文章编号：2095-2163(2020)04-0113-07 中图分类号：TP391 文献标志码：A

基于依存句法的句子相似度计算方法

胡雨晴，纪明宇，王晨龙
（东北林业大学 信息与计算机工程学院，哈尔滨 150040）

摘 要：本文针对句子长短不一而影响句子相似度计算的问题，提出了一种基于依存句法的句子相似度计算方法。根据依存句法分析，找出句中各语法成分对句子语义表达的重要程度，制定句子特征提取规则、提取长句的主要特征和补充短句的重要语义信息。利用卷积神经网络中的 2 种卷积模式，扩张卷积和反卷积学习句子特征，并将句子特征进行全连接降维，得到句子的相似度计算结果。此算法在包含中、英文 3 个公共数据集上进行了验证，结果证明了该方法的有效性。

关键词：句子相似度；句子特征；依存句法；扩张卷积；反卷积

Method of sentence similarity calculation based on dependency parsing

HU Yuqing, JI Mingyu, WANG Chenlong
(School of Information & Computer Engineering, University of Northeast Forestry, Harbin 150040, China)

【Abstract】 Dealing with the problem that different sentence lengths affect similarity calculation, this paper proposes the method of sentence similarity calculation based on dependency parsing. Firstly, according to dependency parsing analyses the semantic expression importance of each grammatical component in a sentence, making a rule of sentence feature extraction, which extracts the major features of the long sentence and to supplement the important semantic information of the short sentence. In addition, it learns sentence features by dilated convolution and deconvolution that are convolution neural network two models, and reduce full join dimension the sentence features to get the result of sentence similarity calculation. Finally, the method is verified on three data sets selected in this paper.

【Key words】 sentence similarity; dependency parsing; sentence features; dilated convolution; deconvolution

0 引 言

近年来，文本语义的相似度计算广泛应用于机器翻译、信息检索、对话系统等领域[1]。语义的相似度是指用于比较语义实体（如词语、句子、或定义为知识库的概念和实例）之间的语义相似性或相关性的方法[2]。语义实体中的句子在实际应用场景中最为常见，国内外学者对于句子的相似度计算提出了许多方法。

句子的特征提取是句子相似度计算的核心，目前主要有两类句子相似度计算方法：人工提取句子特征和利用神经网络提取。第一类方法：以句中的关键词、词频、语义成分等句子特征，定义计算句子相似度的计算公式。例如，Gunasinghe 等人统计文档的词频作为句子特征，并且利用余弦距离度量相似度[3]。此类方法存在特征稀疏的问题，语义度量不够准确。第二类方法：利用神经网络自动提取句子级或词语级的语义特征。如 Palangi 等人利用循环神经网络模型，将句子中的词语特征依次提取出来[4]。Zhuang 等人引入注意力模型到循环神经网络中更多的获取句子的语义信息[5]。而近年来越来越多的学者将常用于图像处理的卷积神经网络用来处理文本数据。Kim 等人提出 Text-CNN 方法，首次利用卷积神经网络结构学习句子的局部特征[6]。He 等人对卷积网络进行了改进，提出了 2 种卷积方式和 3 种池化方式，从多个角度学习句子特征[7]。

但是，应用神经网络的方法存在一些不足：神经网络的输入端需要固定长度的句子，而文本中的句子大多长短不一，对于长句通常使用直接截取的方法，句子会丢失一部分语义信息导致语义特征提取不全面；短句的处理方式是直接用零来补充缺失的向量部分，这样会使句子所含的重要语义信息过

基金项目：东北林业大学大学生创新创业训练计划项目资助（201910225175）；中央高校基本科研业务费专项资金资助（2572015CB32）；国家自然科学青年科学基金（61806049）。

作者简介：胡雨晴（1999-），女，本科生，主要研究方向：数据挖掘、自然语言处理；纪明宇（1980-），男，博士，副教授，主要研究方向：机器语言、自然语言处理；王晨龙（1994-），男，硕士研究生，主要研究方向：自然语言处理、机器语言。

通讯作者：纪明宇 Email：962750672@qq.com

收稿日期：2020-02-07

少[8]。而且,仅仅使用基本的神经网络结构学习句子语义特征不够充分。

针对上述问题,本文在依存句法的基础上对长句的重要特征进行提取、对短句中重要的语义信息进行补充。为了提高卷积神经网络学习句子特征的能力,本文基于 Text-CNN 模型,引入了扩张卷积和反卷积的结构,来获取多个层面的句子语义特征。

1 基于依存句法的长短句特征提取

1.1 英文依存句法分析

依存句法(Dependency Parsing)是指通过分析语言单位内成分之间的依存关系揭示其句法结构[9]。StanfordCoreNLP 是由斯坦福大学开发的关于自然语言处理的工具包,其中包括句子的依存句法分析、分词、词性还原等功能。其工具将句子的语法结构以句法结构树的形式展现,并对句中的每一个词语的语法成分和词性进行了标注。

本文利用 StanfordCoreNLP 对句子进行依存分析后的词语标注,制定出提取长短句特征的规则。对于长句,采用去掉一部分语义较弱的词语来削减句子长度;而对于短句,则增强语义信息较为重要的词语来增加句长。如,在英文句子中常见的限定词“the”、“to”及介词“in”、“for”等等出现频率较高,而包含的语义信息又过少,则可在长句中删减。而句中的名词短语、动词、副词、人称代词等句中语义重要的词语应在短句中进行补充。

应用上述长短句提取规则,作用于本文选用的 2 个公开实验数据集 MSRP 和 STS,其长短句数量变化情况见表 1。由表 1 可知,MSRP 数据集的句长小于 12 的 句子从 821 条减少到 484 条,长度超过 20 的句子减少了 2405 条。STS 数据集的短句子过多而且大多只有 5 个单词,短句子的语义信息过少不利于提取出句子的特征,通过本文提出的依存句法提取规则,补充短句语义信息使短句减少了 151 条。2 个数据集所减少的长句和短句都分别在句长适中的区间相应的增加了 2 742 条和 2 261 条。

表 1 依存句法规则下英文长短句变化情况

Tab. 1 Changes of English short and long sentences under dependency rules

方法	MSRP 数据集			STS 数据集		
	<12	12-20	20<	<5	5-15	15<
应用规则前	821	6 311	4 470	570	13 990	2 694
应用规则后	484	9 053	2 065	419	16 251	584
数量变化	-337	+2 742	-2 405	-151	+2 261	-2110

1.2 中文依存句法分析

中文的依存句法分析,则利用哈尔滨工业大学提供的语言技术平台(Language Technology Platform, LTP)[10]。LTP 中常见的依存关系类型和对应标注见表 2。

表 2 LTP 中依存关系类型及标注

Tab. 2 Dependency types and annotations in LTP

关系类型	标注	关系类型	标注
核心关系	HED	定中关系	ATT
主谓关系	SBV	状中关系	ADV
动宾关系	VOB	动补结构	CMP
并列关系	COO	介宾关系	POB

句法分析的目的是构造句子的句法结构树(Syntactic Structure Tree)[11]。句法结构树是由词语、词性标注、语法成分和表示依存关系的依存弧组成的[12]。句法结构树与句子的结构有着密切关系。本文将长度超过 10 个字的句子定义为长句子,长度小于 5 个字的句子定义为短句子,5 到 10 个字之间的句子定义为标准句。图 1 是本文数据集中的几个常见用句,对 3 种句子的 LTP 句法结构描述。

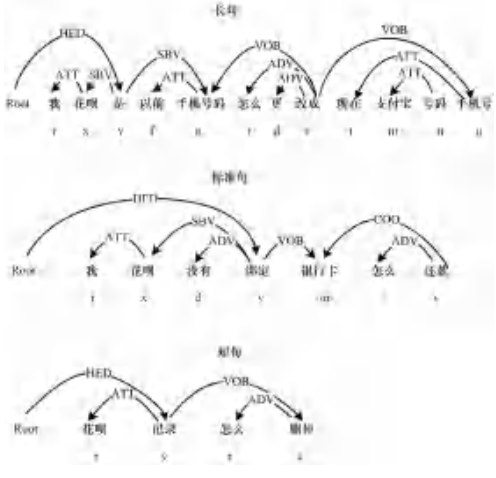


图 1 LTP 句法结构树实例

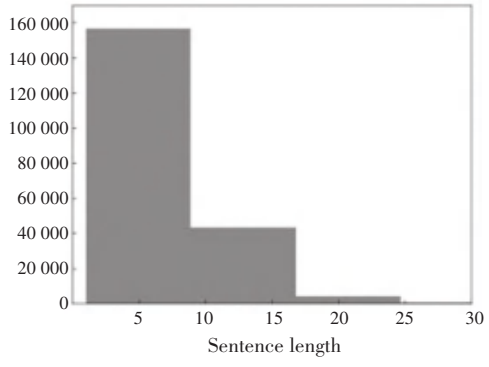
Fig. 1 An example of lip syntax structure tree

由图 1 可以看出:长句、标准句和短句中的核心成分(HED),主语(SBV)、宾语(VOB)与其它成分之间的依存弧最多,即依存关系最多。如果缺少这 3 类成分句子的语义表达将会受到严重影响。其

次，依存关系较多的是修饰主语的定语成分（ATT）和修饰谓语的状语成分（ADV）。其它语法成分的依存关系较弱，如并列关系（COO）、动补关系（CMP）等。可见，依存关系越多的语法成分对句子的语义表达越重要，将长句中依存关系较多的语法成分提取出来，可以在保留长句重要语义信息的同时压缩句长；补充短句子中依存关系较多的语法成分，可以增加短句长度，并提高短句中依存关系较多语法成分的出现概率，使句子的语义表达更充分。

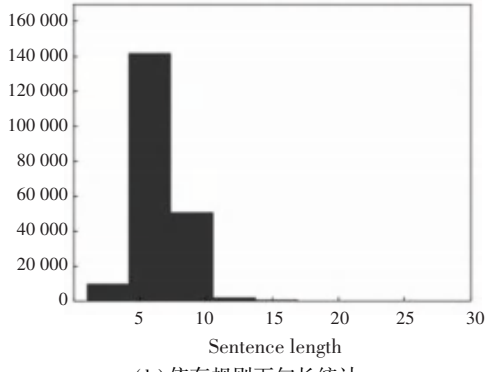
基于此，本文提出的长短句提取规则如下：对长句而言，将句中标注为 SBV、HED、VOB 的主谓宾等核心词全部提取；句中标注为 ATT、ADV、CMP（动补）修饰主谓宾的助词选择性提取；其它标注不进行提取。对于短句中存在修饰词，周围大多是被修饰的主语、谓语等依存关系较多成分的情况，采用复制短句中标注为 ATT、ADV、VOB 等修饰词的前一个词到当前词之后的方式，进行补充语义。

应用上述长短句提取规则，作用于本文实验数据集的 102 477 对句子，相应的长短句数量变化情况如图 2 所示。



（a）句长统计

（a）Sentence length statistics



（b）依存规则下句长统计

（b）Statistics of sentence length under dependency rules

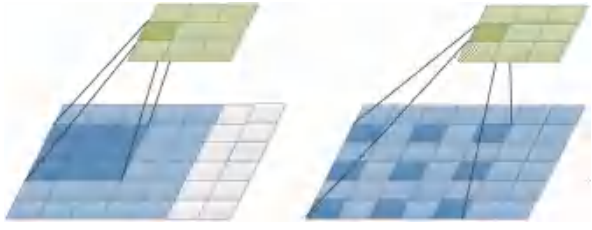
图 2　运用依存规则后的句长对比

Fig. 2　Sentence length comparison after using dependency rule

由图 2（a）可知，智能数据集的句长主要分布在 3 到 10 个词语之间，但是词语少于 5 个词，包含语义信息太少，不利于句子相似度的比较。在本文的依存句法提取规则下，数据集中的句长小于 5 和大于 10 的句子大幅度的减少，如图 2（b）所示。可见，数据集中过长或过短句子在依存句法规则下大量减少，便于神经网络的输入。

2　扩张卷积和反卷积学习句子特征

扩张卷积（Dilated Convolution）又称空洞卷积，由国外学者 Yu 等人[13]提出，在图像分割、目标检测等领域被广泛应用。扩张卷积和普通卷积相比，除了卷积核的大小以外，还有一个扩张率参数表示扩张的大小。利用这种结构，在保持参数个数不变的情况下增大了卷积核的感受野，同时可以保证输出特征映射的大小保持不变。普通卷积和扩张卷积的对比如图 3 所示。



（a）普通卷积　　（b）扩张卷积

（a）Convolution　　（b）Dilated Convolution

图 3　扩张卷积和普通卷积对比

Fig. 3　Comparison between convolution and dilated convolution

图 3 中绿色区域代表卷积核，深蓝色区域为感受野。同样是 3x3 的卷积核，扩张卷积可以获得更大的感受野，学习到更远的特征信息。国外学者 Bai 等人[14]提出的时间卷积网络 TCN 中也用到了扩张卷积的思想。本文利用扩张卷积相比普通卷积具有更大的感受野，可以同时获取更远的上下文信息，并能保持计算量变化较小。利用扩张卷积代替文献[6] Text-CNN 方法中的普通卷积部分，得到融合扩张卷积的 Text-CNN 模型结构如图 4 所示。

反卷积（Deconvolution）也称转置卷积，是由 Zeiler 等人[15]提出，常用于图像处理的场景分割、生成模型等领域。最近流行的生成对抗网络（GAN）生成器模块，采用的就是反卷积操作，将低分辨率图像生成高分辨率图像[16]。反卷积操作是卷积的逆过程，即卷积层的反向传播就是反卷积层的前向传播，其模型结构如图 5 所示。

从图 5 可以看出，2×2 的输入，通过 3×3 步长为 1 的过滤器，反卷积得到 4×4 的输出，反卷积时空白的部分用 0 补充。通过反卷积可以对短句子的信息

进行扩充提取,利于短句子在相似度比较时可以考虑到更多的信息。

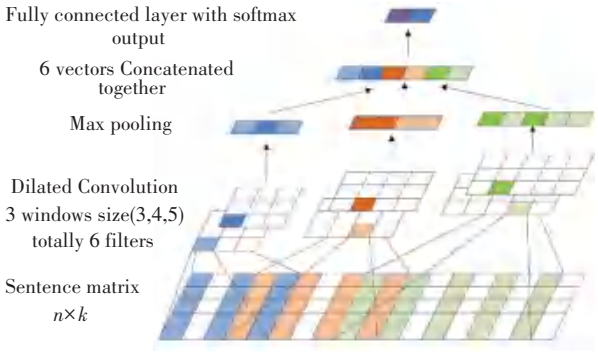


图 4 融合扩张卷积的 Text-CNN

Fig. 4 Text-CNN that blends dilated convolution

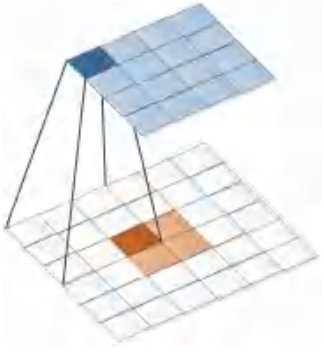


图 5 反卷积模型结构

Fig. 5 Deconvolution structural model

3 实验与结果分析

3.1 实验语料准备

本文选用 2 个英文公开的数据集和一个中文智能客服数据集进行实验。3 个数据集分别是微软提供的 MSRP 数据集[17]、2012-2017 年 SemEval 跨语言语义文本相似性(cross-lingual Semantic text Similarity, STS)任务的数据集[18]和蚂蚁金服提供的智能客服数据集。3 个数据集的统计信息见表 3。

表 3 3 个数据集的统计信息

Tab. 3 Statistics of three data sets

数据集	训练集	测试集	正样本	负样本	总计
MSRP	4 640	1 161	3 900	1 901	5 081
STS	6 900	1 726	2 853	5 773	8 626
智能客服	81 981	20 496	18 685	83 792	102 447

对于英文数据集,卷积神经网络之前的词嵌入层,选用斯坦福大学利用 Glove 语言模型,在 2014 年维基百科和 Gigaword 语料预训练得到400 000个英文词汇的 100 维词向量[19]。对于中文的数据集,词嵌入层选用 Li S 等人基于上下文特征(单词、n-gram、字符等),在中文维基百科语料预训练得到 2 129 000个中文词汇的 300 维词向量[20]。

3.2 评价指标

文本语义相似度计算方法的评价指标通常包括准确率、召回率、精确度、F1 值和评价模型稳定的 ROC(Receiver Operating Characteristic)曲线,以及曲线面积值 AUC(Area Under ROC Curve)。模型的 ROC 曲线越靠近左上角,模型的面积 AUC 越大,表示模型的稳定性越高。

3.3 实验结果

为了验证本文方法的有效性,将文献[6]的 Text-CNN 方法、文献[7]的 Mp-CNN 方法、融合了反卷积的 Text-cnn 的 De-CNN 方法和本文提出的融合扩张卷积 Text-cnn 的 Dilated-CNN 方法,上述 4 种方法结合依存句法长短句提取规则 Dependency 的方法,进行实验对比。表 4 为几种方法针对准确率、召回率、精确度和 F1 值等评价指标的实验对比。

表 4 主要评价指标对比

Tab. 4 Comparison of main evaluation indexes %

方法	MSRP 数据集				STS 数据集				智能客服数据集			
	准确率	召回率	精确度	F1 值	准确率	召回率	精确度	F1 值	准确率	召回率	精确度	F1 值
Text-CNN[6]	0.769 4	0.913 3	0.757 8	0.835 2	0.763 5	**0.654 3**	0.818 6	**0.704 7**	**0.454 5**	**0.607 0**	0.795 5	**0.519 8**
Mp-CNN[7]	0.748 1	0.887 9	0.723 6	0.812 0	0.713 4	0.557 6	0.779 6	0.626 0	0.456 8	0.607 8	0.796 6	0.521 6
Dilated-CNN	**0.795 4**	0.887 6	0.771 0	**0.839 0**	0.767 5	0.634 0	0.815 4	0.694 4	**0.438 8**	**0.718 5**	0.781 1	**0.544 9**
De-CNN	0.764 1	0.892 3	0.742 4	0.823 2	0.709 6	0.588 5	0.784 2	0.643 4	0.463 9	0.527 2	0.802 7	0.493 5
Text-cnn+Dependency	0.781 7	0.903 5	0.765 5	0.838 2	0.717 0	**0.724 8**	0.814 4	**0.720 9**	0.445 7	0.609 5	0.790 6	0.514 9
Mp-cnn+Dependency	0.765 7	0.875 1	0.736 0	0.816 8	0.644 9	0.623 2	0.761 9	0.633 8	0.459 1	0.591 0	0.798 4	0.516 7
Dilated-CNN+Dependency	**0.831 5**	0.911 0	0.816 0	**0.869 4**	0.751 7	0.652 6	0.813 8	0.698 6	**0.490 3**	**0.689 4**	0.812 7	**0.573 0**
De-CNN+Dependency	0.799 7	0.885 8	0.774 1	0.840 6	0.710 4	0.637 2	0.794 1	0.671 8	0.451 8	0.615 6	0.793 7	0.521 2

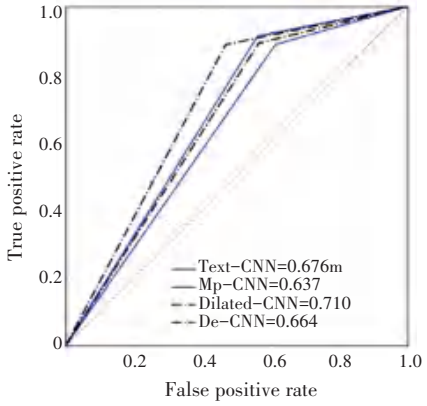
3.3.1　准确度和 F1 值的对比分析

由表 4 可见，Dilated−CNN 方法在 MSRP 数据集上的准确率和精确度分别比其它 Text−CNN 方法高 2.60% 和 1.32%，而召回率低 2.57%。在 STS 数据集上两方法的准确率和召回率相差无几。在中文智能客服数据集上准确率降低了 1.57%，但召回率提升了 11.15%，最终的 F1 值提高了 2.51%。表明 Text−CNN 引入扩张卷积结构后，可以学习到更远距离的词语语义，在长句较多的 MSRP 数据集，准确率会得到提高。而在 STS 短句较多的数据集变化不大。

Dilated−CNN 结合依存句法长短句提取规则 Dependency 后，相比 Dilated−CNN 方法在 MSRP 数据集上的准确率提高了 3.61%，F1 值提高了3.04%。而其它方法在结合依存句法提取规则后，准确率和 F1 值都有一定的提升。在 STS 数据集的表现上准确率有所降低，但召回率却有所提高，最终的 F1 值提升不明显。在智能客服数据集上，结合了依存句法长短句提取规则的 Dilated−CNN 的表现最优，相比 Dilated−CNN 准确率和 F1 值分别提高了 5.15% 和 2.81%。表明在扩展卷积的基础上再引入依存分析，长短句得到语义补充后准确率和 F1 值在 Text−CNN 的方法上再次得到提高，并且应用到智能客服中文数据集也有明显的提升。

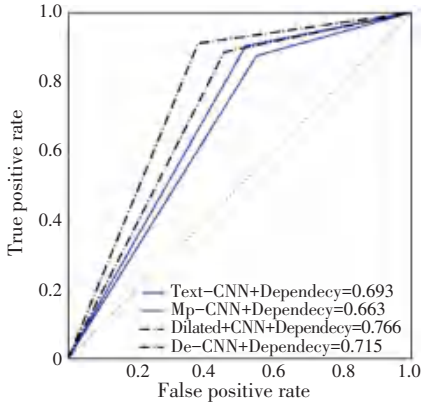
3.3.2　模型稳定性 ROC 曲线的对比分析

上述方法在 ROC 曲线表现方面的实验对比结果如图 6、图 7、图 8 所示。



（a）ROC 曲线对比

（a）ROC curve comparison

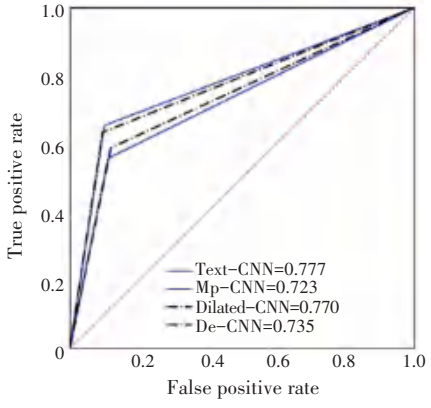


（b）依存规则下 ROC 曲线对比

（b）ROC curve comparison under dependency rules

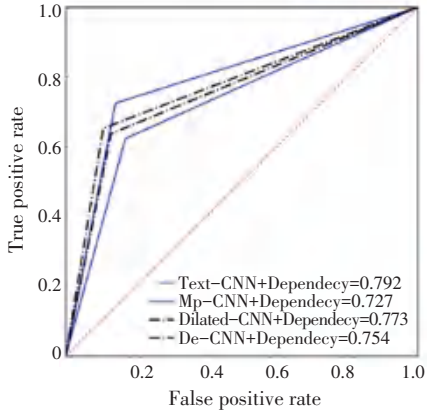
图 6　MSRP 数据集 ROC 曲线对比

Fig. 6　ROC curve comparison of MSRP data set



（a）ROC 曲线对比

（a）ROC curve comparison

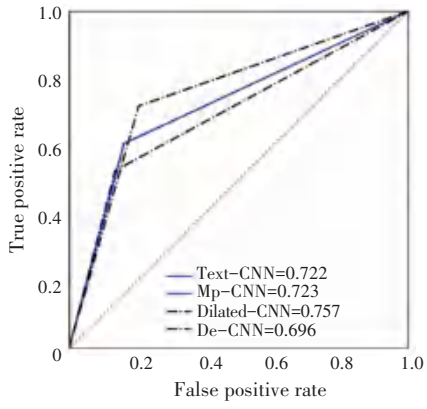


（b）依存规则下 ROC 曲线对比

（b）ROC curve comparison under dependency rules

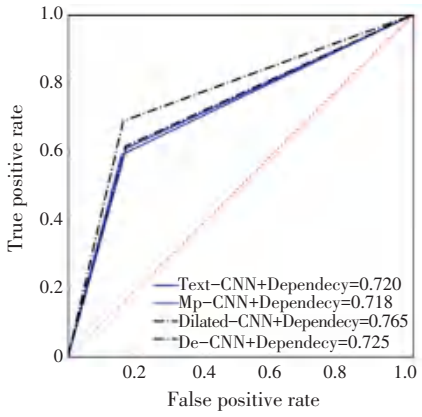
图 7　STS 数据集 ROC 曲线对比

Fig. 7　ROC curve comparison of STS data set



(a)ROC 曲线对比

(a) ROC curve comparison



(b)依存规则下 ROC 曲线对比

(b) ROC curve comparison under dependency rules

图 8 智能客服数据集 ROC 曲线对比

Fig. 8 ROC curve comparison of intelligent customer service data set

从图 6(a)可以看出,在 MSRP 数据集下 Dilated-CNN方法的 ROC 曲线最靠近左上角,而且面积值最高为 0.710,De-CNN 方法的面积值为 0.664,较低于 Text-CNN 方法的 0.676。图 6(b)表明在依存规则下,以上几种方法的 ROC 曲线表现均有提升,而且结合依存句法后的 Dilated-CNN 和 De-CNN的 ROC 面积值为 0.766 和 0.715,高于Text-CNN 和Mp-CNN方法。

由图 7(a)可知,在 STS 数据集下,Dilated-CNN 和 Text-CNN 方法相差不大。图 7(b)在依存规则下各方法的 ROC 曲线面积值提升不明显,而各方法的曲线较原来相比均更靠近上方,表明召回率都得到了提高。

由图 8(a)可知,在智能客服数据集下 Dilated-CNN 方法的曲线最靠近左上方,且面积值最高为 0.757。图 8(b)在依存规则下 Dilated-CNN 和 De-CNN 的 ROC 曲线面积值,相比未用依存规则之前分别提高了 0.08 和 0.29。表明模型的稳定性在引入依存分析后得到一定提升。

4 结束语

本文提出了一种基于依存句法的句子相似度计算方法,在中英文 3 个公共数据集上验证了该方法的有效性。在准确度和 F1 值的表现上本文提出的模型方法较其它方法有所提升;而且在模型稳定性上也有较好表现。未来,将进一步研究中文的弱语法性和口语表达多样性等影响句子相似度计算的问题。

参考文献

[1] OMAR N A, KASIM S, FUDZEE M F, et al. A review of semantic similarity approach for multiple ontologies [J]. International Journal of Information and Decision Sciences, 2018, 10(3): 212-221.

[2] HARISPE S, RANWEZ S, JANAQI S, et al. Semantic Similarity from Natural Language and Ontology Analysis [J]. Synthesis Lectures on Human Language Technologies, 2015, 8(1): 254.

[3] GUNASINGHE U, DE SILVA W A M, DE SILVA N, et al. Sentence similarity measuring by vector space model [C]//2014 14th International Conference on Advances in ICT for Emerging Regions (ICTer). IEEE, 2014: 185-189.

[4] PALANGI H, DENG L, SHEN Y, et al. Deep sentence embedding using long short-term memory networks: analysis and ap-plication to information retrieval [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2016, 24(4): 694-707.

[5] ZHUANG W, CHANG E. Neobility at SemEval-2017 Task1: An Attention-based Sentence Similarity Model. [J]. Meeting of the Association for Computational Linguistics, 2017: 164-169.

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

[7] HE H, GIMPEL K, LIN J J, et al. Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Net-works[C]// empirical methods in natural language pro-cessing, 2015: 1576-1586.

[8] 庞亮,兰艳艳,徐君. 深度文本匹配综述[J]. 计算机学报,2017, 40(4):985-1003.

[9] NIVRE J. Dependency Parsing [J]. Language & Linguistics Compass, 2010, 4(3):138-152.

[10] 江腾蛟,万常选,刘德喜,等. 基于语义分析的评价对象-情感词对抽取[J].计算机学报,2017,40(3):617-633.

[11] 邵帅,刘学军,李斌. 基于关键句分析的微博情感倾向性研究[J]. 计算机应用研究,2018,35(4):982-987.

[12] 甘丽新,万常选,刘德喜,等. 基于句法语义特征的中文实体关系抽取[J]. 计算机研究与发展,2016,53(2):284-302.

[13] YU F,KOLTUN V. Multi-Scale Context Aggregation by Dilated Convolutions [J]. arXiv: Computer Vision and Pattern Recognition, 2015.

[14] AI S, KOLTER J Z, KOLTUN V, et al. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Se-quence Modeling.[J]. arXiv: Learning, 2018.

(下转第 123 页)