NLP&CC2013 中文微博情绪分析评测大纲

1. 评测对象

本次评测面向中文微博中的情绪分析，对整条微博的情绪判别和分类，以及情绪句的识别和分类的算法和技术进行评测。

2. 任务设置

本次评测设置了如下两个任务：

- 1) 微博情绪识别与分类（必选任务）
- 2) 情绪句识别和分类（可选任务）

2.1 微博情绪判别及分类（必选任务）

对于输入的整条微博，本任务要求判断出该微博是否包含情绪。对包含情绪的微博，要求判别其情绪分类输出为 **anger** 愤怒、**disgust** 厌恶、**fear** 恐惧、**happiness** 高兴、**like** 喜好、**sadness** 悲伤、**surprise** 惊讶中的一种，即单分类输出。需要注意的是，一条微博中可能包含多个个体的不同情绪，情绪分类应以博主的主要情绪为准。评测使用的情绪分类体系基本沿用大连理工建立的情感词汇本体。（详见徐琳宏，林鸿飞，潘宇，任惠，陈建美. 情感词汇本体的构造. 情报学报, 27(2):180-185, 2008.）

输入数据格式为 xml，例如：

```
<weibo id="1">
  <sentence id="1">难道是花样男子? </sentence>
  <sentence id="2">只有一小块屏幕还是看得很开心</sentence>
</weibo>
```

提交格式为

id: 评测任务序号 微博情绪分类为 1

team-tag: 队伍结果标识

run-tag: 提交结果组号（同一队伍可以提交至多 4 组结果，其中 Open/Close 测试分别提交至多 2 组结果）

run-type: Close 测试标记为 C，Open 测试标记为 O，详细说明见第 4 节评测方法部分。

weibo-id: 微博 id

emotion-tag: 微博情绪标识，包含情绪标记为 Y，无情绪标记为 N

emotion-type: 博主主体情绪的类型: anger、disgust、fear、happiness、like、sadness、surprise 中的一种。无情绪标记为 none.

针对上例，提交结果如下

```
1 team_name 1 C 1 Y happiness
```

评价标准:

微博情绪判别任务使用正确率 (Precision)、召回率 (Recall) 和 F 值 (F-measure) 来评价各个队伍对于这一子任务的识别结果, 其计算方法如下:

$$\text{Precision} = \frac{\#system_correct(emotion = Y)}{\#system_proposed(emotion = Y)}$$
$$\text{Recall} = \frac{\#system_correct(emotion = Y)}{\#gold(emotion = Y)}$$
$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

其中#gold 是人工标注结果的数目, #system_correct 是提交结果中与人工标注匹配的数目, #system_proposed 是提交结果的数目。

在微博情绪分类任务中, 使用宏平均 (Macro) 和微平均 (Micro) 的准确率 (Precision)、召回率 (Recall)、F 值 (F-measure) 作为评价指标, 具体计算方法如下:

$$\text{Macro_Precision} = \frac{1}{6} \sum_i \frac{\#system_correct(emotion = i)}{\#system_proposed(emotion = i)}$$
$$\text{Macro_Recall} = \frac{1}{6} \sum_i \frac{\#system_correct(emotion = i)}{\#gold(emotion = i)}$$
$$\text{Macro_F-measure} = \frac{2 \times \text{Macro_Precision} \times \text{Macro_Recall}}{\text{Macro_Precision} + \text{Macro_Recall}}$$
$$\text{Micro_Precision} = \frac{\sum_i \#system_correct(emotion = i)}{\sum_i \#system_proposed(emotion = i)}$$
$$\text{Micro_Recall} = \frac{\sum_i \#system_correct(emotion = i)}{\sum_i \#gold(emotion = i)}$$
$$\text{Micro_F-measure} = \frac{2 \times \text{Micro_Precision} \times \text{Micro_Recall}}{\text{Micro_Precision} + \text{Micro_Recall}}$$

其中#gold 是人工标注结果的数目, #system_correct 是提交结果中与人工标注匹配的数目, #system_proposed 是提交结果的数目。i 为七类基本情绪之一。

2.2 情绪句识别与分类 (可选任务)

针对微博文本中的每一个句子, 本任务要求判断出该句是情绪句还是非情绪句并判别其情绪类别。由于人的情绪表达往往是复杂的, 一段情绪表达的文本经常含有多种情绪, 因此本次人工标注对于待标注句子采取了 top-2 情绪标注的策略, 即利用人工标注的方式标注微博每个句子的前两种主要情绪, 其中 emotion_1 为主要情绪, emotion_2 为次要情绪。如果此句子只有一种情绪, 则只标注 emotion_1。

在此体系下, 情绪分类的任务包括情绪句 top_2 情绪的识别, 即参赛队伍在识别出情绪句的主要情绪 (emotion_1) 的前提下, 进一步识别出情绪句中可能的次要情绪 (emotion_2)。对应的, 自动分析结果提交格式为:

id: 评测任务序号 情绪句情绪识别与分类任务为 2

team-tag: 队伍结果标识

run-tag: 提交结果组号 (同一队伍可以提交至多 4 组结果, 其中 Open/Close 测试分别提交

至多 2 组结果)

run-type: Close 测试标记为 C, Open 测试标记为 O, 详细说明第 4 节评测方法部分。

weibo-id: 微博 id

sentence-id: 句子 id

emotion-tag: 情绪句标识, 是情绪句为 Y, 非情绪句为 N

emotion-1-type: 该句子主要情绪的类型: anger 愤怒、disgust 厌恶、fear 恐惧、happiness 高兴、like 喜好、sadness 悲伤、surprise 惊讶中的一种。如无主要情绪标注为 none。

emotion-2-type: 该句子次要情绪的类型: anger 愤怒、disgust 厌恶、fear 恐惧、happiness 高兴、like 喜好、sadness 悲伤、surprise 惊讶中的一种。如无次要情绪标注为 none。

例如对下面的句子:

<weibo id="2">

<sentence id="1">厦门午后的阳光很好, 海边的风很舒服, 可是即将要离去, 有点淡淡的忧伤, 但还是很开心。</sentence >

</weibo>

此句中主要情绪是幸福和喜悦、次要情绪是悲伤。则提交结果为:

2 team_name 1 C 2 1 Y happiness sadness

评价标准:

情绪句识别任务的评价标准基本与情绪微博识别任务标准一致。使用正确率 (Precision)、召回率 (Recall) 和 F 值 (F-measure) 来评价。

情绪句情绪分类任务为多标签分类问题, 因此采用多标签分类平均精度(Average Precision)作为评价指标。具体的计算方法如下:

$$\text{AveragePrecision} = \frac{1}{2n} \sum_{i=1}^n \sum_{k=1}^2 \frac{|(\text{emotion}_k \in Y_i | \text{rank}(x_i, \text{emotion}_k) \leq \text{rank}(x_i, y))|}{\text{rank}(x_i, \text{emotion}_k)}$$

其中, $\text{rank}(x_i, y)$ 表示 x_i 句子情绪识别结果中 y 情绪的位置。

3. 评测数据集

本次评测数据来自新浪微博。评测数据全集共约 40000 条微博。数据采用 xml 格式, 已经预先切分好句子。数据格式为 xml 格式, 编码为 Unicode (utf-16)。

4. 评测方法

本次评测为离线评测。参赛单位应当采用自动的方法, 针对微博进行分析。参赛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。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 20% 左右。

为了对各参赛单位的算法及资源、知识库等各方面因素对分析精度影响进行比较客观的评价和分析。本次评测分别进行 Close 封闭资源测试和 Open 开放资源测试。其中 Close 封闭资源测试要求各参赛单位只允许使用下列资源建立和训练系统: 1. 训练数据严格只限使用评测单位提供的测试样例; 2. 情绪词表只限使用大连理工大学开发的情绪本体库 (可免费获得) 和哈尔滨工业大学开发的情绪表达常识库 (建设中, 将免费提供) 3. 句法分析限使用 Stanford Parser 4. 词汇语义知识库限使用同义词词林和 HowNet. 分词、词性标注系统不限制。

Open 测试则允许参评单位利用任何现有语言资源开发和训练系统，并用于测试结果生成。评测将分别评估 **Close** 和 **Open** 测试所取得的系统性能。