南京農業大學
NANJING AGRICULTURAL UNIVERSITY

数字人文下的机器翻译

目录
CONTENT

- 基本知识
- 在数字人文研究中的应用
- 古英和古白机器翻译实现

南京農業大學

NANJING AGRICULTURAL UNIVERSITY

机器翻译的基本知识

1、机器翻译

◆机器翻译就是利用计算机自动将一种语言翻译为另外一种语言，其一直以来都是自然语言处理的核心问题，深度学习兴起之后，研究水平达到了新的高度。

◆对于以传统典籍为对象的数字人文研究来说，古文机器翻译能够自动地“解其言、知其意”，古汉语到现代汉语的语内翻译有利于中华传统文化的教育、普及和传播，古汉语到英语的语间翻译则有利于向海外弘扬中华传统文化。。

2、平行语料库

◆平行语料库，也叫双语语料库，由源语言和目标语言一一对应的翻译句对构成。

◆从机器学习的角度来看，平行语料库是机器翻译模型所需训练和测试数据集的来源，其规模和质量是影响机器翻译效果的重要因素。

◆以传统典籍为对象的数字人文研究需要机器翻译技术帮助古文的自动理解，因而对于大规模古籍平行语料库有较高的需求，其中还可区分为古白平行语料库和古外平行语料库。

3、基于规则的机器翻译

◆基于规则的机器翻译方法将包含语句的多重转换。

◆首先对源语言句子分别从词汇、句法和语义三个层次进行分析和转换，接着借助转换规则、知识库和中介语等资源得到对应目标语言的分析和转换结果，最后再反向从语义、句法和词汇三个层级进行生成，得到最终的目标语言句子。

◆基于规则的机器翻译需要人工编写准确和完整的转换规则，需要耗费大量的人工，同时由于规则对于特定语言的局限性，导致方法的迁移能力较差。

4、统计机器翻译

◆统计机器翻译是指使用统计机器学习模型实现的机器翻译方法，主要区别于早期基于规则的诸多方法。

◆统计机器学习模型使用语料库来分别进行训练和测试，通过语言模型表示源语言和目标语言的句子，再使用解码算法将两种语言的句子进行匹配。

◆统计机器翻译可以看作找出一个模型最优的模型，使得对于所有的源语言句子S都能在目标语言中找到匹配的句子T，这样一个模型一般可以建模为S和T的条件概率P（S|T），而最优的模型就是条件概率最大的情况。

5、神经机器翻译

◆神经机器翻译基于一种更加直接的端到端（Seq2seq或编码器到解码器）思路，利用深度学习中表示学习的特性，将源语言和目标语言的建模直接交给神经网络，从而大大提高了机器翻译的性能。

◆具体来说，编码器将输入的源语言句子表示为向量形式，包含输入序列的全部信息；解码器将这种表示重新转换，将源语言句子的向量作为隐藏输入，预测和生成目标语言中的词语并构成句子。端到端的神经机器翻译方法关键在于编码器和解码器的选择，根据深度学习中常见的模型，编码器和解码器有CNN、LSTM、GRU、Transformer等形式。

6、OpenNMT模型

◆OpenNMT是目前较为常用的神经机器翻译模型，由哈佛大学自然语言处理实验室开发，可以实现序列到序列、语音到文本等多项自然语言处理任务，并取得显著成效，已用于多个研究和行业应用。

◆该模型在哈佛大学开发的seq2seq-attn的基础上发展而来，具有较高的效率、可读性和可推广性。其在基本神经机器翻译模型外，还增加注意力机制、门控单元、多层神经网络堆叠、输入反馈、正则化等多项先进技能，在机器翻译任务上具有良好表现。OpenNMT的编码器和解码器均采用transformer结构。

7、迁移学习与机器翻译

◆迁移学习的主要思想是把源域的知识，迁移到目标域，使得目标域达到更好的学习效果。

◆通俗来讲，就是运用已有的知识来学习新的知识。传统的机器学习方法需要针对特定语言，通过大量相应的语料进行模型训练，然后将模型应用到特定的任务中。

7、迁移学习与机器翻译

◆传统方法的实现一般需要收集大规模的语料，但是仅针对翻译任务来说，目前除了像汉英平行语料丰富的语言外，很多语言都存在着平行语料资源匮乏问题。在这种标注数据缺乏的情况下，迁移学习可以更好的利用小规模数据达到理想的效果。

◆迁移学习按照学习方式的不同有多种划分，对于低资源语言翻译，通常采用基于参数的迁移，即源域和目标域的任务之间共享模型参数，以此来解决神经机器翻译中资源不足问题的先河。一般来说，迁移语言的相似性越高，效果越好。

8、BLEU值

◆BLEU（双语评估替换）是一种非常有效的以单一数字指标评估机器翻译结果的方法，通过对比连续多个词是否出现在参考译文中，对机器翻译的结果进行自动评估。

◆在实际应用中，通常采用四元BLEU评分，即依次比对单个、连续两个、连续三个和连续四个词在参考译文中出现的比例，作为其评分的准确率。为了避免因句子长度短、词语正确而带来的评分过高问题，设置惩罚值，当机器翻译句子长度小于参考译文长度时，根据其长度差异对其进行惩罚，句子长度越短，则得分越低。

8、BLEU值

◆BLEU分数的计算方法如公式所示，其中 为惩罚值，句子越长，惩罚值越小。 为n元准确率的得分， 代表n元准确率的权重。BLEU的值域为[0,1]，值越大表示翻译效果越好，大部分语言的BLEU分数在0.2-0.5。一般展示计算结果为BLEU实际分数乘以100。

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \, log p_n\right)$$

南京農業大學
NANJING AGRICULTURAL UNIVERSITY

机器翻译在数字人文研究中的应用

应用现状

- 从现有研究来看，神经网络算法在机器翻译任务中取得显著成效，并在古文信息处理任务中发挥着重要作用。
- 目前神经机器翻译的语言主要集中在英语、德语、法语等印欧语系的语言上，由于上述语言具有丰富的语料资源，算法和模型多针对上述语言开发和设计。神经机器翻译的算法逐渐成熟，能够实现语言之间的翻译，具有一定的研究基础。
- 在汉语机器翻译方面，研究对象多集中于英汉、俄汉等语言，也有部分少数民族语言与汉语互译的研究，但对于古代汉语自动翻译的研究较少。

应用现状

◆相较于人工翻译耗时长与成本高，利用机器自动翻译成本较低，可以在较短时间翻译较大批量的文字，具有较高的科学研究价值和实用价值。

◆在古文信息处理方面，大部分研究是从字、词的角度出发，挖掘古文中的信息，如古文自动断句、古文词性自动标注以及古文中人名地名实体识别等。以字词为单位的研究只能够提取古文中的单个信息点，很难将其连接成线。对于句子、段落、篇章等更长文本的古文信息处理仍较少，难以从中挖掘出完整的信息。

南京農業大學

NANJING AGRICULTURAL UNIVERSITY

典籍的古英和古白机器翻译实现

古英机器翻译·语料准备

◆典籍数据来源于“中国哲学书电子化计划”（https://ctext.org/confucianism/zhs）网站。

◆备用网址：http://ctext.cn/

古英机器翻译·语料准备

English 繁体

- [本站介绍]
  - 简介
  - 字体试验页
  - 协助
  - 常见问答集
  - 使用说明
  - 工具
  - 系统统计
  - 数位人文
- 先秦两汉
  - 儒家
  - 墨家
  - 道家
  - 法家
  - 名家
  - 兵家
  - 算书
  - 杂家
  - 史书
  - 经典文献
  - 字书
  - 医学
  - 出土文献
- 汉代之后
  - 魏晋南北朝
  - 隋唐
  - 宋明
  - 清代
  - 民国
- 简介说明
- 相关资料

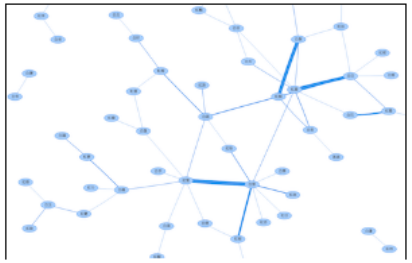
百諸家子 中国哲学书电子化计划

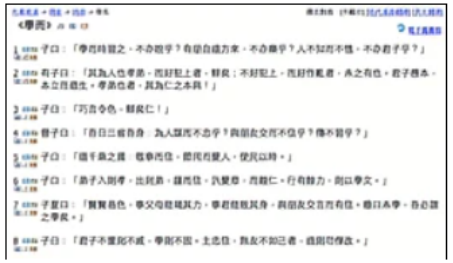
简体字版

欢迎

欢迎您来到中国哲学书电子化计划首页。中国哲学书电子化计划是一个线上开放电子图书馆，为中外学者提供中国历代传世文献，力图超越印刷媒体限制，通过电子科技探索新方式与古代文献进行沟通。收藏的文本已超过三万部著作，并有五十亿字之多，故为历代中文文献资料库最大者。 欢迎参阅先秦两汉、汉代之后或维基区资料库目录，或参考系统简介、常见问答集、使用说明和相关工具。若欲寻找特定著作，可使用书名检索功能一并检索本站各种主要原典资料。

This site is available in Chinese and English, as well as simplified and traditional Chinese characters - you can switch between these at any time using the links at the top left of the page.






本站所提供的资料和服务都不收费，因此网站所需要的资金全来自捐款。若您愿意捐款补助，请参閱相关说明。感谢您的支持。

最新消息

日期	内容
2016-10-10	**哈佛燕京图书馆历代中文文献已收录** 通过哈佛燕京图书馆的支持，本站电子图书馆已收录燕京图书馆五百多万页历代中文文献的影印资料，其中包括中文善本特藏项目中的高质量影印本。本站字符识别技术已将资料加以文字化，现已放入本站维基区，并实现了资料的全文检索。希望将来能够与其它收藏中国古籍的图书馆合作，以提供更丰富更全面的资料。
2015-07-02	**统一码8.0升级，增加新汉字字形** 统一码标准最近推出新版本，增加了数千新的罕见字、异体字。本站字典功能现已支持这些新字的查询，使用者安装花园字体的最新版本即可显示这些新增字形。许多新增字形属于"CJK扩展E区"；您可以参考字体试验页面以确认您的系统是否支持这些字。

完整的更新记录存放于"最新消息"讨论区。请注意：上述最新消息只包括最主要的技术升级项目。若想要参考文献内容的更新记录，登入后可另外参考维基修改记录的各种项目。

最新讨论主题

日期	内容

古英机器翻译·语料准备

◆选取《论语》、《礼记》、《战国策》、《尚书》、《道德经》、《左传》、《史记》、《孙子兵法》、《论衡》、《周易》、《孝经》、《商君书》、《墨子》、《庄子》、《孟子》和《公孙龙子》共十六部历史典籍，共得到40799个古英文平行句对。

◆经过数据清洗，包括通过正则匹配去除奇异字符、删除任意一方存在缺失的句对、去重等操作，最终得到40633个古英平行句对，构建完成古英历史典籍平行语料库。

古英机器翻译•数据预处理

◆进行古文分词和英文分词。由于英文单词之间本身存在空格，将标点（.!?’\-”）作为分隔符实现英文分词。

◆划分训练集 验证集 测试集

表 9-1 数据样例

序号	古文	英文
1	斯人也而有斯疾也！	That such a man should have such a sickness !
2	老子曰：“幸矣，子之不遇治世之君也！	Laozi replied , ' It is fortunate that you have not met with a ruler fitted to rule the age .
3	夏，楚人既克夷虎，乃謀北方。	Having occupied the tribe of Yi in summer , the state of Chu planned to expand its territory northward .

古英机器翻译•模型构建与参数调整

◆参数调整在onmt目录下的opts.py文件。可根据实验语料平均句子长度调整，模型源端最大序列长度设置为50，目标端最大序列长度设置为150，词典大小为50000，每个batch中训练样本的数量为4096。

◆编码器和解码器的隐层维度为512，层数为12层，其中Transformer自注意力机制头数为标准头数8个。注意力机制和前馈神经网络中增加了dropout层，p值设置为0.3，共迭代50000次，每迭代2000次进行一次验证。使用了Adam优化器，学习率设置为0.001，损失函数是CrossEntropy。

古英机器翻译•模型构建与参数调整

◆使用pip安装OpenNMT-py，并找到安装的源码。或者直接从OpenNMT的GitHub网址（https://github.com/OpenNMT/OpenNMT-py）直接下载包。并将上述三集数据放入data目录下。

◆将数据进行词典构建和模型输入前预处理。创建preprocess.py文件，调用后得到以pt结尾的文件。

```
python preprocess.py --train_src=data/src-train.txt --train_tgt=data/tgt-train.txt --valid_src=data/src-valid.txt --valid_tgt=data/tgt-valid.txt --save_data=data/dataset
```

古英机器翻译•模型训练

◆模型保存在data/model目录下。

```
python train.py --data data/dataset --save_model data/model/model
```

◆调用翻译文件将上述训练得到的model使用测试集进行测试。模型预测文件保存在pred.txt下。

```
python translate.py -model data/data/model.pt -src data/src-test.txt -tgt data/tgt-test.txt -output data/pred.txt -replace_unk -verbose -gpu 3
```

古英机器翻译·结果评估

◆将模型预测和测试集标准翻译计算bleu指标。

```
perl tools/multi-bleu.perl data/tgt-test.txt < data/pred.txt
```

```
PRED AVG SCORE: -0.6139, PRED PPL: 1.8476
GOLD AVG SCORE: -5.8529, GOLD PPL: 348.2350
BLEU = 10.11, 39.2/15.6/9.4/6.7 (BP=0.721, ratio=0.753, hyp_len=46905, ref_len=62266)
```

古白机器翻译·python环境准备

◆打开windows命令提示符，输入：

pip install OpenNMT-py

◆等待该第三方模块安装完毕。将路径切换至OpenNMT所在路径，输入：

python setup.py install

古白机器翻译•语料准备

◆此处源语言为文言文，目标语言为白话文，需要构建一定规模的源语言和目标语言数据集。共需准备六个文档，分别为训练集源语言和目标语言文档train_src.txt和train_tgt.txt，测试集源语言和目标语言文档test_src.txt和test_tgt.txt，以及验证集源语言和目标语言文档val_src.txt和val_tgt.txt。为了更直观地解释语料格式，将训练集的部分展示如下。

◆train_src.txt格式如下：

```
1  飞 怒 ， 令 左 右 牵 去 斫 头 ， 颜 色 不 变 ， 曰 ：
2  晋 侯 使 郄 乞 告 瑕 吕 饴 甥 ， 且 召 之 。
3  四 人 相 谓 曰 ： “ 郁 成 王 汉 国 所 毒 ， 今 生 将 去 ， 卒 失 大 事 。 ”
4  命 大 师 陈 诗 以 观 民 风 ， 命 市 纳 贾 以 观 民 之 所 好 恶 ， 志 淫 好 辟 。
5  燕 、 赵 、 韩 、 魏 闻 之 ， 皆 朝 于 齐 。
6  璋 遣 刘 璝 、 冷 苞 、 张 任 、 邓 贤 等 拒 先 主 于 涪 ， 皆 破 败 ， 退 保 绵 竹 。
7  醯 酒 浇 于 清 ， 汁 献 浇 于 醯 酒 ；
8  是 疵 为 赵 计 矣 ， 使 君 疑 二 主 之 心 ， 而 解 于 攻 赵 也 。
9  昔 金 天 氏 有 裔 子 曰 昧 ， 为 玄 冥 师 ， 生 允 格 、 台 骀 。
10 威 王 六 年 ， 周 显 王 致 文 武 胙 于 秦 惠 王 。
```

古白机器翻译·语料准备

◆与其对应的train_tgt.txt格式如下：

1 张飞大怒，命令身旁的兵卒将严颜拉出去砍头，严颜面不改色，说道：
2 晋惠公派遣郄乞告诉瑕吕饴甥，同时召他前来，
3 四个骑兵互相商议说：“郁成王是汉朝所恨的人，如今若是活着送去，突然发生意外就是大事。
4 命令各诸侯国的太师一一演唱当地的民歌民谣，从而了解民风习俗。命令管理市场的官员呈交物价统计表，从而了解百姓喜欢什么物品，讨厌什么物品。
5 燕、赵、韩、魏四国听到这件事，都来齐国朝见。
6 派刘璜、冷苞、张任、邓贤等人到涪县抵御先主，都被击败，只好退守绵竹。
7 对于盎齐以下三齐，因其较清，不用过滤，只须用清酒冲淡一下就行了。至于郁色，用盎齐来冲淡。
8 可是郄疵在为赵国谋划，以便使贤君怀疑韩、魏两国，进而瓦解三国攻赵的盟约。
9 从前金天氏有后代叫做昧，做水官，生了允格、台骀。
10 威王六年，周显王把祭祀文王、武王的福肉送给秦惠王。

古白机器翻译·语料准备

◆值得注意的是，源语言和目标语言句子每一行是一一对应的。因为OpenNMT首先对词语进行处理，需要对语料进行分词，此处分词直接以字符为单位，让OpenNMT自行寻找各个字之间的相关关系。在训练中，也可以采用自己的分词方法，对语料进行分词。

古白机器翻译·设置配置文档

◆以yaml为后缀的文档为配置文档，需要在其中配置数据存储路径等的相关信息。gu_bai.yaml配置文档部分代码如右：

◆其中词表、语料的命名可以不同，只要路径和配置中路径对应即可。若在多个GPU上训练，可以在world_size处更改GPU个数，在gpu_ranks处设置GPU编号。

```
## 样本存储路径
save_data: gu_bai/run/example
## 词表存储路径
src_vocab: gu_bai/run/example.vocab.src
tgt_vocab: gu_bai/run/example.vocab.tgt
# 是否覆盖文件夹中的现有文件
overwrite: False

# 语料选项
data:
    corpus_1:
        path_src: gu-bai/train_src.txt
        path_tgt: gu_bai/train_tgt.txt
    valid:
        path_src: gu_bai/val_src.txt
        path_tgt: gu_bai/val_tgt.txt
```

古白机器翻译·程序运行

◆将路径切换至OpenNMT所在路径，首先生成词表，输入：

python build_vocab.py -config gu_bai.yaml -n_sample 10000

◆等待词表生成。其中config为配置文档的路径，n_sample为训练数据的句子个数。词表生成后，即可开始模型的训练，输入：

python train.py -config gu_bai.yaml

◆等待模型训练完成。在这过程中，可以在设置的路径中查看定时保存的模型
模型训练完成后，对测试集进行翻译，输入：

python translate.py -model gu_bai/run/model_step_1000.pt -src gu_bai/test_src.txt -output gu_bai/pred_1000.txt

古白机器翻译·程序运行

◆等待翻译完成。其中model为选择进行翻译的模型，此处选择的是第1000步时保存的模型，可以选择任意模型进行翻译，根据语料规模的不同，相对效果也不同，在训练时可以自行尝试选择最优模型。src为源语言测试集的路径，即待翻译文档存放在哪里。output为模型预测存放的路径，即模型给出的翻译存放在哪里。

◆采用bleu值进行效果评估，输入：

```
perl tools/multi-bleu.perl gu_bai/test_tgt.txt gu_bai/pred_1000.txt
```

◆等待评估完毕。其中前一个参数为参考翻译存放的路径，后一个为机器翻译存放的路径。

古白机器翻译•代码解析

◆本例中使用的代码来自于OpenNMT-py开源代码。

a）.build_vocab.py用于生成词表

b）.train.py用于训练翻译模型采用bleu值进行效果评估，输入：

c）.translate.py用于翻译文本