西安交通大學

XI'AN JIAOTONG UNIVERSITY

面向对象 Java 编程

课程报告

一．背景介绍

随着近年来互联网技术的迅速发展，各式各样的数据信息爆炸增长。文本数据是大数据时代人们普遍使用的数据形式，而人类的对话数据是文本数据的主要组成部分。大数据时代的到来催化了自然语言处理和深度学习技术的发展和应用，互联网每天都能产生大量的人类对话数据用于训练学习，提取关键信息，了解需求并作出反馈，因此能够感应用户情感需求并提供帮助的聊天机器人技术已经成为当前人机交互开发的重要方向。

1966 年，世界上第一个聊天机器人 EKLIZA 诞生于麻省理工学院，ELIZA 的主要功能是对用户提出的心理问题进行反馈。ELIZA 的出现激起了行业内对聊天机器人的开发浪潮。我们所熟知的苹果 Siri，微软小冰，Cortana 是现在客户端和电脑端较为流行的聊天机器人，可用于简单的信息查询和人机对话。其中微软小冰更加注重人工智能在你和人类情商维度的发展，强调人工智能情商，而非任务完成在人机交互中的基础价值。如今，微软小冰已是跨平台人工智能机器人，可以在微博、京东、微信、QQ、网易云音乐等平台见到小冰的身影。此外，近期小冰加入什么值得买 APP 后，在 APP 评论中的表达与正常人的表达方式几乎无异，甚至能理解一些流行用语。

二．概要设计

2.1 需求分析

本次课题主要是围绕心理健康问题设计聊天机器人作为用户的一种便捷的咨询方式。随着现代生活节奏的加快，学习、工作和生活的压力日益增大，越来越多的人出现较严重的心理问题，大多数人迫于环境压力得不到有效的心理咨询，通过与聊天机器人的交流能够了解到心理健康的重要性，进而进行更深入治疗。

因此，专注于心理健康的聊天机器人的设计开发便显得十分重要。

2.2 整体架构

1）数据库建立：在关于心理健康方面的网络、论坛或文献等对数据进行爬取、解析和整理，建立一个问答数据库。

2) 自然语言处理: 对于输入的问题进行分词解析，在数据库中进行相似度匹配，寻找最准确的答案；

3）机器人获取问题与回复：解析用户的问题，根据语言处理将得到的回答反馈给用户。

4）聊天机器人的搭建：选取框架搭建服务，用户界面设计；

5）集成测试：对各个环节之间进行沟通对接，对软件接口的规范提出要求，将不同的模块归总集成并对其进行测试；

2.3 功能设计

在此次聊天机器人项目中，本组负责自然语言处理部分。首先，我们需要对用户输入的问题进行分词处理，提取关键属性，向量化处理后进行情感分析。对于用户输入的问题我们采取两种机制产生针对该问题的回答。一个是生成机制，以 Seq2Seq 为框架的循环神经网络，根据用户输入的问句生成答句，可以满足聊天机器人的日常简单对话功能。另一个是基于检索式机制，在建立好的心理健康相关数据库中搜索、匹配出能够正确回答用户问题的答案，保证对于用户的心理问题能够得到有效的回答。在实际操作过程中，两种机制同时计算，当检索结果有效，能精准匹配上已有问答库的数据时，返回检索式结果；当输入的新问句与我所有的问句库的相似度都低于设定阈值时，则输出生成式结果。

三．语料库预处理

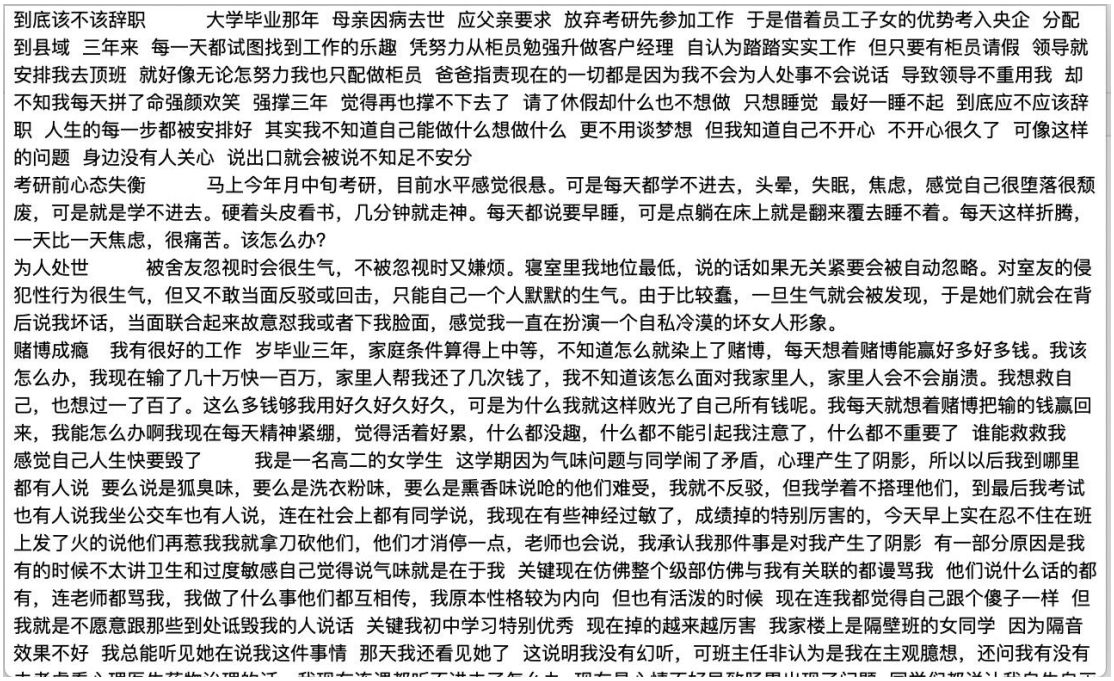
我们与第一组数据库建立的两个小组进行对接，得到了刘书航组和王欣蕾组爬取的语料库作为检索式机制的训练语料库。另外，我们选取中文对话语料库作为生成机制的训练语料库。对于语料库的预处理包括以下步骤：

a) 去除无用的标点以及所有的数字和字母

b) 对文本进行分词

c) 去除停用词

d) 问答对（‘Q’, ’A’）



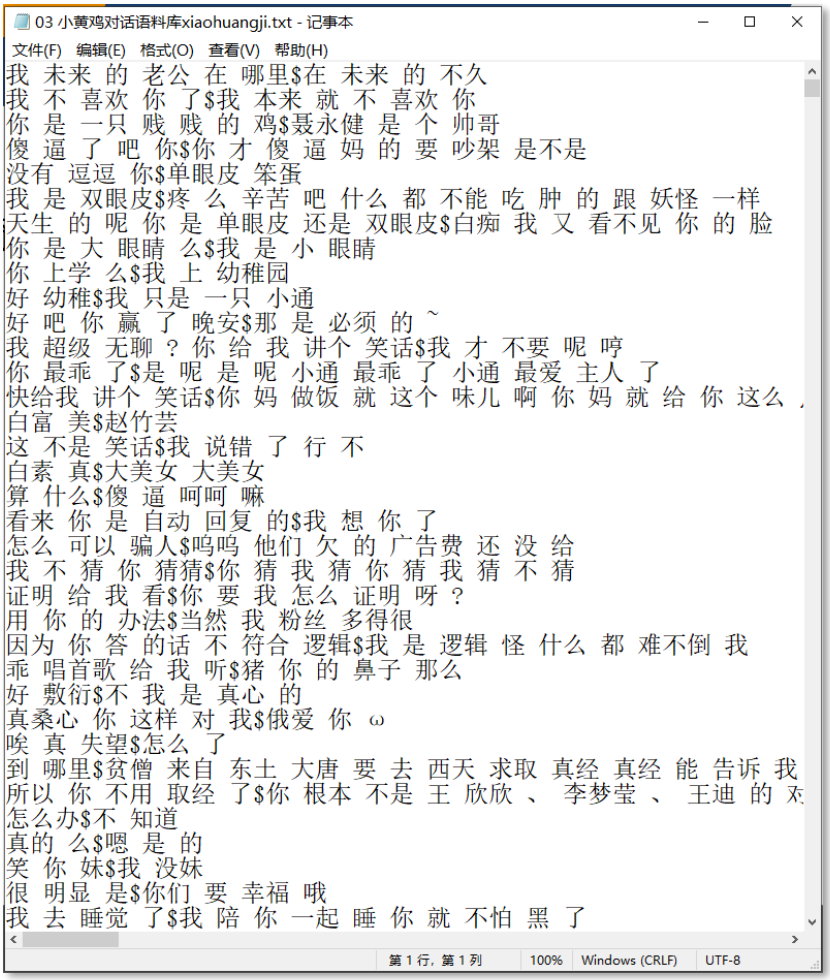
到底该不该辞职　　大学毕业那年 母亲因病去世 应父亲要求 放弃考研先参加工作 于是借着员工子女的优势考入央企 分配到县域 三年来 每一天都试图找到工作的乐趣 凭努力从柜员勉强升做客户经理 自认为踏踏实实工作 但只要有柜员请假 领导就安排我去顶班 就好像无论怎努力我也只配做柜员 爸爸指责现在的一切都是因为我不会为人处事不会说话 导致领导不重用我 却不知我每天拼了命强颜欢笑 强撑三年 觉得再也撑不下去了 请了休假却什么也不想做 只想睡觉 最好一睡不起 到底应不应该辞职 人生的每一步都被安排好 其实我不知道自己能做什么想做什么 更不用谈梦想 但我知道自己不开心 不开心很久了 可像这样的问题 身边没有人关心 说出口就会被说不知足不安分

考研前心态失衡　　马上今年月中旬考研，目前水平感觉很悬。可是每天都学不进去，头晕，失眠，焦虑，感觉自己很堕落很颓废，可是就是学不进去。硬着头皮看书，几分钟就走神。每天都说要早睡，可是点躺在床上就是翻来覆去睡不着。每天这样折腾，一天比一天焦虑，很痛苦。该怎么办?

为人处世　　被舍友忽视时会很生气，不被忽视时又嫌烦。寝室里我地位最低，说的话如果无关紧要会被自动忽略。对室友的侵犯性行为很生气，但又不敢当面反驳或回击，只能自己一个人默默的生气。由于比较蠢，一旦生气就会被发现，于是她们就会在背后说我坏话，当面联合起来故意怼我或者下我脸面，感觉我一直在扮演一个自私冷漠的坏女人形象。

赌博成瘾　我有很好的工作 岁毕业三年，家庭条件算得上中等，不知道怎么就染上了赌博，每天想着赌博能赢好多好多钱。我该怎么办，我现在输了几十万快一百万，家里人帮我还了几次钱了，我不知道该怎么面对我家里人，家里人会不会崩溃。我想救自己，也想过一了百了。这么多钱够我用好久好久好久，可是为什么我就这样败光了自己所有钱呢。我每天就想着赌博把输的钱赢回来，我能怎么办啊我现在每天精神紧绷，觉得活着好累，什么都没趣，什么都不能引起我注意了，什么都不重要了 谁能救救我

感觉自己人生快要毁了　　我是一名高二的女学生 这学期因为气味问题与同学闹了矛盾，心理产生了阴影，所以以后我到哪里都有人说 要么说是狐臭味，要么是洗衣粉味，要么是熏香味说呛的他们难受，我就不反驳，但我学着不搭理他们，到最后我考试也有人说我坐公交车也有人说，连在社会上都有同学说，我现在有些神经过敏了，成绩掉的特别厉害的，今天早上实在忍不住在班上发了火的说他们再惹我我就拿刀砍他们，他们才消停一点，老师也会说，我承认我那件事是对我产生了阴影 有一部分原因是我有的时候不太讲卫生和过度敏感自己觉得说气味就是在于我 关键现在仿佛整个级部仿佛与我有关联的都谩骂我 他们说什么话的都有，连老师都骂我，我做了什么事他们都互相传，我原本性格较为内向 但也有活泼的时候 现在连我都觉得自己跟个傻子一样 但我就是不愿意跟那些到处诋毁我的人说话 关键我初中学习特别优秀 现在掉的越来越厉害 我家楼上是隔壁班的女同学 因为隔音效果不好 我总能听见她在说我这件事情 那天我还看见她了 这说明我没有幻听，可班主任非认为是我在主观臆想，还问我有没有

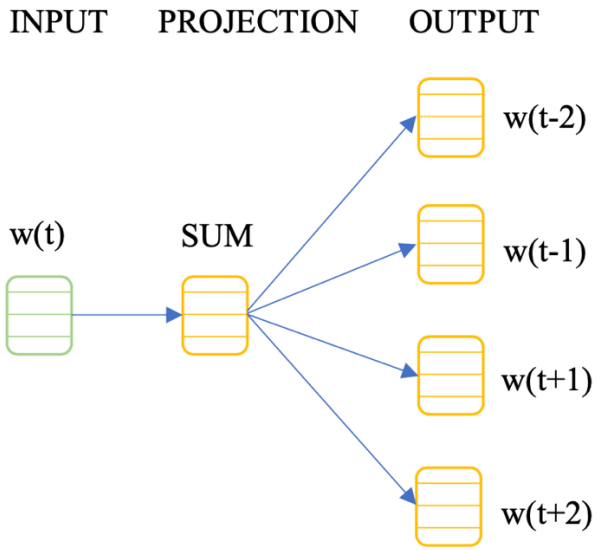


四．模型建立

4.1 检索式机制

在检索式机制中，我们采用基于负采样的 Skip-gram 模型获取词向量，能有效提高训练速度并保证词向量的质量。

Skip-gram 模型和 CBOW 模型均包含输入层、投影层和输出层，但 Skip-gram 模型与 CBOW 不同的是，Skip-gram 模型是在已知当前词w_t的前提下，预测其上下文w_{t-2}， w_{t-1}， w_{t+1}， w_{t+2}（假设窗口大小为 2）。



但是，skip-gram 模型中，中心词w_c生成背景词w_o的概率使用了 softmax

$$P(w_o|w_c) = \frac{exp(u_o^T v_c)}{\sum_{i \in \mathcal{V}} exp(u_i^T v_c)}$$

考虑到背景词可能是词典中的任意一个词并体现在 softmax 的分母上，在利用梯度下降计算时，每次运算的开销都词典大小相关。因此考虑采取负采样法，提高训练速度。我们假设中心词w_c生成背景词w_o有以下两个相互独立时间联合组成来近似：

1) 中心词w_c和背景词w_o同时出现在该训练数据窗口；
2) 中心词w_c和第 i 个噪声w_i不同时出现在该训练数据窗口（噪声词w_i按噪声词分布$P(w)$随机生成）。

我们使用 sigmoid 函数来表示中心词w_c和背景词w_o同时出现在该训练数据窗口的概率：

$$P(D = 1|w_o, w_c) = \sigma(u_o^T v_c)$$

那么，中心词w_c生成背景词w_o的对数概率可以近似为

$$logP(w_o|w_c) = logP(D = 1|w_o, w_c) + \prod_{k=1, w_k \sim P(w)}^{K} (1 - P(D = 1|w_o, w_c))$$

目标函数为

$$min(-logP(w_o|w_c))$$

经过负采样后，梯度下降计算的复杂度由原来的O(|$\mathcal{V}$|)变为O(K)。

将词转化为词向量后，还需将词向量转化为句向量，此时我们使用的是 SIF 嵌入法。嵌入法的计算步骤如下：

1）得到初步句向量。遍历语料库中的所有橘子，假设当前句子为 s，通过如下计算得到当前句子 s 的初步句向量：

$$v_s = \frac{1}{|s|}\sum_{w\in s}\frac{a}{a+P(w)}v_w$$

其中 a 是可调参数，|s|是句子 s 中词语的个数，P(w)是词语 w 的词频处于所有词语词频之和。上述过程是词向量加权求平均的过程。

2）主成分计算。将全体初步句向量进行主成分分析，计算出全体初步句向量的第一主成分 u。

3）得到目标句向量。

$$v_s = v_s - uu^T v_s$$

由上述模型得到了语料库中所有问题的句向量，构成问题矩阵。对于用户新输入的问题，将其转化为句向量后，计算该句向量与问题矩阵中的句向量之间的相似度，从而得到已知问题中与其最相近的问题，匹配得到相应的答句反馈给用户。

4.2 生成机制

4.2.1 文本向量化

将文本数据向量化是自然语言处理问题中的关键。经过大量资料查询之后发现，常用的文本向量化的方法是 One-hot 表征稀疏向量，这种方法简单直观，但是在实现过程中是利用单词在字典中的索引作为词向量，这在实际的操作过程中

会丧失大量的信息，如词与词之间的相关信息。因此，我们考虑使用 GloVe 方法来表征词向量。

GloVe 方法基于全局词汇贡献的统计信息来学习词向量，从而将统计信息与局部上下文窗口方法的优点都结合起来，同时还减少了计算量。通俗来讲，GloVe 方法是将词向量序列中出现在当前词周围的所有词的频率作为一个向量来表征当前词，这样对词向量中的任意两个元素相除之后，就可以根据比值区分相关单词和不相关单词，也能区分相关单词的具体相关性。

GloVe 使用了词与词之间的共现信息。我们定义 X 为共现词频矩阵，在给定窗口大小后，其中元素x_{ij}为窗口中词 j 出现在词 i 环境下的次数。那么

$$P_{ij} = P(j|i) = \frac{x_{ij}}{x_i}$$

为词 j 出现在词 i 环境的概率，这一概率也称为共现概率。其中x_i为所有出现在词 i 附近的词的个数$x_i = \sum_k x_{ik}$。我们称词 i 和词 j 分别为中心词和背景词，在这里使用v和$v^{\sim}$来表示中心次和背景词的词向量。利用有关词向量的函数 f 来表达共现概率比值：

$$f(v_i, v_j, v_k^{\sim}) = \frac{P_{ik}}{P_{jk}}$$

由于共现概率比值是一个标量，我们可以使用向量之间的内积把函数 f 的自变量进一步改写，得到

$$f((v_i - v_j)^T v_k^{\sim}) = \frac{P_{ik}}{P_{jk}}$$

由于任意一对词共现的对称性，满足以下两个性质：

a) 任意词作为中心词和背景词的词向量应该相等，即对任意 i，$v_i = v_i^{\sim}$。

b) 词与词之间共现次数矩阵 X 应该对称，即对任意词 i 和 j，$x_{ij} = x_{ji}$。

为满足上述性质，我们令

$$f((v_i - v_j)^T \tilde{v_k}) = \frac{f(v_i{}^T \tilde{v_k})}{f(v_j{}^T \tilde{v_k})}$$

且$f(x) = exp(x)$，那么

$$exp(v_i{}^T \tilde{v_k}) = P_{ik} = \frac{x_{ij}}{x_i}$$

可得

$$v_i{}^T \tilde{v_k} = log(x_{ij}) - log(x_i)$$

为了保证对称性，当 i 与 j 互换时，上式仍然满足，可以将替换$log(x_i)$成两个偏移项之和$b_i + b_k$。因此，对于任意一对词 i 和 j，用它们的词向量表达共现概率比值最终可以简化为表达它们共现词频的对数

$$v_i{}^T \tilde{v_j} + b_i + b_j = log(x_{ij})$$

上式中共现词频时直接由训练数据统计得到，为了学习词向量和相应的偏移项，我们希望上式等号两边越接近越好，给定词典大小$\mathcal{V}$和权重函数$f(x_{ij})$，我们定义损失函数为

$$\sum_{i,j=1}^{V} f(x_{ij})(v_i{}^T \tilde{v_j} + b_i + b_j - log(x_{ij}))^2$$

对于权重函数$f(x)$，当$x < c$（如 c=100）时，令$f(x) = (x/c)^\alpha$（如$\alpha = 0.75$），反之令$f(x) = 1$。此时，损失函数的计算复杂度与共现词频矩阵中的非零元素的数目呈线性关系，可以从共现词频矩阵中随机采样小批量非零元素，使用随机梯度下降迭代词向量与偏移项。当学习完所有的词向量后，我们使用一个词的中心词向量与背景词向量之和作为该词的最终词向量，可以提高鲁棒性。

$$v_i = v_i + \tilde{v_i}$$

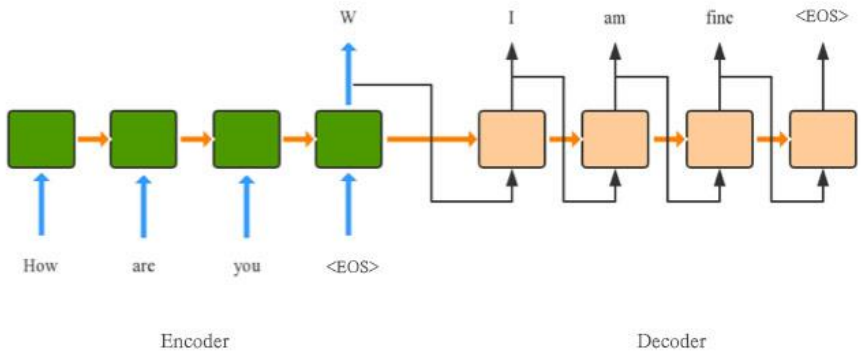
以下是选取词语计算词语之间共现概率比值，得到如下结果：

k =	哥们	闺蜜	分手	电影
P(k|男朋友)	2.62E-04	4.89E-04	5.02E-03	4.36E-05
P(k|女朋友)	2.18E-05	9.54E-03	6.55E-03	1.21E-05
P(k|男朋友)/P(k|女朋友)	12.04	0.05	0.77	3.60

上述结果可以得知，哥们出现在男朋友附近的概率远大于出现在女朋友附近的概率，闺蜜出现在女朋友附近的概率远大于出现在男朋友附近的概率。这样的结果是符合实际情况的，说明 GloVe 能够保留文本中单词之间的相关性，便于后续的训练学习生成答句。

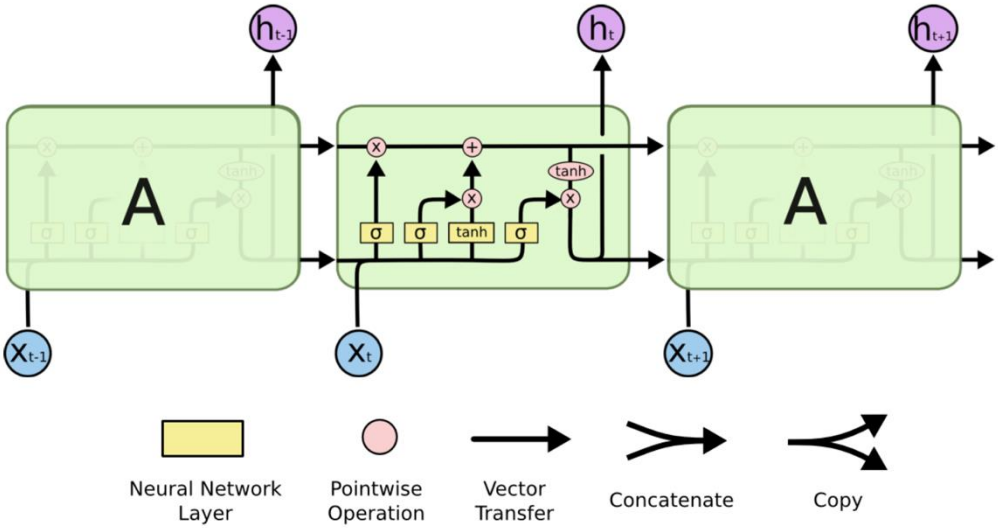
4.2.2 模型建立

Seq2Seq 是一个 Encoder-Decoder 结构的神经网络，它的输入是一个序列(Sequence)，输出也是一个序列(Sequence)。



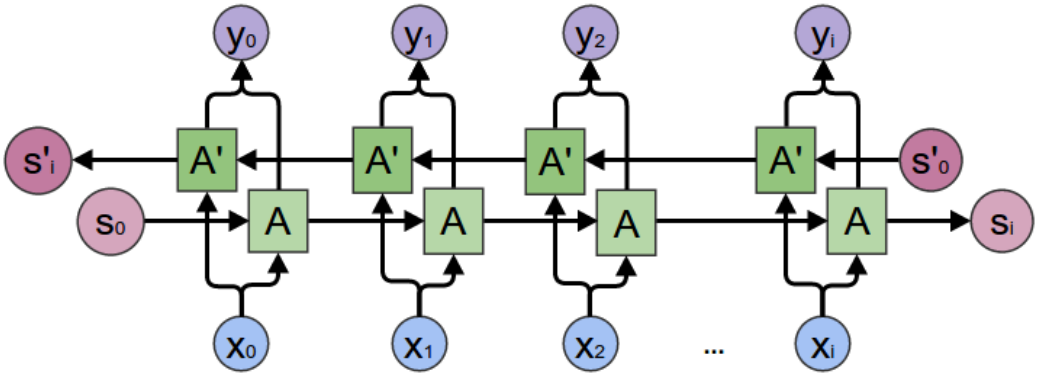
聊天机器人除了能够回答用户问题以外，还需要能够对对话过程中的内容具有一定时间长度的记忆性，从而使用户感到准确并有兴致地聊下去。因此，在 Seq2Seq 框架中选择 GRU（LSTM 的一种变体）作为处理器。LSTM 是忘记一部分记忆，再保留一部分当前的输入信息，从而更新为新的记忆，从而得到输出。GRU 减少了需要训练的参数个数，在忘记信息时，遗忘了多少信息就从当前输

入的信息中保留多少信息。



1）编码

双向 RNN 适用于当前输出不只依赖于之前的序列元素，还可能依赖之后的序列元素这种情况。

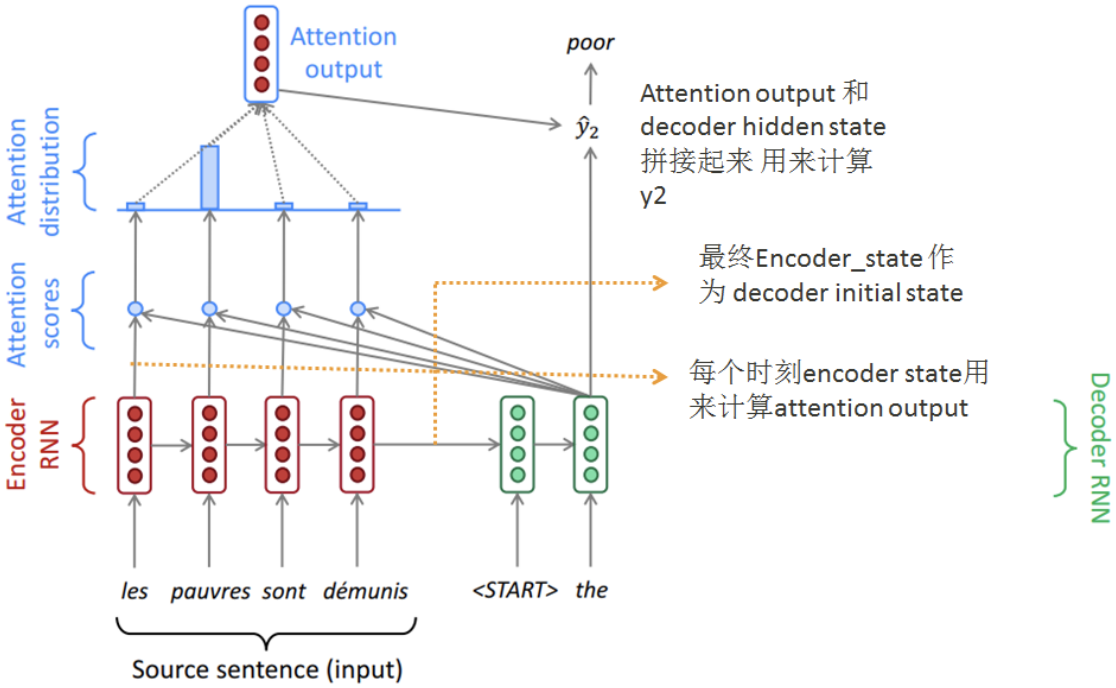


由于在实际应用过程中，一个以心理健康为主题的聊天机器人应该不仅仅拥有匹配心理健康相关的问答的功能，还应当具有基础的聊天功能，从而使得用户体验更佳。双向 RNN 就保证了基础聊天功能的实现。

2） 解码

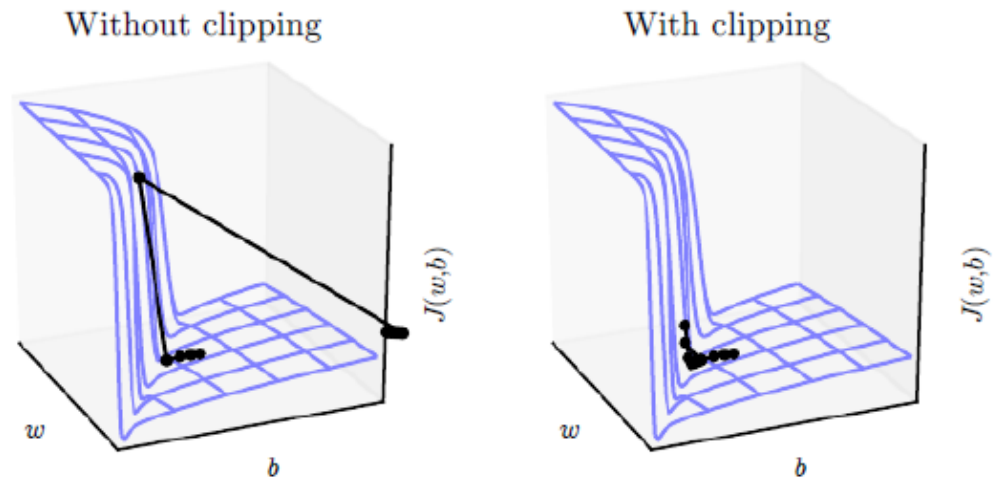
在一般的解码过程中，我们引入了 Attention 机制。Attention 模型就是基于 Encoder-Decoder 框架下的一种模拟人类注意力直觉的一种模型。在产生下一预

测结果词向量时，对编码前每个输入词向量的计算关注程度，输出答案词向量序列。在编码过程中，每个词向量转换为了中间词向量，Attention 机制计算了每个词向量的得分，进而得到每个词向量的权重，得到 Attention 机制的输出，此时将 Attention 机制的输出和解码部分的隐藏层信息相结合的到最后的词向量序列，在字典中匹配相应的单词构成答句反馈出来。

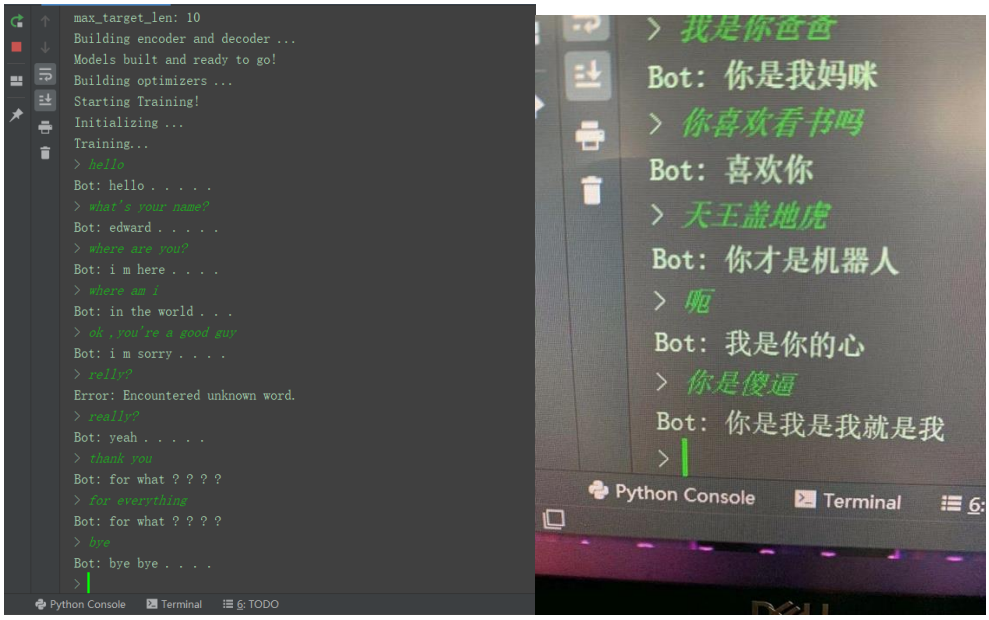


五．训练迭代

在训练迭代过程中会出现梯度爆炸的情况，我们采取的措施是设定梯度上界从而进行正常的梯度下降进行迭代。



三．结果展示



五．总结

1） 语料库质量较差。问答内容冗杂，难以提取有效信息。在实现检索式机制时，针对语料库中的原始问句得到的效果很差，于是采取根据主题进行检索。

2）心理健康这一主题涉及到的知识十分广泛，需要系统而详细的心理健康相关语料库来训练。与普通的客服机器人不同，普通的客服机器人具有很强的针对性，而心理健康聊天机器人很难根据用户的个体情况做出针对性的

解答

3) 目前的网络模型对于硬件性能要求很高，实用性较低。

4) 我们的模型可以保证日常聊天及心理健康咨询，提升用户体验。

5) 在程序方面，更新优化时只需要更换 4 个数据模型文件即可，通过 demo.py 函数把两种方法整合，加入了系统自检过程，通过自检后只需要加载一次模型后即可以开始问答处理。

6) 与第四组之间的对接还未实现。