面向分级读物的多尺度难度向量分类方法

·技术领域

本发明属于自然语言处理中的明确性分析领域。具体涉及一种面向分级读物的多尺度难度向量分类方法。

·背景技术

难度向量分类的任务是，给定一篇文本，通过对文本进行分析，给出该文本的难度值或判断该文本适合哪一水平的读者。应用在教育领域，可以为分级语料和课本素材的选取提供参考依据，对句子的理解难度、复杂性有定量的度量。在通用文本领域如新闻文本，本专利也可对新闻阅读难度、专业性做分析。本难度向量可对文本的理解难度、复杂性做一个较为准确的度量,为句子简化和提炼提供重要的依据，同时也为教育领域分级语料的挑选提供参考。在如今自然语言处理的不断发展下，句子难度分析也有着重要的实践和应用价值。

在难度向量的特征提取方面,国内外使用的模型任务可分为基于可读性公式、基于分类、基于排序的方法。可读性公式综合特定文本因素输出文本难度分数，目前主要作为机器学习分类的特征之一。基于分类与排序的方法所使用的特征，主要可分为词汇特征与句法特征。Louis 等人首先提出了句子明确性分类问题，考虑了句子长度,词语极性,词性等基础特征应用在 WSJ 新闻语料库分类上[1]。Junyi Jessy Li 等人在原先基础上通过引入词聚类，tf-idf 数值为出现次数较少的单词提供了更多信息，提升了泛化性，并使用半监督方法扩充了语料，开源了 speciteller 项目。Jorge 等人扩充了难度特征个数达到 89 个，包括词语音

节数、句法树等其他语言学特征[3]。这些词语特征反映了词语的复杂程度，句法特征反映了句法使用的复杂程度。在构造完句子的特征表示后，使用机器学习分类器如决策树、神经网络等进行分类。目前存在的问题是为了获得丰富的句子特征，构造特征、模型学习需要花费较多的时间，使用的特征大多局限于词汇与句法级别，对句子信息的提取不够全面。考虑到这一点，本专利提出的多尺度难度向量提取方法先构造了词语搭配特征，上下文特征，主题特征等丰富了特征表示，结合之前研究中效果最突出的特征，获得一个轻量、全面的句子难度向量，再输入到分类器如梯度提升树（GBDT）中，可以达到很好的分类效果。

·参考文献

[1]Annie Louis and Ani Nenkova, “Automatic identification of general and specific sentences by leveraging discourse annotations. ”, In Proceedings of International Joint Conference on Natural Language Processing (IJCNLP).pp.605 - 613,2011.
[2]Junyi Jessy Li and Ani Nenkova,“Fast and Accurate Prediction of Sentence Specificity”, In AAAI Conference on Artificial Intelligence (AAAI).pp.2281 - 2287, 2015.
[3] Jorge Alberto Wagner Filho, Rodrigo Wilkens and Aline Villavicencio.“Automatic Construction of Large Readability Corpora”, in Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC).pp. 164 - 173,2016.

范立础,胡世德.橡胶减震挡块:中国,96230918.4 [P].1997-08-27 增加和中文专利的比较 后面的例子写成教育 高年级 低年级 和马老师讲的就是 1.标题 2.例子上 3.diss 了其他专利 4.说下其他专利大概做了啥

·发明内容

为了解决现有技术的上述问题，本发明提供了一种多尺度的难度向量提取方法，丰富了原有特征表示并进行简化，达到了更好的效果。为了实现上述目的，

本发明采用如下技术方案：

1. 一种基于多尺度的难度向量提取方法，包括以下步骤：
   1）若原始文本数据为 web html 文本，需预先进行清洗，再进行分句，分类每一个样本。
   2）将文本切分为句子$\mathrm{s}=(\mathrm{s}_1,\mathrm{s}_2,\ldots\mathrm{s}_T)$，每个句子切分为词$\mathrm{w}=(\mathrm{w}_1,\mathrm{w}_2,\ldots\mathrm{w}_T)$。
   3）将文本句子输入到特征提取器$\mathrm{m1}$中，$\mathrm{m1}$提取先前研究提出的特征中 15 个效果最为显著的基础特征，将句子中每个词的词法特征求和并用句子长度做归一化得到词法向量e_w，与句法特征e_s拼接，得到每个句子的基础特征表示$e_t^{basic}\in\mathbb{R}^1$。
   4）将文本句子输入到特征提取器$\mathrm{m2}$中，$\mathrm{m2}$提取本专利新提出来的 6 个特征：句子困惑度、句子主题类型、单词学会年龄、句法树宽度、前后句子相似度、前后句子使用词汇的重叠度，将以上的词法特征求和并用句子长度做归一化得到词法向量e_w，与句法特征e_s拼接，得到每个句子的多尺度特征表示$e_t^{multiscale}\in\mathbb{R}^1$。
   5）将句子的基础特征表示e_t^{basic}和多尺度特征表示$e_t^{multiscale}$拼接起来，获得句子最终的特征表示$e_t^{all}\in\mathbb{R}^1$。
   6）将特征表示输入梯度提升树（GBDT）中，根据模型输出结果和训练数据真实结果训练模型，利用训练好的模型在测试数据上得到最终结果的准确率 accuracy 值，并计算特征重要性排序。
2. 所述步骤 3）中，首先为语料库建立词频字典。遍历每一个单词，获取每一个单词的以下信息：词频（出现频次）、音节数（发音时的音节数）、单词字符个数、词极性（情感词典分数）、idf 值（逆文档概率）、词向量（word2vec 向量）、词聚类（词向量的聚类标签）、单词含义个数，将以上信息求和并用句子长度做归一化得到词法向量e_w。对每一个句子，统计以下信息：句子长度、特殊符号个数（标点、数字等）、停用词个数、句子中特性词个数（名词、形容词、动词、连词），获得句法特征e_s。将词法特征e_w与句法特征e_s拼接得到每个句子的基础特征表示$e_t^{basic}\in\mathbb{R}^1$。其中部分特征使用工具 speciteller 获取，单词相关特征使用 nltk 库获取，句法相关特征使用 spacy 包获取。
3. 所述步骤 4）中，首先使用在英国国家语料库（British National Corpus）上预训练的 SRILM 工具产生每个句子的困惑度，衡量每一个 n-gram 窗口搭配在语言模型中的困惑程度。困惑度体现了词语搭配的好坏程度。遍历每一个单词，通过外部词典获取词语学会年龄，求和并用句子长度做归一化得到词

法向量e_w。词语学会年龄体现为越晚学会的单词会相对复杂。对每一个句子，统计一下信息：词语学会年龄、句子困惑度、句法树宽度（构建依存句法分析树）、前后句子相似度（词向量余弦相似度）、前后句子使用词汇的重叠度、主题模型（隐含狄利克雷分布），获得句法特征e_s。其中句法树宽度体现了句法结构的复杂程度。前后句子相似度和前后句子用词重叠度是引入了上下文信息。主题模型体现在若语料库中包含不同话题和风格的文章会影响难度向量，如 Louis 等人曾提到新闻类文章总体为写得更加笼统，难度相对简单[1]。将词法特征e_w与句法特征e_s拼接得到每个句子的多尺度特征表示$e_t^{multiscale} \in \mathbb{R}^1$。

4. 所述步骤 5）中，将句子的基础特征表示e_t^{basic}和多尺度特征表示$e_t^{multiscale}$拼接起来得到最终的难度向量e_t^{all}，这两者的特征提取可以同时遍历句子、遍历单词获得。
5. 所述步骤 6）中，将向量输入到分类器梯度提升树（GBDT）中，训练模型后获取准确率，并且计算每个特征对节点分裂的收益，节点分裂时收益越大，该节点对应的特征的重要度越高。通过计算特征重要性排序，可以获得每个难度特征对模型的重要程度，也可以根据语料情况进行动态调整。

本发明采取以上步骤，具有以下优点：简化了特征表示，只需要 21 个向量就能体现句子难度，引入了多尺度特征丰富了难度特征表示，增强了模型泛化性；结合新使用的上下文信息构建了对句子级别和文章级别都适用的难度向量表示系统，在句子级别和文章级别的两个数据集都获得了较好的效果；分类器使用梯度提升树，训练速度快，可以获得特征重要性排序。

• 附图说明

图 1 为模型流程图，体现构建难度特征流程。

图 2 为模型概要图。

图 3 为分类效果表，在句子级别和篇章级别语料库的训练效果。

图 4 为特征消解实验表，分析难度特征对结果的影响。

图 5 为特征重要性示意图。

• 具体实施方式

下面结合附图对本发明的实施例进行详细说明。所描述的实例旨在便于理解，对本发明的适用性不起任何限定作用。图 1 是本发明的流程图，如图 1 所示，所述方法包括以下步骤：

1）若原始文本数据为 web html 文本，需预先进行清洗，再进行分句，分类每一个样本。中文语句可以用 jieba 分词，但不限于此。在这里以英文数据为例，如图 2 所示将句子“<p>‘And it was only 10 rubles for all this,’ she said. ‘ I’m taking it back for the girls at the factory to try.’<p>”去除 html 标签后切分为两个句子“ ‘And it was only 10 rubles for all this,’ she said.”和“I’m taking it back for the girls at the factory to try.”，句子内再进行分词，如第一句可分为该句的词语列表：[Andit, was, only, 10, rubles, for, all, this, she, said]。

2）遍历句子$\mathrm{s} = (\mathrm{s}_1, \mathrm{s}_2, \ldots \mathrm{s}_T)$中的每个单词$\mathrm{w} = (\mathrm{w}_1, \mathrm{w}_2, \ldots \mathrm{w}_T)$，获取基础特征和多尺度特征中单词级别的属性。以第一句“‘And it was only 10 rubles for all this,’ she said.”为例，其中部分特征使用工具 speciteller 获取，该句的复杂度得分是 0.11（0 到 1），单词相关特征使用 nltk 库获取，如词频累计值为 4.00，单词平均含义数为 3.91，单词平均音节数为 1.00，获取单词级别属性e_w [0.11, 4.00, 3.91, 1.00...]。

3）遍历样本中的每个句子$\mathrm{s} = (\mathrm{s}_1, \mathrm{s}_2, \ldots \mathrm{s}_T)$，获取基础特征和多尺度特征中词语搭配级别、句子级别、主题级别的属性。以新提出的特征为例：通过 SRILM 工具产生示例句子的句子困惑度为 85，主题模型超参数主题个数为 3 的情况下产生的主题向量为[0.08, 0.82, 0.08]，句法树宽度为 0.45，前后句子词汇重叠度为 2，前后句子词向量的余弦相似度为 0.93，由此获得句法特征e_s。

4）将词法特征e_w和句法特征e_s拼接起来，最终的难度向量e_t^{all}。

5）将难度向量e_t^{all}和难度标签输入梯度提升树（GBDT）中训练，获取最佳模型，计算模型在测试集上分类准确率。

6）计算特征重要性排序，可以获得每个难度特征对模型的重要程度，也可以根据语料情况进行动态调整。