计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(11):232−236 [doi: 10.15888/j.cnki.csa.007556]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

Transformer 及门控注意力模型在特定对象立场检测中的应用①

何孝霆[1,2], 董　航[3], 杜义华[1]

[1](中国科学院 计算机网络信息中心, 北京 100190)
[2](中国科学院大学, 北京 100049)
[3](清华大学 工业工程系, 北京 100084)
通讯作者: 杜义华, E-mail: yhdu@cashq.ac.cn

摘　要: 立场检测旨在通过意见持有者的表达来判断其是支持还是反对给定对象. 准确地检测立场不仅需要对表达内容进行信息提取, 而且还需要针对特定的对象进行立场匹配. 本文将 Transformer 结构与门控注意力模型应用在特定对象立场检测中. 该模型可以有效利用推文中独特的标签短语信息, 同时结合门控注意力机制形成推文与对象的匹配信息, 从而更好地判断该推文对该对象的真实立场. 此外, 该方法将情感分类作为辅助任务, 可以更充分地将情感信息纳入立场判别当中, 提高模型的表现. 实验结果表明, 该模型在 SemEval-2016 数据集上表现优于最新的深度学习方法.

关键词: 立场检测; Transformer; 注意力机制; 词片模型

引用格式: 何孝霆,董航,杜义华.Transformer 及门控注意力模型在特定对象立场检测中的应用.计算机系统应用,2020,29(11):232–236. http://www.c-s-a.org.cn/1003-3254/7556.html

Transformer and Gated Attention Model on Target-Specific Stance Detection

HE Xiao-Ting[1,2], DONG Hang[3], DU Yi-Hua[1]

[1](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](Department of Industrial Engineering, Tsinghua University, Beijing 100084, China)

Abstract: Stance detection tells whether the expressions of opinion holders are in favor of or against the given objects. To accurately detect stance, the information of the expressed contents must be extracted, alongside a stance match for specific objects. In this study, the Transformer structure and gating attention is applied to specific object stance detection. By effectively utilizing the tag phrase information of the posts and the matching information between posts and objects, which are a result of gating attention mechanism, it delivers a better judgment over the post's authentic stance regarding the object. Moreover, this approach takes emotional classification as an auxiliary task to fully include emotional information into stance detection for better performance. Experimental results show that the model is superior to the latest deep learning method on the SemEval-2016 dataset.

Key words: stance detection; Transformer; attention mechanism; Wordpiece

随着社交媒体的迅速发展, 用户对各种对象 (例如政客和宗教) 的丰富意见很容易地得以传播. 这些意见可以帮助优化管理系统, 并可以洞察重要事件. 例如在总统选举中, 从社交媒体上的内容中判别用户对总统

① 基金项目: 中国科学院战略性先导科技专项 (C 类) (XDC02060100)
Foundation item: Category C Strategy Priority Research Program of Chinese Academy of Sciences (XDC02060100)
收稿时间: 2020-01-07; 修改时间: 2020-01-22; 采用时间: 2020-03-11; csa 在线出版时间: 2020-10-29

候选人的立场可以更好地预测民意, 对政治走向进行判断. 立场检测任务旨在确定人们对特定对象是赞成, 反对, 还是持中立态度.

针对对象的立场检测与情感极性分类是不同的问题. 情感极性分类是不针对特定对象的, 而立场检测是针对于特定对象. 但这个对象不一定必须在推文中出现, 因为可以通过隐式提及对象或谈论其他相关对象来表达针对特定对象的立场. 此任务的主要挑战是分类器作出的决定必须要针对特定对象. 来自 SemEval-2016 的特定对象立场检测数据集的训练数据示例可以在表 1 找到. 同时可以看到推文是由用户生成, 简短且嘈杂. 并具有独特的特征, 如短语标签. 已有工作并未充分利用这些推文独特的特征.

表 1 针对特定对象的立场检测示例

对象	推文	立场
Donald Trump 唐纳德·特朗普	#DonaldTrump my tell it like it is but his comments speaks to a prejudice and cold heart. #唐纳德·特朗普 我这样说, 但他的评论表达了偏见和冷漠的心.	反对
Hillary Clinton 希拉里·克林顿	I love the smell of Hillary in the morning. It smells like Republican Victory. 我喜欢早上希拉里的气味. 闻起来像共和党胜利.	反对

受词片模型[1] 和 Transformer[2] 在语言建模任务的有效性所带来的启发[3]. 我们将 Transformer 与门控注意力应用于特定对象立场检测任务, 同时将情感预测作为立场检测的辅助任务. 具体来说, 该方法首先通过 Wordpiece 模型将原始文本拆分成词片序列, 文本中所有的短语标签会在此步被有效拆分成单词组合; 随后将词片序列输入 Transfomer 进行编码; 紧接着, 门控注意力被用来识别与给定对象相关的重要单词. 此外, 取中间编码预测情感, 并将立场检测和情感预测的损失整合到最终损失中. 上述模型的表现在 SemEval-2016 数据集上得到了验证. 总体来说, 我们的贡献概括如下:

首先, 我们将 Wordpiece 和 Transformer 结构应用于立场检测模型, 该模型借助建模与编码可分离短语来改善在立场检测任务上的表现; 其次, 我们将门控注意力应用于“感知”特定对象, 在细粒度语义层面, 使得模型可以根据对象对文本进行自适应编码; 最后, 情感得分预测任务的加入进一步提升了立场检测的效果.

1 针对特定对象的立场检测模型

立场检测任务即判断给定文本对于特定对象的立场, 类似于进行情感极性的分类任务. 但是, 与情感分类不同, 在给定的句子中可能未明确提及立场检测的对象. 考虑以下示例推文: @realDonaldTrump is the only honest voice of the @GOP and that should scare the shit out of everyone!(译文: @唐纳德·特朗普是@GOP 唯一的诚实声音, 应该吓到所有人！). 我们进行立场检测的对象是 Hillary Clinton/希拉里·克林顿, 我们观察到即使对象希拉里·克林顿并未出现在此推文中, 我们仍可以推断出, 意见持有者不太可能赞成希拉里·克林顿. 因此识别对象信息对于立场检测至关重要. 此前的立场检测研究工作包括基于特征工程[4]、卷积神经网络[5,6]、循环神经网络[7] 的工作, 但是他们没有考虑针对特定对象的问题. 为了解决这个问题, 有工作已经提出了几种针对特定对象的注意力机制方法[8–10], 将对象信息嵌入句子表示中.

在社交媒体中, 往往存在短语标签. 如#NoHillary, 我们称这类短语标签为可分离短语, 因为这些短语标签往往可以被拆分成若干个单词, 如“No Hillary”. 这些短语标签对立场检测往往很有帮助. 通过准确识别并理解特定对象及短语标签, 可以更有效的判断推文立场. 但已有方法并未有效的利用这些短语标签.

情感信息也被证明对立场检测任务有帮助. 例如 Sobhani 等[11] 发现情感特征与其他特征结合使用时, 对立场检测任务有帮助. Sun 等[10] 提出一个层次注意力模型用来学习情感信息的重要性, 而不是直接将情感特征整合进向量表示中. 后来 Sun 等[12] 提出了一个联合模型, 可以同时确定立场和情感.

为了有效利用可分离短语、特定对象以及情感信息, 我们利用 Transformer 和门控注意力构造一个针对特定对象立场检测的模型. 模型主体结构如图 1 所示. 该模型包含带有词片的 Transformer 编码器和细粒度对象注意力两部分, 除了标准的立场检测任务外, 还附加了情感预测辅助任务帮助立场检测.

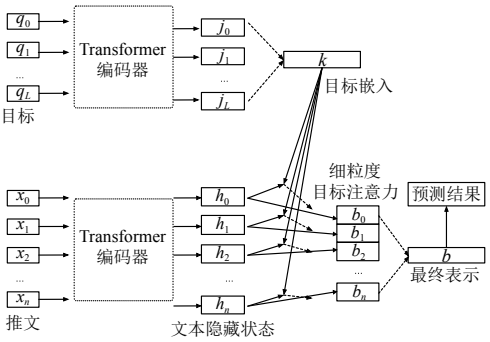

图 1 模型主体结构

1.1 带有词片的 Transformer 编码器

给出示例推文, @Reince This is very credible! Good work! America is desperately in need of good leadership. #Vote- GOP #NoHillary. (译文: @Reince 这是非常可信的！干得好！美国迫切需要良好的领导才能. #反对希拉里.) 我们进行立场检测的对象是 Hillary Clinton. 如果我们不考虑标签#NoHillary 就很难推断出正确的立场标签. 因此无法分离的连续短语会导致重要对象信息的丢失. 在分离连续短语后, 我们需要知道 No 指向的是 Hillary 从而判断推文作者反对希拉里.

为了能够有效将标签中的连续短语分割成有意义的单词, 我们使用无监督词片模型 (Wordpiece), 用来分离连续短语. 而为了更好地捕捉到句子内部的依赖关系, 我们使用 Transfomer 结构代替此类任务中通常使用的循环神经网络结构, 并提升了模型表现. 接下来将介绍具体过程.

首先, 我们给定原始推文序列 s, 并对 s 应用词片模型, 生成确定性分段, 即词片序列.

词片模型的一种主要实现方式为字节对编码 (Byte-Pair Encoding, BPE) 算法[1]. 该算法首先将原始文本视为字母组成的符号序列, 每次合并最频繁的相邻符号对, 并将合并后的相邻符号对作为新的符号. 直到达到指定的合并次数. 如我们的原始文本是 Jet makers feud over seat width with big orders at stake. 应用词片模型后可以形成词片序列 J et_ makers_ fe ud_ over_ seat _ width_ with_ big_ orders_ at_ stake_. 此时“_”是一个特殊字符, 代表单词结束. 再如例子#NoHillary, 在应用词片模型后会被分成“#”、“No”和“Hillary” 3 个符号.

在对 s 使用词片模型进行切分后, 得到输入词片序列$s=\{s_0,s_1,s_2,\cdots,s_L\}$. 接下来将其送入嵌入层, 每个单词都是一个向量表示$X=\{x_0,x_1,x_2,\cdots,x_n\}$, 其中 n 是句子长度. 下一步使用 Transformer 进行特征提取. Transformer[2] 作为编码器的结构如图 2 所示. 在时间 t, Transformer 输出的隐向量通过所有时刻的输入决定.

$$h_i=E_{\text{Transformer}}(x_0,x_1,x_2,\cdots,x_n)$$

同样的, 对于特定对象 q, 同样对其应用词片模型得到词片序列$q=\{q_0,q_1,q_2,\cdots,q_L\}$, 其中 L 是对象短语的长度, 并同样使用 Transformer 进行特征提取.

$$j_i=E_{\text{Transformer}}(q_0,q_1,q_2,\cdots,q_n)$$

最终对象向量 k 是 Transformer 输出在对象短语上所有词的平均.

$$k=\frac{\sum_{i=1}^{L}j_i}{L}$$

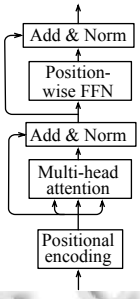

图 2 Transformer 编码器

1.2 细粒度对象门控注意力

当人类被要求标记一条推文对特定对象的立场时, 他们很可能将有关对象的信息牢记在心, 并更多的关注与对象相关的部分. 注意力机制首先应用于机器翻译任务, 允许神经网络自动为源句子中与预测对象的有关的词分配权重, 并屏蔽不相关的标记. 注意力机制已应用于问答、生成、情感分析等任务.

我们将注意力机制应用于该模型, 使模型能自动计算推文中词的权重, 从而反映出不同词在特定对象时的重要性. 在标准注意力机制中, 推文向量被表示为隐藏状态的加权和, 因而在这样的模型下文本表示和对象表示没有直接交互作用. 而直觉上, 人类只会关注与对象相关的部分词, 例如文本中的某个词可能暗示了对象的反对立场, 但这个词与其他对象无关. 标准注意力机制所使用的 Softmax 归一化权重间接使得词与词之间存在了关系, 这与我们的直觉不符. 词的权重得分应该是独立的, 即独立考虑每个词与对象的相关性.

为了独立考虑每个词语对象的相关性, 我们通过引入隐藏状态和对象向量表示之间的交互, 即使用门控结构将当前词的注意力拓展到更细粒度的语义级别. 针对特定对象的推文隐藏状态表示为:

$$b_i=a_i\odot h_i$$

注意力权重a_i用来确定h_i对最终对象的重要性. 这是通过一个门控结构计算得出的:

$$a_i=\sigma(e_i)$$

其中, a_i的计算可以有多种选择, 如内积注意力或多层感知机注意力. 在本研究中e_i使用了多层感知机注意力来计算, 即将输入通过含单隐藏层的多层感知机变换.

$$e_i=V^T(\tanh(W_hh_i+W_kk))$$

为了得到最终的句子表示, 借鉴文献 [13] 的做法, 使

用对每个词的向量表示取平均来作为最终句子的表示.

$$b=\frac{\sum_{i=1}^{n} b_i}{n}$$

在得到最终的句子表示 b 后, 首先通过多层感知机进行变换, 多层感知机输出的维度为可能的立场类别数. 将输出送入 Softmax 层, 转换为概率分布$\hat{y}$. 由于模型的所有部分都是可导的, 因而我们可以使用标准的反向传播以端到端的方法进行训练. 我们使用多类别交叉熵作为损失函数, 该损失函数定义如下:

$$L_{\text{main}}=\sum_{N}\sum_{i\in 1,\cdots,z}-\widehat{y_i}\log(y_i)$$

其中, N 是训练数据集, z 是立场类别数.

1.3 情感预测辅助

先前的研究表明, 情感信息对于立场检测任务是有帮助的[11]. 为此我们同时加入情感得分的预测来改善立场检测任务. 我们对训练集中的每个推文进行情感打分, 标注情感得分为–0.5 到 0.5, 其中–0.5 代表最消极, 0.5 代表最积极. 例如对推文“Hillary is our best choice if we truly want to continue being a progressive nation.” (译文: 如果我们真正想继续成为一个进步的国家, 希拉里是我们的最佳选择.) 标注得分 0.41, 代表比较偏向于积极.

预测情感得分将作为模型的辅助训练任务. 具体而言, 参照文献 [14] 的方法, 对原始推文序列通过 Transformer 输出的隐向量h_i取平均作为情感表示向量.

$$r=\frac{\sum_{i=1}^{n} h_i}{n}$$

将 r 送入多层感知机进行变换, 输出维度为 1, 即预测的情感得分 u. 我们使用均方误差作为损失函数.

$$L_{\text{aux}}=\sum_{N}\frac{1}{2}(p-u)^2$$

合并立场检测主任务和情感预测辅助任务的损失:

$$L=\lambda L_{\text{main}}+(1-\lambda)L_{\text{aux}}$$

其中, λ是超参数, 用来调整两个任务的权重.

2 实验与分析

2.1 数据集和实验设置

我们使用 SemEval-2016 任务 6.A 来测试本文模型的性能. 该数据集包含有 5 个不同的对象: “Atheism/无神论” (“A”)、“Climate Change is a Real Concern/气候变化是一个真正的问题”(“CC”)、“Feminist Movement/女性主义运动”(“F”)、“Hillary Clinton/希拉里克林顿”(“H”) 和“Legalization of Abortion/堕胎法律” (“LA”). 表 1 显示了这些对象在数据集中的分布. 每条推文有立场标签 (“支持”、“反对” 和“无关”), 情感得分使用两个目前最好的第三方标注服务, 通过 Amazon Comprehend 及 Azure 文本分析服务分别进行标注, 并取两者的平均值作为最终的情感得分.

我们采样了约 15% 的训练数据作为验证集以调整参数. 词嵌入层使用预训练 BERT. Tansformer 中单元数设置为 256. Dropout 层, 比率为 0.5. 使用 Adam 作为优化器, 学习速率设置为 1e–4. 立场检测主任务的全连接层维度为 128, 情感预测辅助任务的全连接层维度为 64. λ设置为 0.75. 此外, 我们将 L2 正则用于损失函数, 并将正则化参数设置为 0.01.

2.2 评估指标

Macro $F1$ 被用于评估本文模型的性能. 标签“支持”和“反对”的 $F1$ 得分计算如下:

$$F_{\text{支持}}=\frac{2P_{\text{支持}}R_{\text{支持}}}{P_{\text{支持}}+R_{\text{支持}}};\ F_{\text{反对}}=\frac{2P_{\text{反对}}R_{\text{反对}}}{P_{\text{反对}}+R_{\text{反对}}}$$

其中, P 指准确率, R 指召回率. 然后计算 Macro $F1$:

$$F_{\text{avg}}=\frac{F_{\text{支持}}+F_{\text{反对}}}{2}$$

值得注意的是, “无关”标签不会在训练中被丢弃. 但是评估中不考虑标签“无关”. 因为在此任务中我们仅对“支持”和“反对”标签感兴趣.

我们平均每个对象的F_{avg}得到Mac F_{avg}.

2.3 结果

首先使用消融实验来确定本文模型中每个组件对立场检测的重要性.

(1) WT-all 是带有情感预测辅助任务的模型.

(2) WT-main 跟 WT-all 相比不带有情感预测辅助任务.

表 2 是在 SemEval-2016 数据集上立场检测的性能比较结果, 表 2 的前两行数据显示了该消融实验的结果. 在所有对象的表现上, 带有情感预测辅助任务的模型都优于不带该辅助任务的模型. 这可以表明情感预测辅助任务的有效性.

其次, 我们将本文模型与以下基准方法进行比较 (所有基准方法的实验结果均来自原始论文):

(1) SVM[4]: 通过单词和字符的 n-gram 进行训练, 超越了 SemEval-2016 竞赛中的最佳模型.

(2) JOINT[12]: 利用情感信息来改进立场检测任务.

(3) TAN[8]: 基于注意力的 LSTM 模型.

(4) AS-BiGRU-CNN[9]: 在基于注意力的 LSTM 模型之后加入 CNN 以提取对象特征.

(5) HAN[10]: 利用层级注意力机制建模各种语言特征.

(6) TGMN-CR[15]: 使用注意力机制和记忆机制提取重要信息进行立场检测.

表 2 F_{avg} 性能比较 (%)

模型	A	CC	F	H	LA	Mac F_{avg}
WT-all	67.02	**54.16**	**59.42**	**67.15**	59.12	**61.37**
WT-main	65.83	53.82	59.01	66.77	57.93	60.67
SVM	65.19	42.35	57.46	58.63	66.42	58.01
JOINT	66.78	50.60	59.35	62.47	61.58	60.16
TAN	59.33	53.59	55.77	65.38	63.72	59.56
AS-BiGRU-CNN	66.76	43.40	58.83	57.12	65.45	58.31
HAN	**70.53**	49.56	57.50	61.23	66.16	61.00
TGMN-CR	64.60	43.02	59.35	66.21	66.21	59.88

表 2 其余数据显示了比较的结果. 我们可以看到, WT-all 在“Climate Change is a Real Concern/气候变化是一个真正的问题”(“CC”)、“Feminist Movement/女性主义运动”(“F”)、“Hillary Clinton/希拉里克林顿”(“H”) 这 3 个对象中优于所有基线模型. 在“Atheism/无神论” (“A”) 对象上也可以取得可比较的结果. 该模型 Mac F_{avg} 比 JOINT 模型高出 1.21%, 证明了模型的有效性.

3 结论与展望

在本文中, 我们将 Transformer 结构与门控注意力应用于特定对象立场检测. 使用 Wordpiece 拆分可分离短语, Transformer 用于建模文本语义, 门控注意力用于建模文本与对象的关系. 此外, 我们还加入了情感预测任务作为辅助任务, 以充分利用文本中的情感信息来提升表现. 在基准数据集上的实验结果表明, 在 Macro F1 分数的评价体系下, 我们的模型比其他模型具有更好的性能. 在未来的工作中, 我们可以进一步考虑对象间的关系: 在立场检测中, 对象之间的相互关系往往是有帮助的, 即推文可能通过谈论其他相关对象, 来推理对某一特定对象的立场, 比如支持“特朗普”意味着反对“希拉里”. 因而, 如何捕捉到对象之间的关系, 从而利用此类关系进一步提升立场检测应用的表现, 这需要更深入的研究.

参考文献

1 Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. arXiv: 1508.07909, 2015.

2 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of Advances in Neural Information Processing Systems. Long Beach, CA, USA. 2017. 5998–6008.

3 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv: 1810.04805, 2018.

4 Mohammad S, Kiritchenko S, Sobhani P, *et al*. Semeval-2016 task 6: Detecting stance in tweets. Proceedings of the 10th International Workshop on Semantic Evaluation. San Diego, CA, USA. 2016. 31–41.

5 Vijayaraghavan P, Sysoev I, Vosoughi S, *et al*. Deepstance at semeval-2016 task 6: Detecting stance in tweets using character and word-level cnns. arXiv: 1606.05694, 2016.

6 Wei W, Zhang X, Liu XQ, *et al*. Pkudblab at semeval-2016 task 6: A specific convolutional neural network system for effective stance detection. Proceedings of the 10th International Workshop on Semantic Evaluation. San Diego, CA, USA. 2016. 384–388.

7 Zarrella G, Marsh A. Mitre at SemEval-2016 task 6: Transfer learning for stance detection. arXiv: 1606.03784, 2016.

8 Du JC, Xu RF, He YL, *et al*. Stance classification with target-specific neural attention networks. Proceedings of the 26th International Joint Conference on Artificial Intelligence. Sydney, Australia. 2017. 3988–3994.

9 Zhou YW, Cristea AI, Shi L. Connecting targets to tweets: Semantic attention-based model for target-specific stance detection. Proceedings of the 18th International Conference on Web Information Systems Engineering. Moscow, Russia. 2017. 18–32.

10 Sun QY, Wang ZQ, Zhu QM, *et al*. Stance detection with hierarchical attention network. Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, NM, USA. 2018. 2399–2409.

11 Sobhani P, Mohammad S, Kiritchenko S. Detecting stance in tweets and analyzing its interaction with sentiment. Proceedings of the 5th Joint Conference on Lexical And Computational Semantics. Berlin, Germany. 2016. 159–169.

12 Sun QY, Wang ZQ, Li SS, *et al*. Stance detection via sentiment information and neural network model. Frontiers of Computer Science, 2019, 13(1): 127–138. [doi: 10.1007/s11704-018-7150-9]

13 Shen DH, Wang GY, Wang WL, *et al*. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. arXiv: 1805.09843, 2018.

14 Tang DY, Qin B, Feng XC, *et al*. Effective LSTMs for target-dependent sentiment classification. arXiv: 1512.01100, 2015.

15 Wei PH, Mao WJ, Zeng D. A target-guided neural memory model for stance detection in Twitter. Proceedings of 2018 International Joint Conference on Neural Networks (IJCNN). Rio de Janeiro, Brazil. 2018. 1–8.