自动完形填空系统构建

问题描述

在语义连贯的句子中去掉一个词语，用空格取代，要求在给出的对应备选答案中，系统自动选出一个最佳的答案，使语句恢复完整。

相关语料

1、 Training data：

课程会提供一份未标注训练语料，供同学选使用，也可自行搜集补充但需要说明。

2、 Development set：

提供一份含有 240 句话的语料及答案，供同学自行测试结果，根据结果调整优化自己的算法。

3、 Test set：

提供一份含有 800 句话的测试语料，每句话有一个空格和 5 个备选答案。该语料不提供答案，同学提交测试结果，由助教统一评测。

(相关语料将在课程网站公布，可以在课程网站上进行下载。)

评测方法

准确率＝正确填空句子的个数/全部句子的个数

题目要求

要求通过语料训练出相关模型，对测试语料进行预测，对每句话提供一个系统认为正确的选项。无统一标准方法，自行设计模型，鼓励同学积极创新。

提示：模型的构建可以简单也可以复杂，方法不限，鼓励创新。例如可以基于 N 元模型建立一个朴素的系统；除所提供的训练语料外，也可以使用自行整理搜集的词典和语料资源。

作业要求

1、可分组进行，但每个小组的规模不能超过 2 人(即≤2)

2、实现相关程序，可用 c/c++或者 java 语言完成。可参考网上源代码，但必须重新实现，要求程序代码完整，有必要的说明文档和 Makefile 等文件；

3、提供测试语料的预测结果，输出文件以“题目号+选项+英文单词”形式输出，中间用空格或制表符间隔，每个答案占一行(可参考 development set)。即：

1 choice1 answer1

…… ……

…… ……

800 choice 800 answer800

4、撰写实验报告以及 PPT。实验报告以小论文的形式，要有必要的参考文献等信息，将使用的方法讲解清楚；PPT 用于在课堂上报告实验成果；

5、将预测答案、实验报告、PPT 及源程序提交到助教用以评分。

6、作业提交截止时间：2013 年 12 月 15 日。