章鱼输入法拼音引擎关键技术分析报告



北京百文宝科技股份有限公司

2016年6月

# 第一章 概述

拼音输入法的引擎是输入法的核心，主要功能是把用户输入按键信息转换为汉字，其转换率的高低是用户体验的重要指标。

章鱼输入法拼音引擎是专门为智能手机开发的，相对于传统非智能手机在功能上需要有较大的改进。 改进的目标主要有以下三点：

* 整句输入
* 中英文混合输入
* 智能纠错（只针对26键键盘布局，9键必要性不大）

为了解决着三个问题，需要引入中英文统一的语言模型（Language Model）；为了保住运行的效率，需要把着三个问题在同一框架下考虑。

本文将首先介绍语言模型，再对这三个问题分别介绍。

# 第二章 语言模型

语言模型（Language Model）在自然语言处理（Natural Language Processing）中占有重要的地位，在输入法中的得到了广泛的应用。目前主要采用的是n元语言模型（n-gram model），这种语言模型基于大规模语料库的统计，构造简单直接。

一个语言模型通常构建为字符串的概率分布，这里的反映的是字符串作为一个句子出现的频率。

对于一个由个基本单元（这里的基本单元可以是字、词或短语等，为方便起见，以后我们只用词来通指）构成的句子,其概率的计算公式可以表示为

在上式中，产生第个词的概率由已经产生的个词决定的。一般的，我们把前第个词称为第个词的历史。为了解决计算规模的问题，实际使用时，我们仅取个词作为历史来解决。上式近似为

这个公式称为元语言模型。当时，我们称之为一元语言模型，当时，我们称之为二元语言模型，以此类推。其中