问题分类

问题分类旨在分析问题并将其分配到表示其预期答案类型的适当类别，因此它能够限制系统仅对特定信息源的可能答案的搜索，以制定适当的策略来找到响应，并缩小可能的回答候选的数量。

基于机器学习的问题分类方法先从问题的数据集中提取各种词法，句法和语义特征，然后应用例如朴素贝叶斯（NB），决策树（DT），Winnows的稀疏网络（SNoW）和支持向量机（SVM）、最大熵等各种监督学习方法来构建分类器，训练的分类器可以用来预测问题的类型。根据研究[1,2,3,4]，在上述方法中具有线性内核功能的支持向量机是最有效的方法。

基于深度学习的问题分类方法主要是CNNs和RNNs，同时随着word2vec和GloVe等词向量模型的发展，使CNNs和RNNs的输入能在保留语义距离信息的同时便于NLP各项任务的计算。递归神经网络（RNN）因其具有复现的特征有助于突破以往前馈神经网络遇到的词依赖边界限制问题，已经被广泛应用于语言建模。然而，RNN具有的梯度消失问题使其难以训练，因此提出了其改进的具有门控存储单元的RNN网络例如LSTM、GRU。LSTM具有长距离记忆的特性被应用于生成句向量（Ravuri和Stolcke，2015a; Tang等人，2015; Tai等人，2015）。当LSTM被应用于对句子建模时，句子结尾词的存储单元携带整个句子的信息，将此处输出向量作为句向量（Ravuri和Stolcke，2015a）。或者，一个句子用它各个词的LSTM输出向量的平均值表示（Tang等人，2015）。卷积神经网络（CNN）最初是为图像处理而开发的（Lecun等人，1998），是目前计算机视觉领域的最优模型。它们首先由Collobert等人（2008; 2011）应用于使用最大时间合并方法来聚合卷积层向量的自然语言处理任务；Kalchbrenner（2014）的论文中提出将CNN的k-max池化用于句子建模，通过将CNN与不同的特征映射通道和预训练的单词向量整合，实现了不同基准数据集上的显着查询分类性能（Zhang and Wallace，2015; Kim，2014），显示了CNN对句子分类的潜力，同时因为字符随机出现的特性和分类任务字长有限的特性，使得CNN能很好的拟合问题查询任务。

1. Zhiheng Huang, int Marcus and Zengchang Qin. Question classi cation using head words and their hyper- nyms. EMNLP ’08, In Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2008; p. 927–36.

2. Zhiheng Huang, int Marcus and Celikyilmaz Asli. Investigation of question classi er in question answering. EMNLP ’09, In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 2009; p. 543–50.

3.Krishnan Vijay, Das Sujatha and Chakrabarti Soumen. Enhanced answer type inference from questions using sequential models. HLT ’05, In Proceedings of the con- ference on Human Language Technology and Empirical Methods in Natural Language Processing. 2005; p. 315–22.

4.Zhang Dell and Lee Wee Sun. Question classi cation using support vector machines. SIGIR ’03, In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. 2003; p. 26–32.