**Learning Writing Pattern**

**Project report  
Name: CHEN Yifei Stu #: 20476334**

目录

项目描述2

前期调研2

机器翻译的思路2

Computational Journalism(NLG)的思路3

总结3

# 项目描述

Learning writing pattern 首先需要对金融新闻做摘要(summarization)并根据其subject进行分类，找出subject为stock的新闻，然后提取它的词法和句法方面的特征，学习它的写作模式，还需要建立股票交易情况和文章描述情况之间的联系，从而训练出一个模型，接受特定日期，输出描述当天股票市场信息的文章。

# 前期调研

这段时间我主要对目前主流的NLP领域的一些热门问题进行了调研，并与我们的project goal进行对比，试图找出比较适合需求的系统结构和技术。经过阅读一些相关论文，我感觉目前主要有两个思路：

## 机器翻译的思路

早期的MT主要是基于规则的，需要人为地定义一些两门语言之间转换的规则，但是人为的干预在系统里是一个不确定因素，而且规则总是不完备的。到后来SMT（统计机器翻译）产生，但是仍然没有摆脱人为的干预（需要设计特征），而且预处理环节比较多。现在比较新的技术就是运用DL的NMT（神经网络机器翻译），而它的特点是直接构建源语言到目标语言的映射，省去了中间预处理的过程。在源语言和目标语言两端各自构建一个Encoder和Decoder RNN来实现特征的提取和学习，其中还涉及注意力机制、柱搜索算法。类似的，我们也可以采用这种sequence to sequence 的方法来提取词汇运用和句子结构方面的写作模式特征，然后运用映射的方式将raw data中的趋势表示成文章或者专业性的描述语句，这部分主要是参考的极客学院[[1]](#endnote-1)上面的一些介绍。

## Computational Journalism(NLG)的思路

Computational Journalism有几个应用场景，claimBuster[[2]](#endnote-2)主要是根据输入的辩论资料或者新闻来看它是否有说谎的情况，这需要在一个fact的database中进行检索；还有一个是factWatcher[[3]](#endnote-3)，主要是根据输入的fact tuple生成相应的事实新闻。而这个fact-to-statement的转化过程，实际上和我们的transaction-to-information的转化过程类似，所以这也是可以借鉴的一个思路。

Computational Journalism的fact-to-statement用到了NLG方面的技术，它主要是根据输入的raw data生成描述性的语句。在Rivindu Perera的survey[[4]](#endnote-4)中，NLG主要有以下几个流程模块： content determination, lexicalization, aggregation和realization。Content determination对应于特征提取，在本项目中，stock transaction的所有信息表项都对股票市场的信息产生影响，因此我们我们不需要对content进行筛选。而lexicalization模块中用到的词库可以从training data中训练得到，Reither et al[[5]](#endnote-5) 提出的基于set of data-to-word rules的方法可以给本项目以启发。同样的，aggregation和realization中的grammar rule也需要我们从原来的financial news中学习得到，这里我们可以借鉴OpenCCG [[6]](#endnote-6)(Combinatory Categorial Grammar)。

# 总结

在学习writing pattern时，我们首先需要对金融新闻进行分类，可以运用类似情感分析的技术进行分类筛选出有关stock的新闻，再对其进行特征提取和模型构建。和学长的讨论后，我觉得与NMT类似的深度学习框架存在数据格式不匹配、调参过程比较复杂等缺点，而NLG的record-to-statement架构比较适合本项目，因此下一步就是研究框架下的各个功能模块并进入代码实现阶段。

1. **极客学院**，**http://wiki.jikexueyuan.com/project/deep-learning/machine-translation.html ;** [↑](#endnote-ref-1)
2. **N Hassan ,F Arslan ,C Li ,M Tremayne,** “**Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster”, Acm Sigkdd International Conference on Knowledge Discovery & Data Mining, 2017 :1803-1812;** [↑](#endnote-ref-2)
3. **N Hassan , A Sultana , Y Wu, G Zhang, C Li** ，**“Data in, fact out: automated monitoring of facts by FactWatcher”,《Proceedings of the Vldb Endowment》 , 2014 , 7 (13) :1557-1560;** [↑](#endnote-ref-3)
4. **Rivindu Perera, Parma Nand, “recent advances in natural language generation: a survey and classification of the empirical literature“, Computing and Informatics, Vol. 36, 2017, 1–32, doi: 10.4149/cai 2017 1 1** [↑](#endnote-ref-4)
5. **Reiter, E.—Sripada, S.—Hunter, J.—Yu, J.—Davy, I, “ Choosing Words in Computer-Generated Weather Forecasts”, Artificial Intelligence, Vol. 167, 2005, No. 2,pp. 137–169, doi: 10.1016/j.artint.2005.06.006.** [↑](#endnote-ref-5)
6. ***White, M, “CCG Chart Realization from Disjunctive Inputs”, Proceedings of the Fourth International Natural Language Generation Conference (INLG ’06), Sydney, Australia, 2006, pp. 12–19, doi: 10.3115/1706269.1706274.*** [↑](#endnote-ref-6)