NLPCC 2016 Shared Tasks Stance Detection in Chinese Microblogs 系统描述文件

— Scau_SDCM（华南农业大学）

1 系统流程

系统流程如图 1 所示。

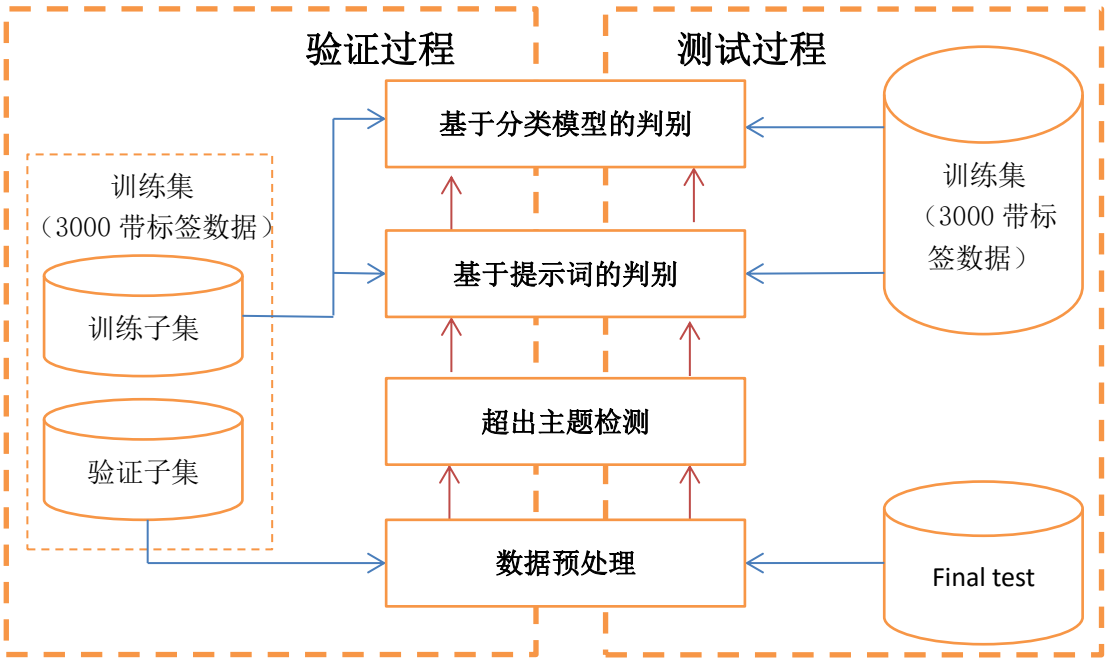

图 1 系统流程

主要包括两个流程：

（1）验证过程：将 3000 带标签的训练数据按 7：3 分成训练子集（2100）和验证子集（900），基于验证选择最后阶段的分类模型。基于训练子集提取提示词进行基于提示词的判别，并进行随机森林(Random forests，RF)和卷积神经网络(Convolutional neural network，CNN)两个分类模型的训练，采用验证子集进行模型选择。

（2）测试过程：采用 Final test 数据，进行超出主题检测，并基于训练集提取的提示词进行基于提示词的判别，最后基于验证过程选出的分类模型（RF）进行立场判别。

2 采用方法

采用了三阶段方法：超出主题检测+基于提示词的判别+基于分类模型的判别。

3 具体过程

数据预处理：将句子进行分词处理，并去除标点符号和中文 stopword（比如“的”,“和”等）,将阿拉伯数字统一替换成标记 NUMBER,将繁体中文转为简体，移除微博 URL。 句子预处理后，将不包含任何词的句子直接判为 None，检出 42 条（占 0.28%）。

第一阶段：超出话题检测，因为时间关系只针对“春节放鞭炮”和“开放二胎”两个主题做了检测，根据一些主题特征词，如“春节放鞭炮”主题中的“鞭炮”、“爆竹”等，以及“开放二胎”主题中的“胎”、“生”等，制定了主题特征词表，将不包含任何特征词的判为 None。为了避免误判，主题词表不大，因此在 final test 中只检出 470 条（占 3.13%）。

第二阶段：基于提示词(Cue phrase)的判别，采用了置信度（Confidence）和支持度（Support）并重的标注，从训练语料（3000 带标注微博）中提取提示词，提取方法采用了 1-gram 和 2-gram 的做法，2-gram 类似于 skip-gram 的做法，不考虑连接性，只考虑搭配性。并给予了不同的置信度不同的支持度要求，即高置信度（如 100%）的提示词给予相对低的支持度要求（如 0.17%,对应频次 5）。在 Final test 中，根据提示词给出了 8793 条（占 58.62%）的判别。

第三阶段：基于分类模型的立场判别，在训练集上进行验证，尝试了多种不同分类模型，最后选择了 RF 和 CNN 的作为主要模型，通过在训练集上进行验证选择模型（选择了 RF）。在 Final test 中对经过了第一阶段超出主题检测和第二阶段提示词判别的基础上，对剩余的微博进行判别。

4 使用的资源

（1）jieba 分词工具：https://github.com/fxsjy/jieba
（2）Keras 1.0.4: 神经网络的框架， https://github.com/fchollet/keras
（3）OpenCC 0.2: Open Chinese Convert，https://github.com/BYVoid/OpenCC
（4）scikit-learn 0.17.1: 机器学习工具类，
https://github.com/scikit-learn/scikit-learn

5 展望

在 Final test 中，我们队总体上取得了 0.6666 的 F-score(FAVOR) 和 F-score(AGAINST)的宏平均效果。

进一步提升的设想：根据对训练集上的验证以及 Final test 上的判别的观察，一方面可以增加对主题的研究，提高第一阶段超出主题检测的判别率；其次，不论是基于提示词的判别还是基于分类模型的判别，对 None 的判出率都很低，在 Final test 中，仅给出了 1552 例 None 的判别（其中还包含了 42 例是预处理和 470 例是超出主题检测的结果），占 10.35%。而根据对训练集的统计，None 占 20.04%，差距仍然巨大，因此这是进一步提升整体判别效果的关键方向之一。

6 部分主要参考文献

[1] Scornet E. Random Forests and Kernel Methods [J]. IEEE Transactions on Information Theory, 2015, 62(3):1485-1500.

[2] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 26(2): 123–140

[3] Kim Y. Convolutional neural networks for sentence classification[C]// Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 2014: 1746–1751.

[4] Johnson R and Zhang T. Effective use of word order for text categorization with convolutional neural networks[C]//Proceedings the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL-HLT, 2015: 103-112.