SemEval-2015 Task 3:
Answer Selection in Community Question Answering

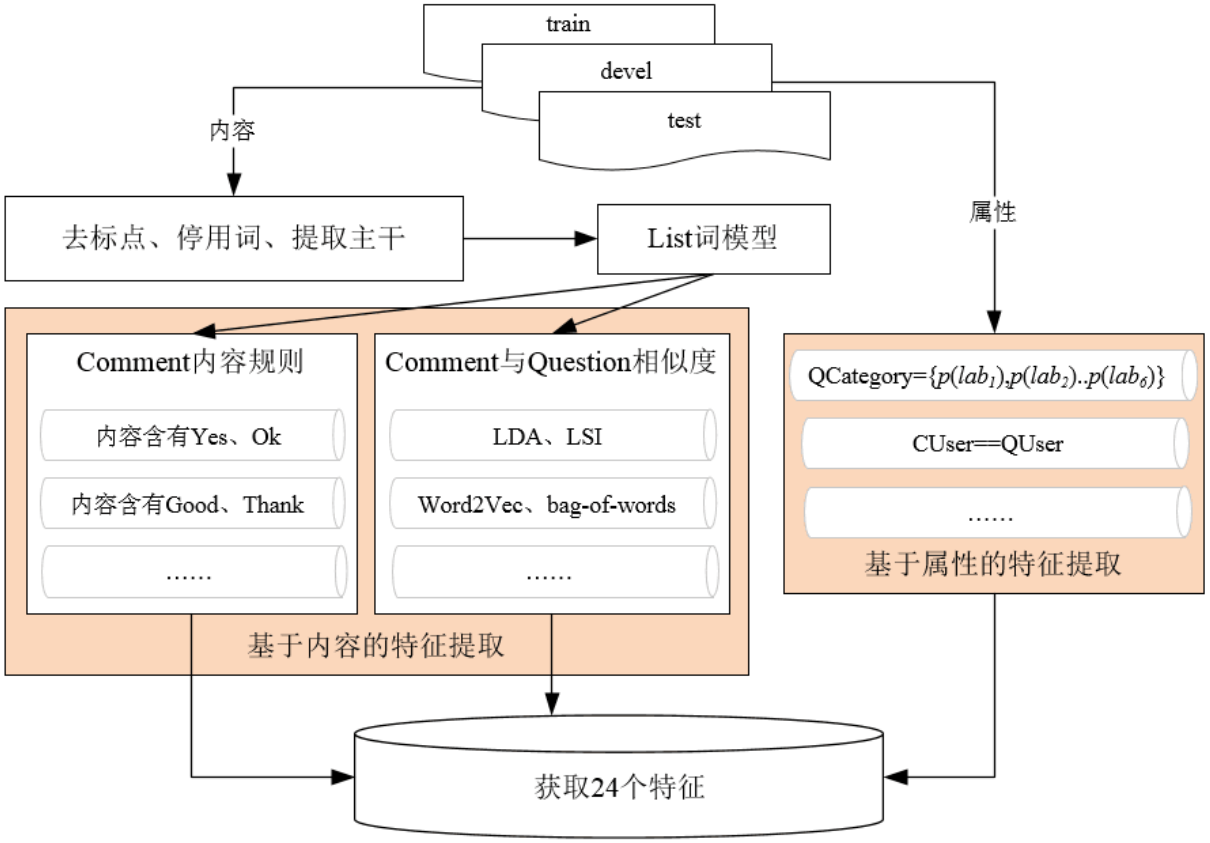
组员：李宏强、王东升

概述:

本实验针对SemEval_2015_Task_3中English的数据集，完成TaskA任务和TaskB任务。本文首先从数据集中提取每个Question和Comment的属性信息和内容信息。分别采用基于规则和基于文本相似度的方式提取了24个特征，使用SVM，GDBT，RandomForest三种分类器进行分类。

GitHub: https://github.com/pku601/NLP_QA

特征提取:



属性信息包括：QID，QCATEGORY，QUSERID，QTYPE，QGOLD_YN，CID，CUSERID，CGOLD，CGOLD_YN。

内容信息包括：QSubject，QBody，CSubject，CBody。

属性信息特征提取:

基于属性信息提取的特征

属性	描述	维度	取值
cuserEqualQuser	Qusertion用户是否等于Comment用户	1	Bool
qCategoryProbility	Question对应的所有Comment的CGOLD	6	List

内容信息特征提取:

基于相似度方式提取的特征

属性	描述	维度	取值
LdaSimilarity	Question内容与Comment内容LDA相似度	1	Float
LsiSimilarity	Question内容与Comment内容LSI相似度	1	Float
BowsSimilarity	Question内容与Comment内容Bows相似度	1	Float
Word2VecSimilarity	Question内容与Comment内容Word2Vec相似度	1	Float

内容信息特征提取:

文本基于规则方式提取的特征

属性	描述	维度	取值
hasURL	内容是否含有URL	1	Bool
hasEmail	内容是否含有Email	1	Bool
hasYes	内容含有Yes的个数	1	Int
hasNo	内容含有No的个数	1	Int
hasSure	内容含有Sure的个数	1	Int
hasCan	内容含有Can的个数	1	Int
hasNeither	内容含有Neither的个数	1	Int
hasGood	内容含有Good的个数	1	Int
hasSorry	内容含有Sorry的个数	1	Int
hasOk	内容含有Ok，Okay的个数	1	Int
HasThank	内容含有Thank，Thanks的个数	1	Int
startWithYes	内容是否Yes开始	1	Bool
wordNums	内容含有的单词数	1	Int

标签制定及建模:

对于TaskA任务，标签的数量为6，标签编号从0到5，分别对应Good，Bad，Potential，Dialogue，Not English，Other。

对于TaskB任务，只提取了QTYPE为YES_NO的样本，并且尝试了以下两种方式:

1. 对Comment进行建模
2. 对Question进行建模

标签制定及建模:

1. 对Comment进行建模，标记Comment标签数量为3，编号从0到2，分别对应Good_Yes，Good_No，Unsure。其中Unsure包括Good，Unsure，Bad，Potential，Dialogue，Not English，Other。通过模型对Comment的标签进行预测，最后通过判断Good_Yes和Good_No的数量。如果Good_Yes数量较多，则这些Comment对应的Question的标签为Yes；如果Good_No数量较多，则这些Comment对应的Question的标签为No；如果相等，Question的标签则为Unsure。
2. 对Question进行建模，标记Qustion的标签数量为3，编号从0到2，分别对应Yes，No，Unsure。Question的属性为Quesiton的所有Comment的属性的平均值。通过实验，采用第一种方式得到的MACRO-averaged F1较低，约为32%，于是本文决定采用第二种方式处理TaskB。

模型训练:

本文采用了三种分类模型，分别为：

- SVM
- GBDT
- RandomForest

实验结果:

Task A (without Dialog)

Model	Dataset	Macro Precision	Macro Recall	Macro F1	Accuracy
SVM	Devel	45.29%	46.70%	44.51%	65.71%
	Test	44.84%	45.43%	**43.35%**	**64.32%**
GBDT	Devel	78.56%	49.26%	47%	67.72%
	Test	45.9%	48.84%	**46.9%**	**68.12%**
Random forest	Devel	48.89%	48.10%	46.42%	64.92%
	Test	48.93%	47.96%	**46.74%**	**65.89%**
Baseline	Test			**22.36%**	**50.46%**
Rank one	Test			**57.29%**	**72.67%**

实验结果:

Task B

Model	Dataset	Macro Precision	Macro Recall	Macro F1	Accuracy
SVM	Devel	55.08%	42.92%	42.97%	44.12%
	Test	48.20%	46.69%	**47.23%**	**58.62%**
GBDT	Devel	53.71%	50%	50.35%	50%
	Test	59.25%	62.7%	**59.14%**	**65.59%**
Random forest	Devel	57.94%	55.42%	55.71%	55.88%
	Test	53.38%	53.31%	**52.86%**	**65.52%**
Baseline	Test			**25.0%**	**60%**
Rank one	Test			**63.7%**	**72%**

实验总结分析:

- 总结
  - 从模型上看，相比于SVM和RandomFores ，GBDT的效果比较好。
  - 从指标Macro F1和Accuracy看，本实验的最好效果比Baseline效果好，与Rank one相比还有较大差距。
- 改进方案:
  - 本实验只是用了传统的机器学习方法，没有采用深度学习算法。后期可以采用深度网络模型，如CNN进行实验。
  - 本实验只随机选取了模型的一些超参数，没有设置细粒度的分析超参数的选择方案。后期可以对超参数进行细粒度调整。
  - 本实验只选择了24个特征，后期可以添加更多特征进行实验。

参考文献:

1. AlessandroMoschitti, PreslavNakov LluısMarquez WalidMagdy, James Glass, and Bilal Randeree. "Semeval-2015 task 3: Answer selection in community question answering." SemEval-2015 269 (2015).
2. Barrón-Cedeno, Alberto, et al. "Thread-Level Information for Comment Classification in Community Question Answering." ACL (2). 2015.

THANKS