On the Naturalness of Software

Abstract

Natural languages like English are rich, complex, and powerful. The highly creative and graceful use of languages like English and Tamil, by masters like Shakespeare and Avvaiyar, can certainly delight and inspire. But in practice, given cognitive constraints and the exigencies of daily life, most human utterances are far simpler and much more repetitive and predictable. In fact, these utterances can be very usefully modeled using modern statistical methods. This fact has led to the phenomenal success of statistical approaches to speech recognition, natural language translation, question answering, and text mining and comprehension. We begin with the conjecture that most software is also natural, in the sense that it is created by humans at work, with all the attendant constraints and limitations—and thus, like natural language, it is also likely to be repetitive and predictable. We then proceed to ask whether (a) code can be usefully modeled by statistical language models and (b) such models can be leveraged to support software engineers. Using the widely adopted n-gram model, we provide empirical evidence supportive of a positive answer to both these questions. We show that code is also very regular, and, in fact, even more so than natural languages. As an example use of the model, we have developed a simple code completion engine for Java that, despite its simplicity, already improves Eclipse’s completion capability. We conclude the paper by laying out a vision for future research in this area.

像英语这样的自然语言是丰富、复杂和强大的。莎士比亚和艾瓦亚尔等大师对英语和泰米尔语等语言的高度创造性和优雅的运用，肯定能带来愉悦和鼓舞。但在实践中，由于认知的限制和日常生活的迫切需要，大多数人类的话语都要简单得多，重复得多，而且可以预测。事实上，这些话语可以非常有用的建模使用现代统计方法。这一事实导致统计方法在语音识别、自然语言翻译、问题回答、文本挖掘和理解方面取得了惊人的成功。我们首先假设大多数软件也是自然的，因为它是由工作中的人类创建的，带有所有随之而来的约束和限制——因此，就像自然语言一样，它也可能是重复的和可预测的。然后，我们接着询问(a)代码是否可以由统计语言模型有效地建模，(b)这些模型是否可以用于支持软件工程师。利用广泛采用的n-gram模型，我们提供了实证证据，支持这两个问题的积极答案。我们展示了代码也是非常规则的，事实上，甚至比自然语言还要规则。作为模型的一个例子，我们为Java开发了一个简单的代码完成引擎，尽管它很简单，但是已经提高了Eclipse的完成功能。最后，我们展望了该领域未来的研究方向。