【[2005]Sequence Modelling or Sentence Classication in a Legal Summarisation System】

We describe a set of experiments using a wide range of machine learning techniques for the task of predicting the rhetorical status of sentences. The research is part of a text summarisation project for the legal domain for which we use a new corpus of judgments of the UK House of Lords. We present experimental results for classification according to a rhetorical scheme indicating a sentence’s contribution to the overall argumentative structure of the legal judgments using four learning algorithms from the Weka package (C4.5, navie Bayes, Winnow and SVMs). We also report results using maximum entropy models both in a standard classification framework and in a sequence labelling framework. The SVM classifier and the maximum entropy sequence tagger yield the most promising results.

我们用一系列的机器学习技巧来描述一组实验, 用于预测句子的修辞状态。这项研究是一个文本总结项目的一部分, 为法律领域, 我们使用了新的判决语料库英国众议院。我们提出的实验结果分类根据一个修辞方案, 表明一个句子的贡献, 整个辩论结构的法律判决使用四学习算法从 Weka 包 (C4.5, 朴素贝叶斯, Winnow和 svm)。我们还在标准分类框架和序列标签框架中使用最大熵模型报告结果。SVM 分类器和最大熵序列标注器产生最有希望的结果。

The rhetorica status classification task is the focus of this paper.

修辞状态分类任务是本文的重点。

【[2008]Experiments with sentence classification.pdf】

We present a set of experiments involving sentence classification, addressing issues of representation and feature selection, and we compare our findings with similar results from work on the more general text classification task. The domain of our investigation is an email-based help-desk corpus. Our investigations compare the use of various popular classification algorithms with various popular feature selection methods. The results highlight similarities between sentence and text classification, such as the superiority of Support Vector Machines, as well as differences, such as a lesser extent of the usefulness of features selection on sentence classification, and a detrimental effect of common preprocessing techniques (stop-word removal and lemmatization).

我们提出了一系列的实验, 包括句子分类, 处理表示和特征选择的问题, 我们将我们的发现与更一般的文本分类任务的工作结果进行比较。我们调查的领域是一个基于电子邮件的服务台语料库。我们的调查比较了各种流行的分类算法与各种流行的特征选择方法的使用。结果突出了句子和文本分类之间的相似性, 如支持向量机的优越性, 以及差异, 如在句子分类中特征选择的用处较小的程度, 以及有害常用预处理技术的效果 (停止字删除和 lemmatization)。

【[2011]Word cooccurrence features for text classification.pdf】

In this article we propose a data treatment strategy to generate new discriminative features, called compound-features (or c-features), for the sake of text classification. These c-features are composed by terms that co-occur in documents without any restrictions on order or distance between terms within a document. This strategy precedes the classification task, in order to enhance documents with discriminative c-features. The idea is that, when c-features are used in conjunction with singlefeatures, the ambiguity and noise inherent to their bag-of-words representation are reduced. We use c-features composed of two terms in order to make their usage computationally feasible while improving the classifier effectiveness. We test this approach with several classification algorithms and single-label multi-class text collections. Experimental results demonstrated gains in almost all evaluated scenarios, from the simplest algorithms such as kNN (13% gain in micro-average F1 in the 20 Newsgroups collection) to the most complex one, the state-of-the-art SVM (10% gain in macro-average F1 in the collection OHSUMED).

在本文中, 我们提出了一种数据处理策略来生成新的判别特征, 称为复合特征 (或 c 特征), 用于文本分类。这些 c 功能由在文档中共同发生的术语组成, 不受文档中术语的顺序或距离的限制。该策略在分类任务之前, 以增强具有判别性 c 特征的文档。它的想法是, 当 c 特征与 singlefeatures 一起使用时, 他们的词袋表示法固有的歧义和噪声减少了。我们使用两个术语组成的 c 特征, 在提高分类器的有效性的同时, 使它们的使用在计算上可行。我们用几种分类算法和单标签多类文本集合来测试这种方法。实验结果在几乎所有评估的场景中都得到了提高, 从最简单的算法, 如 kNN (20 新闻组集合中微平均 F1 的13% 增益) 到最复杂的, 最先进的支持向量机 (在宏平均 F1 中的10% 增益集合 OHSUMED)。

【[2010]An\_unsupervised\_approach\_to\_sentence\_classification.pdf】

一个紧密相关但相当不同的问题是, 将适当的类标签分配给单个句子, 而不是整个文档。从语法上来说, 句子通常被归类为诸如声明性、祈使句、疑问句、EXCLAMATIVE、交际、信息等类, 还有进一步的子类。其他结构化的句子类包括少校 (有主语和谓语), 次要的 (没有一个有限的动词; 例如, 越多越快乐), 周期性 (意思是不完整的直到最后的子句或词组; 例如, 沉默和软, 慢, 下降雪。等。语义上对句子 (根据句子的目的) 进行分类是一项难度更大的任务, 并且越来越受到语言学家和 NLP 研究人员的关注 [31], [27], [18], [2], [3], [12], [15], [28], [7], [8], [21]。这一领域的大多数工作都使用了监督学习方法 (例如, 使用 SVM、决策树、基于最大熵的分类器、天真的贝叶斯等), 除了 [11] (半监督) 和 [26], [10] (基于知识的)。句子分类已被应用到诸如摘要、信息抽取、IR、自动本体创建 [9] 和文本蕴涵 [29] 等任务中。在一些实际应用领域, 如生物医学论文、法律判决、产品评论、客户投诉、电子邮件等, 都使用了句子分类。句子类也更依赖于域 (表 1)。