聊天数据收集及清理建议

白子薇（ZiweiBai）
邢超（chao xing）

2017/1/11

FreeNeb

1.摘要

聊天系统一直是自然语言处理中一个很热门的方向，但是常常因为语料缺乏而无法进行，本文就聊天数据的收集及处理提出一些建议。

私人的聊天数据难以获取，但网上有一些公开的聊天数据集是可以获取到的。我们首先选取了诺亚方舟实验室公开的微博数据。在聊天系统的构建中，常常会用到微博数据。我们把某人发布的微博称为 post，微博下的回复称为 response。 但是，直接从微博中爬取的 post-response 对并不规整，我们需要对原始的微博数据进行处理，以期我们的聊天系统获得更好的效果。

除此以外，一些聊天机器人的数据库虽然不公开，但是我们可以通过与他交互获取一部分对话数据。

我们还从 github 上或 CSDN 下载中获取了一些小型公开语料。但是这类语料容易获取，不需要复杂的清洗流程，因此在后面不再介绍。

2.微博初步清洗

1.冗余信息过滤

由于在博主发微博时或其它人在评论时，可以加上话题，表情，网页链接，当前定位，或者@某人等，我们得到的句子中包含了许多冗余信息，这些冗余信息会影响整个句子的连贯性，因此我们要过滤掉这些信息并保持剩下句子的完整性。

冗余信息大致分为三种：一对符号之间的信息，回复@某人时的信息，包含链接的信息。

1.一对符号之间的信息

在语料库中，存在这样的句子：

#百度下午茶#原来飞机发动机产生的风这么大！
在巴黎关于中银 visa 卡的惨痛教训【全文】alink
记不记得，电影 2012 里的美国总统是黑人。（转）

这类型句子虽然在微博中比较常见，但是对于聊天系统而言，#XXX#、【XXX】、（XXX）、「XXX」、『XXX』，等符号内的信短语属于冗余信息，与符号外的句子不具有语义上的连贯性。我们真正用到的是符号之外的信息。因此我们需要用规则将这类微博中#XXX#、【XXX】、（XXX）「XXX」、『XXX』内的内容过滤掉。

2. 回复@某人时的信息

在语料库中，还存在这样一类句子

回复@糊八圈:我还没看完，你看完了

3. 包含链接的信息

在微博中，经常在句子末尾附上一些文章、网页、音乐、位置的链接，这些链接信息对于我们是无用的，因此我们需要删除微博中包含链接的部分。具体有'我在 http' 和 '[a-zA-Z]*http*' ，'alink' ， '0 网页链接' 等。如：

“好奇号”着陆火星精剪视频，科幻大片一样。alink
今天好开心啊，我在 http
Tom 老婆 Charm 做的美味晚饭 0 网页链接
寻找华人“下南洋”的足迹。http:

将红色部分删除后，句子更加符合我们的需求。

2.句子删除

在经过上一步的冗余信息过滤后，仍有一些句子是不符合我们要求的，这些句子包含一些我们不希望出现的信息，但是删除这些信息后句子的完整性被破坏。对于包含此类信息的 pair，我们直接放弃。需要直接放弃的情况大约有两类：

- 句子包含@
  - @leilei_蝈儿
  - 今天和@雷颐去篦街吃了小龙虾，开森
- 包含数字与字母的句子
  - 2005-2007 那些疯狂的魔兽岁月，算半个高端吧
  - 七天养成一个好习惯，52 个星期后你就会脱胎换骨
  - t.cn/zlvHFRT

3.微博进阶清洗

初步清洗之后，所剩的基本都是通顺、完整、无冗余信息的句子。但是对于有的 post 和 response，关联性并不是很强，
比如一些微博中，博主除了文字还附带发了照片，对应的 response 是针对照片的评论，而非针对文字的评论。这样的 post 和 response 关联性不大，不适合作为聊天数据，需要删除。因此，我们将 post 包含“照片”、“图片”这样的 pair 也删除。
最后将包含英文、数字或其它特殊字符的数据删除，只保留包含中文和标点的数据。并对所剩数据进行清洗，删除数据中的标点符号，只保留中文。
不同的模型对数据的要求也有所不同，在实际使用时，我们可能还要对数据进行进一步处理，这与我们的需求及模型特征有关。

4.程序使用

处理微博数据的脚本为：run_deal.sh
使用示例：sh run_deal.sh
得到的结果同一 post 对应的 response 在同一行，用 ‘\t’ 隔开
处理完成之后的数据是未分词的，需要进行分词，才能进行下一步。分词词表已经上传。

5.聊天机器人数据爬取

“聚合数据”提供了图灵机器人的免费 api（https://www.juhe.cn/docs/api/id/112），我们可以通过此 api 对图灵机器人进行爬取。教程中有不同调用 api 的示例，每次给 api 发送一句话，返回对应这句话的回复。我们可以将我们获得的公开数据集中的 Q 或处理好的 post 发送给 api，获取对应的回复。

为了增大数据量，我们可以在获取到图灵机器人的回复后，将回复再次发送给机器人，实现机器人-机器人的对话。为了防止陷入循环，我们设置 5 轮之后就终止对话，发送一个新的 Q 或 post，开始新一轮对话。

注：

1、 频繁的请求会导致请求被拒绝，因此，每次请求被拒绝后需要暂停一个小时左右的爬取。

2、 图灵机器人要求输入不得超过 60 个字符，因此，每次需要判断发送给 api 的字符串长度是否符合要求。

3、 我们发送给图灵机器人的 post 是微博数据，所以爬下来的对话数据也要使用微博数据处理脚本进行处理。