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

## 课题研究背景

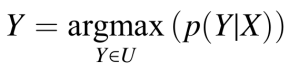
早期的对话系统的用途主要是帮助用户用自然语言完成某项任务，比如技术支持（Technical Support），预订机票、预订餐馆的座位、查询航班等。这一类系统又被称为面向任务的系统（Task-Oriented），其实现技术包括关键词匹配、规则和模板以及对话状态追踪（Dialogue State Tracking）等等，往往需要大量人工标注的数据。这些系统只能处理特定领域内的对话，不能回答开放性问题，用途局限于特定领域[55，58，66]。

随着在线聊天的流行，社交媒体和互联网论坛积累了大量的聊天语料数据，具有代表性的社交媒体和论坛有Twitter，Reddit和微博。大量的数据使人们可以构建数据驱动的（Data-Driven），开放领域（Open-Domain）的对话系统[49]。这种系统能根据对话的上下文和用户的提问产生语义相关的回答，用途有娱乐、语言学习工具和陪伴[56]等等。该领域主要考察二人对话，两人聊天的历史记录称为上下文（context），记为c；当前说话的人说出的话语称为消息（message），记为m；另外一个人对该消息的回复称为响应（response），记为r。c，m，r三者的关系如图1.1所示。系统的输入是c，m，输出是r，也就是把对话的上下文和消息映射为响应，这个问题被称为对话响应生成（Conversation Response Generation）。



图 1.1 上下文，消息和响应的关系

对话系统又可以分为生成式对话系统（Generative System）和检索式对话系统（Retrieval-Based System）[39，55]。如果一个系统能生成训练集里没有的句子，就把它称为“生成式系统”[36]。生成式系统以输入句子为条件，把条件概率最大的句子作为输出。由于搜索空间过于庞大，在实际中通常采用某种启发式搜索方法，如集束搜索（Beam Search），贪婪搜索（Greedy Search）和随机取样（Random Sample）。设X为输入句子，Y为输出句子，U是全部句子的集合，生成式模型的一般表示为:



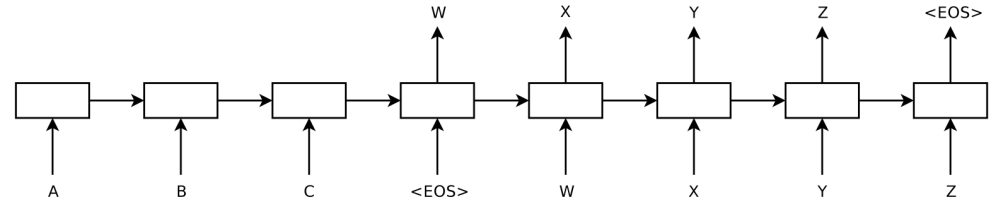
检索式系统根据输入句子从一个数据库D中检索输出句子。D通常由人类撰写的语句组成，并且足够大，使得输出句子不容易重复。系统通过某种打分机制，如词频-逆文档频率（Term Frequency-Inverse Document Frequency，TF-IDF）对数据库中的候选句子进行打分，并把得分较高的候选作为输出:



可见，生成式系统和检索式系统的根本区别在于获得输出句子的机制不同。这两种方法各有优劣：检索式模型的输出没有语法错误并且可以限制输出的范围 [39] ，但是不能生成新句子；生成式的模型能针对输入生成个性化的输出且系统高度灵活 [58] ，但是容易生成过短的句子 [23] 。在实际环境中，通常将它们作为模型联合体（Model Ensemble）使用 [62] ，而检索式系统也经常作为生成式系统的基线系统与之比较 [58,64] 。

生成式模型的流行得益于自然语言处理领域发展的一系列基础技术，包括为单词提供平滑特征的词嵌入（Word Embedding））[5，41，46] ；能对变长序列建模的循环神经网络语言模型（Recurrent Neural Networks Language Model，RNNLM） [42] ；易于训练，能避免梯度消失问题的循环门单元，如长短期记忆单元（Long Short-Term Memory，LSTM） [22]和门循环单元（Gated Recurrent Unit，GRU） [8] ；以及基于编解码器结构和RNN的序列到序列框架（Sequence to Sequence Framework，Seq2Seq） [8，66] ，图1.2是Seq2Seq的模型结构图。

Seq2Seq框架在自然语言处理的多项任务上都超过了之前的方法，因此被广泛应用到对话生成领域。最早把Seq2Seq用到对话生成领域的是Vinyals等人[70]，他们在OpenSubtitles[35]上训练的模型能回答简单的常识问题，并且比基于规则的系统Clever-Bot[[1]](#footnote-0)获得了更高的人类评价得分。Li等人提出了一系列基于Seq2Seq框架的对话系统，包括利用最大互信息（Maximum Mutual Information，MMI）增加输出多样性的目标函数 [25]，在解码器端（Decoder）加入说话人身份信息（Speaker ID）以达到输出的人格一致性（Persona Coherence）[26]以及利用对抗生成网络（Generative Adversarial Networks，GAN）[18]使系统输出和人类输出难以分辨[31] 等等。Serban 等人把 Sordoni 等人提出的用于查询建议（Query Suggestion）的层级循环编解码器（Hierarchical Recurrent Encoder-Decoder，HRED） [63] 应用到对话生成领域，提出了能捕捉对话的层级结构的 HRED 模型 [56] 。基于 HRED，Serban 等人又提出了利用随机潜变量（Stochastic Latent Variable）增加对话多样性的 Variational Hierarchical Recurrent Encoder-Decoder，即 VHRED [57] ；以及加入了多层次抽象信息的多精度循环网络（Multiresolution Recurrent Neural Networks，MultiRNN） [54] 。在国内， Shang 等人研究了微博数据集上的短文本对话生成问题（Short-Text-Conversation，STC），提出了以 GRU 为门单元的编解码器模型（Neural ResponseMachine，NRM） [58] ，并在人类评价上取得了比检索式系统和翻译式系统 [49] 更好的成绩。 Chen Xing 等人向 Seq2Seq 加入了从预训练 LDA 模型中获取的主题词（Topic Words）并提出了 Topic-Aware Seq2Seq [73] 。由于篇幅有限，不能一一介绍。



## 课题研究意义

生成式对话系统目前有两个难题:一是评价系统生成的响应的困难性 [17, 30, 36, 37, 58] ;二是如何避免生成单调的响应 [25, 28, 49, 56, 64] 。对第一个难题的一个普遍解释是,和机器翻译相比,对话生成的输出具有更高的熵(Entropy)。在对话的场景下,给定一个消息,往往存在大量合理的响应,这些响应和参考响应(Ground Truth)的语义可能完全不一样,并且彼此有着各异的语义。

表1.1 合理的响应不唯一[67]

|  |  |
| --- | --- |
| Query: | 为什么不领养一个? |
| Why not adopt one? |
| Groundtruth: | 我也想领养,但是不好找 |
| I hope so, but it’s hard to find one |
| Candidate 1: | 恩, 我已经在朋友中找了 |
| I’m looking for one from my friends |
| Candidate 2: | 巴比伦是稀有品种 |
| Pappilon is a rare breed |
| Candidate 3: | 可以哈, 谢谢你的建议 |
| OK, thank you for your advice |

如表 1.1所示,合理的响应不唯一,这种固有的多样性决定了评价响应生成系统的困难。Liu 等人在 [36] 中研究了两类指标,分别是基于词重叠的指标(Word Overlap Based)和基于词嵌入的指标(Embedding Based),并且发现这些指标在非技术性的 Twitter 数据集上和人类评价只有弱相关性,在技术性的 Ubuntu Dialogue Corpus [38] 没有相关性。Liu等人的工作为对话领域的自动评价指标的研究开辟了新的方向。

Lowe 等人以自动化的图灵测试(Automatic Turing Test)为灵感,提出了一个以VHRED 为编码器,根据消息、参考响应和模型响应预测人类评分的模型,称为自动化对话评价模型(Automatic Dialogue Evaluation Model,ADEM) [37] 。他们在新的数据和Liu 的数据上评估了 ADEM 模型,发现它和人类评价的相关性在句子水平和系统水平都达到了很高水平。Kannan 等人在 [23] 中初步尝试了把 GAN 的方法应用到对话系统的自动评价中。他们训练了一个鉴别器(Discriminator)来区分一个响应是来自系统还是来自人类,并且发现鉴别器能捕捉到基于 Seq2Seq 的模型倾向于生成短句子和通用句子的缺点。在国内,Tao 等人提出了基于神经网络和词嵌入的 RUBER 指标 [67] 。该指标线性的结合了带参考的指标(Referenced Score)和不带参考的指标(Unreferenced Score)并采用无监督的负采样(Negative Sampling)来训练模型,在中文数据集豆瓣网 2 上取得了很高的人类评价相关性。

本课题延续 Liu 等人的工作,对自动评价指标进行深入研究。Liu 等人已经研究了评价指标与人类评价的相关性,并且发现了它们的弱点;本文进一步研究了评价指标之间的相关性,以及模型的性能在不同数据集上的一致性。虽然,目前大多数评价指标尚不能和人类评价一样准确的衡量系统的性能,但是对它们性质的研究将有助于理解现有指标的弱点,进而有助于自动评价生成式对话系统的发展。

## 课题研究内容

本课题以 [57] 的实验中使用的三个模型 LSTM,HRED,VHRED为基线,扩展了[36] 中考察的两类指标,并在三个具有代表性的公开数据集:Ubuntu Dialogue Corpus,OpenSubtitles 和 LSDSCC [74] 上进行了实验。尽管没有人类监督信号,但是我们对指标之间和模型之间的一致性分析还是取得了有意义的结论。

## 论文组织结构

本文的组织结构如下:第 2 章相关工作介绍了生成式对话领域中自动化指标的使用情况和研究现状;第 3 章研究方法介绍了我们的实验方法以及实验涉及的指标、模型和数据集;第 4 章实验结果与讨论详细展示了我们的实验配置、实验数据以及结论;第 4.4 章结论总结了本课题的研究成果,并展现了若干未来的研究方向。

# 相关工作

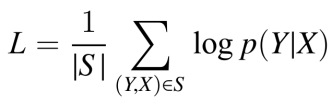
## 生成式模型

### 定义

一个生成式模型定义了给定输入序列 *X = x1 , x2 , · · · , xn* ,任意输出序列 *Y = y 1 , y 2 , · · · , y m*的条件概率:

1-prob

模型的训练目标就是在数据集 S 上最大化给定 X,Y 的对数概率(Log Probability):



从这个角度来看,语言模型 [5, 42] (Language Model)和编解码器(Encoder-Decoder)都属于生成式模型,因为它们都定义了条件概率 p(Y|X)。生成式模型把一个长度可变的序列 X 映射到另一个长度可变的序列 Y,且 X 和 Y 的长度可以不相等。循环神经网络(RNN)为这个问题提供了自然的解决方案。RNN 的基本思想是:序列由有序的元素组成,每一个时刻(Time Step)输入并输出一个元素,同时更新内部的隐层状态(Hidden State)。在时间轴上展开的 RNN 和一般的前馈神经网络(Feed Forward Neural Networks)很像,不过每一个时刻的权重矩阵都是共享的。这个共享的权重矩阵 A 又被称为循环矩阵(Recurrent Matrix)。循环矩阵的作用是保存输入序列的顺序信息,并把当前时刻的输入序列编码成一个定长向量。图 2.1 1[[2]](#footnote-1) 描绘了简化的 RNN 结构。

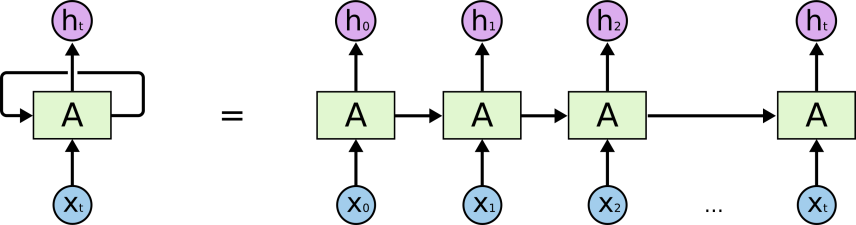
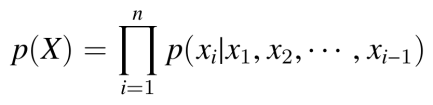


图 2.1 RNN 的一般表示和展开表示

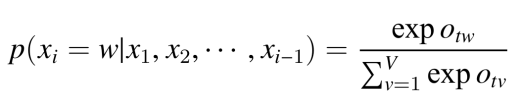
根据是否使用了某种门单元,RNN 可分为普通 RNN [42] ,LSTM [22] 和 GRU [8] 。根据是否对反向序列(Reversed Sequence)建模,RNN 可分为单向 RNN(Unidirectional RNN)和双向 RNN(Bidirectional RNN) [52] 。由于普通 RNN 受到梯度消失的影响,目前学界普遍采用 LSTM 或者 GRU;尽管后者受到梯度爆炸的影响,但是可以通过梯度剪裁(Gradient Clipping) [70] 解决。采用多层 RNN 组成的深度神经网络比单层 RNN 能获得更好的性能 [70] 。

### RNN 语言模型

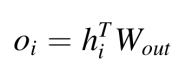
RNN 语言模型可以给出序列 *X = x 1 , x 2 , · · · , x n* 的概率分布:



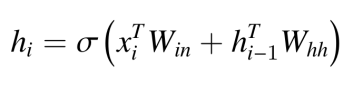
RNN 语言模型[[3]](#footnote-2)通过神经网络中的参数来估计公式 2.3 乘积中的一项:



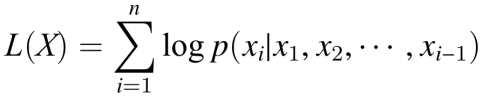
*o t* 是 RNN 的在 t 时刻的输出向量,V 是词汇表的大小。公式 2.4 的右边本质上是对一个长度为 V 的向量进行 softmax 运算。t 时刻的输出向量是由输出矩阵 W out 和 t 时刻隐层状态 h i 相乘得到的:



而 h i 则是当前输入 x i 和上一时刻的隐层状态 h i−1 在输入矩阵 W in 和循环矩阵 W hh 分别作用后再相加的结果:



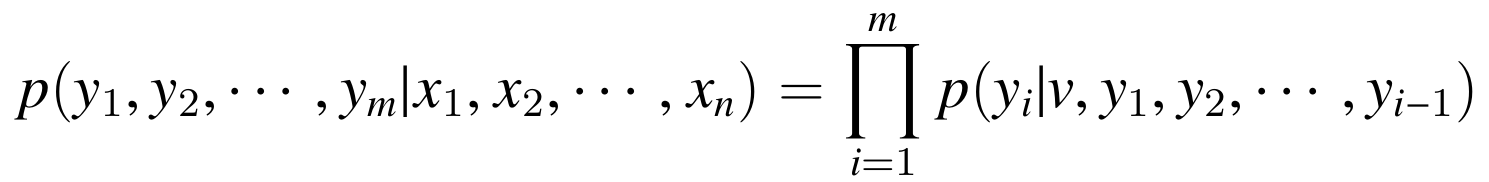
RNN 语言模型在训练时最大化训练集上的句子的对数概率:



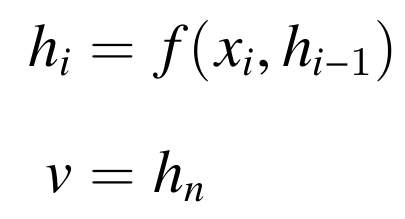
在预测时,对模型输入消息 m,从模型的给出的概率分布中用某种搜索方法,如 Beam Search 可得出响应 r。

### Seq2Seq框架

Seq2Seq 框架使用两个拥有独立参数的 RNN 分别作为编码器和解码器。尽管编解码器不一定都使用 RNN [2] ,本文仅关注使用 RNN 的 Seq2Seq 变体。首先,编码器把输入序列 X 编码成一个定长向量 v。该向量又称为思考向量(Though Vector),是编码器完全读取输入序列后的隐层状态(Last Hidden State)。接着,解码器以 v 为初始隐层状态(Initial Hidden State),像一个 RNN 语言模型一样对输出序列进行预测。整个过程可以描述为:编码器把输入序列 X 变换成某种压缩编码 v,再由解码器把 v 还原为另一个序列 Y,Seq2Seq 把公式 2.1 作了如下转化:



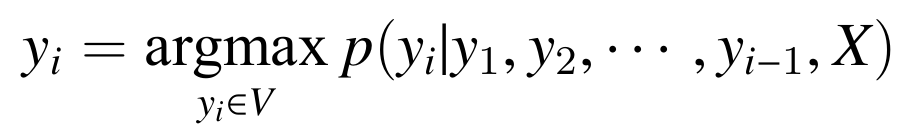
其中 v 是以 f 为门单元的编码器的最后一个隐层状态:



编码器和解码器以同一个目标函数同时训练。为了更好的处理长序列,Seq2Seq 一般引入注意力机制(Attention Mechanism) [2, 40] ,使输入序列的信息不必全部通过固定长度的向量 v 传递。解码器能自动关注和当前输出最相关的输入部分,实现输入序列与输出序列的对自动对齐(Alignment)。注意力机制使传统 Seq2Seq 模型对较长输入也具有鲁棒性。

### 解码算法

生成式模型仅仅定义了条件概率 p(Y|X),在推理阶段(Inference),需要采用某种启发式搜索算法从概率分布中生成输出 Y,这个过程又称为解码(Decode)。最简单的搜索算法是贪心搜索(Greedy Search):在每一时刻都输出条件概率最大的单词:



因为各个 y i 的概率都不是独立的,而是受之前输出的单词的影响,贪心搜索不能保证得到概率最大的输出序列。随机取样(Random Sampling)在每一时刻都从模型生成的全体词汇的概率分布中随机选取一个单词。这样就使输出就带有不确定性,在一定程度上增加了输出的多样性。Serban 等人曾发现随机取样的输出能避免单调响应的问题,并且能产生多样化的,话题相关的输出 [56] 。最为常用的方法是集束搜索(Beam Search),它在生成整个句子的过程中维护一个大小为 B 的列表,称为集束(Beam)。算法开始时,集束初始化为模型生成的 B 个概率最高的单词。在每一个时刻开始时,集束中都有 B 个部分生成的句子,它们称为候选 Y c 。在每一个时刻,算法对集束中的每一个候选都生成B 个概率最大的下一个单词 w i+1,c ,从而形成 B × B 个部分生成的句子 Y ex ,称为扩展的候选集。从扩展的候选集中,只保留前 B 个概率 p(Y ex |X) 最大的句子。算法不断迭代直到在某次对候选集的扩展中,某些候选中产生了句子结束符号(End-of-Sentence,EOS),于是概率最大而且完成了的候选句子将作为输出。

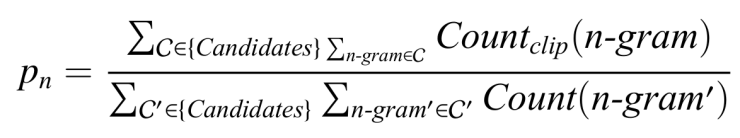
为了增加模型输出的多样性,学者们提出了许多改进的解码算法。 Li 等人在 [27] 中提出在标准集束搜索中,对来自相同父节点的候选加以惩罚,即鼓励来自不同父节点的候选。他们在 [25] 中提出用最大互信息(MMI)对标准集束搜索生成的候选列表进行重新排序,从而提高候选输出的反向概率(Backward Probability),即给定输出 Y,输入 X的条件概率 p(Y|X),使输出对输入更有针对性。他们在 [28] 中提出了一种名为随机贪心取样(Stochastic Greedy Sampling)的解码算法,以求在随机取样和贪心搜索之间找到一个平衡点。与传统的随机取样不同,他们的算法只在条件概率最高的前 K 个候选单词中取样,参数 K 控制了随机取样和贪心搜索之间的比例:K 越大,算法越接近随机取样,K 越小,算法越接近贪心搜索。这些改进的解码算法在不同程度上提高了响应的多样性。

## 自动化评价指标

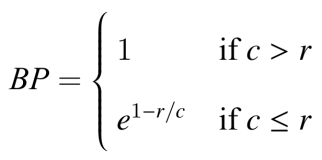
### 评价指标简介

机器翻译领域已有大量和人类评价相关性较高的指标,例如 BLEU [45] ,NIST [13] ,METEOR [4] ,BEER [65] ,CHRF [48] ,TER [60] 等等。然而,适用于开放领域的,面向闲聊的对话系统的指标要少得多;在考察本领域对自动指标的使用情况之前,先对各种指标作一个简要介绍。

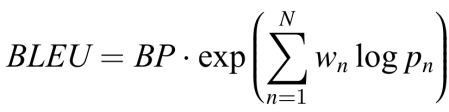
BLEU,Bilingual Evaluation Understudy [45] 是 Papineni 在 2002 年提出的,用于机器翻译的自动评价指标。它是一个系统层面的评价指标,即评价一个系统在整个测试集上的性能。BLEU 指标只有一个参数 N,表示要计算的各阶 n-gram 准确率的最大值;例如 N = 4 表示要计算 1-gram 到 4-gram 的准确率。准确率指的是系统输出和参考输出之间的 n-gram 重叠数占系统输出总的 n-gram 的比例。BLEU 由在整个数据集上计算的各阶 n-gram 准确率的几何平均值(Geometric Mean)和简短惩罚系数(Brevity Penalty)相乘得到。引入简短惩罚系数的原因是,较短的系统输出句子的准确率较高,需要矫正。n-gram 准确率的计算公式为



Candidates 为系统输出的句子集合, Count clip (n-gram) 为截断的 n-gram 共现数, Count(n-gram ′ )是 Candidates 中的总 n-gram 数。简短惩罚系数 BP 的计算公式为：

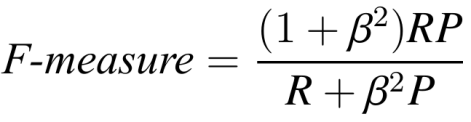


其中 c 是模型输出句子的长度,r 是参考输出句子的长度。BLEU 的最终公式为:



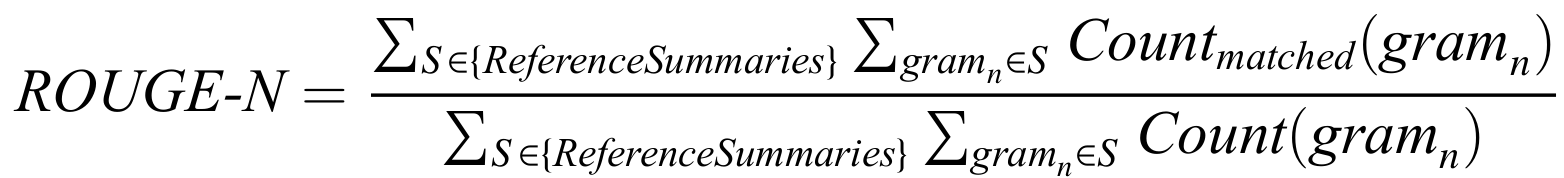
实际使用中一般取 N = 4,w n = 1/N。原始的 BLEU 容易在句子层面给出 0 分,人们提出了各种平滑处理 [7] 。本文在使用 BLEU 时也采用了一种平滑处理。

ROUGE,Recall-Oriented Understudy for Gisting Evaluation [34] 是一种基于召回率的自动摘要领域的指标。它有多个变体:ROUGE-N,ROUGE-L,ROUGE-W,ROUGE-S,以及 ROUGE-SU,分别使用了不同的计数单元(Counting Unit),如 n-gram 共现数、最长公共子序列(Longest Common Subsequence,LCS)和二元跳词(Skip-Bigram)等等。这些指标的基础是信息检索领域常用的 F-measure,即准确率 P 和召回率 R 的加权调和平均值:



β 控制准确率和召回率的相对重要性。以下无特殊说明时,当指标是句子层面的时候,n是系统输出的句子的长度,m 是参考输出的句子的长度;当指标是摘要层面的时候,n是系统输出的摘要的总单词数,m 是参考输出的摘要的总单词数。

ROUGE-N 利用了 n-gram 共现数,其公式为:



摘要层面的 ROUGE-N 具有相同的形式。

句子层面(Sentence Level)的 ROUGE-L 的公式为:

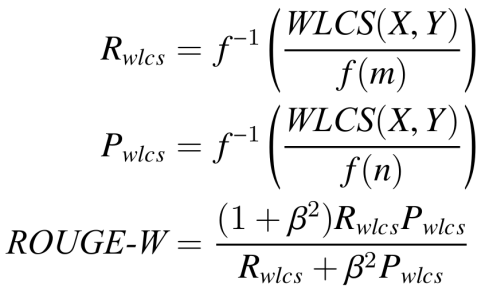
# L-s

LCS 是计算两个序列的最长公共子序列的长度的函数。摘要层面的 ROUGE-L 的公式为:

# L-sum

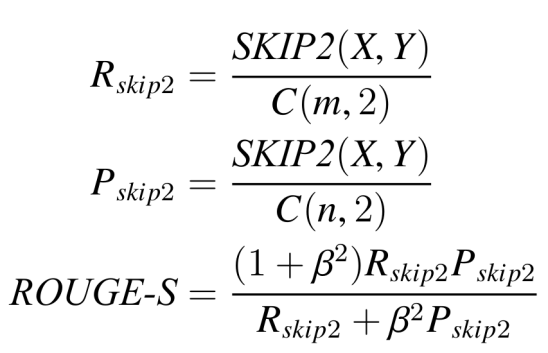
μ 是系统输出的摘要句子的数量,LCS ∪ (r i , C) 计算了参考句子 r i 和候选摘要C(由多个句子组成)的 LCS 的并集。

句子层面的 ROUGE-W 的公式为:



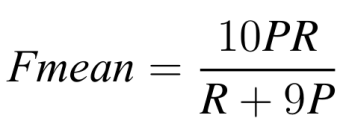
其中 WLCS 是一个计算两个序列的加权 LCS 的算法,该算法奖励较长的连续的 LCS。摘要层面的 ROUGE-W 与摘要层面的 ROUGE-L 类似。

二元跳词是句子中保持句中顺序的一对单词,两个单词之间可以有任意数量的其他单词。基于二元跳词的句子层面 ROUGE-S 定义为:

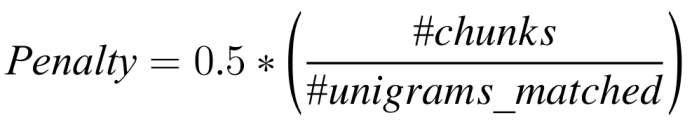


C(·, ·) 为组合数。摘要层面的 ROUGE-S 相当于把摘要看作首尾相连的句子来计算。ROUGE-SU 是 ROUGE-S 加入了 Unigram 的扩展。

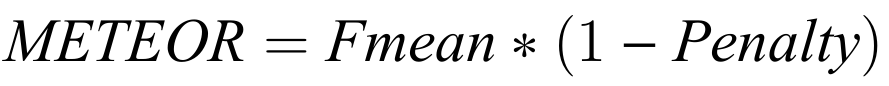
METEOR, Metric for Evaluation of Translation with Explicit ORdering [4] 是针对 BLEU的一些弱点提出的机器翻译的指标。与 BLEU 相比,METEOR 在句子水平上与人类评价有更好的相关性。METEOR 首先计算系统输出和参考输出之间的 Unigram 匹配,这些匹配由多个可配置的模块组成,包括 Exact,Porter-stem,WordNet-synonymy,分别表示严格匹配,Porter 词根匹配和 WordNet 同义词匹配。接着,METEOR 在 Unigram 匹配上计算一个对齐,并得到基于 Unigram 匹配的准确率和召回率,进而得到二者的加权调和平均值:



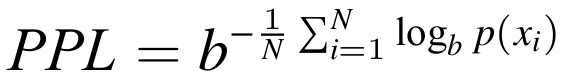
METEOR 还加入了对较短的 n-gram 匹配的惩罚系数:



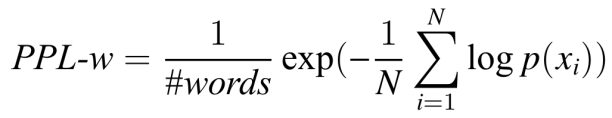
#unigrams\_matched 是所有匹配的 Unigram 的数量;一个 Unigram 匹配越短,#chunks 就越大。METEOR 的最终公式为:



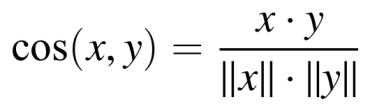
困惑度(Perplexity,PPL)是一种衡量统计语言模型性能的指标。困惑度 P 可以形象的表述为:一个语言模型在预测一个词的时候,平均需要从 P 个词中等可能的选出一个, 因此,困惑度越低,语言模型在选择一个词时就越不“困惑”。困惑度的计算公式为:



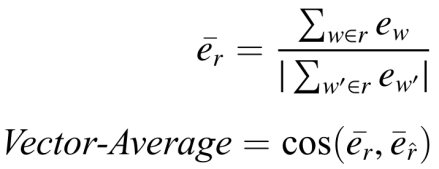
b 是常量,通常取自然对数、10 或者 2。N 是测试集的样本数,x i 是一个样本,在语言模型中它是一个句子,在生成式模型中它是一对输入输出序列 (X, Y)。p(x i ) 是模型赋给样本的概率。一个好的模型应该给测试集的样本赋予较高概率,所以好的模型 PPL 较低。实际在自然语言处理中使用的 PPL 还要除以文本中的总单词数,得到平均每个词的困惑度(Perplexity per-word):



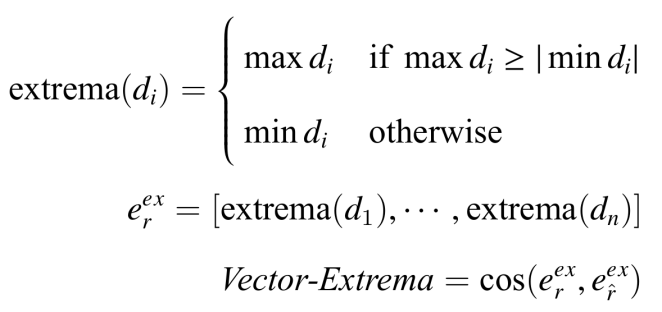
词嵌入(Word Embedding)指标是一类建立在分布式假设 [20, 21] (Distributed Hypothesis)上,用分布式语义(Distributed Semantic)来衡量两个句子的相似程度的指标,常被用于句子文本相似性(Sentence Textual Similarity)和学生输入自动打分 [51] 等任务中。这类指标一般用某种组合方式从单词的向量表示得到句子的向量表示,再用余弦相似度(Cosine Similarity)测量两个句子向量的相似程度 [44] :



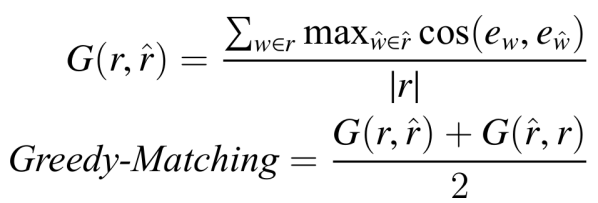
最常见组合方式是对单词向量取平均值,这类似于词袋表示(Bag-of-Words),对应的指标就是向量平均值(Vector Average):



r 是参考输出, r̂ 是系统输出,w 是句子中的一个单词。另外一种组合方式被称为向量极值(Vector Extrema),它把单词向量每个维度上最极端的值作为句子向量在该维度上的值 [15] :

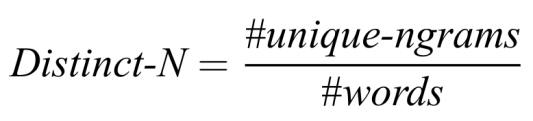


*eexr* 是参考输出的向量极值表示,e r̂是系统输出的向量极值表示*,[·, ·]* 表示连接多个标量形成一个向量。最后一种方法是贪心匹配(Greedy Matching),它得名于边加权的二部图(Weighted Bipartite Graph)的最大匹配问题 [51] :把两个句子的单词看做二部图的节点,任意两个节点之间有一条边,边的权重定义为两个单词的余弦相似度,一个匹配定义为一对点,问如何构造一个匹配的集合,使其权重之和最大。贪心匹配给出了一种贪心算法:



除了上述三种组合方法外,还有很多其他方法 [44] 。

Distinct-N 是 Li 等人提出的衡量句子层面的 n-gram 多样性的简单指标 [25] 。它测量了一个句子不重复的 n-gram 数量除以句子的长度,公式为:



### 评价指标使用情况

生成式对话系统并不像机器翻译领域有许多和人类评价相关性高的指标。对话的研

究者们很早就意识到了这一点 [49, 56, 58, 70] ,所以他们常常使用人类评价作为核心的评判标准,有时用一些指标来衡量模型某一方面的性质,但很少把指标作为评价系统的主要手段。此外,展示和对比不同模型的响应输出也是这个领域常见的分析手段,通常称为定性分析(Qualitative Analysis),表 2.1 展示了模型 GoogleChatbot 对哲学问题的响应。

表 2.1 定性分析的例子[70]

|  |  |
| --- | --- |
| **Human:** | what is the purpose of life ? |
| **Machine:** | to serve the greater good . |
| **Human:** | what is the purpose of living ? |
| **Machine:** | to live forever . |
| **Human:** | what is the purpose of existence ? |
| **Machine:** | to find out what happens when we get to the planet earth . |
| **Human:** | where are you now ? |
| **Machine:** | i’ m in the middle of nowhere . |
| **Human:** | what is the purpose of dying ? |
| **Machine:** | to have a life . |
| **Human:** | what is the purpose of being intelligent ? |
| **Machine:** | to find out what it is . |
| **Human:** | what is the purpose of emotions ? |
| **Machine:** | i don’ t know . |

Ritter 等人首次尝试了数据驱动的,面向闲聊的对话响应生成 [49] 。因为不清楚面向任务的指标能不能用于评价生成式模型,他们就用了基于 Amazon Mechanical Turk 的人类评价。他们利用人类评价的数据考察了 BLEU 在这方面的适用性,并发现系统的BLEU 得分非常低,和人类评价的相关性也不是很高,因此他们认为 BLEU 不能直接应用到本领域。

Shang 等人在评价他们的 NRM 模型时分析了几种指标在本领域的适用性 [58] 。他们认为 BLEU 并不适用,因为合理的响应的范围实在是太大了,参考响应不可能完全覆盖到;而常用于语言模型的指标困惑度(Perplexity,PPL)也不适用,因为它不能测量响应的自然程度及其对消息的相关程度。最后他们选择了人类评价。

Sordoni 在 DCGM 的评价时采用了 BLEU 和 METEOR 两种自动指标 [64] 。为了处理庞大而且多样的响应空间,他们用信息检索(Information Retrieval,IR)的方法从数据集中挖掘潜在的合理响应,并让人类评估员对其合适度(Appropriateness)打分,从而构造了一个多重参考评测集(Multiple-Responses Benchmark Dataset)。在这样的评测集上,他们发现 BLEU 对系统的排名和人类评价非常一致。这种构建多重响应测评集的方法也见于 LSDSCC [74] 的测评集的构造方法,以及 DeltaBLEU [17] 指标的设计思路。

Serban 等人在 HRED 模型的评价中使用了困惑度和单词分类错误(Word Classification Error,WCE) [56] 。低的困惑度表示模型对数据的概率分布的拟合程度好。Serban 等人认为困惑度是适用的,因为在存在多个合理输出的情况下,困惑度总是衡量生成单一参考输出的概率,这意味着它能可靠的衡量模型的质量。WCE 计算模型的输出中正确预测的单词占整个数据集单词数量的比例,这里的“正确预测”要求单词在句子中的顺序也要正确,所以它是一个比困惑度更严苛的指标。尽管使用了自动化指标,Serban 等人指出:这些指标与他们想要衡量的语法正确度(Grammatical Correctness)和语义连贯性(Semantic Coherence)之间的相关性并不明朗。值得注意的是,他们并没有使用人类评价。

Serban 等人在 VHRED 模型的评价中主要使用了基于词嵌入的指标 [57] 和人类评价。Serban 等人认为,虽然这些基于词嵌入的指标和人类评价的相关度不高,但是它们可以用于测量话题相似性(Topic Similarity),也就是模型的输出和参考输出的语义内容是否相近。模型的输出和参考输出可能没有 n-gram 重叠,但是语义的距离总是可以衡量的。

Vinyals 等人在 [70] 的评估中使用了困惑度,定性分析和人类评价。他们展示了模型在不同话题的对话中的输出,定性的分析了模型的优缺点。他们认为这些测评方法都存在明显的弊端,而设计出能快速评价对话模型的好指标仍有待学界研究。

### 公开的对话数据集

本领域在数据集的采集方面已经取得了很大进展,表 2.2 列举了一些常见数据集,大规模是指对话数量在 1M 以上的,中等规模是指对话数量在 50K–1M 之间的[[4]](#footnote-3)。Sina Weibo Corpus 是一个中文数据集,其他都是英文数据集。Twitter 相关的数据集由于隐私保护政策的原因,不能公开原始数据。数据集 [9–11, 16, 32, 76, 77] 都具有丰富的元信息,如时间戳,对话双方身份信息等等。数据集 [32, 64, 74] 有人类标注信息,如情感、话题和方面(Aspect)等。元信息和人类标注信息提供了额外的表征,对构建更智能的对话系统很有帮助。本文把注意力集中在三个数据集上:Ubuntu Dialogue Corpus,OpenSubtitles,LSDSCC,因为它们分别代表了三个常见的对话领域,分别是:技术支持,电影字幕和在线论坛讨论。

表 2.2 数据集一览

|  |  |  |  |
| --- | --- | --- | --- |
| 名称 | 规模 | 领域 | 是否公开 |
| Ubuntu Dialogue Corpus [38] | 大 | 技术支持 | 是 |
| Twitter Corpus [49] | 大 | 短对话多领域闲聊 | 否 |
| Twitter Triple Corpus [64] | 大 | Twitter Corpus 的扩展版本 | 否 |
| OpenSubtitles [64] | 大 | 电影字幕(Subtitles) | 是 |
| LSDSCC [74] | 中 | Reddit 论坛电影板块 | 是 |
| Supreme Court Corpus [10] | 中 | 美国高等法院辩论 | 是 |
| Wikipedia Talk Pages Corpus [11] | 中 | 维基百科编辑者在线讨论 | 是 |
| Tennis Corpus [16] | 中 | 网球比赛赛后新闻发布会 | 是 |
| Parliament Corpus [77] | 中 | 国会讨论 | 是 |
| Conversations Gone Awry  Corpus [76] | 中 | 维基百科讨论页面吵架集锦 | 是 |
| Movie Dialogs Corpus [9] | 中 | 电影对白 | 是 |
| Movie-DiC [3] | 中 | IMSDB 电影对白 | 否 |
| MovieTriples [56] | 中 | Movie-DiC 的扩展版本 | 否 |
| SubTle [1] | 中 | 电影字幕 | 否 |
| DailyDialog [32] | 中 | 日常对话 | 是 |
| Sina Weibo Corpus [71] | 中 | 新浪微博短对话闲聊 | 是 |

Ubuntu Dialogue Corpus 是 Lowe 等人从 Freenode IRC 网络的 Ubuntu 板块的聊天日志[[5]](#footnote-4)中获取的技术性两人多轮对话。这个数据集中包含了大量技术符号,比如路径、命令、URL,还有笔误(Typo),缩写(Abbreviation)和俚语(Slang)。它的对话数量多达930,000(接近 1M),庞大的数据量为数据驱动的模型提供了极佳的试验场。它的多轮特性为模型提供了更长的上下文,有助于能利用多轮对话的模型生成更有意义的响应。

OpenSubtitles 是 the Open Parallel Corpus,OPUS 工程的一部分。它是从一个人们可以自由上传和下载电影字幕的网站[[6]](#footnote-5)的数据库中获取的,庞大而充满噪音的开发领域数据集。由于 OPUS 的目的是收集机器翻译所用到的双语文库(Parallel Corpus),OpenSubtitles 也是一个用于机器翻译的双语文库,它里面并没有对话数据集所需要的轮换信息(Turn-Taking),而且也没有区分旁白、独白和对白。Sordoni 等人在 [70] 中把OpenSubtitles 用到对话领域,他们把相邻的两个句子视为消息和响应,并且每一个句子既是消息又是响应。Li 等人在 [25] 中也采取了把相邻句子分别作为模型的输入和输出的方法。尽管充满噪音,OpenSubtitles 仍然是目前最大规模的电影字幕数据集。致谢

我要感谢我的毕业设计指导老师荣文戈副教授。我和荣老师是在大三下的《研究方法论》课程认识的,也正是这个课程让我选择了荣老师作为我的毕设导师。荣老师带我走进了学术写作的殿堂,在他的《研究方法论》上,我第一次用 LATEX 写作了第一篇带有学术性质的文献综述,那篇比较各种词嵌入模型的“处女作”至今让我记忆犹新。在毕设一开始,我因为考研、重修课程等事情无法立刻投入工作,荣老师对此表示了极大的理解,对此我深表感谢。在中期报告前的一个月,荣老师给了我关键的支持:他不但给了我两台学院的服务器,而且还在紧缺的实验室工位中给我安排了一个。在此后的毕设工作中,他不断根据我的工作汇报的反馈为我的毕设提供切实有效的指导。没有荣老师的帮助和指导,我的毕设工作可能完全走不上正轨。荣老师温文尔雅的风度和关爱学生的作风是我十分赞赏的。能在他的指导下完成毕设是我的荣幸。

我要感谢我们计算机学院对毕设工作的高度重视和大力支持。学院早在去年 10 月就召开了毕设动员大会,不就之后就组织了全学院的开题报告。据我所知,全校范围内恐怕没有比我们学院更早进行本科生毕设开题的了。得益于此,我们比别的学院多出几乎半年时间准备毕业设计,虽然惭愧的是,我在蹉跎中耗费了大段光阴。我非常感谢高小鹏院长等学院领导对我们毕设事宜的重视。高院长在毕设动员大会上指出了毕设过程中可能会遇到的种种挫折,告诫我们要对自己负责,要多和导师沟通等等,是我在毕设过程中一直受用的。开题报告、中期报告和毕设答辩等环节有赖于学院的全体教职工的辛勤付出。感谢他们不惜牺牲宝贵的科研时间来为我们这些初出茅庐的本科生提意见。

我还想感谢 G951 的各位学长学姐,感谢他们在五一劳动节和我一起看电影,这是一段美好的回忆。感谢学生三公寓的楼管阿姨,她总是把我们当成自己的孩子看待,热心帮助我们,关心我们。感谢密码学课程的郭华老师,她在我毕业这年教会我信息安全的重要性。感谢我校的校车司机,他们一天出车十几趟,将满载的师生安全送达。感谢在某个暴雨的深夜,一位不知名的博士学长用他的伞送我回宿舍,陌生人也能让人感到温暖。

最后,我要感谢我的父母:你们的无私和博大的关爱,我将永远无法偿还。

参考文献

1. http://www.cleverbot.com/ [↑](#footnote-ref-0)
2. http://colah.github.io/posts/2015-08-Understanding-LSTMs/ [↑](#footnote-ref-1)
3. 为了简洁起见,我们描述了普通 RNN。LSTM 和 GRU 有着更复杂的数学表达式。 [↑](#footnote-ref-2)
4. 1K = 1024,1M = 1024 K,1G = 1024 M。 [↑](#footnote-ref-3)
5. https://irclogs.ubuntu.com/ [↑](#footnote-ref-4)
6. http://www.opensubtitles.org [↑](#footnote-ref-5)