

基于递归神经网络的散文诗自动生成方法^①

姜力, 詹国华, 李志华

(杭州师范大学 信息科学与技术学院, 杭州 311121)

通讯作者: 姜力, E-mail: jiangli.blue@foxmail.com

摘要: 针对中文散文诗歌的自动生成, 提出一种基于循环神经网络的时序性文本生成方法. 通过现有语料库构建好一个词语集后, 首先给定若干关键词, 在聚类模型生成的词语集基础上进行关键词扩展生成首句. 在确定首句的基础上, 利用上下文模型对已生成内容进行压缩和上文特征获取, 最后将之前上下文内容传递给递归神经网络模型实现后续句子的生成. 该方法中首句生成的过程利用语言模型中的词汇集扩展, 并通过上下文模型获取关联实现上下句的映射关系. 本文采用 BLEU 自动评测方式和人工评测方式, 建立起较为标准的评测系统, 实验结果证实了该方法的有效性.

关键词: 深度学习; 递归神经网络; 卷积神经网络

引用格式: 姜力, 詹国华, 李志华. 基于递归神经网络的散文诗自动生成方法. 计算机系统应用, 2018, 27(8): 259-264. <http://www.c-s-a.org.cn/1003-3254/6318.html>

Automatic Generation Method of Prose Poem Based on Recurrent Neural Network

JIANG Li, ZHAN Guo-Hua, LI Zhi-Hua

(School of Information Science and Engineering, Hangzhou Normal University, Hangzhou 311121, China)

Abstract: Aiming at the automatic generation of poem, a temporal text generation method using recurrent neural network is proposed. After building a word set according to the existing corpus, a number of keywords is given, and the first sentence is generated by expanding the keywords based on the word set constructed by clustering model. On the basis of the first sentence determined, the generated content is compressed by the context model and the feature is extracted, and finally the content of the previous context is passed to the generation model to realize the subsequent sentence generation. In order to achieve the mapping between the upper and lower sentences, the first sentence of the process is a vocabulary expansion, and the context model can be a good grasp of the context. BLEU automatic evaluation method and manual evaluation method are used to establish a more standard evaluation system. The results approve the effectiveness of the method.

Key words: deep learning; recurrent neural network; convolution neural network

散文诗是一种现代文体, 兼有诗和散文特点的一种现代抒情文学体裁. 它融合了诗的表现性和散文描写性的某些特点. 它具有散文的外观和内涵, 给读者美和想象, 但不像诗歌那样分行和押韵, 保留了诗的意象和细节. 国内外诗歌生成的方法可被分类成基于模板

的生成方法、基于遗传算法的方法和基于实例推理的方法. 模板生成方法即给定一个模板, 在满足语法等约束下进行填词作诗. 基于遗传算法的方法, 结合遗传算法和评测模板, 根据语法等信息用遗传算法生成备选作品^[1], 代表系统有 POEVOLVE^[2]和 McGonnagall^[3].

① 基金项目: 浙江省自然科学基金 (LY17D060005)

Foundation item: Natural Science Foundation of Zhejiang Province (LY17D060005)

收稿时间: 2017-08-16; 修改时间: 2017-09-05; 采用时间: 2017-09-11; csa 在线出版时间: 2018-07-28

基于实例推理的方法,通过检索已有诗句根据用户的目标信息对已有诗句进行内容调整. ASPERA^[4]是此类方法系统的代表. 国内周昌乐等^[5]在宋词生成上的研究方法是在给定词牌与韵律模板基础上,用遗传算法来进行宋词的自动生成. He 等人^[6]则将统计机器翻译的方法应用到了格律诗自动生成上. 随着深度学习的发展,为了解决传统表示方法中字词表示缺乏语义依存关系的问题,构建语义单元向量表示的神经网络模型,该模型包含了针对语义单元局部上下文和全局上下文的语义神经网络^[7]. 另外,当前多个深度模型中,基于长短期记忆单元的递归神经网络模型因其有效利用序列数据中远距离依赖信息的能力,故被用于文本序列数据处理,并展示出了在挖掘文本序列语义信息任务上的强大能力^[8].

本文提出利用主题模型和深度学习下的递归神经网络方法来进行生成,使得诗句上下句实现关系映射. 首先获取到样本语料库,使用中文分词技术和主题模型方法,建立词汇集以及实现词的主题聚类. 然后,利用人为给定的关键词,通过首句语言模型根据首句结构选词实现首句的生成. 在获取到首句的基础之上,利用上下文模型进行句子向量的压缩,将压缩后的向量作为输入喂给递归神经网络进行训练. 最后用训练完成的神经网络,实现诗句的自动生成.

通过对大规模诗词数据进行机器学习,将创作散文诗的特征融入到统计概率模型中,由此实现散文诗的辅助创作,为广大散文诗爱好者提供了帮助,对散文诗文学的传承和发扬具有积极意义.

1 散文诗创作背景和基本框架

在考虑如何创作散文之前,先从考虑如何赏析的角度出发. 赏析过程中,可以获取到诗歌的结构、用词和情感表达,而生成作品的过程其实就是赏析结果的逆过程.

如图1所示诗,乡愁的赏析:这首乡愁用简短的七行来进行概括:第一节写乡音的清新缭绕笛声,通过塑造笛声在有月亮的晚上响起的场景从而渲染一种画面感,第二节是抒发情感,写乡情的情感描述,体现乡情缠绵,第三节写乡愁的永恒,是结合前面两节过渡而来,而且最后部分直接点出主旨,乡愁让之前的模糊的表达递进,逐渐鲜明. 以上这是对诗的简单分析,可以发现整个作品都是围绕主题乡愁进行描写,通过与乡愁

接近的很多意象词和情感词来表现和抒发情感. 本文的生成过程可以看成是这个简单分析的逆向过程,即先进行简单的主题定位和意象选取,再通过这个给定的内容生成出文本.

通过以上分析,进而确定整个生成的步骤,相应的散文诗辅助创作系统具体结构设计如下(图2)整个诗句创作的过程可分为以下步骤:

(1) 确定下主题思想或者意境. 这些主题思想通过确切的主题词来体现,比如描述想念家乡的思想,可以直接用“乡愁”或者“思乡”这样的词来表述.

(2) 在第一步确定下主题词后,下一步需要的是通过已建立语料库,找出与主题语义相关的词汇,如“思乡”可能对应了“月亮”、“别离”等词.

(3) 在第二步获取到一系列词汇后,要利用首句生成算法对这些词汇进行排列组合完成首句创作.

(4) 在生成后续诗句过程中,主要利用递归神经网络的时序性实现下一句诗句的生成,在此基础上再重新循环直至完成全部的生成.

乡愁 (席慕容)

故乡的歌是一支清远的笛
总在有月亮的晚上响起
故乡的面貌却是一种模糊的惆怅
仿佛雾里的挥手别离
别离后
乡愁是一棵没有年轮的树
永不老去

图1 《乡愁》

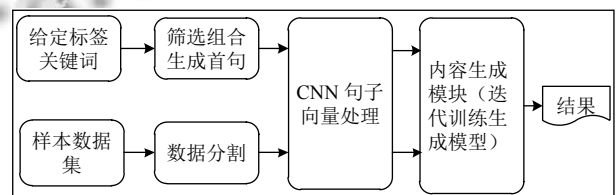


图2 诗句创作过程

2 诗句生成

2.1 中文词处理及数据集建立

在使用模型对语料文本训练之前,必须要对汉语语料文本进行一系列的处理. 首先针对一份文本语料,需要做的是将语料中出现的标点符号与不同于汉语的其他字符进行获取和处理,剔除掉相关的噪声信息^[6]. 另外语料中一般会有冗余的字词. 比如对于语料中出

现的阿拉伯数字,则需要把它转换成汉语中的表达方式,这类的操作是语料的正规化.在对语料正规化之后开始分词,首先根据原先的不同的标点符号对句子进行分割.要知道在文本生成的时候,对标点符号的生成也是很重要的部分,它可以更好的分割句子,运用好标点也能让文本的更富有情感化.

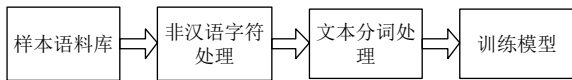


图3 文本样本处理流程

除了分词过程以外,在已有的分词数据集基础上,另外建立了一个主题聚类模型^[9,10],模型的目的是将不同的词汇根据不同意象进行聚类,这样会方便对主题词进行扩展,给定某词后提供多个相关词汇.本文中选取的语料库来源于原创文学交流平台的作品,以及名家作品集中的作品,共计13 187首,在获取这些作品时,平台已经有栏目分类共计19类,如四季(春、夏、秋、冬),伤感类、夜、雪、雨、远山、月亮、童年等等,对于没有确定分类的作品采用人工鉴别给定主题.这些给定类别会作为文档分词后,词汇进行主题聚类的重要依据^[11].

该主题模型基本算法如下.首先模型的输入是所有文档的集合 D ,给定需要的分类个数 k ,那么所有的主题集合就是 T ,每个文档 d 的内容就是一个单词序列 (w_1, w_2, \dots, w_n) , w_i 表示第 i 个单词,设 d 有 n 个单词.文档集合 D 中所有文档中的不同单词组成一个大的词汇表 V .模型的输入是文档集合 D ,在输入之前,对文档内容进行相应的分词,去除停用词,获取词干信息等预处理操作.之后,假设每个文档 d ,对应到不同主题的概率 $(p_{t1}, p_{t2}, \dots, p_{tk})$,其中, p_{ti} 表示 d 对应 T 中第 i 个主题的概率 $p_{ti} = n_{ti} / n$, n_{ti} 表示 d 中对应第 i 个主题的词的个数, n 为 d 中所有词的总数.对于主题 t ,生成不同单词的概率 $\alpha(p_{w1}, p_{w2}, \dots, p_{wn})$,其中 p_{wi} 表示 t 生成 V 中第 i 个单词的概率. $P_{wi} = N_{wi} / N$,其中 N_{wi} 表示对应主题 t 的 V 中第 i 个单词的数目, N 表示所有对应到 t 的单词总数.模型核心公示如下:

$$p(w|d) = p(w|t) * p(t|d) \tag{1}$$

主题作为中间层,通过初始给定的 $(p_{t1}, p_{t2}, \dots, p_{tk})$ 和 $(p_{w1}, p_{w2}, \dots, p_{wn})$,给出了文档中 d 出现词 w 的概率.实际上,利用当前的向量,可以为一个文档中一个单词

计算它对应任意一个主题时的 $p(w|d)$,然后根据这个结果更新该词对应的主题,如果这个更新改变了单词对应的主题,就会影响 θ 和 α 变量.最后的模型输出就是这两个向量.从而可以得到词汇聚类结果.如图下,是给定分类数目50情况下的部分分类结果.

2.2 生成方法

参考之前的工作,认为在生成处理过程中,必须要从相对整体的环境来考虑分析.所以,在生成诗句的时候,每一句诗都是根据上下文来综合衡量的.在生成过程中,需要考虑到是内容的选择,以及字词的表现.每一行诗句的组成是由初始建立的词库中选择的词汇集.

表1 分类结果举例

主题	流水	远山	月亮
词汇	宏声, 浪涌, 殿, 深秋, 凭栏, 群山, 枫, 流烟, 阻	影, 惆怅, 清风, 月照, 玉笛, 冷, 舟 ...	断, 匿, 朦胧... 月夜, 寂静...

2.2.1 生成首句

在生成首句前需要提供一些关键字,根据提供的关键字,将相应关键字进行扩展成一系列相关的词组,再将词组相连组成句子.扩展的词组来源于初始创建的词组分类库.

这里给定的关键字一般是包括一些意象词,情感词.在给定关键词之后,生成第一句的基本实现思想是,利用之前建立的同语义词关系集合,进行词汇扩展.这些词汇是实现生成首句的基本依据.然后,结合模型来对生成结果进行评价,以保证好的生成结果,流程见图4.

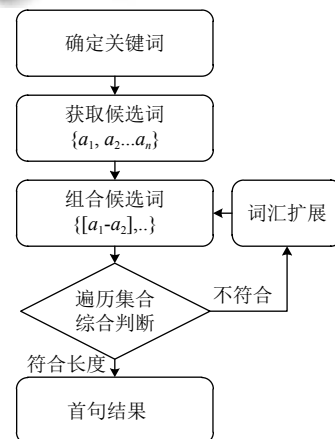


图4 首句组合生成

以下介绍算法流程,利用给定的关键词和生成的主题词关系模型,得到所有和给定关键词主题相近的

所有词集合 $\{a_1, a_2, \dots, a_n\}$, 将获得到的集合中的词进行排列组合, 即可获取到 $\{[a_1 - a_2], [a_1 - a_3], \dots\}$ 这样的词汇组合集合. 利用判断条件, 确定是否符合句子标准. 这个标准从组合结果的总长度, 组合词中的类别主题距离, 以及是否出现重复词等方面进行判断. 如果生成的字数太长就及时结束生成过程, 类别距离是在之前建立的主题-词关系矩阵中计算得到, 具体词之间的关系距离给定一个阈值, 超出的话也会作为不合格组合, 重复性判断是指相同字词不能出现次数过多. 如果达到预定的判定标准后, 就根据词性进行停用词插入, 组成输出的句子. 如果没有达到最终结果, 则需要迭代修改, 继续进行生成. 这个迭代过程, 首先对不符合类别距离和重复性的组合词从组合集合中进行删除, 这样达到精简组合集合的目的, 迭代时使用精简后的集合再次进行排列组合和计算, 直至得到最后结果.

2.2.2 句子上下文模型

在生成了首句之后, 下一步则是要确定的句子结构. 理论上, 其他模型也能实现词组或句子的向量表示. 我们选择了卷积句子模型, 因为它是基于 n -gram 计算, 只需自身的词向量就可以无需借助其他信息, 工具等, 这些方法因会出现误差不适宜用在中文诗歌上. 而基于朴素贝叶斯的概率模型虽然能得到句子的向量表示, 却不能考虑句中字词出现的先后顺序. 卷积句子模型通过按顺序合并相邻向量的方法, 从而计算一个句子内容的连续表示. 另外在图像处理的过程中, 图像的处理方法是转换成向量表示, 使用卷积神经网络获取整个图像向量的特征, 不仅仅可以一定程度上缩小图像表示的维度大小, 而且较好的获取到图像的所有特征信息^[12]. 此处, 因为卷积神经网络在图像处理上的效果很不错, 考虑将卷积神经网络移植到文本处理中, 将整个上文诗句向量看作一个整体, 将这个整体向量进行卷积化处理, 实现上文特征的提取.

该上下文模型包含输入单元 (input units), 输入集是各个句子中提取到的句子特征标记为 $\{k_0, k_1, \dots, k_t, k_{t+1}, \dots\}$, 而输出单元 (output units) 的输出集则被标记为 $\{y_0, y_1, \dots, y_t, y_{t+1}, \dots\}$. 输入集组合成矩阵记为 $A_{m \times n}$, 卷积计算时, 需要给定卷积窗口大小 (此处给定为 $i \times i$ 的矩阵, $i < m$) 和卷积窗口移动步数 $step (step < i)$, 通过卷积窗口从左向右, 从上向下按照 $step$ 大小进行移动计算卷积, 此处对窗口采用平均采样方法, 由此可以得出经过卷积计算后的新矩阵 A' 的

大小为 m', n' .

$$(m - i) / step = m' \quad (2)$$

显然新矩阵的长宽小于原来矩阵的长宽, 并且卷积窗口的平均采用方式对原来矩阵的基本特征改变不大. 如图 5, 句子中的单词向量依次进行卷积计算, 并且采用平均采样的方式得出新的一层的向量矩阵. 以此类推, 最后获取到句子的矩阵结果向量 R .

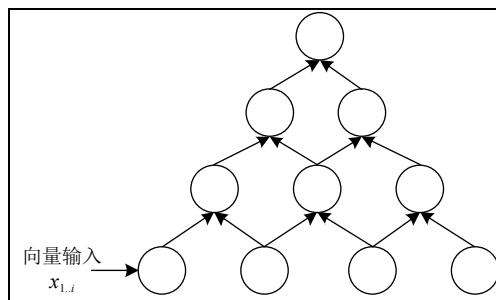


图 5 卷积结构图

2.2.3 诗句生成

这部分工作是在确定了上文向量 R 的基础上, 进行语句的下句生成操作. 将之前获取到的矩阵结果向量作为输入, 在 RNNs 中包含隐藏单元 (hidden units), 将其之后的输出集标记为 $\{s_0, s_1, \dots, s_t, s_{t+1}, \dots\}$, 这些隐藏单元完成了最为主要的工作.

如图 6 所示为该网络结构的展开情况. 举例, 对于一个有 5 个词语的语句, 展开后就是一个 5 层的神经网络, 每层代表一个词. 对网络的计算首先明确如下变量:

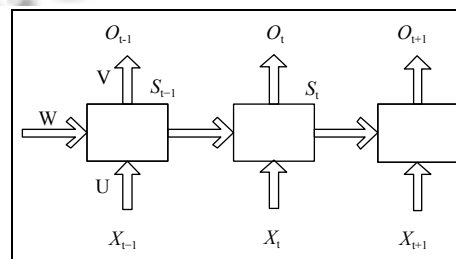


图 6 神经网络结构展开图

- ① x_t 表示第 $t (t=0, 1, 2, 3, \dots)$ 步输入, 即第 t 个句子的向量表示.
- ② 这里设隐藏层的输出 S_t , 表示第 t 步的状态. $S_t = f(W_1(x_t, x_{t+1}) + b)$, f 表示非线性的激活函数, 如 \tanh 或 ReLU , 在 $t=0$ 时, S_{t-1} 设置为零向量计算.
- ③ O_t 是第 t 步的输出, $O_t = \text{softmax}(VS_t)$. 这个输出

只与当前的隐藏层 S_t 有关。

关于该神经网络模型, 训练样本是之前计算所得的上文句子向量集 $R \{r_1, r_2, \dots\}$. 训练过程中, 首先给网络确定权值矩阵 W_1 和偏置向量 b , 对其进行随机赋值初始化, 设置初始学习率为 0.1, 该层接受的输入有两个: (1) 当前时刻输入向量 r_i 记作 x_i ; (2) 上一时刻单元计算结果 S_{t-1} . 其中第一层计算时忽略 S_{t-1} . 具体计算方法见公式 (1)、(2).

$$S_t = \sum_{i=1} w_{1i}x_i + \sum_{c=1} w_{ci}S_c^{t-1} \quad (3)$$

$$S_c^t = f(S_t) \quad (4)$$

以上为输入层向量的处理. 后面隐藏层的计算和 BP 神经网络不同, 它不仅会接受来自输入层处理后的数据 S_t , 而且会接受来自上一时间隐藏层传入的输入 O_{t-1} 输入向量是诗句的上下文相关向量. 这其中包含长距离信息, 是对输入向量的一个补充, 最终概率计算更加准确. 该层的计算如公式 (3)、(4):

$$O(t) = f(UO_{t-1} + WS_c^t) \quad (5)$$

$$y(t) = g(O(t)) \quad (6)$$

其中, $f(z)$ 为 sigmoid 激活函数:

$$f(z) = (1 + e^{-z})^{-1} \quad (7)$$

$g(z)$ 为 softmax 激活函数:

$$g(z) = \frac{e^z}{\sum_k e^z} \quad (8)$$

以上就是隐藏层计算结果, 而输出层和传统的 BP 神经网络一样, 通过激活函数获取模型输出 y' . 而后需要不断迭代计算交叉熵最小值进行模型的调参, 这里使用实际概率结果和输出层得出的概率分布结果进行交叉熵计算, 交叉熵的计算公式如下:

$$J(y, o) = -\frac{1}{N} \sum_{n \in N} y_n \log y'_n \quad (9)$$

其中, y_n 代表实际句子向量分布, y' 代表预测模型中的概率分布. 提出交叉熵的目的, 就是将其作为损失函数, 然后利用梯度下降向着下降最快方向不断更新参数使得损失函数获取到最小值, 在达到一定迭代次数后, 所对应的参数就是我们要求的参数, 见公式 (8)

$$\frac{\partial J}{\partial \theta} = \frac{1}{N} \sum_{n \in N} \frac{\partial (y_n \log o_n)}{\partial \theta} + \lambda \theta \quad (10)$$

参数 θ 代表式 (1) 中待求参数 W_1, b , 参数的优化方法则使用了 L-BFGS 算法. 得到所需参数以后, 也就得到了整个适用于后续句子生成的递归神经网络模型. 利用模型生成输出句子向量, 直至最后得到指定结束符, 即完成创作.

3 实验设计与结果

实验目的是对自动生成的文本进行评测, 有人工评测和自动评测两种方式, 人工评测主要从流畅程度, 情感理解, 意象角度来进行评测. 自动评测是基于已经创建的散文诗的语料库. 关于自动评测方案, 本文的诗句生成方法是由上文生成下句的方法, 此处借鉴用于机器翻译系统^[13]自动评测的 BLEU 评测方式^[14]. 另外, 多种文本生成和翻译^[15]方法的评测中都采用该种评测方案. BLEU 的评测标准是给定上句后生成的下句, 认为能够更贴近于已有的参考下句, 则判断生成的质量更好. 但是因为生成的内容比较多样化, 所以需要多个下句的数据样本加入答案集合. 因为诗句的多样性, 在准备数据集时, 会选取多个下句作为上句的对应结果, 这些句对放入答案集中, 这个筛选过程采用的人工整理. 因为生成的结果总句数不一定, 所以表中给出前 6 句评测结果和最后平均值. 实验结果见, 表 2 通过和其他类型的系统, 如隐马尔科夫模型下的数据结果对比, 该隐马尔科夫的模型的结果均值为 0.1516, 本文系统结果为 0.1806, 高于隐马尔科夫模型下的系统.

表 2 自动评测结果

BLEU	1~2 句	2~3 句	3~4 句
系统结果	0.1432	0.1901	0.1894
BLEU	4~5 句	5~6 句	平均
系统结果	0.2103	0.1702	0.1806

人工评价内容为, 总计调查了 40 位文学专业的学生, 让他们对本文的生成结果进行一些判定, 其判定标准主要是流畅通顺性, 诗句意象, 情感传递等这几个方面. 每个方面有总分为 5 分, 分数越高表示越优秀. 人工评价部分结果见表 3.

如图 7 所示就是其中一段生成的文本, 如表 3 实验调查的结果, 从统计结果发现, 基本生成的作品在流畅通顺性和情感传递上来说是相对有效的,

通过实验发现: 首先对话料库的文本预处理可以很好的剔除无意义词, 提高生成结果. 在初始化参数时

需要考虑,不能直接把它们都初始化为0.因为初始化的不同对最后的结果是有不同影响的,初始化参数与神经网络的激活函数是有关的.初始选择的 tanh 函数,推荐应该使用 $[-1/n, 1/n]$ 之间的随机数作为初始值, n 表示和前一层的连接数.对于初始化参数,将他设置为很小的随机数,有助神经网络的正常训练.

表3 人工评测结果

评分关键词	流畅通顺	意象清晰	情感传递
春天,鸭子,喜悦	3.48	3.93	3.98
蓝天,古村,汽车	3.29	2.84	2.24
海洋,夜晚,孤寂	3.65	4.31	4.21

默默的,
他们在吱呀作响的甲板上,
静静的,
看着这个世界,没有声音,
只有海洋中摇晃出阵阵沉默,
吞噬着他们,
一点,一点,
逐渐渐暗淡在巨浪中,沉沦。

图7 生成的文本结果

4 结论与展望

目前国内外的诗句生成研究基本采用直接基于模板的生成方法,本文提出利用主题模型和深度学习下的递归神经网络方法来进行自动生成,使得诗句上下句关系与自动翻译中的双语关系进行映射,并考虑了该模型下的一些技术难题,实现了这一原理下的诗句生成.最后通过给定的人工评测和自动评测方法,对创作结果进行了判分,从结果来看,本研究对散文诗自动生成和自然语言生成有一定的参考价值.以后的工作需要更多的考虑语句的连贯与通顺性,使之能够又更好的表现.

参考文献

1 路永和,梁明辉.遗传算法在改进文本特征提取方法中的

应用.现代图书情报技术,2014,(4):48-57.

- Kempe V, Levy R, Graci C. Neural networks as fitness evaluators in genetic algorithms: Simulating human creativity. Proceedings of the Annual Meeting of the Cognitive Science Society. Edinburgh, Scotland, UK. 2001. 1221.
- Manurung H. An Evolutionary Algorithm Approach to Poetry Generation[Ph.D. Dissertation]. Edinburgh, UK: University of Edinburgh, 2003.
- Gervás P. An expert system for the composition of formal Spanish poetry. Knowledge-Based Systems, 2001, 14(3-4): 181-188. [doi: 10.1016/S0950-7051(01)00095-8]
- 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.
- He J, Zhou M, Jiang L. Generating Chinese classical poems with statistical machine translation models. Proceedings of the 26th AAAI Conference on Artificial Intelligence. Toronto, Canada. 2012. 1650-1656.
- 李岩.基于深度学习的短文本分析与计算方法研究[博士学位论文].北京:北京科技大学,2016.
- 胡新辰.基于LSTM的语义关系分类研究[硕士学位论文].哈尔滨:哈尔滨工业大学,2015.
- 胡吉明,陈果.基于动态LDA主题模型的内容主题挖掘与演化.图书情报工作,2014,58(2):138-142.
- 刘铭,咎红英,原慧斌.基于SVM与RNN的文本情感关键词判定与抽取.山东大学学报(理学版),2014,49(11):68-73. [doi: 10.6040/j.issn.1671-9352.3.2014.025]
- 邸亮,杜永萍.LDA模型在微博用户推荐中的应用.计算机工程,2014,40(5):1-6,11. [doi: 10.3778/j.issn.1002-8331.1311-0272]
- 刘占一,李生,刘挺,王海峰.利用统计搭配模型改进基于实例的机器翻译.软件学报,2012,23(6):1472-1485.
- 刘群.统计机器翻译综述.中文信息学报,2003,17(4):1-12. [doi: 10.3969/j.issn.1003-0077.2003.04.001]
- 刘欢,邵蔚元,郭跃飞.卷积神经网络在验证码识别上的应用与研究.计算机工程与应用,2016,52(18):1-7. [doi: 10.3778/j.issn.1002-8331.1603-0181]
- 周胜臣,瞿文婷,石英子,等.中文微博情感分析研究综述.计算机应用与软件,2013,30(3):161-164,181. [doi: 10.3969/j.issn.1000-386x.2013.03.043]