文本生成概述

本文主要介绍文本生成的定义、任务、评价指标和实现方法。

参考地址：

<https://www.jiqizhixin.com/articles/2017-05-22>

1. 文本生成的定义

Reiter等人将文本生成定义为：接受非语言形式的信息作为输入，生成可读的文字表述。数据到文本的生成适用于这个定义。

Wan等人将这个概念拓展为包括了文本到文本的生成、数据到文本的生成以及图像到文本的生成的文本生成技术。

1. 文本生成任务

按照输入数据的区别，可以将文本生成任务大致分为以下三类：1）文本到文本的生成；2）数据到文本的生成；3）图像到文本的生成。

* 文本到文本的生成任务又可根据不同的任务分为：文本摘要、古诗生成、文本复述等。
  + - 文本摘要：可以分为抽取式摘要和生成式摘要。抽取式摘要通常包括信息抽取和规划等步骤。
    - 古诗生成
    - 文本复述：Quirk【1】等人使用机器翻译的方法生成复述文本；Max【2】等人采用基于枢轴（pivot）的复述生成方法，以另一种语言作为中间媒介。
* 结构化数据生成文本的任务：Reiter等人将数据到文本的系统分为了信号处理（视输入数据类型可选）、数据分析、文档规划和文本实现这四个步骤。
* 图像到文本的生成任务：在这方面也有很多不同的种类，如image-caption、故事生成、基于图像的问答等。

1. 文本生成方法

文本生成有基于规则、基于规划（Planning-based）以及数据驱动的方法。本文主要介绍基于数据驱动的方法。

可以分为如下两大类：

* 基于语言模型的自然语言生成。一般基于马尔可夫的语言模型在数据驱动的自然语言生成中有着重要的应用。
* 基于深度学习方法的自然语言生成。

1. 模型评价

分为内在评价和外在评价两大方面

* 内在评价：
  + 1）采用BLEU、NIST和ROUGE等进行自动评价
  + 2）人工评价
* 外在评价：关注生成文本在实际应用中的可用性

参考资料：

【1】Quirk C, Brockett C, Dolan W B. Monolingual Machine Translation for Paraphrase Generation[C]//EMNLP. 2004: 142-149.

 【2】Max A. Sub-sentential paraphrasing by contextual pivot translation[C]//Proceedings of the 2009 Workshop on Applied Textual Inference. Association for Computational Linguistics, 2009: 18-26.