

## SemEval\_2015\_Task\_3 实验报告

---

本实验在 SemEval\_2015\_Task\_3<sup>1</sup>数据集上使用规则提取特征并使用 SVM 和随机森林两种方法在提取特征的基础上进行分类。

- 特征提取

本实验参照<sup>ii</sup>文中的思路提取特征。

在 A 任务中，每个 question-comment pair 有 100 维特征。其中关于问题（QBody 字段）与回答（CBody 字段）本身各有 25 维特征，见表 1，在提取这些特征之前先去掉了文中的所有链接，但没有去掉 html 标签。question-comment pair 整体还有 50 个特征，见表 2。

特征描述	形式	维数
标签数（以<>数量中较少的估计）	计数	1
最长单词长度	计数	1
平均单词长度	数值	1
单词数量	计数	1
句子数量（以问号、句号、叹号总数估计）	计数	1
平均每句单词数量	数值	1
大写单词数量	计数	1
命名实体数量	计数	1
no 的数量	计数	1
yes 的数量	计数	1
thank, thanks 的数量	计数	1
please 的数量	计数	1
may, might, could, can, would, will 的数量	计数	1
问号的数量	计数	1
叹号的数量	计数	1
名词、动词、代词、疑问词、外来词的数量	计数	5

---

<sup>1</sup> <http://alt.qcri.org/semeval2015/task3/>

名词、动词、代词、疑问词、外来词的频率	比率	5
---------------------	----	---

表 1 问题和回答文本分别统计的特征

特征描述	形式	维数
问题和回答中的图数和链接数	计数	4
回答是否是第一个或最后一个	布尔	2
当前及前后回答的 USERID 是否和提问者相同	布尔	3
问题是否为是否类问题	布尔	1
问题类型 ( QCATEGORY )	one-hot	27
回答的 title 是否问 “Re : ” + 问题 title	布尔	1
问题和回答相同的 uni bi tri gram 数	计数	3
上数相对于问题长度和回答长度的比值	比率	6
问题和回答的 uni bi tri gram 的 one-hot 向量余弦值	数值	3

表 2 问题和回答联合统计的特征

由于官网的评价脚本中，A 任务只考虑 Good、bad 和 Potential 三个类别，所以在标记标签时，尝试了两种标签标记方式：A-full 是标全 6 类的。而 A-sim 是只标记 3 类的。

在 B 任务中使用的特征和 A 任务中相同，不过是以题为单位。故每道题的特征向量是 A-sim 标记下 svm 训练结果中标记为 good 的选项的特征向量的均值。

### ● 实验设置

在具体实现时，使用的工具 libsvm<sup>2</sup>和 xgboost<sup>3</sup>。分词、词性标注和命名实体识别等使用 nltk<sup>4</sup>。

简单调参后，两个任务均使用线性核 SVM，task-A 的两种标注的参数-c 均为 0.5，task-B 的参数-c 为 1。而在随机森林时，两个任务均使用 multi:softmax 作为目标函数，最大树深为 5，学习率 eta 为 0.2，训练轮数为 50 轮。

评价时，task-A 和官网上评价脚本相同，只考虑 3 类。task-A 和 task-B 的评价指标均为 macro-f1 和 accuracy。

### ● 实验结果

<sup>2</sup> <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

<sup>3</sup> <http://xgboost.readthedocs.io/en/latest/>

<sup>4</sup> <http://nltk.org>

Task-A 在验证集和测试集上的实验结果如表 3 所示，Task-B 在验证集和测试集上的实验结果如表 4 所示。

	macro-F1	Acc	F1-Good	F1-Pot.	F1-Bad
A-full-svm-dev	47.39%	67.42%	76.05%	0.00%	61.66%
A-full-svm-test	49.41%	70.34%	76.70%	0.00%	67.99%
A-Sim-svm-dev	48.46%	69.36%	76.60%	0.00%	68.49%
A-Sim-svm-test	<b>50.56%</b>	<b>72.77%</b>	77.86%	0.00%	73.42%
A-full-rf-dev	48.51%	68.94%	77.12%	0.00%	64.96%
A-full-rf-test	49.57%	70.70%	77.08%	0.00%	68.49%
A-Sim-rf-dev	48.77%	69.85%	77.40%	0.00%	68.79%
A-Sim-rf-test	<b>50.24%</b>	<b>72.27%</b>	77.16%	0.00%	73.17%

表 3 Task-A 上的实验结果

	macro-F1	Acc	F1-Yes	F1-Unsure	F1-No
B-svm-dev	54.32%	55.88%	66.67%	40.00%	50.00%
B-svm-test	<b>43.39%</b>	<b>51.72%</b>	62.86%	46.15%	20.00%
B-rf-dev	52.77%	55.88%	68.75%	45.45%	42.86%
B-rf-test	<b>57.50%</b>	<b>68.97%</b>	82.35%	58.82%	28.57%

表 4 Task-B 上的实验结果

### ● 一点分析

由于本实验采用的机器学习方法并没有针对 macro-F1 做特殊地优化，在 Task-A 上，小类 Pot.的效果很差。导致了本实验 Accuracy 看起来还不错但 macro-F1 明显偏低。

Xgboost 在 TaskB 上的效果莫名地好。不过可能也是由于 Task-B 的数据集较小导致的意外吧。

但是网上下载的数据集的测试集规模和 SemEval-2015 评测报告中写的不一样。

### ● 横向对比

我们将我们的结果与一些相关文章进行对比，Task-A 与 Task-B 的结果分别如表 5、表 6 所示。其中，Hou' s 为参考文献[2]的方法，也是我这里主要参考的方法；Quan' s 为参考文献[3]<sup>iii</sup>的方法，也是当时评测比赛时 Task-A 效果最好的方法（他们当时没做 Task-B）。Belinkov' s 为参考文献[4]<sup>iv</sup>的方法，也是当时评测比赛时 Task-B 效果最好的方法。

	macro-F1	Acc	F1-Good	F1-Pot.	F1-Bad
Task-A-SVM	50.56%	<b>72.77%</b>	77.86%	0.00%	73.42%
Task-A-RF	50.24%	72.27%	77.16%	0.00%	73.17%
Hou' s	56.44%	69.43%	78.87%	<b>17.86%</b>	72.58%
Quan' s	<b>57.29%</b>	72.52%	<b>78.96%</b>	14.36%	<b>78.24%</b>
Belinkov' s	49.54%	70.45%	----	----	----

表 5 Task-A 上的横向比较

	macro-F1	Acc	F1-Yes	F1-Unsure	F1-No
Task-B-SVM	43.39%	51.72%	62.86%	46.15%	20.00%
Task-B-RF	57.50%	68.97%	<b>82.35%</b>	<b>58.82%</b>	28.57%
Hou' s	53.60%	64.00%	80.00%	44.44%	<b>36.36%</b>
Belinkov' s	<b>63.70%</b>	<b>72.00%</b>	----	----	----

表 6 Task-B 上的横向比较

从上面的比较可以看出，我们的(SVM 与随机森林)方法在 Task-B 与 Task-A Accuracy 上超过了这两种方法选取特征时主要参考的工作，但是距离最好的工作还略有差距。

GitHub 地址：

[https://github.com/Erutan-pku/SemEval\\_2015\\_Task\\_3](https://github.com/Erutan-pku/SemEval_2015_Task_3)

<sup>i</sup> Nakov, Preslav, et al. "SemEval-2015 Task 3: Answer Selection in Community Question Answering." *International Workshop on Semantic Evaluation* 2015:269-281.

<sup>ii</sup> Hou, Yongshuai, et al. "HITSZ-ICRC: Exploiting Classification Approach for Answer Selection in Community Question Answering." *International Workshop on Semantic Evaluation* 2015:196-202.

<sup>iii</sup> Quan, Hung Tran, et al. "JAIST: Combining multiple features for Answer Selection in Community Question Answering." *International Workshop on Semantic Evaluation* 2015:215-219.

<sup>iv</sup> Belinkov, Yonatan, et al. "VectorSLU: A Continuous Word Vector Approach to Answer Selection in Community Question Answering Systems." *International Workshop on Semantic Evaluation* 2015.