单位代码 **10006**

学 号 15061084

分类号 TP312

密 级 秘 密

****

毕业设计（论文）

面向生成式对话的多样性评价指标分析

|  |  |
| --- | --- |
| 院（系）名称 | 计算机学院 |
| 专业名称 | 计算机科学与技术 |
| 学生姓名 | 冯聪 |
| 指导教师 | 荣文戈 |

2019年5月

北京航空航天大学

**本科生毕业设计（论文）任务书**

Ⅰ、毕业设计（论文）题目：

面向生成式对话的多样性评价指标分析

An Diversity-Oriented Analysis on Evaluating the Generative Dialogue Models

Ⅱ、毕业设计（论文）使用的原始资料（数据）及设计技术要求：

由三部分组成:学术论文、数据集和项目工程。其中学术论文主要为发表在人工智能和计算语言学等领域的国际顶级期刊的论文。数据集主要为训练对话系统的结构化或者非结构化的对话预料库。项目工程则是实现了某一模型的，用于定量实验的程序。设计要求为理解对话系统中各种评价指标的原理、优点和局限性。

Ⅲ、毕业设计（论文）工作内容：

本课题力求对现有的面向生成的对话系统的评价指标做一个尽可能完备的文献综述。在对话系统领域，由于人类对话的多样性和歧义性，评价系统输出的响应是一个比较困难的问题，也是一个开放的学术问题。我们的工作是把目前所有的评价指标都整理出来，对它们作逐一的考察，在考察现在的评价指标的基础上，我们将分析不同评价指标在不同的数据集上的特点，总结出好的指标应该具有的优点，促进对话系统评估的自动化。

Ⅳ、主要参考资料：

How NOT to Evaluate Your Dialogue System: An Empirical Study of Unsupervised Metrics for Response Generation （Liu et al. 2016）

Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Model （Serban et al. 2016）

A Survey of Available Corpora for Building Data-Driven Dialogue Systems （Serban et al.）

BLEU: A Method for Automatic Evalutation of Machine Translation （Kishore Papineni et al. 2002）

计算机 学院（系） 计算机科学与技术 专业类 150613 班

学生 冯聪

毕业设计（论文）时间： 2018 年 10 月 23 日至 2019 年 6 月 5 日

答辩时间： 2019 年 5 月 27 日

成 绩：

指导教师： 荣文戈

兼职教师或答疑教师（并指出所负责部分）：

系（教研室） 主任（签字）：

注：任务书应该附在已完成的毕业设计（论文）的首页。

**本人声明**

我声明，本论文及其研究工作是由本人在导师指导下独立完成的，在完成论文时所利用的一切资料均已在参考文献中列出。

作者：冯聪

签字：

时间：2019年 6 月

面向生成式对话的多样性评价指标分析

学 生：冯聪

指导老师：荣文戈

摘要

尽管基于序列到序列的生成式对话系统已经能够生成自然而流畅的响应，这类模型普遍存在着生成单调响应（GenericResponse）的倾向。对话系统的目标是生成多样的，有意义的，能引起人们兴趣的对话。为了实现这一目标，学者们提出了各种模型，从不同角度解决单调响应的问题。但是，由于这类系统缺乏好的自动化评价指标，这个领域高度依赖于人类评价，从而使评估系统的代价高昂，规模难以扩大，而且主观性强。为了了解不同的评价指标的优缺点，本文在多个公开数据集上训练了多个生成式对话模型，测定不同指标的句子级别分数，并进行指标之间的相关性分析。从中我们发现，不同的模型、指标和数据集所得的分数存在聚类现象。例如，有些指标的得分在很多数据集和模型的组合中都具有高度相关性。另一方面，当一个模型在不同的数据集上训练，并用某一指标测量时，得分也可能具有高度相关性。通过实验，本文指出:指标并不总是能一致的衡量在不同的数据集上训练的不同的模型，而在不同数据集上训练的模型也不是在所有指标上的表现都一致的好。本文提出的分析方法为学者们提供了研究现有的指标优缺点的新视角，从而有助于他们提出改进方案。

关键词：自然语言处理，深度学习，聊天机器人，对话生成，循环神经网络

An Diversity-Oriented Analysis on Evaluating the

Generative Dialogue Models

Author: Cong Feng

Tutor: Wenge Rong

**Abstract**

Although the Seq2Seq-based generative dialogue systems are able to generate natural and fluent responses, they have been long known for the inclination to generate simple and repeated responses. Towards the goal of generating diverse, meaningful and engaging dialogues, many researchers proposed various methods to address the problem of low-quality responses. However, it has been known that these systems lack good automatic evaluation metrics so the field relies heavily on human evaluation, which is expensive, unscalable and subjective. To better understand the pros and cons of various automatic metrics, we trained a number of open-domain generative models on a few public-available datasets and measured their performances with various metrics. We then analyzed the utterance-level inter-metric correlation on the scores of the trained models. From the experiments we found that scores of various combinations of model, dataset and metric tended to cluster along some axises. For example, for most of the combinations of model and dataset, some metrics tended to highly correlate with one another, which form a cluster. On the other hand, when trained on different datasets, a model’s performance measured by a certain metric may be highly correlates with one another. Although desirable, we empirically conclude that the metrics did not measure all the models trained on different datasets consistently. Similarly, the models did not perform consistently when trained on different datasets or measured by different metrics. Our novel approach provides an new perspective from which researchers can learn the pros and cons of existing metrics and models and thus propose enhancements.

**Key words**: Natural Language Processing, Deep Learning, Chatbot, Dialogue Generation, Recurrent Neural Network

目录

# 绪论

## 节标题

## 节标题

# 章标题1

## 节标题

正文

### 条标题

## 节标题

正文

### 条标题

正文

# 章标题2

## 节标题

### 条标题

# 结论

内容

# 致谢

内容

# 参考文献

# 附录A

# 附录B