TG 网络：一种更有效识别助词“的”用法的模型

昝红英[1], 刘创[2], 段雪敏, 张坤丽, 韩英杰

郑州大学信息工程学院

[1] 通信作者，E-mail: iehyzan@zzu.edu.cn，[2] 通信作者，E-mail: 214674227@qq.com

摘　要：现代汉语中，助词“的”具有出现频率高，用法灵活多变的特点。本文提出一种自动识别“的”用法的神经网络模型（TG 网络），此网络首先采用基于自注意力机制的模型作为第一层特征编码器，门控循环网络作为第二层语义提取器，识别的准确率达到82.8%。实验表明，TG 网络的识别效果优于先前的方法。进一步实验中，通过不同窗口的设置得出，窗口越大模型的效果越好，同时对每一个用法类别进行细粒度的分析。未来期望此模型对更多的虚词进行用法自动识别，并将识别结果应用于其他自然语言处理任务中。

关键词：虚词知识库，TG 网络，自注意力机制，门控循环网络，自然语言处理

TG network: a model that more effectively identifies the use of the auxiliary word “DE”

Hongying Zan[1], Chuang Liu[2], Xuemin Duan ,Kunli Zhang , Yingjie Han

[1] School of Information Engineering, Zhengzhou University, Zhengzhou, Henan 450001, China

[1] E-mail: iehyzan@zzu.edu.cn, [2] E-mail: 214674227@qq.com

Abstract: In the knowledge base of function word usage of "trinity", the auxiliary word "DE" has the characteristics of high frequency and flexible usage. In this paper, a neural network model (TG network) is proposed to automatically recognize the usage of "DE". In this network, the self-attention mechanism is firstly adopted as the first-layer feature encoder and GRU(gated recurrent unit) as the second-layer semantic extractor, and the recognition accuracy rate reaches 82.8%.Experiments show that the recognition effect of TG network is better than that of previous methods. In further experiments, the larger the window, the better the effect of the model is proved by setting different windows. At the same time, the fine-grained analysis of each usage category is carried out. In the future, it is expected that this model will automatically recognize more function words and the recognition results can be applied to other natural language processing tasks.

Key words: the knowledge base of function word, TG network, self-attention mechanism, GRU, natural language processing

1 引言

语法问题长期以来一直是自然语言处理（Natural Language Processing，NLP）的难点之一，由于虚词在句子中承担语法特征与实词之间的语义关系描述，所以对于句子中虚词的研究一直是语法研究的重要内容。其中，虚词中的助词最具有汉语类型学的特点，而且与各类实词关系更为密切。现代汉语中助词“的”的用法多，使用频率高，而且十分灵活，其研究成果对汉语语法研究，对外汉语教学和自然语言处理等都具有重要的参考意义[1]。

文献[1,2]对多年来“的”的研究进行了综合论述，包括对其用法的讨论。俞士汶等[3]提出了构建“三位一体”（虚词用法词典，虚词用法规则库和虚词用法语料库）广义虚词用法知识库的思想。此后，昝红英等[2, 5-6]和张坤丽等[7]构建了现代汉语广义虚词用法知识库((Chinese function word usage knowledge base, CFKB))，其中包括助词“的”的虚词用法词典，规则库以及用法标注语料库。韩英杰等[8]在构建助词用法知识库中，探讨了基于规则的助词用法自动标注。由于“的”的用法复杂，文献[7-8]指出，采用规则的方法，识别效果非常不理想，而助词“的”又是出现频次最高的虚词,因此对其用法的自动识别成为研究重点。之后刘秋慧[9]等人分别使用基于规则，基于 CRF 模型，基于深度学习中 GRU 模型对“的”进行自动识别研究。David Chang 等人[10]、Zhang 等人[11]分别将虚词用法知识应用到机器翻译任务中，

本文根据 CFKB 中“的”的用法规则，共有 39 类，在《人民日报》2000 年 4 月语料上进行实验，由于“的”出现频率较高，共选择出 64449 个序列。本文提出的 TG 网络首先使用 Transformer 模型[12]中的编码块作为序列第一

层特征编码器，自动提取序列内部语义联系，再配合GRU模型进行再一次的语义提取。实验表明，TG网络能得到更好的识别效果。

2 相关工作

2.1 助词“的”用法描述

根据吕叔湘的《现代汉语八百词》[13]、《现代汉语词典》(第 5 版)[14]和张斌的《现代汉语虚词词典》[15]，结合2000年4月《人民日报》的语料，构建助词用法词典,其中“的”包含 11 个大类用法, 39 种细分用法。通过释义、用法、例句、搭配和合用等属性，对助词“的”的用法进行全方位的描述。表 1 仅列出部分用法规则属性，其中各属性的详细说明见文献[6]。

表 1 助词“的”部分用法描述

ID	释义	用法	例句
u_de5_t2_1a	构成“的”字短语修饰名词。<b>	名词+~+名词。<b>用在名词及其修饰语之间,修饰语为名词。<x><z>	他行李多，我~行李很少<b><z>|以前~不合理负担<z>
u_de5_t2_1b	构成“的”字短语修饰名词。<b>	动词+~+名词。<b>用在名词及其修饰语之间,修饰语为动词。<x><z>	原来安排~时间是星期三,现在改到星期五<b><z>|唱~声音太小<b><z>
u_de5_t2_1ba	构成“的”字短语修饰名词。<b>	动词+名词+~+名词。<z>用在名词及其修饰语之间，修饰语为动宾结构的短语。<x><z>	打电报~费用<b>|中国永远是维护世界和平与稳定~重要力量。<r>
u_de5_t2_1bb	构成“的”字短语修饰名词。<b>	动词+形容词+~+名词，或者介词短语做补语。<z>用在名词及其修饰语之间，修饰语为动补结构。<x><z>	进不去~人便继续在电梯门口静候。<r>|丁玉珍把冲好~照片交给了孔玲。<r>
u_de5_t2_1bc	构成“的”字短语修饰名词。<b>	用在名词及其修饰语之间，修饰语为状中结构的短语。状语的成分包括形容词、时间词、副词、名词、名词短语、介词短语、助动词和方位词短语,中心语是动词或动词短语。<x><z>	与中共同心协力，共同奋斗~各民主党派。<r>

2.2 神经网络模型的相关研究

2014年Yoon Kim尝试用卷积神经网络（CNN）对文本进行分类[16],而后为了防止训练中的过拟合情况，Hinton提出了dropout方法[17]。由于语言具有天然的序列性，循环神经网络能够有效地解决序列性问题，所以循环神经网络在自然语言处理中的应用也得到了广泛的应用。但循环神经网络本身存在梯度爆炸或梯度消失的问题，为了解决这一现象又提出长短记忆神经网络(LSTM)[18]等模型。作为LSTM模型的拓展，GRU模型参数较少、结构更简单，并且在小数据集上结果较好[19]。2017年，Ashish Vaswani等人[12]提出的Transformer模型，这是一种基于注意力机制的特征提取器，可以解决循环神经网络不能并行等缺点，同时能有效获取序列的全局信息。不同的神经网络模型都有自身的特点和缺点，针对不同任务，利用各自模型的特点，设计不同的模型也是如今的主流方法，本文提出的TG网络正是将transformer模型和GRU模型结合到一起，达到较好的效果。

3 TG 网络结构

3.1 模型总览

模型主体结构如图 1 所示，输入序列包括词信息和词性信息，首先序列经过 Transformer[12]编码块进行序列编码，寻找模型内部结构，再将编码后的序列送入 GRU 模型中进行进一步的语义提取，最后经过多层感知机进行输出。在 3.2 节对 Transformer 编码层进行简述，3.3 节将重点阐述编码层中的核心部分，即多头注意力机制。在最后的 3.4 节，介绍如何对编码层的输出进行拼接，再次作为语义提取层的输入。

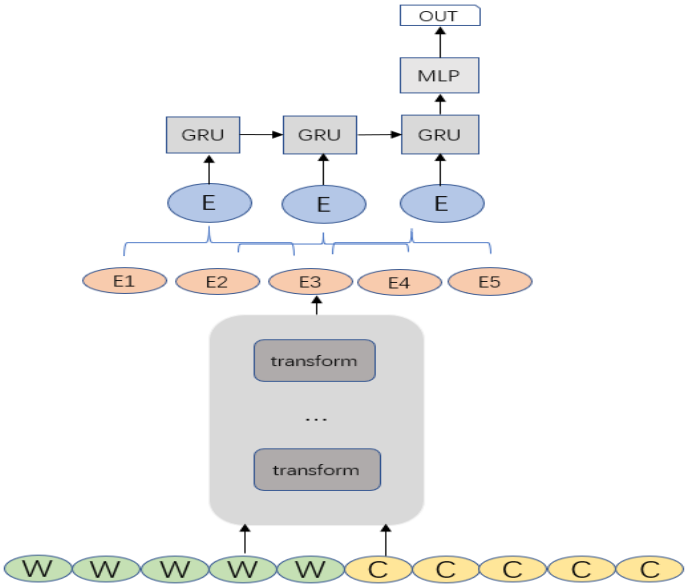

图 1 模型结构图

3.2 编码层

编码层采用 Transformer 模型中的编码块。N 代表有几个相同的模块，本实验中 N 为 1。每个编码层包含两个子层。首先是多头自注意力机制，第二层是全连接的前馈神经网络层。模型在这两个子层的每个周围都使用一个残差连接[20]，然后进行层归一化[21]，所以每一个子层输出都是 LayerNorm（x+sublayer(x)），其中 sublayer(x)是每一个子层的输出。

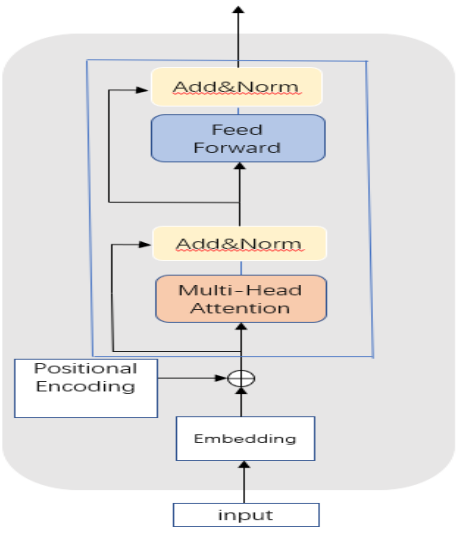

图 2 编码层结构

3.3 注意力

注意力机制可以有效的提取信息中的重要部分，例如在一段文本中，通过注意力计算，文本中不同词会得到不同的权重，权重越大表示该词在文本中的重要性越高。在本文中，注意力方程可以描述为将一个 Q 和一组 K，V 对映射到输出，其中 Q,K,V 三者向量表示。输出是这些值的加权和。当需要计算序列自身的内部关联时，Q=K=V,三个输入均为输入序列，这种机制的好处是序列中每一个位置的信息都与其他位置的信息进行关联计算，有效的得到序列中的全局信息。

3.3.1 点积注意力方程

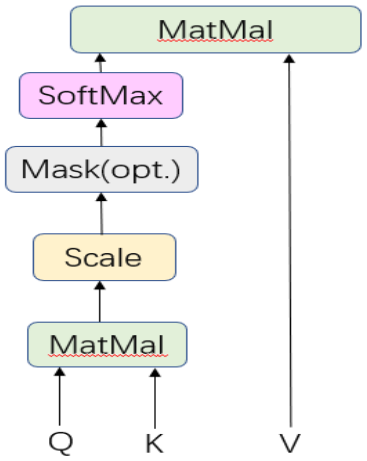

图 3 “Scaled Dot-Product Attention”结构图

注意力计算有不同的方式，Transformer 模型中采用的这种注意力方程被称为“Scaled Dot-Product Attention”[22]，结构如图 3 所示。输入由 Q,K,V 组成，其中，Q,K 具有相同的向量维度 d_k。首先计算 Q 和 K 的点积，再除以 d_k 的平方，经过一个 softmax 函数后得到 V 的权重。$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$ (1)

3.3.2 多头注意力

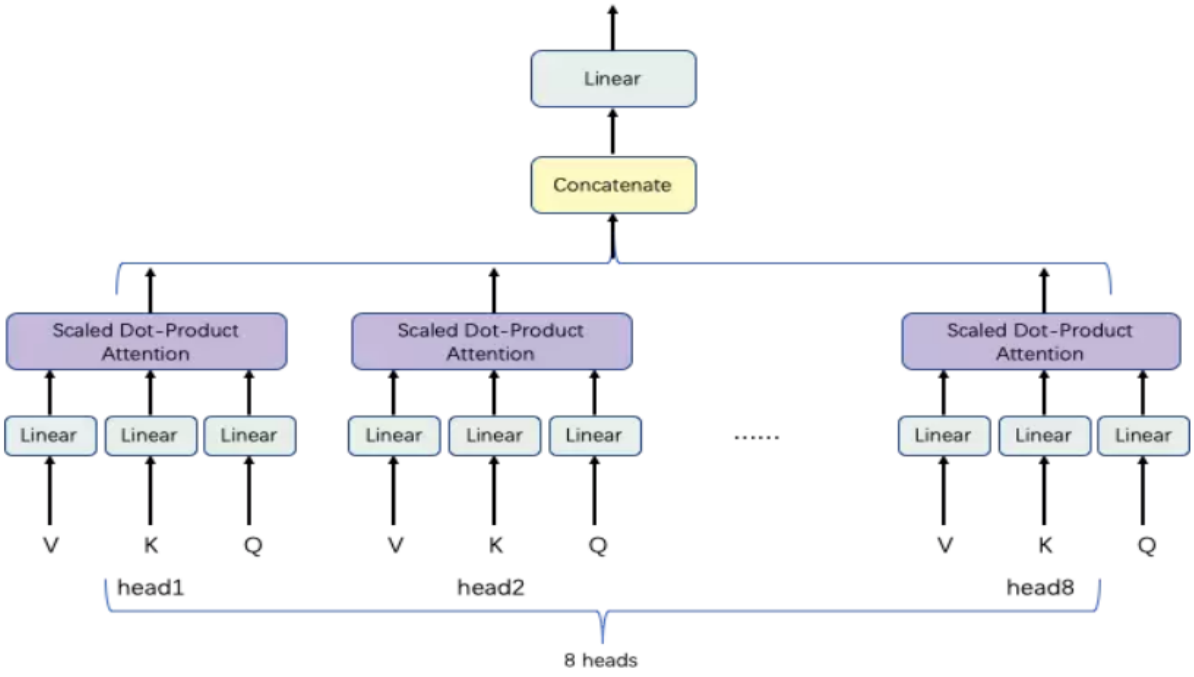

图 4. 多头注意力模型示意图

多头注意力是通过 h 个不同的线性变换对 Q,K,V 进行投影，最后将不同的注意力结果拼接起来。如图 4，h 的大小设置为 8，这表示将输入序列的词向量维度分为 8 份，分别计算各自的注意力，然后将计算的结果进行拼接。

$$MultiHead(Q,K,V) = Concat(head1,\dots,headh)Wo \quad (2)$$

其中，Wo 为可训练参数，$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$。

3.4 语义提取层

经过第一个编码层后得到已经获取内部语义信息的序列，设置一个大小为 3 的窗口，按序列顺序将每三个元素进行拼接，重新得到一个拼接后的序列，再将新序列作为输入放入 GRU 模型中[9]，如图 5 所示。最后将 GRU 的最后一个隐层输出放入多层感知机进行分类。

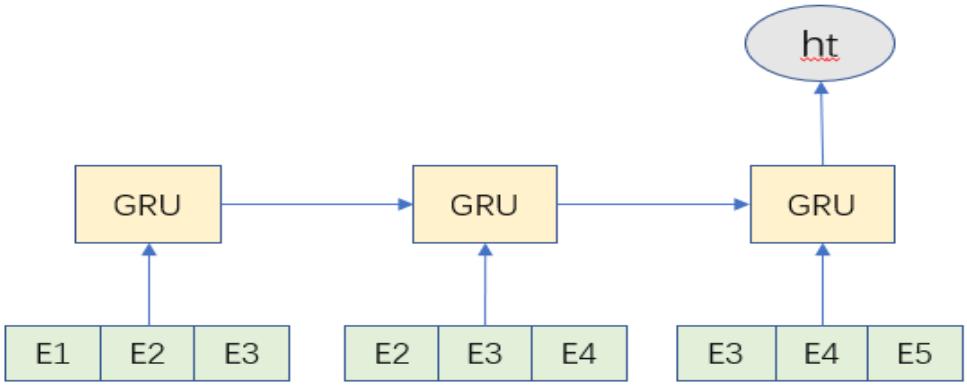

图 5 语义提取层结构

4 实验结果及分析

实验选用《人民日报》2000 年 4 月虚词用法标注语料作为实验数据。训练集，测试集随机按 9：1 的比例进行划分。“的”的用法共分为 39 类。由于一句话中可能存在不止一个“的”，所以在实验中首先设置一个窗口大小，选择“的”前后的词，例如大小为 3 的窗口意味着选取“的”前面三个词和后面三个词。本文实验分别在不同的窗口大小（3，5，7）下进行。

实验中，批处理大小设置为 64，向量维度 d_k 设置为 128，transformer 编码块的层数设为 1，head 数设置为 8。实验选用 Adam 作为优化器，学习率进行随机设置。损失函数使用交叉熵函数。表 2 为实验结果对比。前三行是刘秋慧等人[9]将基于规则，基于 CRF 模型，基于 GRU 模型三种方法应用到助词“的”识别的实验结果，第四行是本文单独基于 Transformer 编码器模型对助词“的”用法识别的结果，最后一行是本文提出的 TG 网络在窗口大小为 7 时的实验结果。选用准确率（Precision）、召回率（Recall）,F 值（F-Measure）作为评判标准。实验表明，本文提出的方法会有效提升用法识别效果。

表 2 实验对比

模型	P/%	R/%	F/%
基于规则	70.9	22.7	34.4
基于 CRF 模型	76.8	78.3	77.5
基于 GRU 模型	81.2	81.5	81.3
基于 Transformer 模型	72.6	72.6	72.6
基于 TG 模型（窗口为 7）	**82.8**	**82.8**	**82.8**

从实验结果中不难发现，后两个基于深度学习的模型效果明显好于基于规则和基于 CRF 模型的效果。与传统方法相比，深度学习模型可以自动提取特征，减少人工选择特征的工作量。而序列经过 Transformer 模块进行编码后，通过其内部的注意力机制，序列自动获取内部特征，取得序列自身的内部联系，再使用 GRU 网络进行语义信息提

取更能准确的得到序列中的重要信息，从而得到更准确的分类结果。

同时，通过实验发现，窗口大小的不同同样会对识别效果带来不同的影响，窗口设置较大会得到更长的序列，这也意味着在模型提取特征时将得到更多的信息，而实验结果也证实，更长的序列将得到更好的识别效果。如表3所示。

表3 基于TG模型不同窗口的实验对比

窗口大小	F/%
基于TG模型（窗口为3）	77.8
基于TG模型（窗口为5）	81.6
基于TG模型（窗口为7）	**82.8**

由于不同用法在语料中出现频率不同，为了进一步了解每个用法类别的识别效果，本文在窗口为7时，选取测试集中用法频率至少为1次的类别，分别计算其P，R，F值，如表4所示。实验结果显示，“的”用法频率分布不均，训练集出现多的用法更容易获得较好的准确率，说明模型的效果同样受到数据集大小的影响。

表4 助词“的”用法细粒度识别结果（窗口为7）

ID	P/%	R/%	F/%	训练集出现频次	测试集出现频次
u_de5_t2_1c	95.0	95.0	95.0	5374	536
u_de5_t2_1a	89.7	92.5	91.1	24171	2781
u_de5_t2_1bc	74.7	66.9	70.6	4630	509
u_de5_t2_1e	81.4	75.9	78.5	2418	272
u_de5_t2_1bd	53.5	53.7	53.6	1761	239
u_de5_t2_1ca	86.8	92.2	89.4	1800	175
u_de5_t2_1cb	79.3	92.0	85.2	735	81
u_de5_t2_1g	74.2	73.5	73.8	6758	720
u_de5_t2_5a	72.1	75.0	73.5	988	134
u_de5_t2_1ba	66.7	69.0	67.8	4052	436
u_de5_t2_1b	81.4	81.1	81.3	2356	237
u_de5_t2_1bb	75.7	54.9	63.6	503	57
u_de5_t2_2d	39.1	51.4	44.4	350	28
u_de5_t2_2cd	30.0	33.3	31.6	119	16
u_de5_t2_2cc	46.3	42.2	44.2	351	40
u_de5_t2_2ca	66.7	41.7	51.3	230	33
u_de5_t2_2b	75.0	57.7	65.2	178	14
u_de5_t2_1be	89.8	91.7	90.7	429	53
u_de5_t2_2a	62.5	50.0	55.6	99	9
u_de5_t2_2ce	75.0	75.0	75.0	44	5
u_de5_t2_1d	100.0	60.0	75.0	114	13
u_de5_t2_2cb	100.0	12.5	22.2	33	6
u_de5_t2_3c	28.6	22.2	25.0	72	5
u_de5_t2_9	0.0	0.0	0.0	4	1
u_de5_t2_2c	0.0	0.0	0.0	173	15
u_de5_t2_3b	0.0	0.0	0.0	14,	1
u_de5_t2_2ba	76.5	86.7	81.3	156	16
u_de5_t2_5b	0.0	0.0	0.0	17	4
u_de5_t2_4b	0.0	0.0	0.0	41	3

5 总结和展望

本文重新设计一种端对端的神经网络来识别助词“的”的用法，对比先前方法，取得了更好的结果，

这为虚词用法识别算法提供了又一种思路，也证明了针对设计更复杂的网络能更好的提取序列的信息，从而更有效的进行识别，为虚词用法知识库的应用奠定了基础。

下一步工作是：1）完善网络设计，使用更好的网络作为特征提取器；2）针对部分用法较灵活，出现频率较高的虚词进行用法识别实验；3）在自然语言处理应用中，深度学习虽然有着独特的优势，但由于它缺乏可解释性，在不同任务中存在着泛化能力不足的缺点，引入外部知识已成为解决这一问题的重要途径之一，如何更好的将虚词的用法知识融入到自然语言处理任务中也将是后续的重点工作之一，可以预见的是，虚词用法知识的融入能更好的提升自然语言的理解和推理能力，与现有深度学习算法形成互补。

参 考 文 献

[1] 徐阳春. 虚词“的”及其相关问题研究. 北京: 中国社会科学出版社, 2006

[2] 吕叔湘. 现代汉语八百词. 北京: 商务印书馆, 1980

[3] 俞士汶, 朱学锋, 刘云. 现代汉语广义虚词知识库的建设. 汉语语言与计算学报, 2003, 13(1): 89－98

[4] 昝红英, 朱学锋. 面向自然语言处理的汉语虚词研究与广义虚词知识库构建. 当代语言学, 2009, 11 (2): 124–135

[5] 昝红英, 张坤丽, 柴玉梅, 等. 现代汉语虚词知识库的研究. 中文信息学报, 2007, 21(5): 107－111

[6] Zan Hongying, Zhang Kunli, Zhu Xuefeng, et al. Research on the Chinese function word usage know-ledge base. International Journal on Asian Language Processing, 2011, 21(4): 185–198

[7] 张坤丽, 昝红英, 柴玉梅, 等. 现代汉语虚词用法知识库建设综述. 中文信息学报, 2015, 29(3): 1－8

[8] 韩英杰, 昝红英, 张坤丽, 等. 基于规则的现代汉语常用助词用法自动识别. 计算机应用, 2011, 31 (12): 3271–3274

[9] 刘秋慧等 助词“的”用法自动识别研究. 北京大学学报(自然科学版) 第 54 卷 第 3 期 2018 年 5 月

[10] David Chang, Pi-Chuan, Jurafsky, Dan, Manning, Christopher D. Disambiguating "DE" for Chinese-English machine translation[C]// The Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2009:215-223. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. CoRR, abs/1706.03762, 2017a.

[11] Zhang K, Xu H, Xiong D, et al. Improving Chinese-English Neural Machine Translation with Detected Usages of Function Words[C]// National CCF Conference on Natural Language Processing and Chinese Computing. Springer, Cham, 2017:741-749

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. CoRR, abs/1706.03762, 2017a

[13] 吕叔湘. 现代汉语八百词. 北京: 商务印书馆, 2006

[14] 吕叔湘. 现代汉语词典. 北京: 商务印书馆, 2007

[15] 张斌. 现代汉语虚词词典. 北京: 商务印书馆, 2006.

[16] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.

[17] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

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

[19] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [EB/OL]. (2014–12–11)[2017–06–01]. https://arXiv.org/abs/1412.3555

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.

[21] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.

[22] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473, 2014.