络的新闻与案件相关性计算方法。通过计算文本中句子与标题的相似度选取与新闻标题最相关的句子表征文档，去除新闻文本中的冗余句子，利用非对称孪生网络建模，考虑到案件要素蕴含案件的关键语义信息，将案件要素作为监督信息融入到非对称孪生网络中对新闻文档和案件描述进行编码，解决新闻和案件在结构和语义上不平衡的问题，最终实现新闻与案件的相关性判断。实验表明建立的模型相比基线模型准确率提升了 2.5%。

**关键词：**非对称孪生网络；案件要素；相关性分析

**中图分类号：**TP391 **文献标识码：**A

# Correlation Analysis of News and Cases Based on Unbalanced Siamese Network

ZHAO Chengding[1,2], GUO Junjun[1,2], YU Zhengtao[1,2], HUANG Yuxin[1,2], LIU Quan[1,2], and Song Ran[1,2]

(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China;
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, Yunnan 650500, China)

**Abstract :**The correlation analysis of news and cases is a key point on news comments analysis，which is to predict whether the news and the case are correlative, and is similar to the text similarity calculation task. Siamese network is one of the most effective method to text similarity calculation task, and it can model two texts with similar structures and semantics. But siamese network may loss accuracy due to the unbalance of texts and the redundancy of news text. To solve these problems, we proposed unbalanced siamese network. Because the news headline contains main information, we compressed the news text with title removing redundant information. Because the case element contains the main semantic information of the case, we encoded the new text and the case by unbalanced siamese network using case element as supervisory information. Finally we predict the correlation between news and case texts. Experiment results show that the proposed model improved accuracy by 2.5% compared to baseline, and the method we proposed solved the problems what traditional siamese network faced.

**Key words:** Unbalanced siamese network; case elements; correlation analysis

## 0 引言

自然语言处理技术在法律领域已得到了充分的重视，国家也正在推行智慧法院的建设。早期面向法律智能的自然语言处理集中在法律文本的处理。近年来，司法部门对于新闻舆情也逐渐重视起来。通过自然语言处理技术处理与法律相关的新闻数据，有利于提升相关部门的工作效率。新闻与案件的相关性分析旨在判断一条新闻

**收稿日期**：2019-06-16；**定稿日期**：2019-08-07
**基金项目**：国家重点研发计划（2018YFC0830105，2018YFC0830101，2018YFC0830100）；云南省高新技术产业专项(201606)

数据与某案件是否相关，是新闻舆情处理的重要环节，更深层的案件领域舆情分析研究，都是以与案件相关的新闻舆情作为基础进行的。新闻与案件的相关性分析可以理解为文本相似度计算任务，即通过计算某新闻文本与案件描述之间的相似度是否高于相关阈值，判定新闻于案件是否相关。

文本相似度的计算是自然语言处理中的一个研究热点。现有的文本相似度计算方法主要是通过将文本映射到向量空间中，再对向量的相似度（如余弦相似度）进行计算的[1][2]。孪生网络是一种基于共享权重对两个待比较对象进行空间映射的网络，在文本相似度计算任务中表现良好，Neculoiu 等人基于孪生网络对文本相似度进行计算，取得了显著的效果[1]。也有研究者通过其他方式对待比较文本进行空间映射或特征提取从而计算文本相似度[2][3]。近年来，孪生网络一直是文本相似度计算的热门方法。

目前计算文本相似度的一般方法是不适用于新闻与案件相似度计算任务的，主要问题在于两方面：首先，由于孪生网络共享参数的特性，目前的文本相似度研究都是基于平衡语料进行的，例如两个词语或两个句子之间的相似度。但是，新闻文本与案件的描述文本是不平衡的，这里的不平衡是指结构和内容上的不平衡，直接对其进行语义表征很难捕获到关键的案件语义信息；第二，新闻文本具有较多的冗余信息。这里的冗余是对于该任务而言的，新闻中一般包含事实描述和观点描述两部分，而新闻与案件是否相关主要需要的是事实部分的内容。李兰君等通过 TextRank 算法对长文本进行压缩，避免了数据稀疏的问题[2]。但由于 TextRank 算法是无监督的，很容易抽取出新闻中与案件无关的关键句，例如作者的观点句等，因此对于本任务是不适用的。表 1 展示了一条有关“昆山反杀案”的新闻的部分内容，分析表中新闻信息可知，新闻文本的内容主要有冗余信息、事实描述和观点描述三个部分。判断其与案件相关与否，主要的依据是其中的事实描述部分，而观点描述和冗余信息都是无关的。表 2 展示了“昆山反杀案”案件描述，从案件描述中可以获取到案件要素信息。分析表 1 和表 2 易知，新闻文本与案件描述无论在内容、篇幅还是结构上，都存在较大差异，具有不平衡的特点，传统孪生网络是不适用的。

**表 1 新闻文本举例**

| | |
|---|---|
| 标题 | 昆山“反杀”事件现场：白衣男一直握着刀 警察来才松手 |
| 冗余信息 | …一般情况下，于某某早上八点左右就到公司上班，比规定时间早一小时，而且经常加班到晚上九十点。… |
| 事实描述 | …一黑衣男子返回宝马车，从车内取出一把刀冲向骑车男，多次做出挥刀动作，并和骑车男子发生肢体接触，在此过程中，刀掉在地上，被白衣骑车男子抢到。… |
| 观点描述 | …网友的同情不能妨碍司法公正，却表达着对弱者的同情、对黑暗势力的抵制，和内心中对维护社会安全、稳定秩序持有的正义感。… |

**表 2 案件描述举例**

| | |
|---|---|
| 案件 | 昆山反杀案 |
| 案件描述 | 2018 年 8 月 27 日晚上 21 时 35 分，江苏昆山市开发区震川路、顺帆路路口发生一起刑事案件。昆山一轿车与电动车发生轻微交通事故。双方争执时车内一名男子刘海龙拿出刀，砍向骑车人于海明，之后长刀不慎落地，于海明捡起长刀反过来持刀追赶该刘海龙，刘海龙被砍伤倒在草丛中。 |
| 案件要素 | 2018 年 8 月 27 日，江苏昆山市，砍伤，刘海龙，于海明 |

针对以上问题，本文提出了一种非对称的孪生网络结构，利用新闻标题压缩新闻内容的方法，在保留新闻主要信息的同时压缩了新闻文本，去除了冗余信息。在对新闻文本进行语义编码时，利用文本中的要素作为监督信息增强了新闻文本的案件语义信息。由于新闻的标题具有主题性和事实性，实验表明本文提出的方法在新闻文本与案件的相关性计算任务中获得了更好的结果。

本文的创新包括以下两个方面：一、提出了利用新闻标题压缩新闻文档的方法，在保留新闻关键信息的同时压缩了文本；二、提出了非对称孪生网络结构，将案件要素作为监督融入到神经网络的编码中，解决了新闻文本和案件描述不平衡的问题。

## 1 相关工作

### 1.1 文本相似度计算

文本相似度计算方法分为 2 大类：基于字符串统计的方法和基于机器学习的方法。

基于字符串的方法从字符串匹配度出发，以字符串共现和重复程度为相似度的衡量标准。如编辑距离、汉明距离、余弦相似度、最长公共子串、N-gram 等。该方法原理简单、易于实现，现已成为其他方法的计算基础。但不足的是将字符或词语作为独立的知识单元，并未考虑词语本身的含义和词语之间的关系。

基于机器学习的方法利用从语料库中获取的信息计算文本相似度。基于语料库的方法可以分为：基于词袋模型的方法和基于神经网络的方法，一般以待比较相似度的文档集合为语料库。根据考虑的语义程度不同，基于词袋模型的方法主要包括向量空间模型、潜在语义分析、概率潜在语义分析和潜在狄利克雷分布。Zhao 等人通过句法关系，独特内容相似，长度和字符串的特征预测了句子相似性[3],BJERVA 等人通过 WordNet、词重叠等特征构建了一个预测句子相似度的算法[4],Severyn 等人将语义关系结合到支持向量机中从而建立更精确的句子的分类器[5]。

基于神经网络方法与传统方法的不同之处在于计算文本表示的方式。词向量是经过训练得到的低维实数向量，维数可以人为限制，实数值可根据文本距离调整，这种文本表示符合人理解文本的方式。句子相似度预测主流的神经网络方法有 RNN，CNN 等。

RNN 模型有处理序列信息的能力,因此在自然语言处理中较为常用。特别地，LSTM 在处理具有长距离依赖的序列时更具有优势。Mueller 等人使用 LSTM 作为孪生网络的编码器进行语义编码[6]；Tai 等人提出了一种 tree-LSTM 的方法，对待比较文档通过 tree-LSTM 进行编码进而对编码结果进行相似度计算[7]。CNN 在提取句子局部特征上具有良好的特性，Kim 等人通过 CNN 进行句子分类发现 CNN 比 RNN 能提取到更细粒度的句子特征[8]。He 等人在孪生网络中利用卷积和池化操作提取句子中定长间隔的特征，他们利用卷积核来分析不同大小窗口上的词向量，并通过池化操作来得到句子的特征从而进行句子相似度计算[9]。

对于文本相似度计算的具体任务来说，文本过长或变长是一个常见的问题，已有学者对此进行研究。Gong 等人通过 LDA 主题模型对文档提取隐含的主题，在主题层面上对两文本计算相似度，从而解决变长文档的问题[10]。但该方法只适用于同一领域且主题分布相对接近的两文档计算，而新闻和案件不光是长短差距明显，并且新闻文档的主题不一定是描述案件事实的主题，还可能是观点的主题，因此该方法对于案件任务是不适用的。

### 1.2 孪生网络

Chopra 等提出的孪生网络最早应用于人脸识别，是由一组具有相同参数的网络作为基础构成的神经网络结构[11]，相同的参数带来的对称性使得孪生网络对于具有相同结构的文本具有良好的建模能力。在基于神经网络计算平衡文本相似度的方法中，目前效果最好的模型是基于孪生网络的方法。许多研究者将不同的网络结构与孪生网络相结合探究其用于文本相似度计算的能力，Huang 等人最早将多层感知机和孪生网络结合起来应用到文本的相似度计算中[12]，通过将查询内容和文档映射到语义空间后再通过语义空间上的向量计算余弦相似度从而得到文本的相似度。Shen 等人通过用卷积网络替代多层感知机减少了训练的参数，并增强了模型捕获窗口内连续语义的能力[13]。但是无论是多层感知机还是卷积网络，都无法捕获整个句子的语义信息，因此对于文本的相似度计算效果都不尽如人意。由于长短时记忆网络在对于文本处理上的优势，Neculoiu 等人将双向长短时记忆网络应用到孪生网络，在文本相似度计算的任务上取得了当时最好的效果[1]，之后的研究者对于孪生网络在文本相似度任务的应用大都是基于此开展的，李兰君等[2]通过引入层级注意力机制对文档级的输入进行建模，并通过 TextRank 算法对文档进行压缩，避免了长文档的数据稀疏问题。对于新闻和案件的相关性而言，由于新闻中一般包含事实描述和观点描述两部分，而新闻与案件是否相关主要需要的是事实部分的内容，因此李兰君等的方法不适用抽取出新闻文本中的对于相关性计算的关键句。由于孪生网络本身对称的特性，对于输入不平衡的文本之间的相似度计算不能有效地建模。针对上述问题，我们提出了非对称孪生网络，利用文本中的要素作为监督信息改进新闻文本编码的方法和基于新闻标题的文档压缩方法。

## 2 基于非对称孪生网络的新闻与案件相似度计算模型

基于非对称孪生网络的新闻与案件相似度计算模型分为文档压缩层，词嵌入层，文档编码层和预测层组成，关键步骤如图 1 所示（压缩层未画出）。输入一篇新闻文本 $D=\{w_1,w_2...w_n\}$ 和案件的描述 $C=\{e_1,e_2...e_n\}$，其中 $w_i$ 代表文本中的词，$n$ 代表文本的长度，$e_i$ 代表组成案件的要素，$m$ 代表案件描述的长度，$e\in E$，$E$ 为案件要素的集合。$D'_e=\{D_{e1}...D_{ei}\}$ 表示 $D$ 中包含的要素集合，$e_i\in E$。

设压缩层将文档表示为压缩后的关键句集合为 $D'$，通过词嵌入层将文档和要素嵌入为词级的向量空间表示，再经过文档编码层将上下文信息编码到语义空间，最终通过预测层将文档和案件的空间语义表征计算为最终预测结果。输出 $p$ 表示文本 $D$ 与案件 $C$ 的相似度，当 $p>0.5$ 时预测新闻与案件是相关的。

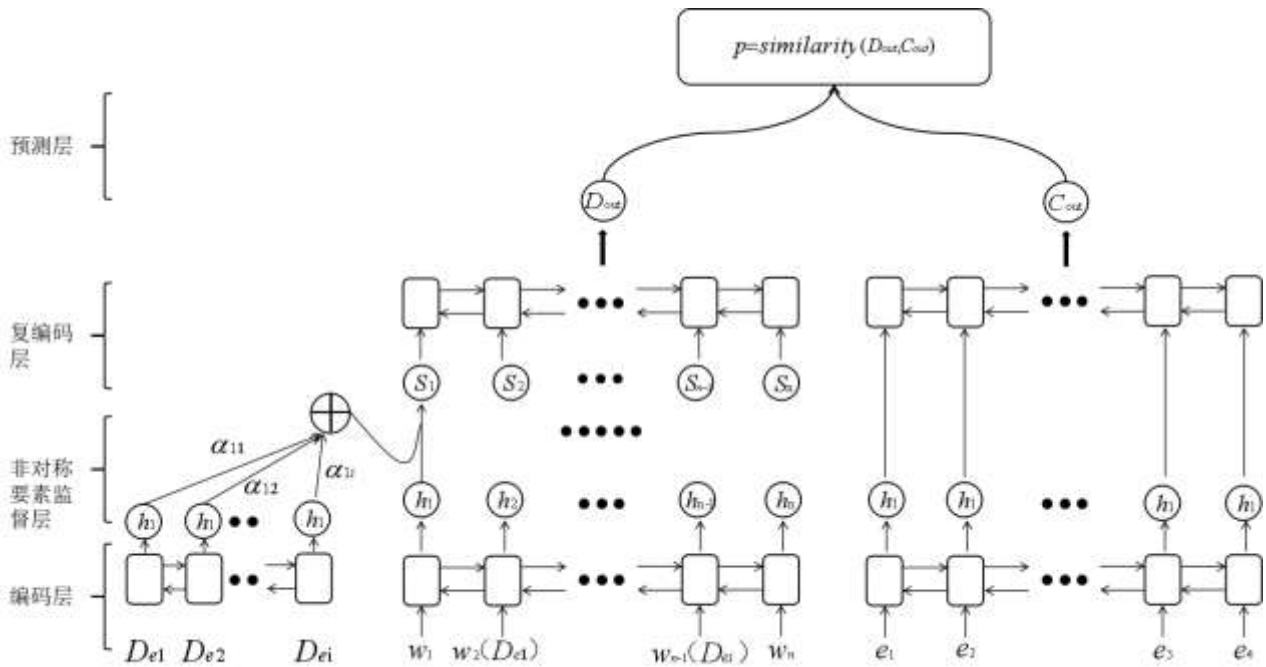


图 1 基于非对称孪生网络的新闻与案件相似度计算模型图

### 2.1 基于标题的文档压缩层

针对新闻文本具有较多冗余信息的问题，本文提出了基于标题的文档压缩方法。新闻中一般包含事实描述、观点描述和冗余信息三部分，而判断新闻与案件是否相关主要依据的是事实部分的内容。由于 TextRank 算法是无监督的，很容易抽取出与案件无关的关键句，因此[2]中的文档压缩方法对于本任务是不适用的。

由于新闻的标题具有主题性和事实性，因此基于标题进行压缩可以满足任务的需求。我们通过计算文档中每个句子与标题的相关性，从中选出相关性较高的几个句子 $D'=\{S_1,S_2...S_k\}$ 作为文档 $D$ 的表示，其中 $S_n$ 表示抽取出来的句子。

设新闻标题为 $S_t$，对于新闻文本 $D=\{S_1...S_n\}$，其中 $S_i$ 表示新闻文本中的句子。计算每个句子与标题的相关性评分 $Score(S_t,S_i)$，本文选取 $ROUGE1$ 的 $f1$ 值作为计算方法，计算压缩句子的公式如下：

$$Score(S_t,S_i)=ROUGE(S_t,S_i) \quad (1)$$

$$D'=\underset{Score(st,si)}{\operatorname{argmax}}(Si...Sj) \quad (2)$$

公式(1)(2)通过计算每个句子与标题的相关性评分，将新闻正文的句子按得分高低排列，选出其中前几句作为压缩内容。公式中 $D'$ 即为选出的与标题最相关的句子集合，作为接下来计算

的输入。

### 2.2 词嵌入层

设 $D'=\{w_1, w_2 \ldots w_k\}\in R^V$ 表示压缩后的文档 $D'$，其中 $w$ 表示词，$k$ 为 $D'$ 的长度，$V$ 为词表的大小。通过预训练的词向量矩阵 $M\in R^{V*d}$，将 $D'$ 中的每个词转化为 $d$ 维的向量 $w'\in R^d$。同理，将案件描述 C 中的词嵌入为 $C'\in R^d$，而 $D$ 中的要素 $De'$ 嵌入为词向量 $E'$。

### 2.3 非对称文档编码层

非对称文档编码层的作用为提取整个文档的语义信息。传统孪生处理新闻文本和案件描述时面临文本不对称的问题。由于案件要素对于新闻中案件相关语义具有指导作用，本文对新闻文本构建了一个非对称的要素监督层，通过案件要素监督选择新闻文本中与案件相关的语义信息，从而解决新闻与案件语义不平衡的问题。

#### 2.3.1 语义编码层

将 $C'$，$D'$，$E'$ 经共享参数的双向 LSTM 转化为 $Ce'$，$De'$，$Ee'$，目的是提取文档和句子的上下文语义信息。词向量通过双向 LSTM 编码的公式如下：

$$i_t = \sigma(W_{ixt} * W_d + U_{iht-1}) \quad (3)$$

$$f_t = \sigma(W_{fxt} * W_d + U_{fht-1}) \quad (4)$$

$$o = \sigma(W_{oxt} * W_d + U_{oht-1}) \quad (5)$$

$$\tilde{c}_t = \tanh(W_{cxt} * W_d + U_{cht-1}) \quad (6)$$

$$c_t = i_t \circ \tilde{c}_t + f_t \circ c_{t-1} \quad (7)$$

$$h_t = o_t \circ \tanh(c_t) \quad (8)$$

$$H_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \quad (9)$$

其中 $i_t$ 是输入门，决定 LSTM 存储什么信息，$f_t$ 是遗忘门，决定 LSTM 丢弃什么信息，o 是输出门，决定这个 LSTM 单元要输出什么信息，∘代表点乘操作。通过不同的门控机制，LSTM 将输入的词向量 $w_d$ 转化为与上文相关的语义向量 $h_d$。本文将双向 LSTM 的每一个时间步上的输出作为语义编码层 1 的结果。即经过语义编码层 1，$D'$ 编码为 $D_h\in\{h_1, h_2 \ldots h_k\}\in R^U$，其中 $U$ 表示 LSTM 隐层维度大小。

同理，$C'$ 和 $E'$ 也分别编码为双向 LSTM 的输出集合 $C_h$、$E_h$。

但是，$D_h$ 中含有许多冗余信息，与案件相关的有效信息只有少部分。因此需要对 $D_h$ 进行语义选择，本文在非对称的要素监督层中利用 $E_h$ 中的要素信息增强新闻中的案件语义。

#### 2.3.2 非对称的要素监督层

由于新闻文本中具有许多与案件无关的信息，因此对于新闻文本，本文通过注意力机制建立 $D_h$ 与 $E_h$ 的联系，对于要素相关性高的词增强语义重要性，相关性低的词降低语义重要性。具体方法是：

$$Score(h_d, h_e) = h_d{}^T W_a h_e \quad (10)$$

$$a_t(s) = align(h_t, \bar{h}_s) = \frac{\exp(Score(h_t, \bar{h}_s))}{\sum_{s'} \exp(Score(h_t, \bar{h}_{s'}))} \quad (11)$$

$$E_{weights} = a_t \circ E_e \quad (12)$$

$$D_{weighted} = \alpha * E_{weight} + (1 - \alpha) * D_{e'} \quad (13)$$

通过公式(10)-公式(13)，使用注意力机制计算 $D_h$ 与 $E_h$ 每个词的联系 $E_{weighted}$，并通过 $E_{weighted}$ 将 $D_h$ 最终转化为经过要素监督的 $D_{weighted}$，其中 $Wa$ 为用于计算 $h_d$，$h_e$ 的注意力矩阵，$scores$ 表示 $h_d$ 和 $h_e$ 之间的相关性得分。$\alpha$ 为用于分配要素部分和文档本身部分的编码在新的向量表征中所占比重的权重。$Wa$，$\alpha$ 都是可训练的参数。

#### 2.3.3 语义复编码层

论文[1][2]中为了在相同维度下进行计算，对语义编码的向量进行全连接，从而得到固定维度的向量。本文考虑到要素监督层改变了文档所在的语义空间，因此在上层使用另一层双向 LSTM 进行再次编码，将 D 和 C 再次映射到同一语义空间进行计算，取末状态作为输出。语义编码层 2 的公式如下。

$$D_{out} = BiLSTM_2(D_{weighted}) \quad (14)$$

$$C_{out} = BiLSTM_2(C_{e'}) \quad (15)$$

为简化书写，本节将 2.3.1 中双向 LSTM 的公式(10)-公式(13)简化为 $BiLSTM_2$ 表示。

### 2.4 预测层

文档 $D$ 和案件 $C$ 经过文档编码层分别编码为 $D_{out}$ 和 $C_{out}$。预测层通过计算 $D_{out}$ 和 $C_{out}$ 的距离或相似度从而计算出文档 $D$ 和案件 $C$ 的相似

度。文献[2]中实验证明使用曼哈顿距离作为损失函数表现良好，因此本文选用曼哈顿距离计算语义差异性。

$$Similarity(D_{out},C_{out}) = 1-sigmoid(manhattan(D_{out},C_{out})) \quad (16)$$

公式（16）通过计算向量 $D_{out}$ 与 $C_{out}$ 间的曼哈顿距离，得到两者在语义上的差异，将曼哈顿距离通过 $sigmoid$ 函数映射到(0,1)区间上，来计算出 $D_{out}$ 和 $C_{out}$ 的相似度 $Similarity(D_{out},C_{out})$。

## 3 实验

### 3.1 数据集

通过分析近年来的热门新闻，本文选择了“昆山反杀案”等 13 个热门案件，爬取与案件相关的新闻 4513 条。通过建立新闻与案件相关关系，得到新闻-案件对应数据 4607 对。通过人为校准，选出有效数据 3374 对，其中相关的案件-新闻对 1630 对，不相关数据 1744 对。从中分离出 675 对作为验证集，验证集中相关数据 326 对，不相关数据 349 对。

### 3.2 超参数设置

本文实验中，压缩层取的句子数 $K$ 为 3 句，词嵌入层的词嵌入维度为 300 维，语义编码层的隐层维度为 128 维，语义复编码层的维度为 32 维，Dropout 设置为 0.5，选择 Nadam 作为优化算法，训练批次的 batchsize 设置为 20。

### 3.3 基线模型

本文使用传统孪生网络[1]和李兰君等的层级注意力机制孪生网络(HASM)作为基线模型[2]。传统孪生网络直接用一层 LSTM 将新闻和案件描述进行编码并计算相似度；HASM 使用 TextRank 压缩长文档，并使用层级注意力机制增强各编码层的编码能力，取得了目前法律领域文本相似度计算的最佳效果。

### 3.4 评价标准

本文使用准确率（$p$），召回率（$r$），F1 值（$F_1\_score$)作为评价指标。其中，$F_1_score$的计算方式为：

$$F_1\_score = 2*p*r/(p+r)。 \quad (16)$$

### 3.5 本文方法的有效性分析

分别与如下方法进行比较，验证提出方法的有效性：要素共现方法（利用案件要素的出现与否判定是否相关），传统的孪生网络方法和 HASM。

表 3 不同方法的有效性验证结果

| 评价指标 | p | r | F1_score |
|---|---|---|---|
| 要素共现 | 0.3822 | 0.7874 | 0.5146 |
| 传统孪生网络 | 0.8405 | 0.8070 | 0.8234 |
| HSAM | 0.8750 | 0.7833 | 0.8266 |
| 本文方法 | 0.9002 | 0.8322 | 0.8648 |

分析表 3 可知案件相关与否与要素是否出现不是紧密关系的，因此简单使用要素出现与否无法判定某新闻是否描述了某案件。

可以看出，对称的孪生网络在非对称的数据集上表现并不好，而本文所提出的方法对于非对称数据上的相似度计算较基线模型提高了准确率 2.5%、召回率 5%、F1 值 4%。且较大程度超过了传统孪生网络。

在本文新闻文本的分类任务上，HSAM 并没有本文方法表现良好，造成其效果不佳的原因主要是：本文任务中，比较文档具有不对称性且新闻文本的标题蕴含了文档的主要信息，而本文的非对称编码以及基于标题的压缩更适用于本任务。

### 3.6 各层的有效性验证

为验证本文提出模型各部分的有效性，分别将各部分删除进行比较。特别的，将基线模型中提出的层级注意力机制融入到模型中观察效果，从而分析每个部分是否对于新闻与案件相关性的计算是有效的。

表 4 各层的有效性验证实验结果

| 评价指标 | p | r | F1_score |
|---|---|---|---|
| 本文方法 | 0.9002 | 0.8322 | 0.8648 |
| -要素监督层 | 0.8576 | 0.7886 | 0.8216 |
| -文档压缩层 | 0.8551 | 0.7589 | 0.8041 |
| +层级注意力机制 | 0.8833 | 0.7131 | 0.7891 |

分析表 4 可知，要素监督层和文档压缩层对于新闻和案件的相关性预测是具有实际作用的。不使用要素监督层和文档压缩层，准确率分别下降了 4.3%、4.5%。特别的，当引入基线模型中提出的层级注意力机制时，效果并没有超过本文方法，这是由于，层级注意力机制使得不相关内容融入了要素中的语义，因此当使用层级注意力机制编码后，要素监督层的作用反而消退了，因此，在本任务中，要素监督的作用比[2]中层级注意力机制更为有效。

### 3.7 不同文档压缩方法的有效性分析

为验证基于标题的压缩方法的有效性，本文分别与如下方法进行比较：不进行压缩的方法和

基线模型中利用 TextRank 进行压缩的方法。

表 5 不同压缩方法的相似度计算实验结果

| 评价指标 | p | r | F1_score |
|---|---|---|---|
| 不压缩 | 0.8551 | 0.7589 | 0.8041 |
| TextRank | 0.8676 | 0.7994 | 0.8321 |
| 只用标题 | 0.8750 | 0.7699 | 0.8190 |
| 本文方法 | 0.9002 | 0.8322 | 0.8648 |

分析表 5 可知，无论是 TextRank 算法还是本文的压缩方法对新闻文档进行压缩都使得准确率较不压缩获得了提升。使用 TextRank 在本任务中得到的效果并不理想，主要原因是新闻中一般包含事实描述和观点描述两部分，而新闻与案件是否相关主要需要的是事实部分的内容。由于 TextRank 算法是无监督的，很容易抽取出与案件无关的关键句。实验表明本文利用标题抽取新闻关键内容的方法在本任务上表现效果更好，比 TextRank 算法提高了准确率 3.3%、召回率 3.2%、F1 值 3%。而只使用标题也不能得到最高准确率，这是因为新闻标题也有不描述事实的情况，在这种情况下抽取三句正文更有可能包含案件事实的内容，因此利用标题抽取正文的方法比只用标题的方法得到了更高的准确率。

## 4 总结和展望

本文提出了对新闻和案件建立相关性计算的任务。针对传统孪生网络在计算相似度时面临的新闻文本和案件描述不平衡的问题，本文提出了一种基于非对称孪生网络的解决方法：在处理新闻文本时使用标题进行文本压缩，从而解决了新闻正文内容冗余以及文本过长的问题；在编码时使用案件要素作为监督进行案件相关的语义增强，从而消除了新闻文本和案件描述不平衡的问题，在不对称文本的数据集上提升了新闻文本和案件相关性计算的精度。

进一步研究可以将案件本身的特性结合到相关性的计算中，也可以尝试将一些新模型应用到案件和新闻的相关性计算任务中，例如通过构建新闻和案件的实体关系图，利用图卷积神经网络来提取两者的特征。

## 参考文献

[1] Neculoiu P, Versteegh M, Rotaru M. Learning text similarity with siamese recurrent networks[C]//Proceedings of the 1st Workshop on Representation Learning for NLP. Berlin: ACL Press, 2016: 148-157.

[2] 李兰君，周俊生，顾颜慧，等. 基于改进孪生网络结构的相似法律案例检索研究[J]. 北京大学学报（自然科学版), 2019, 55(1): 84-90.

[3] Zhao J, Zhu T, Lan M. Ecnu: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment[C]//Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014).Dublin: ACL Press, 2014: 271-277.

[4] Bjerva J, Bos J, Van der Goot R, et al. The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity[C]//Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014).Dublin: ACL Press,2014: 642-646.

[5] Severyn A, Nicosia M, Moschitti A. Learning semantic textual similarity with structural representations[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).Sofia: ACL Press, 2013: 714-718.

[6] Mueller J, Thyagarajan A. Siamese recurrent architectures for learning sentence similarity[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix: AAAI Press, 2016: 2786-2792.

[7] Tai K S, Socher R, Manning C D. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing: ACL Press, 2015: 1556-1566.

[8] Kim Y. Convolutional Neural Networks for Sentence Classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: ACL Press, 2014: 1746-1751.

[9] He H, Gimpel K, Lin J. Multi-perspective sentence similarity modeling with convolutional neural networks[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: ACL Press, 2015: 1576-1586.

[10] Gong H, Sakakini T, Bhat S P, et al. Document similarity for texts of varying lengths via hidden topics[C]//56th Annual Meeting of the Association for Computational Linguistics. Melbourne, ACL Press, 2018: 2341-2351.

[11] Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego: IEEE Press, 2005, 1: 539-546.

[12] Huang P S, He X, Gao J, et al. Learning deep structured semantic models for web search using clickthrough data[C]//Proceedings of the 22nd ACM international conference on Information & Knowledge Management. San Francisco: ACM Press, 2013: 2333-2338.

[13] Shen Y, He X, Gao J, et al. A latent semantic model with convolutional-pooling structure for information retrieval[C]//Proceedings of the 23rd ACM international conference on conference on information and knowledge management. Shanghai: ACM Press, 2014: 101-110.

赵承鼎（1994—），硕士研究生，主要研究领域为自然语言处理，信息检索。
E-mail：zcdfuzhu@qq.com

郭军军（1987—），通信作者，讲师，硕士生导师，主要研究领域为自然语言处理，信息检索，机器翻译。
E-mail：guojjgb@163.com

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