摘要 本文是 BeSt 评测的综述。该评测在知识库中现有知识库对象之间添加信念(belief)和情感(sentiment)的准确性。

1 引言

本评测有以下特点：

- 关注源(sourse)，态度，目标(target)。
- 对态度的触发词或语言上的标记不 感兴趣，只关注态度检测本身。
- 包括信念和情感。
- 源：实体，类型为：**人，地缘政治实体，组织**(Person, GeoPolitical Entity (GPE), or Organization)；目标：关系，事件；对于情感，可以是实体。
- 两种情况：
  - 提供黄金实体、关系、事件 。（**EREs**）
  - 提供预测的 EREs，来自系统的结合。
  - 两种情况下，ERE 都包含实体提及(entity **mentions**)和事件提及(event mentions)的文档内共引用(in-document co-reference。(不太理解，根据下文，是指给出实体与对应的实体提及之间的对应关系吗？)
- 评测在英文文本上进行，另外还有中文和西班牙文。

与 13、14 年相比的扩展：

- 情感基础上加入信念。
- 目标可以是实体、关系、事件，不只是实体。
- 在 ERE 标注基础上标注，不需要识别可能的源和目标，以及确定共现。
- 自动评测，使用一个类似训练数据的标记好的黄金标准。

2 数据和基准

- 数据集：

LDC，英文，中文，西班牙文；可以使用其他数据集。论坛数据和新闻。
注意区别对象与对象提及(object mentions)。
细节见 ERE，给提及按对象分组，提供文档内共引用标注。
术语：relation vs relation mentions, **hopper** vs event mentions

- 基准：
  - 情感：全部"neg"。
  - 信念：全部“CB”。
  - 对可能目标的态度：信念：针对每个关系和事件；感情：针对每个实体、关系、事件。

- 用触发(triggers)识别目标提及，如果没有触发，则通过提及的论据(argument)识别。
- 源：全部假定是作者，选择合适的提到作者之处作为作者提及；没有作者提及的则源为 None。

3 任务的概念性描述

(1) **private state** tuples(**PSTs**)，四元组：
(source-entity, target-object, value, provenance-list)

- 值(value)：
  - 情感：(positive, negative)
  - 信念：(CB, NCB, ROB)，CB （确信，包括未来事件），NCB（不确信），ROB（本人态度不明朗）
- provenance 实例：target mention ID+文件名。
- (source-entity, target-object)可能多次出现，两个原因：1)可能多种情感信念，2)ERE 只记录文档内共引用，多文档则可能多次出现相同的源、目标、个人态度。

(2) 任务描述：

- 输入：源文本文件，ere.xml 文件。 ere.xml 文件列出实体提及，关系提及，事件提及，以及它们的文档内共引用。
- 输出：best.xml 文件，引用 ere.xml 文件。其中所有提及(mentions)都出现在 ere.xml 文件中。

(3) 评测参数：

1. 2 个态度：信念，情感
2. 3 种语言：英语，汉语，西班牙语
3. 2 种体裁：论坛(DF)，新闻(NW)
4. 2 种标注情况：黄金 ere.xml，系统自动输出的 ere.xml

则共 24 个任务。

4 评价指标

评分基于 4 元组 PST，赋予 partial credit。
provenance-list 两种评价方式：所有 provenance 实例都要找到，找到 1 个正确 provenance。

详细描述：
1.给定预测的 best.xml 和黄金 best.xml，先都转换为 4 元组：
对于每个 (source-entity, target-object)对，在 best.xml 中查找：

- 4 元组中前 3 个未出现，则创建新 4 元组
- 出现，则把 target-mention 插入 provenance-list

2.首先评价前 3 个域，与黄金 4 元组匹配，按 match-type 顺序等：
(a)未匹配，false positive
(b)都匹配，true positive，分数 1
(c)源、目标、态度类型匹配， true positive， 分数 2/3
(d)值和目标匹配， true positive，分数 2/3
(e)目标和态度类型匹配， true positive，分数 1/3
(f)未匹配以上 4 条的也为 false positive
3.对于 true positive，检查 provenance-list，是否有相同的 target-mention ：
(a)全部 provenance 的情况，召回率-正确率匹配，得到 f 值，对原分数伸缩；若是 0 则变 false positive
(b)单个 provenance 的情况：有一个匹配则 true positive，否则 true positive

最终指标：f-measure
注： 宏平均，微平均的使用

best.xml 中今年没有使用的部分：（结合赛题解释看）

- 信念的极性
- 讽刺
- 对于事件论据的信念
- 信念 NA 值的实例

5 参加人员

4 支队伍

- REDESB：
  - 关注信念分类
  - 贝叶斯方法，各种特征
- CUBISM：
  - 情感、信念分开两个系统
  - 情感：ACA
  - 信念：图
- Columbia_GWU:
  - 提交了全部 24 种情况
  - 情感：2 种方法
    - 第一种假定源是作者，随机森林，词汇特征等
    - 类似于关系抽取
  - 信念：假定源是目标，word-tagger
- CORNPITTMICH:
  - 规则+基于机器学习的方法

- 英：3 个系统，都有 2 个阶段：1)链接预测；2)情感和信念分类
- 中：
  - 情感分类：混合方法，模型(a)LSTM，Weibo 数据,(b)寻找源,(c)规则模型，基于(a)得到最后输出，3 个提交只是(c)的参数不同
  - 信念分类：3 个相同；论坛：规则模型和线性模型；新闻：规则模型

6 结果

见原文 4 张表，给出各队的信念和情感分类的比较与结果。
belief 的基准很高，因为大多数都是 CB。

7 讨论

- **情况数量多** 24 种，需研究影响因素
- **预测的 ERE** 系统自动预测的 ERE 上没有单独的黄金信念和情感标注，需要进行自动预测的 ERE 到黄金 ERE 上的映射。
- **数据问题** 情感不同，尤其是中文情感少

8 结论

略