一款基于 BiLSTM-CRF 的英文选词填空机器作答应用设计

申　鹏　飞[1]，作　者[2]

（1. 智能科学与技术 2102 2115040110；2. 作者详细单位，省市邮编）

摘要：在英文选词填空题目中，文章中的句子空缺通常具有多个选项。条件随机场的前向算法能够计算出填入各选项后的句子的“前向分数”，从而判断出最佳的选项。由百度研究院提出的双向长短时记忆-条件随机场模型是 Transformer 时代之前最成熟的序列标注模型之一，该网络通过结合双向长短时记忆网络和条件随机场，有较强的结构表示能力。本次报告将详细介绍基于该模型的选词填空机器答题应用的设计原理和表现效果。

关键词：完形填空；机器答题；双向长短时记忆网络；条件随机场

Machine Answering For English Fill-In-The-Blank Word Selection Quizzes Based On BiLSTM-CRF Segmentation

Shen pengfei[1]，NAME Name-name[2]

(1. Artifical Intelligence Science and Engineering 2102 2115040110; 2. Department, City, City Zip Code, China)

Abstract: In English Fill-in-the-Blank word selection quizzes, there are multiple options for a blank. The forward algorithm of conditional random field can ‘predict’ the option with the highest forward score, so as to choose the most likely answer. The bidirectional long short-term memory - conditional random field model was one of the most mature sequence annotation models before the Transformer era, which includes a conditional random field layer. This paper will introduce in detail the design and performance of the fill-in-the-blank word selection machine based on this model.

Keywords: Fill-in-the-Blank; Machine Answering; Bidirectional Long Short-Term Memory; Conditional Random Field

1　引言

“词性标注”是自然语言处理的经典任务之一，AI 能根据输入的句子推测出对应词性序列。目前，主流的模型网络都结合了条件随机场（CRF）层。许多学者和工程师基于 CRF 的前向算法提出了基于序列标注的文本生成应用，如 RNN 的诗歌生成应用等。在生成式算法中，模型会将“前向分数”最大的句子作为最终的生成结果。

此外，关于英文选词填空机器答题的应用，一些学者已经提出了一些比较独特的方法，如来自 Google 团队的 MaskGAN 网络。Transformer 框架问世后，互联网上出现了大量的基于 BERT 的中英文选词填空应用，如 Unmask 等。这些网络较 BiLSTM-CRF 各有优缺点。

1.1　OCR 的作用

123123123213123123 引言内容。引言作为论文的开场白，应以简短的篇幅介绍论文的写作背景和目的，以及相关领域内前人所做的工作和研究概况，说明本研究与前人工作的关系，目前研究的热点、存在的问题及作者工作的意义。1、开门见山，不绕圈子。避免大篇幅地讲述历史渊源和立题研究过程。2、言简意赅，突出重点。不应过多叙述同行熟知的及教科书中的常识性内容，确有必要提及他人的研究成果和基本原理时，只需以引用参考文献的形势标出即可。在引言中提示本文的工作和观点时，意思应明确，语言应简练。3、引言的内容不要与摘要雷同，也不是摘要的注释。4、引言要简短，最好不要分段论述，不要插图、列表和数学公式。

2　量的书写规则

正文内容。正文、图表中的变量都要用斜体字母，对于矢量和张量使用黑斜体，只有 pH 采用正体；使用新标准规定的符号[?1]；量的符号为单个拉丁字母或希腊字母；不能把量符号作为纯数使用；不能把化学符号作为量符号使用，代表物质的符号表示成右下标，具体物质的符号及其状态等置于与主符号齐线的圆括号中。

注意区分量的下标字母的正斜体：凡量符号和代表变动性数字及坐标轴的字母作下标，采用斜体字母。

正文中引用参考文献的标注方法，在引用处对引用的文献，按它们在论著中出现的先后用阿拉伯数字连续排序，将序号置于方括号内，并视具体情况把序号作为上角标或作为语句的组成部分。

2.1 单位的书写规则

正文内容。单位符号无例外的采用正体字母。注意区分单位符号的大小写：一般单位符号为小写体，来源于人名的单位符号首字母大写。体积单位升的符号为大写 L。

2.1.1 表格的规范化

正文内容。表格的设计应该科学、明确、简洁，具有自明性。表格应采用三线表，项目栏不宜过繁，小表宽度小于 7.5 cm，大表宽度为 12～15cm 。表必须有中英文表序、表题。表中顶线与栏目线之间的部分叫项目栏，底线与栏目线之间的部分叫表身。表身中数字一般不带单位，百分数也不带百分号，应把单位符号和百分号等归并在栏目中。如果表中栏目中单位均相同，则可把共同的单位提出来标示在表格顶线上方的右端（不加“单位”二字）。表身中同一栏各行的数值应以个位（或小数点），且有效位数相同。上下左右相邻栏内的文字或数字相同时，应重复写出。

表 1 表题

Model	SMO+DNN	PCA+DNN	DNN
Accuracy	0.994	0.938	0.914
Precision	0.995	0.934	0.891
Recall	0.995	0.918	0.882

3 图

正文内容。插图尽可能不用彩色图。小图宽度小于 7.5 cm，大图宽度为 12～15cm 。图必须有中英文图序、图题。函数图只在靠近坐标线处残留一小段标值短线，其余部分省略。加注坐标所代表的量及单位（如 t/s）。标值排印在坐标外侧，紧靠标值短线的地方；标值的有效数字为 3 位。图中量的意义要在正文中加以解释。若有图注，靠近放在图下部，图序、图题的上方。

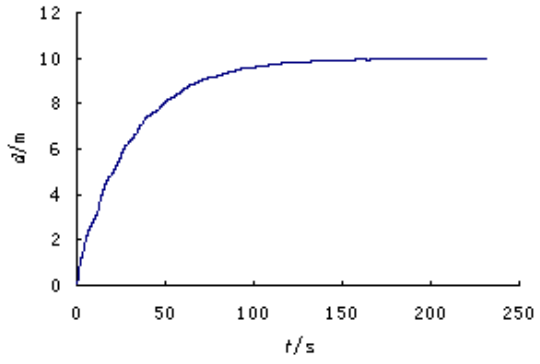

图 1 图题

4 数学符号和数学式的编排规范

正文内容。变量、变动附标及函数用斜体字母表示。点、线段及弧用斜体字母表示。在特定场合中视为常数的参数也用斜体字母表示。对具有特殊定义的函数和值不变的数学常数用正体字母表示。具有特殊定义的算子也用正体字母表示。矩阵符号用大写的黑斜体字母表示，矩阵元素用白斜体字母表示。

公式及公式中的符号说明尽量接排以节省版面。把带有复杂上角标的指数函数 e^t 写成 $\exp t$ 。公式的主体应排在同一水平线上；繁分式的主辅线要分清。长公式在运算符号后回行；长分式转行时，先将分母写成负幂指数的形式，然后转行；矩阵和行列式不能转行。矩阵元素包含式子时，每一列应以中心线上下对齐，行要左右排齐；元素为单个字母或数字时，每列应使正负号对齐。对角矩阵中对角元素所在的列应明显区分，不能上下重叠。

简单的和常识性的运算公式和推导过程不要列写。

$$\begin{cases} O_i = X & i = 1 \\ O_i = f_i(Z_i) & i > 1 \\ Z_i = g_i(O_{i-1}, W_i) \end{cases} \tag{1}$$

$$F_D = \beta \star (1 - A) + (1 - \beta) \star \frac{S}{D} \tag{2}$$

5 结束语

正文内容。结论不应是正文中各段小结的简单重复，它应以正文中的实验或考察得到的现象、数据的阐述分析为依据，完整、准确、简洁地指出以下内容：1）由对研究对象进行考察或实验得到的结果所揭示的原理及其普遍性；2）研究中有无发现例外或本论文尚难以解释和解决的问题；3）与先前发表过的研究工作的异同；4）本文在理论上和实用上的意义及价值；5）进一步深入研究本课题的建议。