基于字符粒度 fastText 的用户意图领域分类

唐梓毅
北京来也网络科技有限公司
tangziyi@laiye.com

摘　要: 用户意图领域分类（domain classification）是人机对话系统的重要组成部分。本文针对 SMP2018 ECDT 评测用户意图领域分类任务的数据特点，采用字符粒度的 fastText 分类模型，并通过引入预训练字向量、领域实体、领域正则等方式优化其表现，最终在测试集上达到了较好的分类结果。实验证明，字符粒度的 fastText 在短文本的用户意图领域分类任务中有一定潜力。

关键词: 短文本分类；fastText；预训练字向量；领域实体；领域正则

1 引言

近年来，人机对话技术受到了学术界和产业界的广泛关注。学术上，人机对话是人机交互最自然的方式之一，其发展影响及推动着语音识别与合成、口语语言理解、对话管理以及自然语言生成等研究的进展；产业上，众多产业界巨头相继推出了人机对话技术相关产品，如个人事务助理、娱乐型聊天机器人等。以上极大地推动了人机对话技术在学术界和产业界的发展。在第七届全国社会媒体处理大会（SMP 2018）上，哈尔滨工业大学和科大讯飞股份有限公司组织了 SMP 2018 中文人机对话技术评测（ECDT），为人机对话技术相关的研发人员提供了一个良好的沟通平台。

在人机对话系统中，用户意图领域分类（domain classification）作为人机对话系统的第一步，通常扮演着重要的角色，其性能好坏直接影响到系统后续模块的运行。由于口语对话场景中用户 query 通常较短，因此领域分类属于短文本分类。传统的短文本分类主要是基于特征工程的统计学习方法，比如支持向量机（SVM）、随机森林（RF）等，但由于短文本特征较少，因此很难取得较好的分类效果。近几年来，基于向量空间表示（vector space representation）以及深度学习（deep learning）的方法的应用，使得短文本分类取得了长足的进步。特别是 Facebook AI 实验室提出的 fastText[1,2]，被证明是一种简单而高效的短文本分类和表征学习的方法。

本文介绍我们参加 SMP2018 ECDT 评测中用户意图领域分类任务的系统，采用了基于字符的 fastText 模型，并针对训练集样本数量有限以及文本长度较短等特点，引入预训练的字向量，以及从样本中抽取领域实体特征、领域正则等，进一步优化领域分类效果。

2 用户意图领域分类系统

2.1 系统总体描述

本文提出的用户意图领域分类系统主要基于字符粒度的 fastText 模型。在此基础上，针对训练集样本数量有限的特点，我们引入了外部语料预训练的字向量，补充必要的语义信息；针对训练集文本长度较短的特点，我们从样本中抽取了一些领域实体特征，丰富了样本的特征以及系统的泛化能力；最后，针对一些容易混淆的样本，我们引入领域正则修正结果，进一步减少系统的误判。

2.2 基于字符粒度的 fastText 领域分类

Facebook AI 实验室已经实验证明，对于中文的 fastText 文本分类，字符粒度的 n-gram 特征优于单词粒度的 n-gram 特征[3]，我们在 SMP2018 测试集上的实验也验证了这点。因此，我们选择了字符粒度输入的 fastText 分类模型作为领域分类方案。

2.3 使用预训练字向量初始化模型

数据稀疏（data sparseness）与未知词（unseen words）是短文本分类中的难点。已经有实验证明，通过预训练的词向量引入额外的语义特征，可以改善数据稀疏与未知词问题，提高短文本分类准确率[4]。因此，我们使用 500 万条语料预训练的字向量初始化 fastText 模型，提高了模型对未知词的覆盖以及语义泛化能力。

2.4 基于领域实体的特征提取

分析数据发现，评测数据中的很多意图包含一些领域实体，因此我们从样本中抽取了一些领域实体，作为额外的特征加入样本训练模型，通过引入外部知识的方式提高了部分样本意图下样本的分类准确率和召回率。

2.4 基于领域正则的结果修正

最后，针对部分极易记忆混淆的样本，我们引入了领域正则，修正一些明显应该属于某意图的样本的分类结果。同样是通过引入外部知识的方式，进一步减少系统的误判。

3 实验与结果分析

3.1 基于字符粒度的 fastText 领域分类

首先，我们以准确率为评价指标，对比字符粒度与词粒度的 fastText 模型，结果如下：

表 1. 字符粒度以及词粒度 fastText 分类效果对比

模型	分类准确率（%）
词粒度 fastText	85.35
字符粒度 fastText	89.30

实验证明，基于字符粒度的 fastText 用于本任务效果更好。进一步用网格搜索的方式调整模型的参数，在最优参数下，得到的模型在测试集上的 F 值为 0.8953

3.2 预训练字向量初始化模型

通过分析实验结果，我们发现有些错误是由于测试样本中存在未知词导致的，比如：“糖醋里脊啊，啊”这句样本，由于没有“里脊”属于训练样本中没有出现的字词，因此没有把这句样本分到 cookbook 意图中。但如果有预训练的字词向量提供“里脊”的语义信息，分类模型就可以通过训练样本中出现的“糖醋排骨”把该样本分到 cookbook 意图下。

基于以上分析，我们使用与训练样本相近的 500 万条语料，预训练了一份字向量，来初始化 fastText 模型，由于未知词造成的错误有所减少，系统的 F1 值提高到了 0.9313。

3.2 基于领域实体的特征提取

进一步分析实验结果，我们还发现，有部分的错误是仅靠字面的语义信息难以得到正确结果的，比如：“一路向西”这句样本，由于仅靠句子的语义信息，模型很难知道《一路向西》是一部电影，因此模型倾向于认为它属于 map 意图，而无法将其正确分到 video 意图下。显然，这种情况下模型需要额外的信息输入。

因此，我们对训练数据进行实体抽取，从中选出了 85 种各意图下可能出现的实体类型，将其作为特征加入训练样本，这样除了字面信息之外，模型还可以借助实体信息对样本进行分类。实验结果是，系统的 F1 值提高到了 0.9382。

3.3 基于领域正则的结果修正

经过以上优化，我们还发现，测试样本中存在一些极易混淆的样本。比如：“翻译背首诗”这句样本，由于“背首诗”的特征较强，因此我们的模型将其判为了 poetry 意图，而非 translation 意图。显然，这种情况下模型也需要输入额外的信息。

基于上述判断，我们引入领域正则，来修正这些模型容易混淆的样本的分类结果，最终系统的 F1 值提高到了 0.9421。

3.4 实验结果总结

经过以上实验，得到不同模型方案的实验结果如下：

表 2. 字符粒度以及词粒度 fastText 分类效果对比

模型	测试集 F1 值
字符粒度 fastText	0.8953
+预训练字向量初始化	0.9313
+领域实体特征	0.9382
+领域正则修正	0.9421

4 结束语

我们在本次 SMP2018 ECDT 用户意图领域分类任务评测数据上的实验结果证明，字符粒度输入 fastText 应用于短文本的用户意图领域分类可以实现较好的结果。并且可以通过引入预训练字向量、领域实体、领域正则的方式，进一步优化分类效果。值得一提的是，fastText 的模型训练速度很快，可以在 1 秒内完成训练，并且可以只用 quantize 等方式压缩内存占用，实际工程应用中潜力较大。

但由于本次评测时间有限，还有一些问题没有来得及解决，比如模型存在对于 chat 样本的误召回问题，分析原因有可能是因为 fastText 本质上相当于线性模型，本身很难对一些样本进行区分。后面有时间的话，可以进一步尝试融合一些非线性模型（如 GBDT）等方法，或者尝试层次化分类，在做 30 个领域分类之前单独做样本是否为 chat 意图的二分类，以解决对于 chat 样本的误召回问题。

References (参考文献)

[1] Bojanowski P, Grave E, Joulin A, et al. Enriching Word Vectors with Subword Information[J]. 2016.

[2] Joulin A, Grave E, Bojanowski P, et al. Bag of Tricks for Efficient Text Classification[J]. 2016:427-431.

[3] Zhang X, Lecun Y. Which Encoding is the Best for Text Classification in Chinese, English, Japanese and Korean[J]. 2017.

[4] Ma C, Wan X, Zhang Z, et al. Short Text Classification Based on Semantics[C]// International Conference on Intelligent Computing. Springer International Publishing, 2015:463-470.