

基于字向量和增强表示 BiLSTM 句子相似度研究

贾 畅,叶 飞,刘帅君,麻之润
(云南农业大学 大数据学院,云南 昆明 650201)

摘 要:目前分词工具在金融领域智能客服中无法对金融相关词汇进行有效切分,且基于单词的模型更容易受到数据稀疏性和词汇表外单词的影响。针对该问题,提出一种基于字向量和增强表示 BiLSTM 的句子相似度计算模型—EBiLSTM。该模型首先通过双向长短时记忆网络 BiLSTM 提取由字嵌入组成的句子的字特征及其上下文表示,然后计算句子对中一个句子与另一个句子的软对齐表示,在此基础上通过句子表示与其对齐表示间的交互来增强最终的句子表示。所提模型可以有效学习到句子对的语义关系,加入增强表示层后通过两个句子的交互可以更好地捕捉两个句子间的语义差异。实验表明,所提模型在真实数据集上,精确率、召回率和 F1 值均优于基于词向量的 CNN 和 BiLSTM 方法,也优于基于字向量的 CNN 和 BiLSTM 方法。

关键词:智能客服;句子相似度;循环神经网络;字向量;句子对齐

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2020)10-0097-04

doi:10.3969/j.issn.1673-629X.2020.10.018

Research on Sentence Similarity Based on Character Vector and Enhanced Representation BiLSTM

JIA Chang, YE Fei, LIU Shuai-jun, MA Zhi-run
(School of Big Data, Yunnan Agricultural University, Kunming 650201, China)

Abstract: Currently word segmentation tools cannot effectively segment financial-related vocabulary in intelligent customer service in the financial field, and word-based models are more susceptible to data sparsity and out-of-vocabulary words. Aiming at this problem, EBiLSTM, a sentence similarity calculation model based on BiLSTM based on character vector and enhanced representation, is proposed. The model first extracts the word features and contextual representation of a sentence composed of words through a bi-directional long-term short-term memory network BiLSTM, and then calculates the soft-aligned representation of one sentence and another sentence in the sentence pair, and then aligns it with the sentence representation. Inter-representation interactions enhance the final sentence representation. The proposed model can effectively learn the semantic relationship of sentence pairs. After adding the enhanced presentation layer, the semantic differences between two sentences can be better captured through the interaction of the two sentences. Experiment shows that the proposed model is better than the CNN and BiLSTM methods based on word vectors and the CNN and BiLSTM methods based on character vectors in terms of precision, recall and F1.

Key words: intelligent customer service; sentence similarity; recurrent neural network; character vector; sentence alignment

0 引 言

智能客服因其商业价值和研究价值,受到了广泛关注。智能客服的本质是通过理解用户提出的服务请求,利用算法在知识库中找到与之相似的问题句子,然后返回给用户合适的答案。近年来,深度学习在自然语言处理领域不断的发展,越来越多的研究人员倾向于使用端到端的神经网络来计算句子相似度^[1-3],上述方法都是英文数据,在使用词向量计算句子相似度时,中文和英文有着天然的区别,中文数据需要先进行

分词,才能进行下一步的工作。目前的分词工具(如:jieba 和 ICTALAS)在大多数实际场景下都能有不错的应用,但是在金融领域智能客服问答系统中,由于用户请求口语化严重、错别字和金融领域的大量特有词汇,现有的分词工具很难进行切分,而分词错误势必会影响到模型的预测。虽然可以制定领域词典来更好地分词,但是往往需要耗费大量的时间,而且很难覆盖所有的词汇。另一方面由于词向量模型的数据稀疏性会导致过拟合,并且 OOV(词汇表外单词)的存在也会影

收稿日期:2019-11-21

修回日期:2020-03-24

基金项目:云南省重大科技专项(2018ZJ001-2)

作者简介:贾 畅(1995-),男,硕士研究生,研究方向为自然语言处理、深度学习。

响模型的训练。

在利用神经网络对句子进行相似度计算方面,多数工作通过 CNN^[1] 或 LSTM^[2-3] 来捕捉句子信息,但是缺少句子对的交互过程。对于计算句子相似度任务来说,这可能会丢失一小部分有用的句子信息。

为了解决上述问题,该文的主要贡献如下:

(1) 利用预训练的字向量作为原始的句子表示,在模型的训练过程中不断进行学习,避免了采用词嵌入技术时,分词错误、数据稀疏和 OOV 导致的模型训练不佳。

(2) 通过双向长短时记忆网络 BiLSTM 提取字嵌入组成的句子的上下文特征表示。

(3) 计算句子对中一个句子与另一个句子的对齐表示,然后通过交互来生成语义差异向量,以增强句子表示。

1 相关工作

传统的句子相似度计算方法有基于词袋特征^[4]、N-gram 重叠^[5]、WordNet 等外部资源^[6]、句法分析特征^[7]等,这些方法侧重于词汇语义,句子的表面形式匹配。

随着深度神经网络在 NLP 任务上的优异表现,使用卷积神经网络和循环神经网络对句子进行建模受到了越来越多的关注。He 等人^[1]利用 CNN 不同粒度和大小的卷积窗口提取句子的不同特征,然后使用多种类型的池化方式,以提取丰富的句子信息;Mueller 等人^[2]通过共享参数的 LSTM 对句子进行建模;Tai 等人^[3]提出了一种 Tree-LSTM 网络结构,用来提取句子更多的句法信息。与此同时,为了增强句子表示,研究者们尝试对通过 BiLSTM 产生的句子表示进行多种交互;冯兴杰等人^[8]在句子相似度任务上提出一种多注意力 CNN 模型,通过两个不同的注意力层来分别捕捉词语间和句子整体的语义信息;Chen 等人^[9]在自然语言推理任务上进行了探索,首先通过词级相似度矩阵计算两个句子的细粒度对齐,然后进行交互,以增加句子的上下文表示中各元素的局部推理信息;Tay 等人^[10]除了计算句子间的软对齐外,还使句子自己进行内部的对齐。

为了避免分词不当对模型训练造成的影响,一些研究集中在字嵌入技术上,如短文本分类^[11]、机器翻译^[12]。陈志豪等人^[13]通过多视角卷积网络对字嵌入表示的句子进行特征提取,在问答匹配模型上取得了优异表现。Meng 等人^[14]在 Bank Question corpus^[15]上进行实验,通过神经网络模型对随机初始化的词向量和字向量参数进行学习更新,最终效果字向量要优于词向量。

受到 Chen 等人^[9]在自然语言推理任务上,通过对 BiLSTM 生成的句子上下文与其对齐表示进行交互以增强推理信息的启发,该文提出了一种基于字向量和增强表示 BiLSTM 的句子相似度计算模型。在真实的金融智能客服数据集上,与基于词向量的模型进行对比,该字向量模型表现更优。

2 模型设计

基于字向量和增强表示 BiLSTM 的问句相似度模型 (EBiLSTM) 框架如图 1 所示。编码层通过共享参数的 BiLSTM 提取句子 S 和 T 的上下文特征信息;增强表示层通过相似度矩阵分别计算句子 S 和 T 的软对齐表示,并且将它们的原表示与对齐表示进行交互以增强句子表示;最大池化层提取所有表示的句子级特征;全连接层通过一个两层的全连接网络对句子级特征进行压缩投影;Softmax 分类层对句子级特征进行归一化,用以预测句子的相似度。

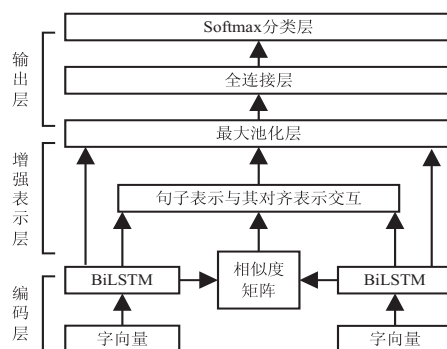


图1 EBiLSTM 模型

2.1 编码层

利用预先训练好的 d 维向量,可以得到句子矩阵 $S = [s_1, s_2, \dots, s_m]$ 和 $T = [t_1, t_2, \dots, t_n]$, m 和 n 为句子 S 和 T 的长度。该文通过 BiLSTM 来捕捉句子中每个单词的上下文表示,LSTM 隐藏层维度大小设置为 u 。给定单词嵌入 x_t ,在 t 时刻隐藏向量 h_t 的计算过程如下:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (2)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (3)$$

$$g_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (4)$$

$$c_t = i_t * g_t + f_t * c_{t-1} \quad (5)$$

$$h_t = o_t * \tanh(c_t) \quad (6)$$

在 LSTM 结构中, σ 是 sigmoid 激活函数, $*$ 是元素相乘,输入门 i 、遗忘门 f 和输出门 o 自适应的控制信息的流动,记忆单元 c 可以记住长距离信息。其中网络参数 $W \in R^{u \times d}$ 、 $U \in R^{u \times u}$ 、 $b \in R^u$ 。

将句子矩阵 S 和 T 输入到 BiLSTM 中,可以得到矩阵 $S_h = [s_{h1}, s_{h2}, \dots, s_{hm}]$ 和 $T_h = [t_{h1}, t_{h2}, \dots, t_{hn}]$,矩

阵 $S_h \in R^{m \times 2u}$, 矩阵 $T_h \in R^{n \times 2u}$ 。

2.2 增强表示层

对于经过 BiLSTM 生成的句子表示 S_h 和 T_h , 通过软对齐的方法来关联句子 S 和 T 之间的相关信息, s_{hi} 为 S_h 的第 i 个向量, 包含有句子 S 中第 i 个字向量及其上下文的特征信息, 而 t_{hj} 为 T_h 的第 j 个向量, 包含有句子 T 中第 j 个字向量及其上下文特征信息, 该文通过两个向量的内积进行两个向量的相似度判断。对于相似度矩阵 M , 其中的元素 M_{ij} 表示 s_{hi} 与 t_{hj} 之间的相似度, 计算公式如下:

$$M_{ij} = s_{hi}^T \cdot t_{hj} \quad (7)$$

如果句子 S 和 T 越相似, 就越可能为句子 S 中的每个部分找到句子 T 中语义对应部分, 反之亦然。这对于评估句子的语义相似性很有帮助。基于这种直观的想法, 应用注意机制对 S_h 和 T_h 进行软对齐, 计算公式如下:

$$s_{hi}^a = \sum_{j=1}^n \frac{\exp(M_{ij})}{\sum_{k=1}^n \exp(M_{ik})} t_{hj}, i = 1, 2, \dots, m \quad (8)$$

$$t_{hj}^a = \sum_{i=1}^m \frac{\exp(M_{ij})}{\sum_{k=1}^m \exp(M_{kj})} s_{hi}, j = 1, 2, \dots, n \quad (9)$$

通过式(8)和式(9)计算得到 S_h 和 T_h 的软对齐表示 $S_h^a = [s_{h1}^a, s_{h2}^a, \dots, s_{hm}^a]$ 和 $T_h^a = [t_{h1}^a, t_{h2}^a, \dots, t_{hn}^a]$ 。

受到 Chen 等人^[15]在自然语言推理中增强推理信息方法的启发, 该文通过计算原始句子表示与对齐表示之间的绝对差和元素乘法, 以度量两个句子间更细微的语义差异, 计算公式如下:

$$s_{ri} = |s_{hi} - s_{hi}^a| \quad (10)$$

$$s_{qi} = s_{hi} \odot s_{hi}^a \quad (11)$$

其中, \odot 表示向量 s_{hi} 与向量 s_{hi}^a 中每个元素按位相乘。

通过式(10)和式(11)计算得到句子表示 S_h 的增强表示: $S_r = [s_{r1}, s_{r2}, \dots, s_{rm}]$ 和 $S_q = [s_{q1}, s_{q2}, \dots, s_{qm}]$, 其中 $S_r \in R^{m \times 2u}$, $S_q \in R^{m \times 2u}$ 。句子表示 T_h 的增强表示 T_r 和 T_q 的计算方法与之相同, 不进行赘述。

2.3 输出层

对于经过长短时记忆网络 BiLSTM 生成的句子表示 S_h 和 T_h , 以及增强表示层生成的增强表示 S_r 、 S_q 、 T_r 和 T_q 采取最大池化的方式提取特征, 分别表示为 S_h^m 、 S_r^m 、 S_q^m 、 T_h^m 、 T_r^m 和 T_q^m , 并对这些向量进行如下操作后进行拼接:

$$M = [S_h^m - T_h^m, S_h^m \odot T_h^m, S_r^m, S_q^m, T_r^m, T_q^m] \quad (12)$$

将拼接后得到的特征表示 M 送入两层的全连接网络, 第一层全连接网络的维度大小为 300 维, 激活函数为 Relu, 将第二层全连接网络的输出经 Softmax 函数归一化后, 用于预测最后句子的相似度。

3 实验与分析

3.1 实验语料准备

实验所用数据为蚂蚁金服的智能客服数据集, 一共 100 000 多个句子对, 正负样本比为 1 : 4。实验随机选用 80% 作为训练集, 其余 20% 作为测试集。

3.2 评价方法

实验选用精确率 (Precision)、召回率 (Recall) 和 F1 值 (F1-Score) 对模型测试结果进行评价, 定义如下:

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \quad (13)$$

$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}} \quad (14)$$

$$F1 = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (15)$$

把相似的句子对定义为正类, 不相似的句子对定义为负类, tp 指正类判定为正类的个数, fp 指负类判定为正类的个数, fn 指正类判定为负类的个数。

3.3 实验设置

为了比较词向量与字向量在该数据集下的实验效果, 分别使用 CNN 和 BiLSTM 神经网络进行实验。分别使用 jieba 和中科院开发的 ICTALAS 分词系统进行分词, 通过 gensim 工具对词向量和字向量进行训练, 词向量和字向量维度都设置为 300 维, 对于字向量训练过程中进行学习更新。使用词向量和字向量时句子最大长度分别设置为 30 和 50。

实验中 CNN 滤波器的窗口宽度设置为 1、2 和 3, 滤波器个数设置为 200, BiLSTM 神经元个数设置为 300。该文训练批次设置为 128, 学习率设为 0.005, 实验中所有超参数都通过五折交叉验证不断调整确定。

3.4 实验结果分析

3.4.1 EBiLSTM 模型有效性验证

使用 ICTALAS 进行分词, 然后进行模型各部分的有效性验证。为了验证增强表示层每一部分的有效性, 将模型分为下面四种情况:

(1) 原特征 (BiLSTM), 即经过双向长短时记忆网络 BiLSTM 捕捉的句子及其上下文表示;

(2) 增强表示一, 即拼接原特征与增强表示层的绝对差特征;

(3) 增强表示二, 即拼接原特征与增强表示层的元素乘法特征;

(4) 最终模型 (EBiLSTM), 即原特征和增强表示层所有特征的拼接。

四种情况的对比结果如表 1 所示。

由表 1 可以看出, 增强表示一与只使用原特征相比, 精确率, 召回率和 F1 值分别提高了 0.98%,

1.13% 和 1.06%。增强表示二与只使用原特征相比,精确率,召回率和 F1 值分别提高了 0.88%,0.37% 和 0.74%。EBiLSTM 与增强表示一和增强表示二相比各项指标有了更高的提升,说明增强表示层的绝对差特征和元素乘法特征对于提升模型性能都有一定的帮助,使 EBiLSTM 模型增强了对句子的语义捕捉能力。

表 1 模型有效性实验

模型	精确率	召回率	F1 值
BiLSTM	45.61	67.07	54.30
增强表示一	46.59	68.20	55.36
增强表示二	46.49	67.44	55.04
EBiLSTM	46.63	69.09	55.68

3.4.2 词向量与字向量模型对比实验

为了比较词向量与字向量的性能,本组实验选取了余弦相似度(Cosine),经典的多视角 CNN 模型和 BiLSTM^[16]模型进行了详细对比,结果如表 2 所示。

表 2 模型对比实验

模型	嵌入方式	精确率	召回率	F1 值
Cosine		21.01	66.00	31.86
CNN	词向量	42.48	61.28	50.18
BiLSTM	(jieba)	41.80	68.22	51.84
EBiLSTM		42.89	68.51	52.76
Cosine		35.55	38.03	36.75
CNN	词向量	43.05	69.01	53.02
BiLSTM	(ICTALAS)	42.95	70.56	53.39
EBiLSTM		42.75	74.56	54.34
Cosine		30.93	47.93	37.60
CNN	字向量	44.39	67.44	53.54
BiLSTM		45.61	67.07	54.30
EBiLSTM		46.63	69.09	55.68

由表 2 可以看出,无论是字向量还是词向量的嵌入方式,双向长短时记忆网络都要比多视角 CNN 效果好,原因可能是 BiLSTM 相比于 CNN 更能抽取句子的整体语义信息。直接用 Cosine 计算效果最差,说明神经网络模型确实捕捉到了句子更深层次的语义信息。

对于词向量嵌入方式,使用 ICTALAS 进行分词要比使用 jieba 分词效果好,原因可能是 jieba 分词粒度更细和产生更多分词错误影响了模型的性能。对于字向量嵌入方式,EBiLSTM 模型相比于多视角 CNN 精确率,召回率和 F1 值分别提高了 2.24%,1.65% 和 2.14%,相比于 BiLSTM 精确率,召回率和 F1 值分别提高了 1.02%,2.02% 和 1.38%。

4 结束语

提出了一种基于字向量和增强表示 BiLSTM 的模

型(EBiLSTM),用于计算智能客服系统中的问句相似度,并且没有使用任何手工特征或基于规则的方法。根据实验可知,EBiLSTM 模型相比其他模型效果更好。

在未来的工作中,将研究词向量与字向量的结合,以更好地捕捉句子对的语义。其次将探索提出的模型在其他语义匹配场景下的适用性,如答案选择和意图识别。

参考文献:

- [1] 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, Portugal: Association for Computational Linguistics, 2015: 1576-1586.
- [2] MUELLER J, THYAGARAJAN A. Siamese recurrent architectures for learning sentence similarity [C]//Proceedings of the thirtieth AAAI conference on artificial intelligence. Phoenix, Arizona, USA: AAAI, 2016: 2786-2792.
- [3] TAI K S, SOCHER R, MANNING C D. Improved semantic representations from tree-structured long short-term memory networks [J]. Computer Science, 2015, 5(1): 36-41.
- [4] JIJKOUN V, DE RIJKE M. Recognizing textual entailment using lexical similarity [C]//Proceedings of the PASCAL challenges workshop on recognising textual entailment. Michigan, USA: ACL, 2005: 73-76.
- [5] WAN S, DRAS M, DALE R, et al. Using dependency based features to take the "para-farce" out of paraphrase [C]//Proceedings of the Australasian language technology workshop. Sydney, Australia: ALTA, 2006: 131-138.
- [6] FERN S, STEVENSON M. A semantic similarity approach to paraphrase detection [C]//Proceedings of the 11th annual research colloquium of the UK special interest group for computational linguistics. Berlin: CLUK, 2008: 45-52.
- [7] MOSCHITTI A. Efficient convolution kernels for dependency and constituent syntactic trees [C]//European conference of machine learning. Berlin: ECML, 2006: 318-329.
- [8] 冯兴杰, 张乐, 曾云泽. 基于多注意力 CNN 的问题相似度计算模型 [J]. 计算机工程, 2019, 45(9): 284-290.
- [9] CHEN Q, ZHU X, LING Z, et al. Enhanced LSTM for natural language inference [C]//Proceedings of the 55th annual meeting of the association for computational linguistics. Vancouver, Canada: ACL, 2017: 1657-1668.
- [10] TAY Y, TUAN L A, HUI S C. Compare, compress and propagate: enhancing neural architectures with alignment factorization for natural language inference [C]//Proceedings of the 2018 conference on empirical methods in natural language processing. Brussels, Belgium: EMNLP, 2018: 1565-1575.

(下转第 186 页)