基于地质领域的中文分词研究

基于地质领域的中文分词研究

摘　要

地质学是研究地球及其演变的一门自然科学。地质学的研究对象为地球的地壳或岩石圈的物质组成、内部构造以及各圈层之间共同演化的过程。我国地质学发展历史由来已久，前人留下了诸多宝贵的地质领域相关文本资料。随着信息时代的到来，如何能够合理高效的利用这些宝贵的地质领域相关文本资料成为了重要的研究方向。为了能够对地质领域相关文本资料进行更好的利用，中文分词这一技术手段将提供重要帮助。

中文分词技术是自然语言文本处理在中文领域的重要研究方向与基础步骤。中文分词就是指要将一个连续的待划分中文文本序列按照已经确定的相关规则将其切分成单独汉语词序列的过程。相较于英文，中文在自然语言处理过程中有着语言学的劣势。中文不像英文一样在词语之间有明确的区分标志。这就导致了汉语言中对于词组的划分较为困难。同时汉语中一个句子可能含有不同的合法划分但语义却有所不同。此外，随着人们对汉语言的传造型开发，汉语中还有许多未登录词在不断的被引入。这些都在技术上对中文分词的发展提出了挑战。国内对于自然语言处理技术的研究相较于国外起步较晚，直到上世纪末我们才有了第一个中文自动分词系统。经过多年的发展与研究，中文分词技术主流技术包括基于字符串匹配的中文分词方法、基于统计的中文分词方法以及基于理解的中文分词方法。

在基于神经网络与深度学习算法的兴起为中文分词技术提供了新的思路。深度学习算法不但能够将地质工作者从对文本特征提取的工作上解放出来，也能特征特征提取过程的准确度。以此为背景，本着提高地质领域中文分词技术的准确率、召回率以及 F-值的目标。本文使用 LSTM 网络模型对地质领域内相关文本进行分词研究。经过训练后结果如下：准确率：0.888，召回率：0.891，F-值：0.889。

关键词：深度学习，中文分词技术，长短期记忆模型，自然语言处理，地质学