镜像图灵测试:古诗的机器识别

薛 扬[1)] 梁 循[1)] 赵东岩[2)] 杜 玮[1)]

[1)](中国人民大学信息学院 北京 100872)

[2)](北京大学王选计算机研究所 北京 100871)

摘 要 古诗伴随着中华文化的历史进程不断发展,有着数千年的灿烂历史,古诗将丰富的情感、有内涵的灵魂和生动的形式完美结合,表现出了中华民族语言的力量."自然语言处理是人工智能皇冠上的明珠",用机器生成语言是机器智慧的核心体现,对机器的语言进行测试是图灵测试的重要内容,用机器生成的中国古代诗词已经可以初步通过图灵测试,在普通人面前得以瞒天过海.本文提出了"镜像图灵测试"框架,其主要设计思想是将图灵测试中的测试者由人更换为计算机,要求测试者在图灵测试的同等条件下对被测试的人和计算机进行识别,若测试计算机不能完成对被测试者的识别,则认为被测试的机器通过了镜像图灵测试.本文以机器生成的古诗和诗人创作的古诗为测试对象,以经过 LDA 主题模型调节的融合自注意力机制和切片 LSTM 网络的模型为测试机,设计了镜像图灵测试实验.实验将古诗分为写景、抒情以及爱国诗三类,为每类诗歌构建了 8 组数据集,共 8 万句古诗,采用了 4 种模型对 24 组数据集进行测试,利用测试机判别诗歌来自诗人还是机器,识别结果可达 80%左右,实验结果显示,镜像图灵测试机可以对机器生成的诗歌进行识别,即机器生成的通过了图灵测试的诗歌并没有通过镜像图灵测试,说明了诗歌作为人类语言文明的结晶,是人脑情感最突出的反应,是诗人全身心的投入后的灵魂映射,在一定意义上是图灵可测的,即如果存在图灵可测的不完备性,那么诗歌这个人类语言的精华所在,就是突破这个图灵不完备性的关隘.本文提出的镜像图灵测试框架为后续图灵测试的研究提供了新的思路与方向.

关键词 镜像图灵测试;诗歌生成;文本分类;切片神经网络;注意力机制

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2021.01398

Mirror Turing Test: Machine Recognition of Ancient Chinese Poetry

XUE Yang[1)] LIANG Xun[1)] ZHAO Dong-Yan[2)] DU Wei[1)]

[1)](*School of Information, Renmin University of China, Beijing* 100872)

[2)](*Wangxuan Institute of Computer Technology, Peking University, Beijing* 100871)

Abstract With the continuous development of Chinese culture and thousands of years of splendid history, ancient poetry combines rich emotions, connotative souls and vivid forms perfectly, showing the power of Chinese language. Language is also an important research direction in the field of artificial intelligence. Using machine-generated language is the core embodiment of machine intelligence. Testing machine language is an important content of Turing test. The ancient Chinese poetry generated by machine has passed Turing test preliminarily and can deceive ordinary people. This paper puts forward the thought model of "Mirror Turing test". Its main design idea is to replace the tester in Turing test with computer, and require the tester to identify the tested person and computer under the same conditions of Turing Test. If the test computer cannot complete the identification of the tested machine, it is considered that the tested machine has passed the Mirror Turing test. Considering that in the field of recognition, the ability of

收稿日期:2019-11-24;在线发布日期:2020-11-17.本课题得到国家社会科学基金(18ZDA309)、国家自然科学基金(62072463)、北京市自然科学基金(4172032)资助.**薛 扬**,硕士研究生,主要研究方向为自然语言处理.E-mail:xueyang@ruc.edu.cn.**梁 循**(通信作者),教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为社会计算、神经网络、自然语言处理.E-mail:xliang@ruc.edu.cn.**赵东岩**,博士,研究员,博士生导师,中国计算机学会(CCF)杰出会员,主要研究领域为自然语言处理、知识图谱.E-mail:zhaody@pku.edu.cn.**杜 玮**,博士,主要研究方向为社会计算、自然语言处理.

computer has surpassed that of human beings, it is more difficult for the poetry generator to pass the Mirror Turing test. In this paper, the machine-generated poetry and the poetry created by poets are taken as the test objects, and the model of integrating self-attention mechanism and slice LSTM network modified by LDA theme model is taken as the test machine, and the Mirror Turing test experiment is designed. In the experiment, ancient poetry is divided into three categories: landscape, lyric and patriotic poetry. Eight data sets are constructed for each category of poetry. Four models are used to test 24 groups of data sets. The test machine is used to determine whether poetry comes from a poet or a machine, and the recognition results can reach about 80%. The recognition accuracy of lyric poetry is relatively low, which shows that when the poetry contains emotion and has more emotion, it will bring more difficulties to the test of the machine. Behind the poetry is the poet's soul, which can't be imitated by the current machine in any case. The research of machine emotion is the most far-reaching in front of artificial intelligence research. The experimental results show that the Mirror Turing test machine can identify the poetry generated by the machine, that is, the poetry generated by the machine that has passed the test The poetry of Turing test cannot pass the Mirror Turing test, which shows that poetry, as the crystallization of human language civilization, is still difficult to be surpassed by machines so far. In the field of poetry generation, Turing test may be incomplete. The problem of Mirror Turing test is the game between the testing machine and the tested machine. Human beings can improve the requirements of Mirror Turing test by constantly improving the testing machine. Considering that the recognition level of the machine can surpass that of human beings, the computer that has passed the Mirror Turing test is far more intelligent than human beings. The Mirror Turing test framework proposed in this paper provides new ideas and directions for the subsequent research on Turing testing.

Keywords　mirror turing test; poetry generation; text classification; sliced neural network; attention mechanism

1　引　言

正如“书读百遍，其意自见”、“天下文章一大抄”，一般来讲，诗人在创作好句名诗的时候，是通过不断学习已有的诗词名句，并结合自身体会感悟来进行写作的. 古代名句中也不乏部分语句相似的作品，例如同为描写春天景色的“野渡无人舟自横”和“野渡舟横，杨柳绿阴浓”除用词相似以外，这两句所描写的意境也有异曲同工之处. 在近代诗词的诗词作品中，模仿古人的用词更是常见，例如“有客自知亡国恨，厌闻人唱后庭花.”则是引用了古代名句“商女不知亡国恨，隔江犹唱后庭花.”的用法. 随着人工智能领域的不断发展，学者也在研究如何让机器模仿人类来使用语言，其中，让机器模仿人类进行写诗就是很有趣的一个研究方向，目前，利用机器生成诗歌的方法主要是通过深度学习的方法，用机器生成诗词也和诗人学习写诗一样，需要多学诗，学好诗，才能有更多的素材，更深的理解. 通过不断地学习已有诗词名句，机器也可以生成看似较好的诗句，对于同样的深度学习算法，影响机器所写的诗词质量的因素通常包括学习的次数，即利用现有诗词进行训练的训练次数和学习的诗词的数量，即训练集的大小. 目前，很多学者利用深度学习的方法让机器通过学习人的写诗习惯来生成诗词，而机器所写的诗词甚至可以通过图灵测试(Turing Test，TT)，即生成的诗句可以以假乱真，让人难以识别.

图灵测试[①]由艾伦·图灵提出，它指测试者(人)与被测试者(一个人和一台计算机)被一个不泄露测试内容以外信息的实体(例如房间)隔离开的条件下，通过一些无智能的装置(如键盘)向被测试者随意提问(见图 1 右半部分). 进行多次测试后，如果有超过 30%的答案，人类测试者不能确定出被测试者是人还是计算机，那么这台计算机就通过了测试，并

① 百度百科. 图灵测试. https://baike.baidu.com/item/图灵测试/1701255?fr=aladdin

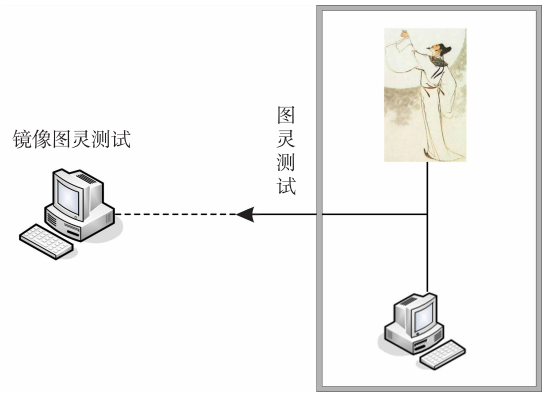


图 1 图灵测试和镜像图灵测试

被认为具有人类的智能.图灵测试是一个思想模型.

目前,利用机器生成中文诗词的各种系统蓬勃发展,包括现代诗歌与古代诗词微软互联网工程院推出的人工智能机器人微软小冰曾于 2017 年 5 月出版了现代诗集《阳光失了玻璃窗》,并于 2019 年出版了第一部人工智能与人类诗人联合创作的诗集《花是绿水的沉默》.清华大学推出的诗歌自动生成系统九歌在 2017 年 8 月参加科学挑战类节目《机智过人》并通过现场观众的图灵测试.华为也推出了人工智能作诗系统"乐府作诗"可以生成唐诗宋词以及藏头诗.

但是很多真正的诗人对于机器写的诗词持负面态度,认为机器写诗不过是语言游戏,无论学习多少句,都无法写出真诗的灵性,写不出泣鬼神的东西.从"魔高一尺,道高一丈"的思想出发,由于目前机器难以模仿出人类的灵魂,在诗歌上是难以造出高水平的"赝品诗人"的,机器诗人也难以造就出流传千古的诗词名句,因此,在人类灵魂的诗词高地,目前是"魔高一尺,道高一丈"的.但是在识别的领域并不是这样,由于深度学习的广泛使用,机器在一些识别领域已经可以超过人类,而本文的研究内容则是利用机器去识别诗歌,判断这些诗歌是出自诗人还是出自已经通过图灵测试的诗歌生成器,即"镜像图灵测试"(Mirror Turing Test,MTT).

镜像图灵测试是指在上述的图灵测试中,测试者(人)也被换成了计算机(见图 1).它要求测试者(计算机)在上述条件下,能够识别被测试者是人还是计算机,如果识别准确率低于 70%(误判率超过 30%,等同于图灵测试的误判率),那么这台计算机就通过了镜像图灵测试.镜像图灵测试是一个思想模型.事实上,如果通过了镜像图灵测试,其智能程度已经超过了人类,有的学者预估这个年代是 2060 年左右(见图 2 阴影部分)①.

"自然语言处理是人工智能皇冠上的明珠"②,

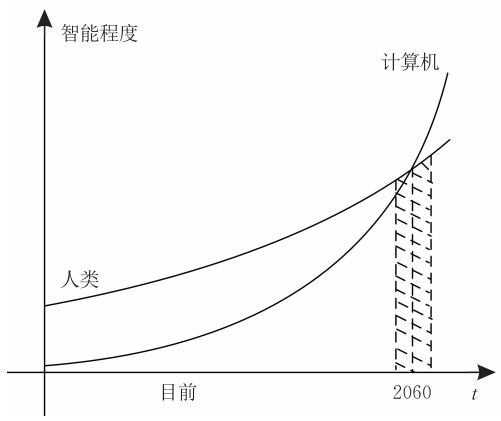


图 2 目前计算机智能和将来能通过镜像图灵测试的计算机的智能

而诗歌是人脑情感最突出的反应,本文以诗词的识别为例,研究了镜像图灵测试.在本文构建的镜像图灵测试实验中,作者假设测试机器的所有提问都是独立的,并不相互依存,同古诗的图灵测试一样,在本文的研究中只考虑单轮镜像图灵测试.

被测试机器用于生成古诗,被测试机器学习生成古诗的数据来源为大量的中国古代诗词(见图 3 右侧),机器通过不断地学习诗人的诗句,可以根据给定的起始句继续向后生成.例如,对于给定的诗句"才饮黄河水,又食酸汤鱼"③,机器可根据这句诗继续生成"故人多感慨,极目独伤情".

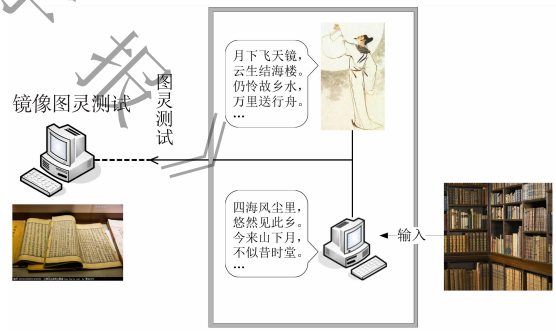


图 3 古诗机器生成和机器识别的数据规模

测试机器(见图 3 左侧)用于测试诗句来自诗人还是生成机器,其数据来源为被测试机器生成的诗句以及和被测试机器生成的数据等量的来自真正诗人的诗句,其中部分诗句用来让机器学习两类诗词的特点,部分诗句让机器来判断来源.例如,识别来自诗人的"才饮黄河水,又食酸汤鱼"和来自机器的"故人多感慨,极目独伤情".

① 因图灵测试而认识图灵社区. https://www.ituring.com.cn/article/468195

② 周明. 语言智能是人工智能皇冠上的明珠. https://zhuanlan.zhihu.com/p/27311604

③ 林鸿飞. 微信朋友圈. 2019-08-07 09:58

2　相关工作

在机器作诗领域，传统的机器作诗方法主要包括基于模板和规则的生成方法、基于遗传算法的评估迭代方法、基于给定写作意图的摘要生成方法以及基于统计机器模型的机器翻译方法.

Zhou 等人[1]利用启发式算法，设计了基于平仄、句法和语义的函数以及基于精英主义和轮盘赌算法的选择策略，提出了宋词生成计算模型，可生成给定词牌的宋词. Yan 等人[2]将诗歌生成看作一个在约束条件下的最优化求解问题，利用摘要生成的方法，不断迭代替换，输出效用最高的结果，在给定写作意图的情况下可以生成符合格式且满足音调节奏要求的诗词. He 等人[3]利用统计机器翻译的方法生成中文古典诗歌，在给定关键词的情况下生成完整的诗句，并提出了一种自动评价诗歌的方法.

这类基于规则或者统计机器学习的方法生成的诗歌的水平与制定的规则有着密切的关系，而且需要大量的规则对生成的文字进行约束，生成的诗歌内容也比较局限，难以在其它场景进行迁移拓展.

随着人工智能在自然语言处理领域的广泛应用，学者也开始探索如何利用人工智能算法解决机器作诗的问题. Zhang 等人[4]首次提出将深度学习算法用于机器作诗的领域，将循环神经网络与古诗生成的韵律等规则相结合，在给定的规则框架中利用神经网络学习下一个字的分布概率，从而可以生成较好的绝句. Yi 等人[5]提出了基于工作记忆模型的诗歌生成方法，在生成诗歌时，保留已经生成诗句的主题和历史信息，通过动态记忆模型生成连贯的古诗，可以生成四行诗和抒情诗，并通过自动评估和人工评估的方法证明了提出模型生成效果比已有的其它模型生成效果更好.

无论是传统的诗歌生成方法还是结合深度学习的诗歌生成方法，即便诗歌生成器已经达到了一个较高的水平，可以生成让普通人真假难辨的诗词，但是即便是水平最高的机器在诗句中学到的也只是字词的分布概率，是根据真正诗人的诗词进行学习并实现文字堆砌的过程，机器没有情感，其生成的诗难免缺乏诗人的灵魂，可以生成好句却无法生成千古名句.

Yu[6]讨论了语言与图灵测试的关系，指出了人类的语言主要功能分为三种，分别为指名、指物和指心，而图灵测试只将智能的表现限定在指名的功能里，也就是说即便机器通过图灵测试也不能说其具有智能. 而语言三指之间的协调才是在不同应用环境下机器未来的发展方向.

就目前的诗歌生成技术水平来看，即便机器诗人有时甚至可以瞒过普通人的眼睛，让普通人难以辨别诗歌是否来自真正的诗人，但是利用机器生成好的诗歌，真正实现诗歌的内外表示协调统一还有很长的路要走，这也是本文构建“镜像图灵测试”的意义所在. 图 4 和图 5 为镜像图灵测试的测试者与被测试者的智能关系图.

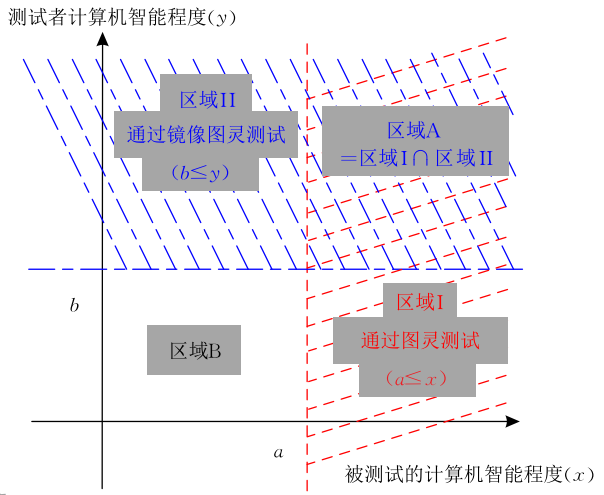


图 4　测试者计算机与被测试者计算机智能关系的区域

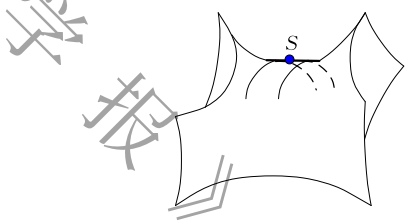


图 5　在图 4 的区域 A 中测试者计算机与被测试者计算机的智能关系

图 4 中，a 表示达到图灵测试最低水平的被测试的计算机智能程度，b 表示达到镜像图灵测试最低水平的测试者计算机智能程度. $a\leq x$ 表示被测试的计算机通过了图灵测试(区域 I)，$b\leq y$ 表示测试者计算机通过了镜像图灵测试(区域 II)，区域 B 表示被测试的计算机和测试者计算机都很“笨”即“瞎子”对“瞎子”的情况，区域 A 表示被测试的计算机和测试者计算机都很“聪明”是“高手”对决的情况. 在图 5 中，不论在鞍面的哪个位置，测试者计算机都高于被测试的计算机，其中典型位置为鞍点 S.

在普通人已经难以对机器生成的诗歌进行识别时，机器在识别领域或许可以超过人类的识别水平，利用机器对机器生成的诗歌进行识别，或许可以在图灵测试的基础上进一步提高机器智能的门槛.

3 模 型

3.1 LSTM 模型

长短期记忆网络[7](Long Short-Term Memory，LSTM)在循环神经网络(Recurrent Neural Network，RNN)的基础上加入了三种“门”的概念，控制信息的通过方式，以此来调节信息的传递，使得神经网络可以学习到时序序列中的长期依赖关系.

对于单向 LSTM，在处理每一个时间点的信息时，利用了遗忘门、输入门和输出门这三个主要的门结构. 对于 t 时刻，三个门结构的输出分别用 f_t、i_t、o_t来表示，输入到每一个 LSTM 单元中的信息包括上一个单元的输出 h_{t-1}和该时刻的输入信息 x_t，即

$$X=[h_{t-1},x_t] \tag{1}$$

对于输入信息的接受、更新以及输出过程见公式，其中 σ 表示 sigmoid 函数，$\boldsymbol{W}$ 为权重矩阵、$\boldsymbol{B}$ 为偏执矩阵，

$$\begin{bmatrix} f_t \\ i_t \\ o_t \end{bmatrix}=\begin{bmatrix} \sigma \\ \sigma \\ \sigma \end{bmatrix}(\boldsymbol{W}X+\boldsymbol{B}) \tag{2}$$

记忆单元的候选状态 $\widetilde{C}_t$的表示见式(3)，$\boldsymbol{W}_c$为权重矩阵，$\boldsymbol{B}_c$表示偏执矩阵. 该时刻的记忆单元状态见式(4)，其中 C_{t-1}为上一时刻的记忆单元状态，

$$\widetilde{C}_t=\tanh(\boldsymbol{W}_cX+\boldsymbol{B}_c) \tag{3}$$

$$C_t=f_tC_{t-1}+i_t\widetilde{C}_t \tag{4}$$

对于每一个 LSTM 单元的输出

$$h_t=o_t\tanh(C_t) \tag{5}$$

在 LSTM 模型单向传播时，每一时刻只能获得在这一时刻前的数据信息，而双向 LSTM 模型[8]将向前传播和向后传播的两个单向 LSTM 模型相结合，可以获取在这一时刻前后的数据，对于 BiLSTM 模型中每一时刻的前后向输出分别为

$$\overrightarrow{h_t}=\mathrm{LSTM}(h_{t-1},x_t,C_{t-1}) \tag{6}$$

$$\overleftarrow{h_t}=\mathrm{LSTM}(h_{t+1},x_t,C_{t+1}) \tag{7}$$

因此，综合前后向 LSTM 输出的结果，BiLSTM 模型的输出为

$$h_t=[\overrightarrow{h_t},\overleftarrow{h_t}] \tag{8}$$

3.2 Sliced-BiLSTM 模型

Yu 等人[9]为提高传统的循环神经网络的运算性能，将传统网络模型进行分解切片，提出了切片循环神经网络(Sliced Recurrent Neural Network，SRNN)，并通过实验证明，该网络模型实现了 RNN 的并行计算，不仅能够提高 RNN 的运算速度，而且构建了更类似于人脑的处理序列的机制，使得传统的循环网络具备了更快速地从序列中获取除字词以外更高层信息的能力，提高了运算精度. 这种模型构建方式在确保精度优先的基础上，利用切片的思想大大提升了模型的计算速度，在本文的诗词识别问题中，由于数据量较大，为了更快更有效地获取高层信息，因此采用切片模型进行实验.

本文以此为启发，在双向 LSTM 模型的基础上构建了切片双向长短期记忆网络模型(Sliced B-directional Long Short Term Memory，Sliced-BiLSTM). 图 6 为 Sliced-BiLSTM 模型结构图.

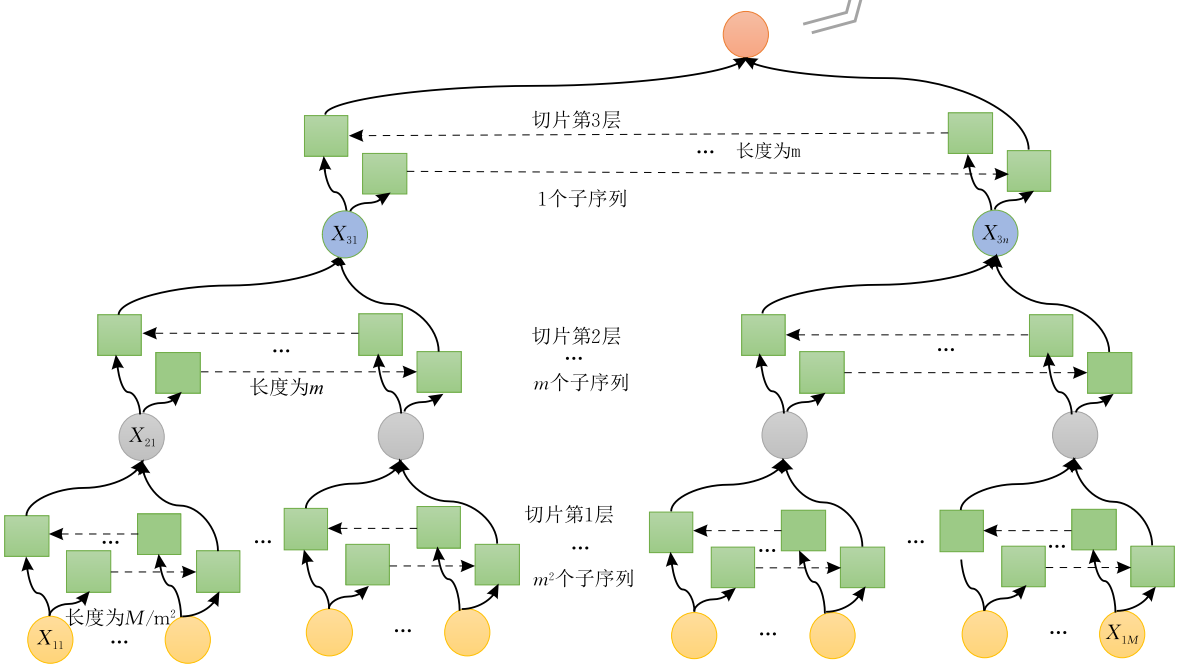


图 6 切片双向 LSTM 结构图

定理 1.　在 Sliced-BiLSTM 模型中，第 p 层序列有 m^{k-p+1} 个子序列，子序列长度均为 m. 其中 k 为切片次数，m 为第一次切分的子序列个数.

推论 1.　在 Sliced-BiLSTM 模型中，对于原序列 $X_1=[X_{11},X_{12},\cdots,X_{1M}]$，经过 k 次切片后的第 p 层序列为 $X_p=[X_{p1},X_{p2},\cdots,X_{p^{m^{k-p+1}}}]$.

推论 2.　在 Sliced-BiLSTM 模型中，对于原序列 $X_1=[X_{11},X_{12},\cdots,X_{1M}]$，经过 k 次切片后的第 p 层序列的第 a 个子序列为 $N_{pa}=[X_{p(am-m+1)},X_{p(am-m+2)},\cdots,X_{p(am)}]$.

对于输入的长度 $len_1=M$ 的序列 $X_1=[X_{11},X_{12},\cdots,X_{1M}]$，对其进行一次切分，切分成 m 个子序列，对于第一次切分后的序列 X_1，每一个序列的长度 $l_1=\dfrac{M}{m}$，原序列 $X_1=[N_{11},N_{12},\cdots,N_{1m}]$，其中，第 a 个子序列 $N_{1a}=[X_{1(al_1-l_1+1)},X_{1(al_1-l_1+2)},\cdots,X_{1(al_1)}]$. 若对输入序列进行 k 次切分，原网络模型变为从第 1 层到第 $k+1$ 层的网络，其中，最小子序列长度为 $l=\dfrac{M}{m^k}$，第一层即原始层中共有 m^k 个最小子序列，则第二层序列的长度为 m^k，对于第 p 层序列，共有 m^{k-p+1} 个子序列，即 $X_p=[X_{p1},X_{p2},\cdots,X_{p^{m^{k-p+1}}}]$，子序列长度均为 m，第 a 个子序列 $N_{pa}=[X_{p(am-m+1)},X_{p(am-m+2)},\cdots,X_{p(am)}]$.

从处理速度来看，对于切片后的 BiLSTM 神经网络，不同子序列之间的运算可以并行进行，若一个 LSTM 神经元的运算时间为 t，则第一层网络的运算时间 $T_1=2\times\dfrac{M}{m^k}\times t$，由于神经网络第 p 层中的序列长度均为 m，因此，第 p 层的运算时间 $T_p=2\times m\times t$. 则对于整个 Sliced-BiLSTM 网络，运算时长为

$$T_s=T_1+(k+1)\times T_p=2\times\left(\frac{M}{m^k}+km+m\right)\times t \tag{9}$$

我们把上述过程写成如下定理.

定理 2.　Sliced-BiLSTM 网络的运算时长为 $T_s=2\times\left(\dfrac{M}{m^k}+km+m\right)\times t$，其中 M 为原序列长度，m 为第一次切分的子序列个数，k 为切分次数，t 为每个 LSTM 神经元的运算时间.

推论 3.　传统 BiLSTM 模型与 Sliced-BiLSTM 模型的运算时间比值为 $\dfrac{T}{T_s}=\dfrac{Mm^k}{M+(k+1)m^{k+1}}$.

对于传统 BiLSTM 模型，运算时长为

$$T=2\times M\times t \tag{10}$$

将传统模型的运算时间与 Sliced-BiLSTM 模型的运算时间进行比较

$$\frac{T}{T_s}=\frac{M}{\dfrac{M}{m^k}+m+km}=\frac{Mm^k}{M+(k+1)m^{k+1}} \tag{11}$$

因此，随着切片次数 k 和切分子序列数量 m 的变化，Sliced-BiLSTM 模型都会对原有模型有着不同程度的速度提升.

3.3　自注意力机制

注意力(attention)机制最初被应用在计算机视觉领域，用来对图片的不同区域赋予权重，增强对于重要区域的关注度，减弱对于无关区域信息的关注度，以模拟人脑对于图片的处理过程，提高了计算机视觉领域的运算效率与准确率.

注意力机制在文本处理中的首次应用是在 2015 年被应用到机器翻译领域[10]，而后注意力机制又被应用到了文本处理中的其它常见领域，包括实体抽取、事件检测以及情感分析等领域. 注意力机制可以学习句子中不同词汇的重要程度和关注度，从而给不同的词汇进行加权. 从目前的应用结果来看，注意力机制在文本处理的多个领域中都取得了较好的成果[11-17].

在本文的诗歌分类问题中，引入注意力机制以增加识别模型的准确度，由于不涉及外部信息，因此，采用自注意力机制(self-attention)对诗歌语句内部信息进行学习，对不同字赋予不同权重. 对于 Sliced-BiLSTM 层的输出 h_t，对于每一句的第 t 个字，可获取注意力权重分布

$$\begin{aligned}a_t&=\mathrm{softmax}(\boldsymbol{W}_2(\tanh(\boldsymbol{W}_1h_t+b)))\\&=\frac{\exp(\boldsymbol{W}_2(\tanh(\boldsymbol{W}_1h_t+b)))}{\sum_{t=1}^{L}\exp(\boldsymbol{W}_2(\tanh(\boldsymbol{W}_1h_t+b)))}\end{aligned} \tag{12}$$

其中，$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 为权重矩阵，b 为偏置，L 为句子长度.

获得经过注意力机制调节的权重后，每一句获得了一个加权后的新向量，自注意力机制层的输出为上一层的输出结果加权求和后的向量，

$$\boldsymbol{H}=\sum_{t=1}^{L}a_th_t \tag{13}$$

3.4　分类器

对于经过自注意力层加权后的向量 $\boldsymbol{H}$，利用 softmax 函数实现分类，输出为 c 类的概率为

$$p_c=\mathrm{softmax}(\boldsymbol{W}_c\boldsymbol{H}+b_c)=\frac{\exp(\boldsymbol{W}_ch_t+b_c)}{\sum_{t=1}^{L}\exp(\boldsymbol{W}_ch_t+b_c)} \tag{14}$$

对于输入分类器的测试集诗词语句序列 M，输出的语句 M 的类别 $\hat{c}$ 为

$$\hat{c}=\arg\max(p_c) \tag{15}$$

因此,对于全部输入的测试集诗句,可以得到标签集合$\{\hat{c}\}$.

本文所使用的模型采用梯度下降的方法对大小为S的训练集进行训练,损失函数为交叉熵函数.

$$loss=-\frac{1}{S}\sum_{i=1}^{S}[c_i\log\hat{c}_i+(1-c_i)\log(1-\hat{c}_i)] \tag{16}$$

其中,c_i为第i句诗句的原始标签,$\hat{c}_i$为第i句诗句的预测标签.

3.5 LDA 调节模型

由于 LSTM 和自注意力机制考虑的主要是上下文的时序关系和句子内部的依赖关系,而机器生成的古诗词,虽然在整体连贯以及语义上都可以解释,但是可能会出现主题不够集中,前后表述内容逻辑性不强等问题,即无法达到诗人的"灵魂"高度,因此选用 LDA 主题模型对上述模型的结果进行部分调节,以期能够提高整体的识别效果.

由于在不同主题的诗词中通常会出现重复的字词,为避免重复的字词出现对模型带来的影响,因此在为数据集构建主题模型前,首先采用 TF-IDF 算法构建 LDA 主题模型的输入向量,即带有 TF-IDF 权重的"诗词-字"矩阵.

对于任意一个句子M中的词x,TF-IDF 算法的计算主要包括词频TF_x和逆文档频率IDF_x两个部分,公式分别为

$$TF_x=\frac{count_{x,M}}{\sum_{i=1}^{M}count_{i,M}} \tag{17}$$

$$IDF_x=\log\frac{count(sentence)}{count(sentence|x\in sentence)} \tag{18}$$

TF-IDF 的值最终可以表示为

$$TF\text{-}IDF_x=TF_x\times IDF_x \tag{19}$$

其中,$count_{x,M}$表示在句子M中x出现的次数,$count(sentence)$表示数据集中全部诗句的数量,$count(sentence|x\in sentence)$表示数据集中包含词$x$的诗句的数量.

由于在诗词中通常会有部分常用字在不同的诗句中经常出现,较高的词频可能会使这一类词的权重过大,利用 TF-IDF 算法对词频矩阵进行加权,可以避免这类影响.

LDA 模型是一种三层的贝叶斯主题模型,属于无监督的学习方法,能够从文本中挖掘主题信息.针对本文中的诗词文本,该模型认为,每一句诗词到主题服从一个多项式分布,每一个主题到字也服从多项式分布.

即对于任意的诗句M、字x和主题t,从诗句到字服从分布

$$p(x|M)=p(t|M)\times p(x|t) \tag{20}$$

LDA 主题模型首先通过统计方法得到$p(x|M)$,并为诗句中的每一个字随机指定一个主题作为初始主题,而后通过吉布斯采样的方法,不断对每一句诗词的主题进行重新采样,直至吉布斯采样收敛,最终可以得到诗词到主题的概率分布.

在得到每一句诗词的主题概率分布后,相当于针对每一句诗词得到了一组维度为主题数的特征向量,利用逻辑回归的方法根据特征向量对诗词进行分类,可以得到一组新的分类结果,根据算法 1 的结果调节算法对原结果进行调节.其中算法 2 为调节算法中的阈值选择算法.

算法 1. 结果调节算法.

输入:未经调节的结果$\{\hat{c}\}$

输出:调节后的结果 new$\{\hat{c}\}$

1. 确定主题数目 $topic\_num=5$
2. 根据 LDA 主题模型,构建训练集和测试集共n条数据的主题分布概率矩阵
   $\{(p_{11},\cdots,p_{1topic\_num}),\cdots,(p_{n1},\cdots,p_{ntopic\_num})\}$
3. 将主题分布概率矩阵作为特征值,利用逻辑回归进行 0-1 分类
4. 取逻辑回归的 *probability* 值构成集合$\{prob[0],prob[1]\}$,其中$prob[0]+prob[1]=1$
5. 根据算法 2 选定阈值x_0和x_1
6. WHILE $\hat{c}\in\{\hat{c}\}$,IF $prob[0]>x_0$,调节结果为 0;IF $prob[1]>x_1$,调节结果为 1,得到最终调节结果为 new$\{\hat{c}\}$

算法 2. 结果调节层阈值x_0、x_1选择算法.

输入:验证集的未调节结果集合$\{c_1,c_2,\cdots,c_t\}$

输出:调节层阈值x_0、x_1

1. 获取原模型验证集的结果集合$\{c_1,c_2,\cdots,c_t\}$
2. 取$x_0\in(0.5,1)$,步长为 0.01,依次遍历
3. 利用算法 1 中的方法计算逻辑回归的 *probability* 值集合$\{prob[0],prob[1]\}$
4. IF $prob_i[0]>x_0$,调节$c_i=0$,更新结果集合$\{c_1,c_2,\cdots,c_t\}$,计算新的精度值 *accuracy*
5. 在遍历中循环步骤 3、4,遍历结束后,获取精度值集合$\{accuracy\}$
6. 取精度最大时的x_0作为最终取值
7. 对于x_1的取值方法和x_0相同,采用控制变量的方法对x_0、x_1进行分别计算

本文使用的最终模型如图 7 所示,输入的数据集中的诗句要首先经过词嵌入层为诗句中的每个字

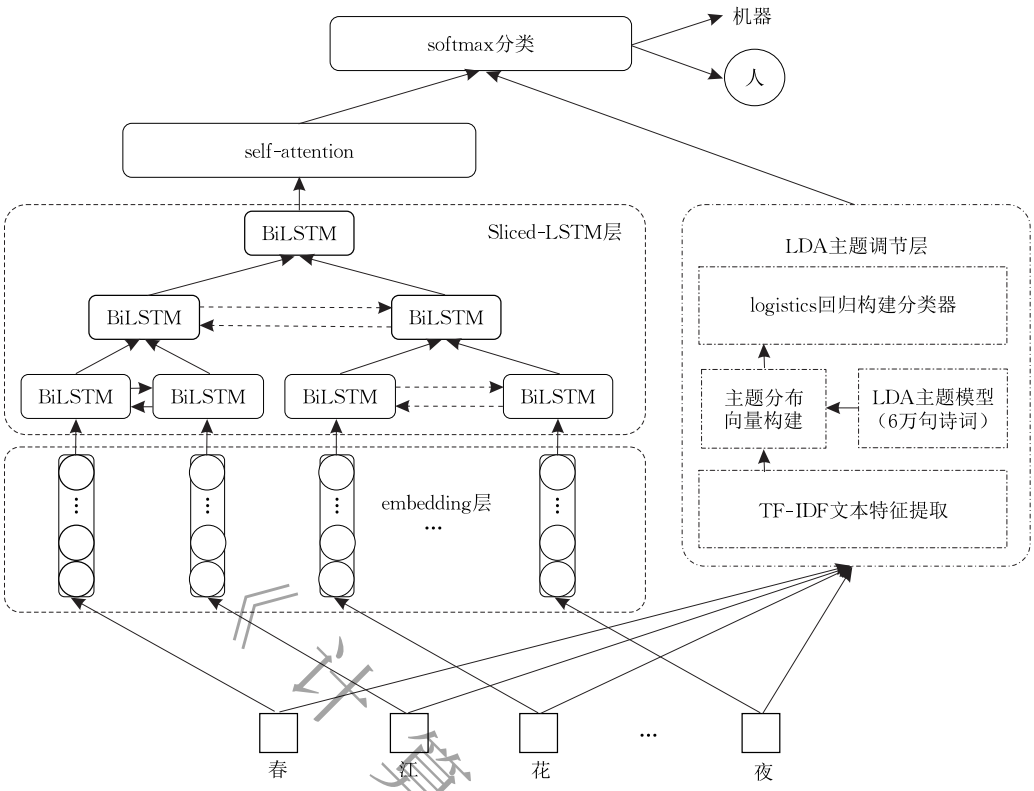


图 7　诗词识别模型图

生成 300 维的向量，将嵌入后的诗句输入三层切片 LSTM 模型中进行训练，而且再输入一层自注意力机制层进行注意力加权，而后将得到的向量利用 Softmax 分类得出一组分类结果，最后再通过 LDA 主题调节模型对分类的结果进行调节，模型输出调节后的结果作为测试的最终的结果.

4　实　验

4.1　评价指标

选用的主要评价指标包括精度(accuracy)、准确率(precision)和召回率(recall). 其中，精度为所有测试结果中机器判断正确的比例，用来判断模型整体的好坏，准确率用来评价查找出的结果是否准确，召回率用来判断查找覆盖范围是否全面.

4.2　数据集构成

本文使用的数据集主要包括古代诗词数据集和机器生成数据集两个模块.

在选取古代诗人所写的诗词时，选取古诗文网中的名句模块为本文古代诗词数据集，并根据诗词主题的不同，将诗词初步分为写景和抒情两个大类，构建两组数据集，每一组数据包含 3500 句诗词名句，不排除两类中有交集部分.

在构建机器生成数据集时，本文选用目前常用的基于 LSTM 的古诗生成算法，选用了包含三万首唐诗的数据集作为训练集供机器进行学习，通过不断对输入文本的下一个字进行预测，并进行迭代训练，构建诗词生成模型. 考虑到训练次数不同的模型的生成效果差异，对于每一组实验，分别选取迭代轮次为 5，15，25，35，45，55，65，75 的共 8 个模型进行生成.

为了保证机器生成的诗词也可以按照写景与抒情两部分进行划分，且诗词的生成风格一致，因此，在生成每一首机器生成的诗词时，随机选取古代诗词数据集中的一句作为机器生成的起始句，利用在起始句后机器所作的诗句来构建数据集. 对于 8 组生成模型，利用每个模型构建两大类机器生成数据集，总计有 16 组机器生成诗词数据，每组数据和古代诗人数据保持同样大小，包含 3500～4000 句机器生成诗词. 在训练机器生成数据集时发现，随着迭代次数的增加，机器生成的诗词也在不断完善，例如，在迭代次数为 5 时，机器生成的诗歌可能会使用一些较为生僻的字词，且有少数标点符号使用不恰当的情况，例如“隔，岁寻人起还书.”但也有生成的稍好的诗词，如“峰峦无一鹤，残月满诗帷.”已经初步达到了以假乱真的地步. 而在迭代次数为 75 时，已经可以生成表面看起来比较合理的古诗，例如“霞下波烟外，溪边风水流.”这一类古诗对于普通人来讲

已经造成了识别上的困扰，人类难以迅速通过字句判断诗词的来源是来自机器还是来自真正的诗人.表 1 和表 2 分别为从本文使用的写景和抒情数据集中随机抽取的部分诗句样例.

表 1 写景诗词数据集举例

诗人写景诗举例	朔风吹雪透刀瘢，饮马长城窟更寒.[18] 两岸青山相对出，孤帆一片日边来.[18] 又有墙头千叶桃，风动落花红簌簌.[18] 云间连下榻，天上接行杯.[18] 又是春将暮，无语对斜阳.[18] 黄云连白草，万里有无间.[18]
机器生成写景诗举例	来处山亭春雪落，远擎山门好看花. 津上渡头春色湿，花开千里路青青. 月明漪雨急风引，云烟晖莽带烟树. 潮残秋夜月，疏杏青帆树. 乱不起朱翠，菡萏委颜色. 霞下波烟外，溪边风水流.

表 2 抒情诗词数据集举例

诗人抒情诗举例	嵩云秦树久离居，双鲤迢迢一纸书.[18] 撩乱边愁听不尽，高高秋月照长城.[18] 正见空江明月来，云水苍茫失江路.[18] 芳草已云暮，故人殊未来.[18] 明朝望乡处，应见陇头梅.[18] 才饮黄河水，又食酸汤鱼.
机器生成抒情诗举例	春风何处思乡市，水阔松篁路客还. 攀折带云芳草色，行人抚泪日中难. 借问风浪言别有，杜陵在泪怜幽红. 迢默居何处，平生处绝境. 我心肠可断，泪痕孤骨中. 青云皆俨动，万丈恨蹉跎.

考虑到机器生成的诗词在情感的强度方面可能和人存在较大的差异，因此选用情感激烈的爱国诗词作为举例，进一步说明镜像图灵测试问题，构建爱国诗词数据集.由于作者难以从大量诗词数据集中主观判定爱国诗词，且在网络资源中的爱国诗词总结数据量较少，因此根据表 3 中的 8 部爱国诗词书籍，人工整理古代诗人爱国数据集 3500 条.并选用上述的诗词生成方法，利用 8 个不同迭代次数的诗歌生成模型，生成同等数据量的爱国诗句作共 8 组作为机器生成数据集.表 4 为从爱国诗词数据集中随机抽取的部分诗句样例，包括诗人所作和机器生成两个部分.

表 3 爱国诗词来源

书名	作者	出版日期
《台湾爱国诗词选》	陈碧笙	1981 年 4 月
《历代爱国诗词选》	薛绥之等	1984 年 4 月
《爱国诗词选讲》	尹贤	1984 年 10 月
《历代爱国诗文选》	左振坤等	1985 年 4 月
《爱国诗华》	木之青等	1986 年 9 月
《中国历代爱国诗词精品》	王启兴等	1994 年 12 月
《历代爱国诗词鉴赏》	蒋学浚	2001 年 12 月
《爱国诗百首》	朱强娣	2010 年 4 月

表 4 爱国诗词数据集举例

诗人爱国诗举例	伏波惟愿裹尸还，定远何须生入关.[19] 弓背霞明剑照霜，秋风走马出咸阳.[20] 半壁江山余涕泪，百年身世感飘零.[21] 疾风冲塞起，沙砾自飘扬.[22] 飞鹊来何意，英雄此日生.[23] 家国嗟何在，乾坤渺一身.[24]
机器生成爱国诗举例	好归相逐随天子，郡指闲庭鬓已添. 七百行军死应接，五胡城汉正难休. 关山乱作相思梦，花落春园溜马尘. 落花过白露，故国到明年. 故河千里雨，蟪蛄北南风. 今日向阳里，南朝飞武燕.

因此，本文共构建了包括写景、抒情与爱国三类诗词在内的机器生成及诗人创作数据集，其中每一类机器生成诗词包含 8 组数据集，在后续的实验过程中，利用“镜像图灵测试”的方法，对已经初步通过图灵测试的诗歌生成数据集进行识别.

4.3 实验过程及结果分析

在实验的预处理环节，考虑到本文所做的识别是对诗句进行识别，因此首先对数据集中的诗词按句进行切分，并去掉过长或者过短的诗句，保持诗句的长度在 8 个字到 64 个字之间.在进行分词时，由于诗词中每一个字都有其独自的含义，且分词错误可能导致对诗句的理解出现偏差，故对诗句按字进行分词，考虑到标点符号可能会影响诗词的情感表达，因此保留原诗中的标点符号.

在数据集的划分方面，由于每一组数据集中的机器和人创作的诗所占比例均等，因此在本文中不存在正负例比例失衡的情况，按照 1∶1∶1 的比例划分训练集、验证集和测试集，利用训练集进行模型训练，利用验证集进行参数调节，利用测试集得到最终的分类结果.

在本文的实验部分，主要采用了四种识别模型，模型一为切片双向 LSTM 模型、模型二为自注意力模型、模型三为添加一层自注意力层的切片双向 LSTM 模型、模型四为在模型三的基础上经过 LDA 调节层调节后的模型.

对于切片双向 LSTM 模型的切片层数的选取，分别对模型进行 1，2，3 次切片，在全部写景数据集中随机抽取 3000 句机器生成诗句和 3000 句诗人写作诗句进行预实验.实验结果见图 8，三组测试的结果相差不大，但是当切片次数为 2 时，精度上升速度最快，精度最高，且波动较小，最先进入稳定状态，因此选择切片次数为 2 的三层网络构建后续模型，考虑到已固定输入诗句的长短在[8，64]内，因此取切片次数 k 为 2，第一次切片长度 m 为 4.

在正式镜像图灵测试之前，对于每一个模型在每个数据集上均进行 300 轮的迭代，图 9 为切片

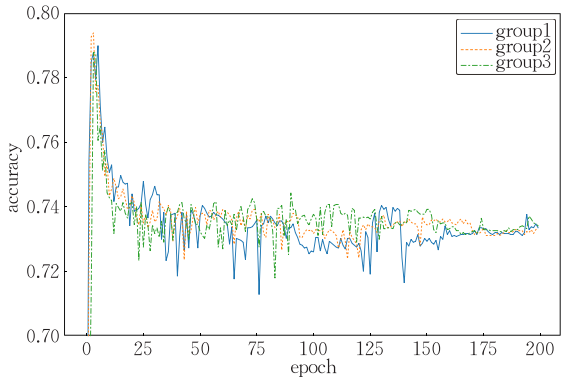


图 8　切片次数为 1,2,3 次的实验结果

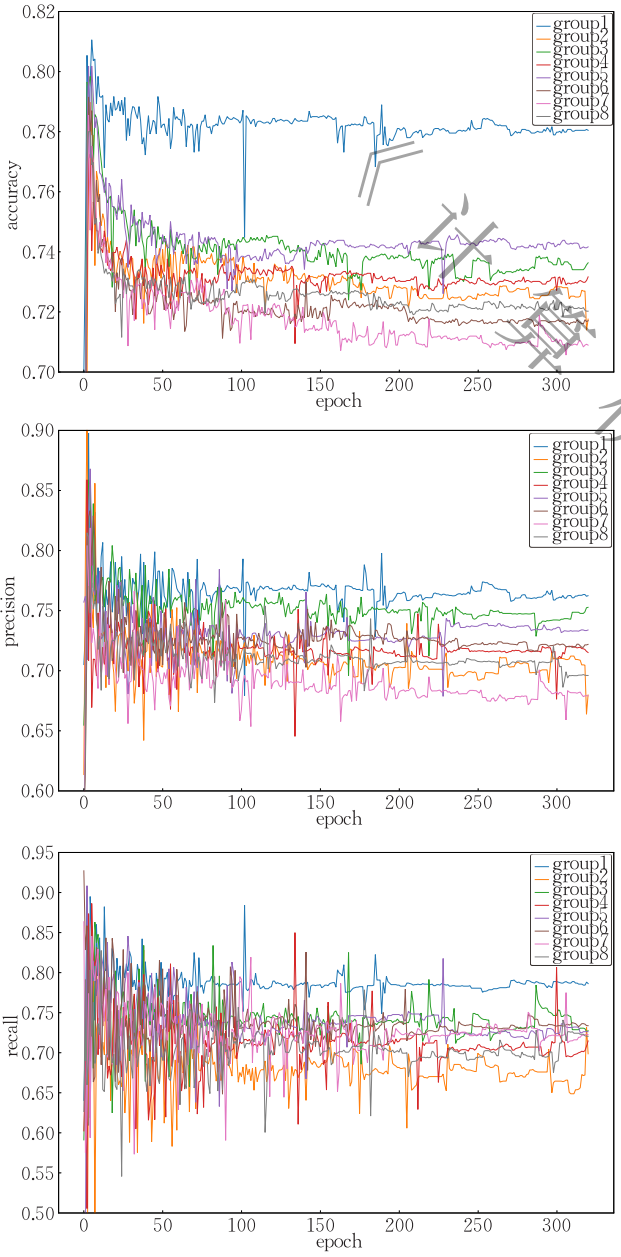


图 9　模型一在 8 组数据集上的实验结果

LSTM 模型在 8 组写景数据集上的不同迭代次数下精度、准确率和召回率折线图，图 10 表示了使用添加自注意力层的切片 LSTM 模型在 8 组写景数据集上的结果图. 整体来看，除第一组数据集的识别精度比其它数据集的识别精度高以外，其它数据集的识别精度相差不大，基本不超过 0.03.

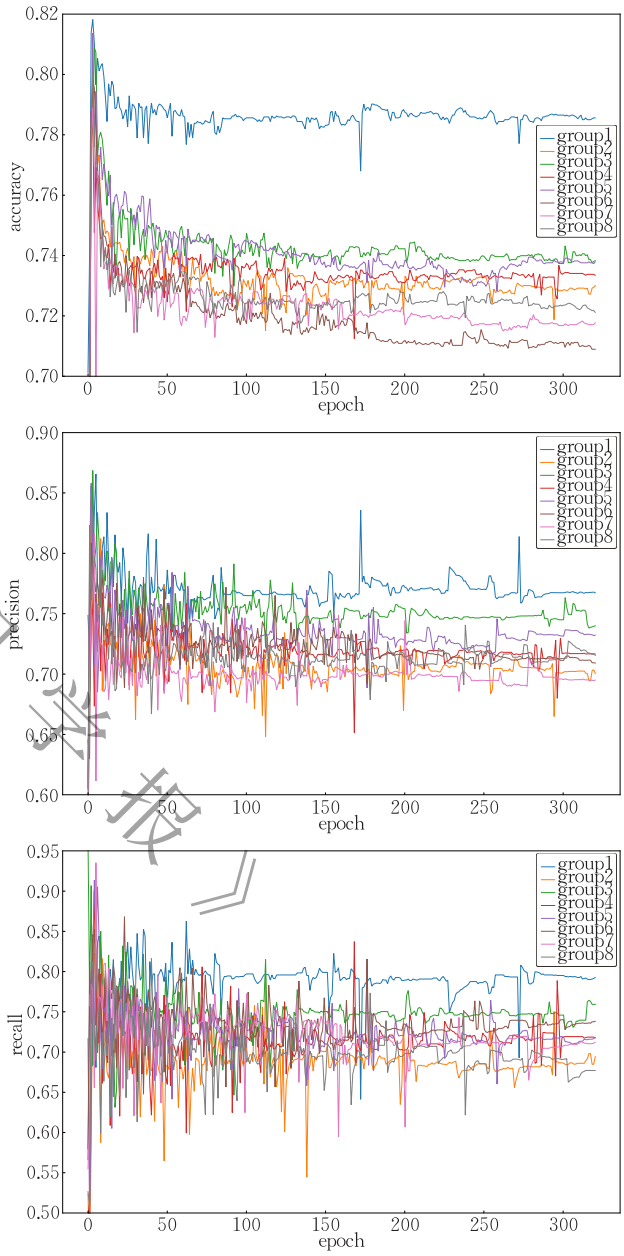


图 10　模型三在 8 组数据集上的实验结果

可以发现，随着迭代次数的变化，精度、准确率和召回率都在不断波动，其中，精度会随迭代次数的增加不断上升，而后再次回落，最终当迭代次数达到一定程度后，逐渐趋于稳定，且波动范围较小. 对于两种模型而言，模型三比模型一的精度更早达到最

大值，也更早进入稳定状态.准确率的波动幅度比精度稍大，整体波动趋势与精度类似，召回率波动幅度最大，可以达到10%左右，最终也逐渐趋于稳定，但是前期波动没有一个固定规律，对于不同的数据集，其波动情况也不相同，在精度已经稳定时，召回率依然可能出现较大范围的波动.

对于四个不同的识别模型在写景类别诗歌的测试集上的具体识别结果，取精度最大时的精度、准确率与召回率作为结果数据，表5中的结果为多次实验后获取的平均值.其中对于8个不同的模型，精度、准确率和召回率的最高值均加粗表示.图9中从左到右分别为不同模型在8组写景数据集上的最大精度、准确率和召回率折线图，四条折线分别表示模型一到模型四的结果曲线.表5中的结果显示，对于写景数据集，本文提出的模型对于机器和人所作的诗的识别精度达到了80%左右，识别的准确率和召回率也达到了一定的高度，整体识别效果较好，甚至可以超过普通人类的识别水平.同时，在利用机器进行识别时发现，若输入的诗句均为人写的或机器写的，机器仍然可以做出有效的判断.

表5 写景诗词在四种模型上的识别结果

		1	2	3	4	5	6	7	8
模型一：Sliced-BiLSTM	精度	0.8105968	0.7872100	0.7984319	0.7725360	0.8018721	0.7915714	0.7854776	0.7800481
	准确率	0.7790834	0.7442681	**0.8535688**	**0.8000000**	**0.8327219**	**0.8282123**	0.7708999	**0.8113349**
	召回率	0.8435115	0.7997473	0.7351077	0.6892151	0.7360571	0.7515843	0.7738629	0.6943204
模型二：Self-Attention	精度	0.8057247	0.7778721	0.7932048	0.7812500	0.7946958	0.7775237	0.7772853	0.7800481
	准确率	0.7659453	0.7522124	0.7954972	0.7613707	0.7942743	0.7777090	0.7382716	0.7667732
	召回率	**0.8555980**	0.7517372	0.8060837	0.7798341	0.7736706	0.7959442	0.7588832	0.7657945
模型三：SBiLSTM+SAtt	精度	0.8182095	0.7954159	0.8088860	0.7878606	0.8137286	0.7981052	0.7860801	0.8001803
	准确率	0.7917415	0.7553444	0.7922307	0.7702448	0.8000000	0.7807018	0.7711918	0.7743309
	召回率	0.8416030	**0.8035376**	**0.8529785**	0.7830249	**0.8171206**	**0.8460076**	0.7751442	**0.8123804**
模型四：LDA调节后的模型	精度	**0.8185140**	**0.7959819**	**0.8108461**	**0.7881611**	**0.8140406**	**0.7994120**	**0.7863810**	**0.8016827**
	准确率	**0.7918660**	**0.7941580**	0.8130019	0.7874294	0.8138142	0.8019671	**0.7856116**	0.8012891
	召回率	0.8190099	0.7966940	0.8094155	**0.7878078**	0.8141762	0.7978180	**0.7857877**	0.8022368

从图11中可以发现，添加了一层自注意力层的切片LSTM模型在精度上要高于单独的切片LSTM模型或自注意力模型，通过调节层的调节，精度有所提高，但整体提高程度不大.对于准确率而言，调节后的模型四的准确率比未经调节模型三的结果更好，即可以认为，模型四相比模型三，保证了在整体精度没有下降的情况下，对准确率进行提升，从而在一定程度上提高了模型的识别水平.

对于其它的两种模型，单独的切片LSTM模型在准确率上有着更好的表现，自注意力模型的准确率较低.从召回率的角度来看，模型三和模型四的召回率整体处于较高的水平，切片LSTM模型的召回率波动较大，受数据集的影响较高，对于第四个数据集，切片LSTM模型的召回率低于70%.对比四个模型在不同数据集上的表现，从精度来看，四个模型基本呈现了相同的波动趋势，在第1,3,5,8个数据集中精度较高，其余数据集中精度较低，由此认为对于识别模型来讲，机器生成数据的不同没有对识别模型的识别结果造成较大的影响，也可以认为识别模型在这一层面上是高于生成模型的.对于准确率，不同模型的波动趋势也大体类似，且和精度的波动趋势类似，但在波动幅度上存在一定差异，模型一的波动幅度相对较大，其余模型波动幅度较小.召回率在不同数据集上的波动幅度比准确率更大，尤其是切片LSTM模型.

表6为四种识别模型在抒情数据集上的结果表现，从表中数据可以发现，对于抒情数据集的识别的准确率在77%左右，整体的识别效果略低于写景数据集，由于诗词中的抒情类别包含思乡、羁旅、爱情、家国情怀等多种类别，甚至有借景抒情的诗词，例如“春风又绿江南岸，明月何时照我还？”表面虽为写景，但却通过景色描写抒发了诗人王安石对于前途的担忧，因此，对于抒情类的诗词，即便是人类也对诗句中所表达的情感也难以界定，需要结合诗人的生平对诗词的情感进行分析.假若对于一句来自不明作者的陌生的抒情诗句，例如“虑澹物自轻，意惬理无违.”如果让人来判断诗词表达的情感以及是否出自真正的诗人，也是比较困难的.因此，对于机器写的抒情诗，若情感表达不够明显，人类也难以辨析其中的情感，很有可能误认为是诗人所作.例如“别后天乡去，重来暮雪红.”、“月自泛酒醒，碧云诗共忧.”这一类机器生成的诗句，从字面来看，语句连贯，上下文也比较通顺，可以认为是比较好的借景抒情的诗句，可以通过用词理解诗句为表达“忧愁”的情感，却不能读出是因何而发的忧愁，由于没有诗人的经历作为依据，对于机器生成的抒情诗，无法结合

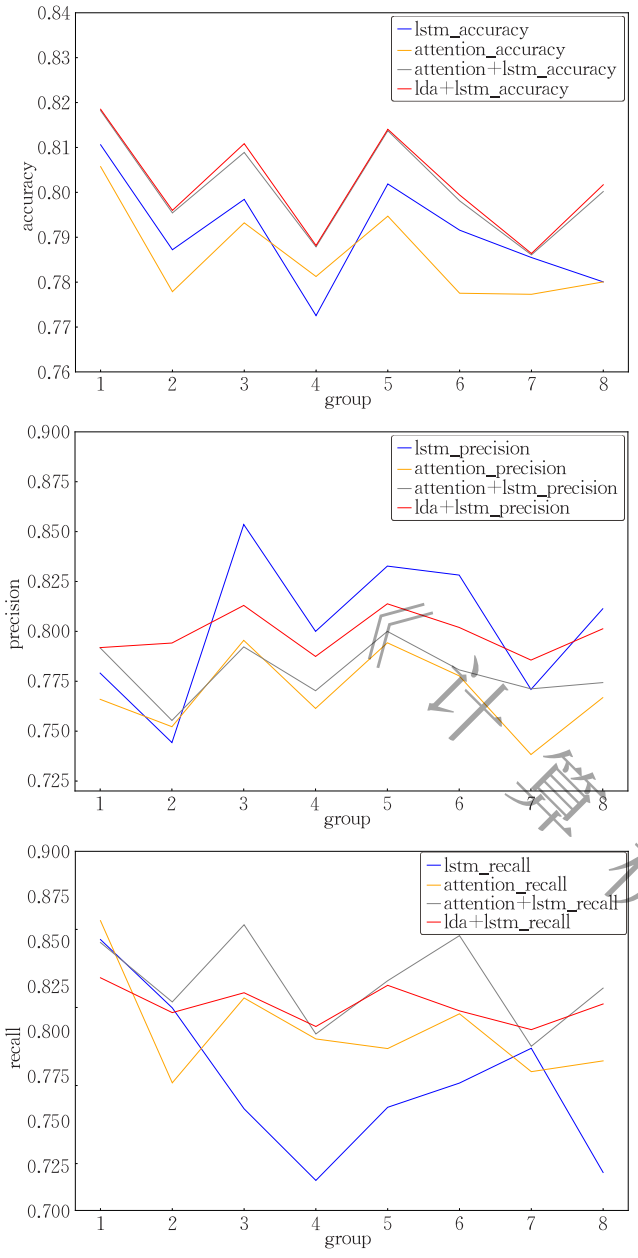


图 11　写景诗词识别中四种模型的精度、准确率和召回率折线图

诗人的经历去理解诗词，在诗中读不出故事、读不出温度. 也就是说机器生成的诗是缺少阅历，缺少灵魂的，而在利用机器进行识别时，对于情感的分析则更有难度，但是从表 6 中的结果整体来看，可以认为机器对于来自写诗机器和诗人的抒情诗句是可以识别的，且识别的效果要高于人类，也就是说，对于人类难以区分的诗词，机器是可以进行识别的，即通过了图灵测试的诗歌生成机器无法通过以机器作为识别器的镜像图灵测试. 当冰冷的诗歌生成器遇见同样冰冷的诗歌识别器，两者的较量避免了人类主观判断，在某种程度上是对现有图灵测试水平的提高.

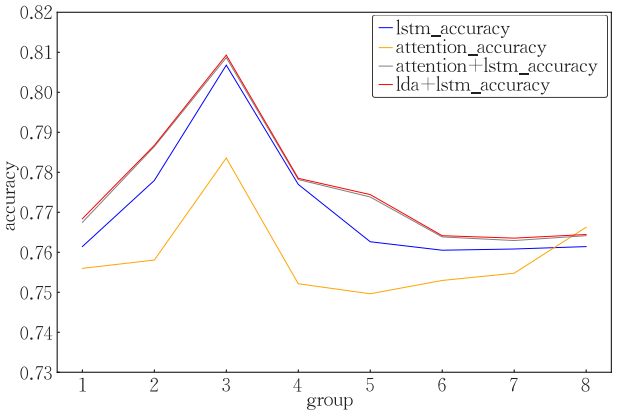
以林鸿飞教授的古诗“才饮黄河水，又食酸汤鱼”为例，原诗的后两句为“西南西北飞度，智能学一路.”①，而机器根据前一句诗生成的诗句为“故人多感慨，极目独伤情”，将这三句诗输入到镜像图灵测试机器中，测试机器认为这三句诗都来自机器. 测试机器能够识别出“故人多感慨，极目独伤情”这句诗为机器所写，即古诗生成机器并没有瞒过测试机器，但是测试机器却误将林鸿飞教授的另外两句诗也判定为机器所作.

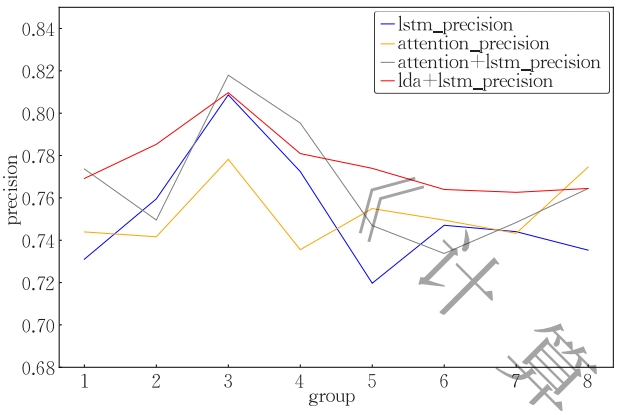
图 12 可以更加直观地观察本文所使用的三种模型在抒情类诗词的 8 组数据集上的识别测试结果，从左到右依次为四个模型在 8 组抒情数据集上的精度、准确率和召回率表现，模型一到模型四分别用四条线条表示. 从精度来看，四种模型均在第三个数据集上有着更好的识别效果，即对于抒情诗词，当迭代次数为 25 时，机器生成的诗词对于镜像识别器来讲更容易识别，但是模型在不同数据集上的识别效果差距不大，整体也呈现一个波动的趋势，说明随着迭代次数的不同，生成器的水平也在不断波动. 从

表 6　抒情诗词在四种模型上的识别结果

		1	2	3	4	5	6	7	8
模型一：Sliced-BiLSTM	精度	0.7614152	0.7779403	0.8067754	0.7769697	0.7626247	0.7605080	0.7608104	0.7614152
	准确率	0.7310261	0.7595270	0.8086501	0.7724510	0.7197014	0.7470665	0.7440515	0.7353123
	召回率	**0.7767742**	0.7462879	0.7920411	0.7414128	**0.8083871**	0.7393549	0.7464516	**0.7670968**
模型二：Self-Attention	精度	0.7559722	0.7580456	0.7835634	0.7521212	0.7496220	0.7529483	0.7547627	0.7662534
	准确率	0.7439265	0.7416332	0.7782052	0.7355425	0.7549435	0.7494894	0.7432522	**0.7745583**
	召回率	0.7309678	0.7153002	0.7792041	0.7336358	0.6896774	0.7103226	0.7283871	0.7079968
模型三：SBiLSTM+SAtt	精度	0.7674629	0.7864248	0.8086575	0.7781818	0.7738131	0.7638343	0.7629271	0.7641367
	准确率	**0.7736510**	0.7495429	**0.8179437**	**0.7953387**	0.7469212	0.7337386	0.7483788	0.7644231
	召回率	0.7122580	**0.7940607**	0.7830552	0.7077122	0.7825807	**0.7787097**	0.7445161	0.7180645
模型四：LDA 调节后的模型	精度	**0.7683701**	**0.7867174**	**0.8092848**	**0.7784848**	**0.7744179**	**0.7641367**	**0.7635319**	**0.7644391**
	准确率	0.7691592	**0.7852925**	0.8096959	0.7809214	**0.7739467**	**0.7639656**	**0.7626210**	0.7644545
	召回率	0.7651027	0.7873460	**0.8087764**	**0.7742143**	0.7749368	0.7649951	**0.7624117**	0.7617453

① 林鸿飞. 微信朋友圈. 2019-08-07 09:58





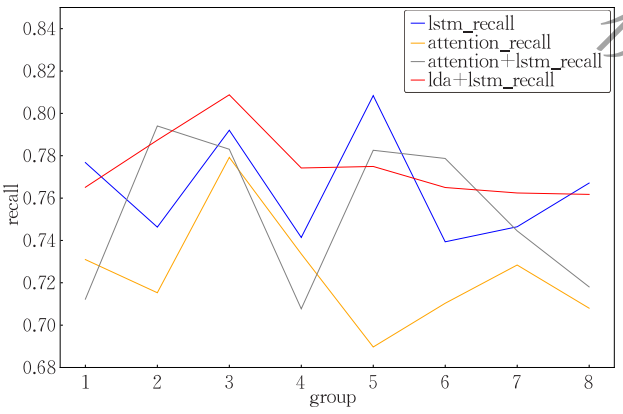


图 12 四种模型在抒情数据集上的精度、准确率、召回率折线图

不同模型的识别效果来看，综合了模型一和二的模型三在识别精度上整体较高，模型四对于模型三的精度提升不明显，但是对于准确率，模型四整体处于较高的水平，相比模型三在部分数据集上有了一定程度提升，同时在波动程度上有所降低，识别效果更加稳定. 在其它三个模型中，模型三的精度相对较高一点，但是波动较大，模型一波动最大，模型二的波动较小，但是整体的精度水平不高. 对于召回率，除模型四以外，四个模型的召回率在不同数据集上的波动较大，模型四和模型一整体召回率水平较高.

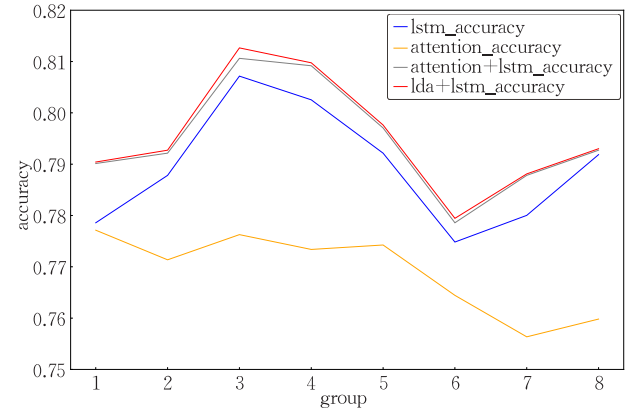
表 7 中的结果为本文使用的四种模型在爱国诗词数据集上的识别结果. 图 13 为四种模型在 8 组爱国数据集中的精度、准确率和召回率波动曲线. 从整体来看，识别的精度在 80% 左右，即测试机器可以较好地区分来自机器和人的诗歌，机器生成的爱国类诗词并没有通过镜像图灵测试.

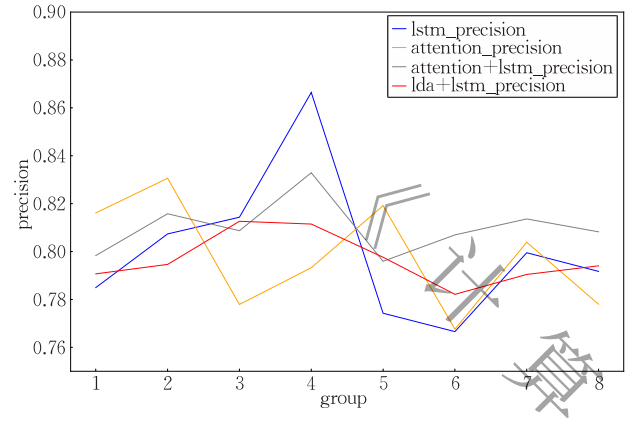
相比较抒情类诗词，爱国类诗词将诗句中的情感细化，只考虑爱国这一种，对于来自真正诗人的爱国诗，其中蕴含的情感普遍较激烈，因为古代诗人常常从战争、生死以及亡国这类事件中抒发自己对于国家的热爱，而机器生成的诗歌只能模仿出表面的场景，难以生成蕴含哲理的句子，例如“人生自古谁无死？留取丹心照汗青.”这类即表达了诗人的爱国情怀又富有哲理的句子机器则无法生成.

从四种模型的识别精度上来看，在第三组数据集上的识别效果最好，在第 6 组数据集上的识别效果较差，说明对于镜像图灵测试机生成模型迭代次数为 55 时生成的诗歌最难识别. 对于四种不同的模型，模型二识别精度较低，模型三、四识别精度较高. 准确率和召回率的波动依旧较大，模型四的波动相对平稳. 从召回率来看，模型三、四的召回率整体水

表 7 爱国诗词在四种模型上的识别结果

		1	2	3	4	5	6	7	8
模型一：Sliced-BiLSTM	精度	0.7785796	0.7878175	0.8071594	0.8025404	0.7921478	0.7748268	0.7800231	0.7918592
	准确率	0.7849463	0.8073566	**0.8143284**	**0.8664311**	0.7742466	0.7665722	0.7995003	0.7917153
	召回率	0.7635096	0.7524695	0.7925625	0.7123765	**0.8210343**	**0.7861708**	0.7437536	0.7884951
模型二：Self-Attention	精度	0.7771363	0.7713626	0.7762702	0.7733834	0.7742494	0.7644342	0.7563511	0.7598152
	准确率	**0.8161226**	**0.8306050**	0.7779083	0.7932331	**0.8191707**	0.7675932	0.8038808	0.7779862
	召回率	0.7117954	0.6780942	0.7693202	0.7356188	0.7001743	0.7542127	0.6740267	0.7228355
模型三：SBiLSTM+SAtt	精度	0.7901270	0.7921478	0.8106236	0.8091801	0.7970554	0.7785797	0.7878175	0.7927252
	准确率	0.7983193	0.8157729	0.8086957	0.8329146	0.7959302	**0.8069498**	**0.8136132**	**0.8082360**
	召回率	0.7728065	0.7513074	0.8105753	0.7704823	0.7954677	0.7286462	0.7431726	0.7640907
模型四：LDA 调节后的模型	精度	**0.7904157**	**0.7927252**	**0.8126597**	**0.8097575**	**0.7976328**	**0.7794457**	**0.7881062**	**0.7930139**
	准确率	0.7906848	0.7946325	0.8126364	0.8114819	0.7976245	0.7821310	0.7904396	0.7940381
	召回率	**0.7903119**	**0.7924637**	**0.8126386**	**0.8095170**	0.7976301	0.7791288	**0.7878153**	**0.7928200**





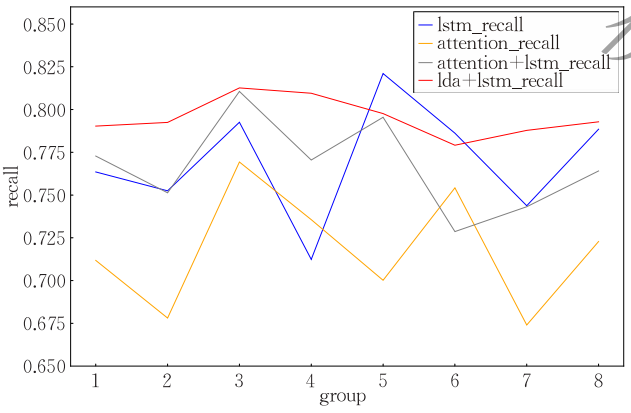


图 13　四种模型在爱国数据集上的精度、准确率和召回率折线图

平较高，单层自注意力机制模型在识别中效果不如其它三种模型.

考虑到本文所用的生成诗歌的方法存在一定的局限性，因此，为提高机器生成的诗歌的质量，本文还通过“九歌”网站在线生成了 3500 句诗句，在生成时随机输入诗歌常用词作为主题词，并对其进行了识别. 图 14 为利用模型三对这 3500 句诗歌的识别结果，识别精度可达 85%左右. 所以说明了，即便是对于“九歌”网站生成的较为成熟的诗歌来讲，机器依旧可以完成识别，也更进一步证明了本文的实验结果，即机器生成的诗歌即使通过了图灵测试，也是难以通过镜像图灵测试的.

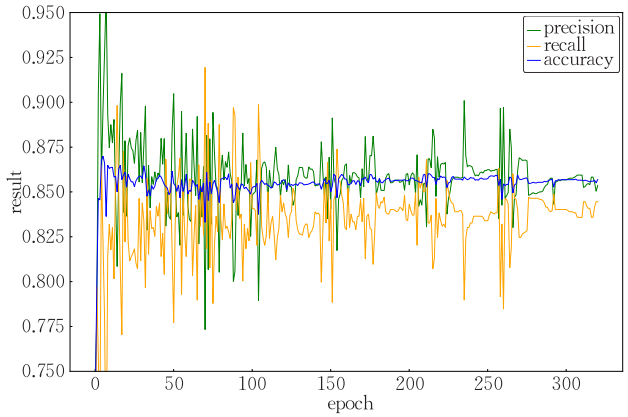


图 14　“九歌”生成诗歌的识别结果

图 15 为整体识别效果最好的模型四在三类数据集上的识别结果平均值，从整体水平上来看，写景数据集的识别效果最好，其次是爱国数据集，抒情数据集的识别效果相对较差. 三类数据集的识别准确率精度均在 77%以上，可以说明对于三类数据集，镜像图灵测试机器均可以进行有效的识别，即三类机器生成的数据集都未能通过镜像图灵测试.

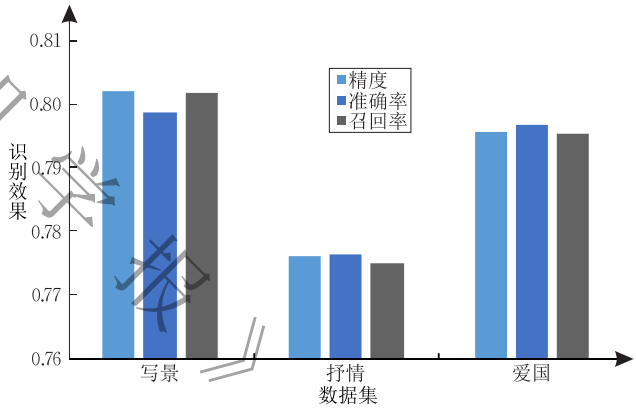


图 15　三类数据集识别结果平均值

对于镜像图灵测试问题，识别效果的好坏不仅与生成诗歌的好坏相关，与测试机器的水平也密切相关，若假定三类诗歌的生成水平一致，则可以说明测试机器对于写景类诗歌的分辨更为擅长，不擅长对于抒情类诗歌的分辨，也意味着当诗句中蕴含了情感并且情感更丰富的时候，会给机器的测试带来更多的困难. 诗歌背后隐藏的是诗人的灵魂，这无论如何是目前的机器无法模仿的，机器情感研究是目前人工智能研究最远的一个前沿阵地.

5　结　论

本文提出了镜像图灵测试这一概念，以不同类型的人写的诗词和机器写的诗词为实验对象，进行

了机器识别测试，其中，对于普通人来说，机器生成的诗歌是难以识别的，但是通过本文的测试发现，机器生成的诗歌并未通过镜像图灵测试，即机器可以对人写的和机器写的诗歌进行测试. 当然，在本文进行的实验中存在一定的局限性，即机器生成的诗歌或许可以有更高的质量，在本文中采用了较为成熟的基于深度学习的文本生成方式去进行诗歌的生成，采用了 LSTM 模型生成文本，同时，还采用了一组“九歌”网站在线生成的诗歌进行了测试.

在未来的研究中，可以考虑采用更多其它的方式去进行文本的生成，例如强化学习模型或者对抗学习模型，或者采用成熟的诗歌生成系统，例如九歌、乐府等，而同理，对于识别器的模型也可以进行相应的提升. 在算法的提升方面，实验是永无止境的，但是镜像图灵测试这一思想却是普适性的，镜像图灵测试或许可以作为人工智能自我突破的又一标准在更多人工智能相关领域得到应用. 作者在未来的研究中将继续致力于镜像图灵测试的理论证明. 在本文中，所研究的问题只涉及了单轮的镜像图灵测试，在未来的研究中，还可以考虑进行有多轮交互的镜像图灵测试问题，对镜像图灵测试的思想进行进一步的拓展.

人类写的诗词是有血有肉的，是有灵魂的. 而通常意义下的图灵测试是无法对灵魂进行探测的，但是写的诗能体会出来，特别是根据上下文环境，因为语言是人脑智慧最突出的体现，自然语言处理是人工智能皇冠上的明珠，诗歌是人脑情感最突出的反应，是诗人全身心的投入后的灵魂映射，是诗人心灵的影子，所以在一定意义上是图灵可测的，或者说，如果存在图灵可测的不完备性，那么诗歌这个人类语言的精华所在，就是突破这个图灵不完备性的关隘.

就图灵测试而言，事实上，一旦众多的机器通过了图灵测试，达到了人类级别的智能，人类社会肯定会发生巨大改变了. 合理利用，可以加速人类社会的发展；利用不好，可能会对人类造成毁灭性灾难. 镜像图灵测试问题是测试机器与被测试机器之间的博弈，人类可以通过不断提升测试机器从而提升镜像图灵测试的要求. 而通过了镜像图灵测试的计算机，超过了人类智能，实际上是对上述的逻辑进行了“平方”式升级，所以它既可以控制和避免这个灾难，也可以使这个灾难进行“平方”操作，因而提前认知它是必要的.

参 考 文 献

[1] Zhou C, You W, Ding X-J. Genetic algorithmand its implementation of automatic generation of Chinese SONGCI. Journal of Software, 2010, 21(3): 427-437(in Chinese)
(周昌乐，游维，丁晓君. 一种宋词自动生成的遗传算法及其机器实现. 软件学报，2010，21(3)：427-437)

[2] Yan R, Jiang H, Lapata M, et al. Automatic Chinese poetry composition through a generative summarization framework under constrained optimization//Proceedings of the International Joint Conference on Artificial Intelligence. Beijing, China, 2013: 2197-2203

[3] He J, Zhou M, Jiang L. Generating Chinese classical poems with statistical machine translation models//Proceedings of the AAAI Conference on Artificial Intelligence. Toronto, 2013: 1650-1656

[4] Zhang X, Lapata M. Chinese poetry generation with recurrent neural networks//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 670-680

[5] Yi X, Sun M, Li R, Yang Z. Chinese poetry generation with a working memory model//Proceedings of the International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 4553-4559

[6] Yu Jian. Language and Turing test. ACTA Automatica Sinica, 2016, 42(5): 668-669(in Chinese)
(于剑. 语言与图灵测试. 自动化学报，2016，42(5)：668-669)

[7] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[8] Schuster M, Paliwal K. Bidirections recurrent neural networks. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681

[9] Yu Z, Liu G. Sliced recurrent neural networks//Proceedings of the International Conference on Computational Linguistics. Santa Fe, USA, 2018

[10] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate//Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015

[11] Shimaoka S, Stenetorp P, et al. An attentive neural architecture for fine-grained entity type classification//Proceedings of the 5th Workshop on Automated Knowledge Base Construction. San Diego, USA, 2016: 69-74

[12] Tang D, Qin B, Liu T. Aspect level sentiment classification with deep memory network//Proceedings of the 21st Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016: 214-224

[13] Xin J, Lin Y, Liu Z, Sun M. Improving neural fine-grained entity typing with knowledge attention//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 5997-6004

[14] Chen P, Sun Z, Bing L, et al. Recurrent attention network on memory for aspect sentiment analysis//Proceedings of the 22nd Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 452-461

[15] Liu Q, Zhang H, Zeng Y, et al. Content attention model for aspect based sentiment analysis//Proceedings of the 27th World Wide Web Conference on World Wide Web. Lyon, France, 2018: 1023-1032

[16] Zeng Yi-Fu, Lan Tian, Wu Zu-Feng, Liu Qiao. Bi-memory based attention model for aspect level sentiment classification. Chinese Journal of Computers, 2019, 42(8): 1845-1857(in Chinese)
(曾义夫，蓝天，吴祖峰，刘峤. 基于双记忆注意力的方面级别情感分类模型. 计算机学报，2019，42(8)：1845-1857)

[17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the Advances in Neural Information Processing Systems. 2017: 6998-6008

[18] Ancient poetry website. https://so. gushiwen. org/mingju (in Chinese)
(古诗文网. https://so.gushiwen.org/mingju)

[19] Yin Xian. Selected Lectures on Patriotic Poems. Lanzhou: Gansu People's Publishing House, 1984(in Chinese)
(尹贤. 爱国诗词选讲. 兰州：甘肃人民出版社，1984)

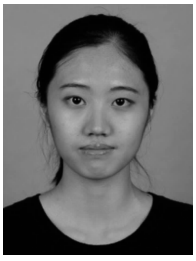
[20] Mu Zhi-Qing, Wang Zhao-Yu. Patriotic Poetry. Guangzhou: New Century Press, 1986(in Chinese)
(木之青，王钊宇. 爱国诗华. 广州：新世纪出版社，1986)

[21] Chen Bi-Sheng. Selection of patriotic poems in Taiwan. Xiamen: Taiwan Institute of Xiamen University, 1981(in Chinese)
(陈碧笙. 台湾爱国诗词选. 厦门：厦门大学台湾研究所，1981)

[22] Zuo Zhen-Kun. Selected Works of Patriotic Poems of Past Dynasties. Tianjin: Tianjin People's Publishing House, 1985(in Chinese)
(左振坤. 历代爱国诗文选. 天津：天津人民出版社，1985)

[23] Zhu Qiang-Di. One Hundred Patriotic Poems. Hefei: Anhui Literature and Art Publishing House, 2010(in Chinese)
(朱强娣. 爱国诗百首. 合肥：安徽文艺出版社，2010)

[24] Jiang Xue-Jun. Appreciation of Patriotic Poems of Past Dynasties. Beijing: Petroleum Industry Press, 2001 (in Chinese)
(蒋学浚. 历代爱国诗词鉴赏. 北京：石油工业出版社，2001)



XUE Yang, M. S. candidate. Her research interests include natural language processing and data mining.

LIANG Xun, Ph. D., professor, Ph. D. supervisor. His research interests include neural networks, social computing and natural language processing.

ZHAO Dong-Yan, Ph. D., professor, Ph. D. supervisor. His research interests include natural language processing and knowledge graph.

DU Wei, Ph. D. Her research interests include social computing and natural language processing.

Background

With the continuous development of Chinese culture and thousands of years of splendid history, ancient poetry combines rich emotions, connotative souls and vivid forms perfectly, showing the power of Chinese language. "Natural language processing is the Pearl on the crown of artificial intelligence", and machine-generated language is the core embodiment of machine intelligence. Testing machine language is an important part of Turing Test. The ancient Chinese poetry generated by machine has passed Turing test preliminarily, and can deceive ordinary people.

The Turing Test was proposed by Alan Turing. It means that when the tester (person) and the tested one (a person and a computer) are separated by an entity that does not disclose information other than the test content, the computer will pass the test and be considered to have human intelligence if the human tester cannot determine whether the tested person is a person or a computer through some non-intelligent devices.

This paper puts forward the frame of "Mirror Turing Test". Its main design idea is to replace the tester in Turing test with a computer. The test machine is required to identify the tested person and computer under the same conditions of Turing Test. If the test machine cannot complete the identification of the tested person and machine, it is considered that the tested machine has passed the Mirror Turing Test. Considering that in the field of recognition, the ability of computer has surpassed that of human beings, it is more difficult for the poetry generator to pass the Mirror Turing Test. The experimental results show that the Mirror Turing Test machine can identify the poetry generated by the machine, that is, the poetry generated by the machine that has passed the Turing Test cannot pass the Mirror Turing Test, which shows that poetry, as the crystallization of human language civilization, is still difficult to be overtaken by the machine so far, in a certain sense, if there is the incompleteness of Turing measurable, then, the essence of poetry, the human language, is the first pass to break through.

This paper is supported by the National Social Science Foundation of China under Grant No. 18ZDA309, the National Natural Science Foundation of China under Grant No. 62072463, and the Natural Science Foundation of Beijing under Grant No. 4172032.