中文微博观点要素抽取评测大纲

1. 评测对象

本项评测任务的对象是面向中文微博观点句的情感要素抽取。

2. 任务设置

本任务要求找出微博中每条观点句（观点句已给出，不需要自动识别，同时非观点句保留）作者的评价对象，即情感对象。每条观点句可能含有若干情感对象，同时判断针对情感对象的观点极性。

注：

1. 只对微博中的观点句进行情感对象的抽取。
2. 情感对象应首先从当前句子中抽取，如果情感对象不存在于当前句子中，再从整条微博中抽取。对于第二种情况，应优先从当前句子的前一句（包括句子中包含的hashtag）开始依次向前寻找情感对象，如果没有再从当前句子的后一句（包括句子中包含的 hashtag）开始依次往后寻找。对于那些整条微博中都没有出现的情感对象(有些情况情感对象是隐含的)的观点句，参赛队伍不必进行抽取。
3. 一个句子中，可以出现多个情感对象，应抽取出每个情感片段所对应的情感对象。“你根本已经不是个人了，你比蛇还冷血，你比畜生还畜生。”，要求抽取出三个“你”。
4. 抽取情感对象时，要求抽取出尽可能完整和明确的对象，例如“ipad 的屏幕很棒！”，要求抽取出情感对象“ipad 的屏幕”，而不仅是“屏幕”。
5. 对于人称代词（你，我，他，它，你们，我们，他们，它们等）单独作为情感对象出现时，需要在该微博范围内（不包括转发、评论信息）尽量进行指代消解（无法指代消解的情况可以采用这些代词作为对象）。例如，“小明就读于北京大学，他是名优秀的学生。”情感对象是“小明”而不能是“他”。

提交格式

id run-tag weibo-id sentence-id target begin-offset end-offset polarity

说明

id：结果序号

run-tag：队伍结果标识

weibo-id：微博 id

sentence-id：句子 id

target：情感对象

begin-offset: 情感对象在整条微博中的起始位置

end-offset：情感对象在整条微博中的终止位置

polarity：对情感对象的观点极性，POS 代表正面，NEG 代表负面，OTHER 代表中性或者无法明确归为正面或者负面的其它情形。

注：

1. run-tag 的格式为“队伍标识_提交结果组号”，队伍标识可自定，组号用于区分同一队伍的多组提交结果。不同字段之间用\t 隔开。
2. 对于从当前某个句子中抽取得到的情感对象，其起始位置与终止位置也要基于整条微博来计算。

例如如下两条微博：

weibo1：

```
<weibo id=”1”>
<sentence id=”1” opinionated="N">渭南城管撕春联事件在成都公交车上的分众传媒广泛报道！</sentence >
  <sentence id=”2” opinionated="Y">渭南城管真变态啊！</sentence >
</weibo>
```

weibo2:

```
<weibo id=”2”>
  <sentence id=”1” opinionated="Y">#iPad3#这么麻烦的东西怎么还有那么多人在用，又是越狱又是破解。</sentence>
  <sentence id=”2” opinionated="N">顺便问一下怎么越狱啊？</sentence>
</weibo>
```

weibo1 中有两个句子，第一句是非观点句，第二句是观点句。weibo2 中有两个句子，

其中第一句是观点句，第二句是非观点句。此任务只需抽取观点句（*opinionated="Y"*）中的情感要素，正确的输出结果为：

1	xyz	1	2	渭南城管	26	29	NEG
2	xyz	2	1	iPad3	1	5	NEG

文件采用 unicode（utf-16）编码，每个字符都占两个字节，任意微博中第一个字符的 offset 为 0，第二个字符的 offset 为 1，以此类推。比如：weibo1 第二句开始位置的“*渭南城管*”这四个字符在整条微博中对应的 offset 分别为 26,27,28,29。评价情感对象时只以 begin-offset 和 end-offset 作为判断依据，target 不参与评价。

评价标准：

本任务采用精确（Strict）评价和宽松（Lenient）评价两种方式，使用准确率（Precision）、召回率（Recall）以及 F 值（F-measure）作为评价标准。

在精确评价中，要求提交的情感对象的 offset 和答案完全相同并且情感对象极性也相同时才算正确。

$$\text{Precision} = \frac{\#system\_correct}{\#system\_proposed}$$

$$\text{Recall} = \frac{\#system\_correct}{\#\text{gold}}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

#gold 是人工标注结果中情感对象的数目，#system_correct 是提交结果中与人工标注匹配的数目，#system_proposed 是提交的情感对象的数目。

在宽松评价中，一个结果包含 4 个参与评测的元素：句子微博 id，句子 id，情感对象区间（由起始位置和终止位置构成）和极性，即 r=(wid, sid, s, p)。我们首先定义两个结果之间的覆盖率 c：

$$c(r, r') = \begin{cases} \dfrac{|s \cap s'|}{|s'|} & if \;\; p = p' \;\&\; wid = wid' \;\&\; sid = sid' \\ 0 & else \end{cases}$$

其中 s 和 s’为两个结果 r 和 r’中情感对象的区间，p 和 p’为对应的极性，wid 和 wid’为微博 id，sid 和 sid’为句子 id。|*| 表示计算区间的长度。

两个结果集合 R 和 R’之间的覆盖率 C 定义为：

$$C\left(R,R'\right)=\sum_{r_i\in R}\sum_{r_j'\in R'}c(r_i,r_j')$$

假设提交的结果集合为 R′，标注结果集合为 R，则精度、召回率和 F 值为：

$$\text{Precision}=\frac{C(R,R')}{\left|R'\right|}$$

$$\text{Recall}=\frac{C(R',R)}{\left|R\right|}$$

$$\text{F-measure}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

其中|*|表示计算集合中元素的个数。

3 评测数据集

本次评测数据来自新浪微博。评测数据全集包括 20 个话题，每个话题采集大约 1000 条微博，共约 20000 条微博。数据采用 xml 格式，已经预先切分好句子。

所有数据文件均采用 unicode（utf-16）编码。读取数据集时，需要进行 XML 实体和字符间的转换，主要包含以下 5 组：(<，<)，(>，>)，(&，&)，('，')，("，")。如果不做转换，会影响情感要素抽取任务中情感对象的位置区间。

4 评测要求

参评单位应当采用自动的方法，针对微博进行情感分析。参评系统应当预先训练模型、调整好所有参数，运行过程中不得有人工干预。本次评测不限制使用各种语义资源。参评单位至多提交 2 组结果。

本次评测为离线评测。参评单位自行处理数据，生成相应结果后提交。答案采用人工标注的方法确定。参赛单位需要处理全部评测数据，但用于实际评测的人工标注数据仅为评测数据全集的 10%左右。