SemEval 2015 Task 3

Answer Selection in Community Question Answering

Hongqiang Li

Peking University, Beijing, China

hongqiang.li@pku.edu.cn

Dongsheng Wang

Peking University, Beijing, China

wangdsh@pku.edu.cn

目录

1 介绍和相关工作

Web 论坛中的社区问答是经典问答的一个演变，在论坛上用户可以相互交流，提问和回答没有太多的限制。这是一种强大的机制，它允许用户自由地问问题并期待得到好的，诚实的回答。

美中不足是一个用户必须浏览所有可能的答案并去理解它们。通常，许多的回答和实际问的问题相关性不强，有些甚至转移了主题。这在比较长的回答中经常出现，随着回答的进行，用户开始互相讨论，而不是回答最初的问题。

选择相关文本段落（即含有良好的答案）的问题已经在问答搜索中得到解决，比如非事实型问答还是段落重新排序问题。通常，搜索引擎应用自动分类的搜索结果页面，导出相对的排序。具体见(Radlinski and Joachims, 2005; Jeon et al., 2005; Shen and Lapata, 2007; Moschitti etal., 2007; Surdeanu et al., 2008; Heilman and Smith, 2010; Wang and Manning, 2010; Severyn and Moschitti, 2012; Yao et al., 2013; Severyn et al., 2013; Severyn and Moschitti, 2013)。

本文针对 CQA 标注问题，采用了以下两种思路：

1.采用传统 QA 基于相似度排序的方式，充分利用问题的答案的信息，寻找问题和答案之间的相似度。我们认为如果问题和答案所说的内容是相似的，那么答案更可信。如果问题和答案都是同一个用户提出的，那么这个问题和答案之间的相似度一定很高。

2.采用基于答案的文本描述信息判断答案是否是合理的回答。我们认为，如果答案的内容比较长，那么答案更可能是一个合理的回答。如果答案中出现 Yes，No 等一些用来描述回答的词，那么这个回答更可能是一个合理的回答。

为了验证我们的想法，本文在 SemEval_2015_Task_3 中 English 的数据集上，实验了 Task 3 的 SubtaskA 任务和 SubtaskB 任务。数据集中的问题分为两类，一类是 General 为题，另一类是 Yes_No 问题。Task3 中的 A 任务针对所有问题，要求给定问题和答案的属性信息和文本信息，要求预测答案的类型，Good，Bad，potential，或者 Dialog。B 任务则是要求预测 Yes_No 问题的标签，Yes，No，或者 Unsure。

本文组织方式如下：第二部分介绍了本文实验的方法，包括特征的选择，标签的制定，模型的构建。第三部分介绍了本文的实验数据以及结果。第四部分对实验结果进行了分析和讨论。第五部分总结了本文的内容，并对探索了未来的工作。

2 方法

本文针对 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(Gradient Boosting Decision Tree)，RandomForest 三种分类器进行分类。

2.1 任务定义

子任务 A：给一个问题（包括短标题和扩展描述）及其回答，将每个回答分为下文其中一类：

a) 绝对相关的（good）
b) 潜在有用的（potential）
c) 不好的或不相关的（bad, dialog, non-English, other）

子任务 B：给一个 YES/NO 类型的问题（包括短标题和扩展描述）和一些回答，基于 Good 的回答判断一下对于整个问题的回答应该是 yes，no 还是不确定。任务 B 只针对英语数据集。

2.2 特征提取

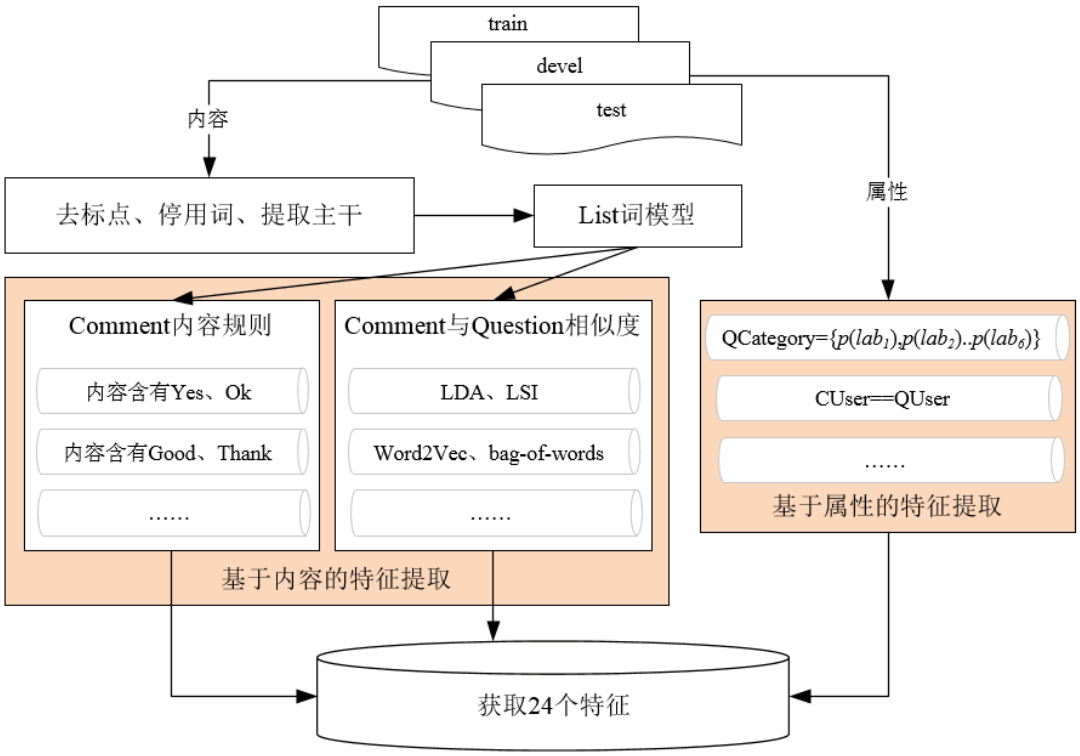

图 1 特征提取流程图

Task3 任务的数据包括三部分，分别为 train, devel 和 test，具体描述见第三部分。本文首先提取了 train，devel 和 test 原始数据的属性信息和内容信息。其中，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。数据详细内容见图 2。

2.2.1 基于内容相似度的特征

我们认为内容信息中，有实际意义的词才能够表示整个句子的信息，并且相同的词可能有不同的表现形式，如单复数，不同时态的情况。因此，对于内容信息，我们执行去标点，去停用词，提取主干的操作，得到词列表。为了试验本文提出的第一种方法，通过文本相似度来判断问题的回答的标签。本文采用了四种相似度计算的模型：LDA（Latent Dirichlet Allocation），LSI（latent semantic index），Word2Vector 以及 BagOfWords 模型。

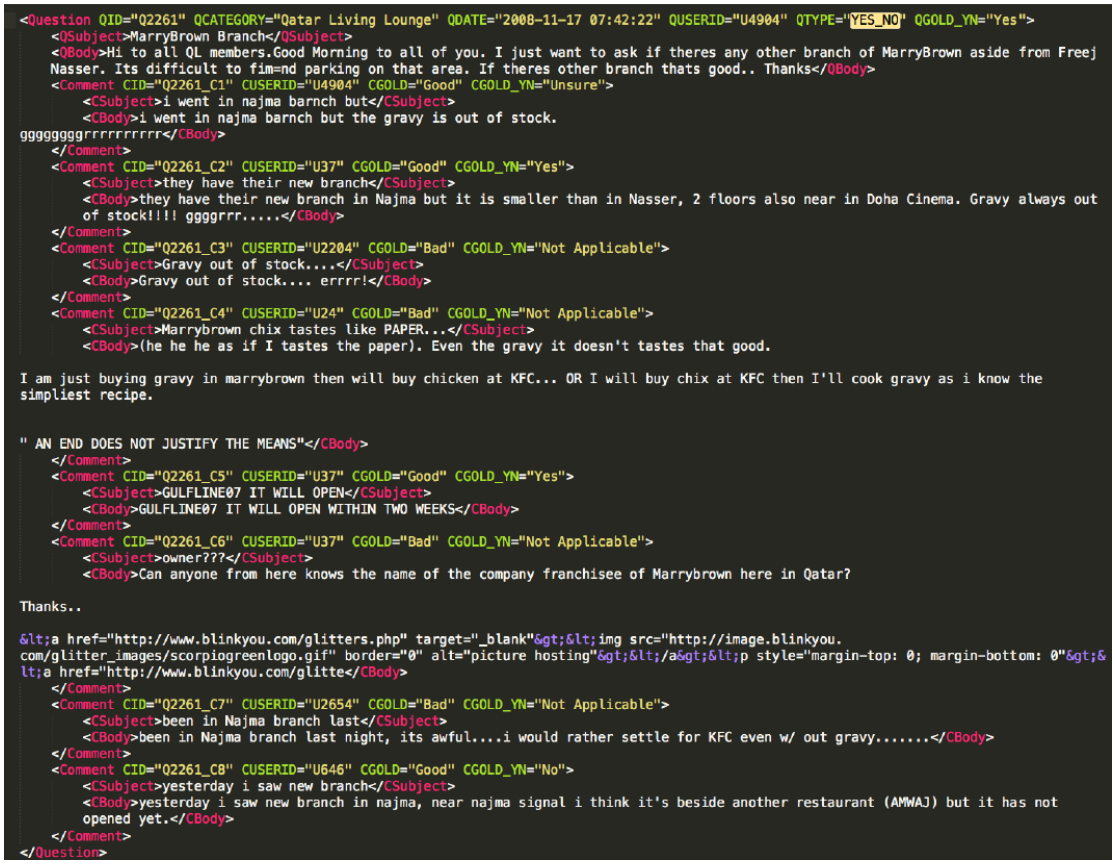

图 2: CQA-QL 语料库中标注的英文问题

由于 LDA 模型，LSI 模型和 Word2Vector 模型都需要使用整体数据集构建，因此本文将 train，devel 和 test 一起作为输入，训练这三个模型。这三个模型都是非监督的模型，LDA 能够设置确定的主题个数，通过训练给出每一段文本的主题分布。LSI 通过 SVD 分解，将文本词向量投影到维度主题大小的向量空间。Word2Vec 模型能够训练每一个词的向量表示，通过对整段文本的所有词的向量求和，得到文本的词向量表示。这三个模型都能通过向量的形式来表示整段文本的语义。最后通过对两段文本的向量求余弦距离，得到两段文本的相似度度量。BageOfWords 模型只需要对两段文本中出现的所有词，构建长度为出现的所有不同词数量的向量，每一维度表示了某一个词出现的频数，采用余弦距离来度量两个文本的相似度。

基于文本相似度提取的信息见表 1.

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

表 1 基于文本相似度的特征

2.2.2 基于内容描述的特征

为了试验本文提出的第二种方法，我们选择了一些关键词来描述答案是否是合理的回答。比如，如果回答的文本描述中出现 URL，Email 这类代表信息的关键词，我们就认为这个回答可能含有有效信息，是一个合理的回答。如果出现 Yes，No，Ok 这类回答的词，该回答可能就是对问题的有效回答。此外，如果回答的长度如果比较长，就说明该回答信息量比较大，也更可能是有效的回答。表 2 显示了基于内容描述的所有特征。

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

表 2 基于内容描述的特征

2.2.3 基于属性信息的特征

问题和答案的属性信息中，如果问题的用户和回答的用户是同一个人，则说明这个问题和答案是相关的。不同类别的问题，可能因为涉及到的内容不一样，导致问题回答的难易程度不一样。于是，我们提取了表 3 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

表 3 基于属性信息的特征

2.3 标签制定与模型构建

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。通过观察实际的数据，发现语料中未出现 Not English 和 Other 的标签。

对于 TaskB 任务，本文只提取了 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

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

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。

TaskA 和 TaskB 任务都是分类任务，因此本文采用了三种常用的分类模型：SVM（Support Vector Machine），GBDT（Gradient Boosting Decision Tree）和 RandomForest。

3. 实验数据和结果

3.1 实验数据

本文只针对英文语料进行了实验。对于英文语料来说，每一个问题有一个标题和描述，以及很多回答组成的列表，具体内容见图 1。表 4 展示了数据集的一些特征。其中，YES/NO 类型的问题约占问题总数的 10%，因为数据量小，所以针对任务 B 使用机器学习来处理会难一些。进一步可以看出，平均每个问题有 6 个回答，具体每个问题，最少回答数为 1，最大为 143。大约有一半的回答是好的，10%的回答是潜在有用的，其它的回答则不好。注意，为了分类，Bad 是一个异构的类，它包括 50%Bad，50%对话和一小部分非英语和其它回答。将 Bad 细分为多个标签的目的是考虑在其它系统中使用。大约 40%－50%的被标为 YES/NO 的回答的 CGOLD_YN 标签是 Yes，剩余部分 No 和 Unsure 各占了一半。然而，在被标为 YES/NO 的问题中的 QGOLD_YN 标签中，Unsure 的数量比 No 的数量多。总体上看，开发与测试数据

集和训练数据集相比，CGOLD 值的标签分布基本相似，但 QGOLD_YN 的标签分布差别比较大。

除了上面的数据集，语义评测举办方还发布了Qatar社区的所有问题和回答的原始文本，包含了超过100万的单词，这对于训练词嵌入、主题模型非常有帮助。

Category	Train	Dev	Test
Questions	**2,600**	**300**	**329**
– *GENERAL*	*2,376*	*266*	*304*
– *YES/NO*	*224*	*34*	*25*
Comments	**16,541**	**1,645**	**1,976**
– *min per question*	*1*	*1*	*1*
– *max per question*	*143*	*32*	*66*
– *avg per question*	*6.36*	*5.48*	*6.01*
CGOLD values	**16,541**	**1,645**	**1,976**
– *Good*	*8,069*	*875*	*997*
– *Potential*	*1,659*	*187*	*167*
– *Bad*	*6,813*	*583*	*812*
– *Bad*	*2,981*	*269*	*362*
– *Dialogue*	*3,755*	*312*	*435*
– *Not English*	*74*	*2*	*15*
– *Other*	*3*	*0*	*0*
CGOLD_YN values	**795**	**115**	**111**
– *Yes*	*346*	*62*	–
– *No*	*236*	*32*	–
– *Unsure*	*213*	*21*	–
QGOLD_YN values	**224**	**34**	**25**
– *Yes*	*87*	*16*	*15*
– *No*	*47*	*8*	*4*
– *Unsure*	*90*	*10*	*6*

表 4 英文语料数据集特征

3.2 实验结果

本文采用了 Train，Devel 和 Test 数据集训练 LDA，LSI 和 Word2Vec 模型提取相似度特征。在 Train 数据集中训练了 SVM，RandomForest 和 GBDT 三个模型，并在 Devel 和 Test 数据集上进行测试。实验统计了 Accuracy，Macro F1，Macro Precision 和 Macro Recall。表 5 显示了 Task A 的结果，在评判的时候，只度量了 Good，Potential 和 Bad 三个标签，Dialog，Not English 和 Other 都被归结为 Bad。表 6 显示了 Task B 的结果，对每个问题给出 Yes，No 和 Unsure 三种标签。此外，实验结果还列举了 Baseline 和 Rank One（参赛者的最好结果）的结果进行对比。

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%**

表 5 TaskA 实验结果

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%**

表 6 TaskB 实验结果

4 分析与讨论

从模型上看，相比于 SVM 和 RandomForest ，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。

通过参考相关论文和实验，本文实验存在以下问题：

1. TaskA 任务中，本文选择了 6 个标签进行标注，而实际评价的时候，只采用了三个标签。因此，如果只对三个标签进行建模，可以极大提高模型的性能。
2. 训练 LDA，LSI 和 Word2Vec 模型的时候，本文选取的语料是 Train，Devel 和 Test 数据集。而在实际比赛的过程中，Test 数据集是未知的。因此，本文可以采用官网额外提供的 Qatar Living 的 100 万原始语料进行训练。

此外，本实验还可以在在以下几个方面改进：

1. 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。
2. 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。
3. 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。

5 结论

本文提出了两种方法：采用传统 QA 基于相似度排序的方式，充分利用问题的答案的信息，寻找问题和答案之间的相似度；采用基于答案的文本描述信息判断答案是否是合理的回答。来解决 CQA 的标注问题，并通过实验进行了验证，实验结果明显高于 Baseline 模型。

但是，本实验还有很大的改进空间。在模型上，还可以采用更多的模型进行实验，比如神经网络，逻辑斯特回归，朴素贝叶斯。在特征上，还有可以增加更多基于内容描述的信息。此外，同一问题的不同回答之间，也存在强烈的依赖关系。比如如果一个问题的回答含有问题提出者的感谢词，比如 Thanks，Thank you，那么这个回答的上几个回答可能是好的回答的可能性就很高。基于回答标签序列进行预测，也可能会获取到更多的特征。

6 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.