2019.8.25汇报-王鹏

1. 论文阅读
    1. 复现实验过程:Is word segmentation necessary for DL of Chinese representations?
2. Web Data Mining
3. CRF的学习使用-达观杯NER比赛
4. Github：https://github.com/RelativeWang/word2vec-study

1 复现实验准备

CTB6.0的下载需要LDC购买，东南和蒙纳士的数据库里都没有，浙大有但是需要浙大的账号，现在打算换其他数据集。

<table>
<tr><th>实验内容</th><th>数据集</th><th colspan="3">方法</th><th>评价指标</th><th>备注</th></tr>
<tr><td rowspan="5">Language Modeling</td><td rowspan="4">CTB
Train:
Validation:
test = 8: 1: 1</td><td colspan="3">用jieba分词
LSTMs用于编码char和word</td><td rowspan="5">Perplexity越小越好
计算方式一般为
$p(w_1, w_2, \ldots,$
通俗解释给定一个词，下一个词的多少种可能性</td><td rowspan="5">Char-only 的混合模型不太明白。是只用分词后的单词吗？</td></tr>
<tr><td rowspan="3">字词混合模型</td><td>1024+1024
2048+2048</td><td>Yin et al., 2016
Yu et al., 2017</td></tr>
<tr><td>1024+2048</td><td>CNNs保持维数一致</td></tr>
<tr><td>Char-only</td><td>分词后只用连续的字代表向量</td></tr>
<tr><td>https://catalog.ldc.upenn.edu/LDC2007T36</td><td colspan="3">还采用了Stanford CWS和LTP包分词，效果类似，但文中没有体现对比实验</td></tr>
</table>

1 复现实验准备

机器翻译的这个实验的模型没有找到模型的代码。

实验内容		数据集		方法	评价指标	备注
Machine Translation（设置：最多3万En，27500Ch，char-base vocab size 4500)	Ch-En	Train	1.25M Sentences Pairs from LDC2002E18等	Standard SEQ2SEQ+ attention SEQ2SEQ+ bag of words as targets when training Encode 阶段词变字 ref	验证集和评估集翻译的准确率	文中还用了BPE字词模型测试，可以自己测一下BLEU(改进的n-gram + brevity penalty) https://www.aclweb.org/anthology/P02-1040
		Validation	NIST2002			
		Evaluated	NIST2003-2006&2008			
	EN-Ch	Same train and test 512 dimensionalities ref				

1 复现实验准备

句子匹配的实验，BQ这个数据集正在申请，LCQMC这个数据集和在github上找到，BiMPM模型，训练有些慢，正在尝试GPU训练。

文本分类模型目前，还没有进行具体实验。

实验内容	数据集	方法	评价指标	备注
Sentence Matching/Paraphrase	BQ句子相似意思不同 LCQMC句子不同意思一样	Jieba分词 用模型 BiMPM	通过valid集合	
实验内容	数据集	方法	评价指标	备注
Text Classification	来自Ref的5个数据集 Description不同即train valid test 比例不同	Wordbased charbased of bi-directional LSTM models	通过valid集合	

2. Web data mining

8.25日看到3.7，朴素贝叶斯分类，这周完成了四节的内容。

3. CRF的学习使用-达观杯NER比赛

CRF是信息抽取的主流模型，目标函数考虑了输入的状态特征函数，同时包含了标签转移特征函数。

这样使得标注过程可以利用到内部特征和上下文特征信息。

HMM是生成模型，CRF是判别模型。

```
# Unigram
U00:%x[-3,0]
U01:%x[-2,0]
U02:%x[-1,0]
U03:%x[0,0]
U04:%x[1,0]
U05:%x[2,0]
U06:%x[3,0]
U07:%x[-2,0]/%x[-1,0]/%x[0,0]
U08:%x[-1,0]/%x[0,0]/%x[1,0]
U09:%x[0,0]/%x[1,0]/%x[2,0]
U10:%x[-3,0]/%x[-2,0]
U11:%x[-2,0]/%x[-1,0]
U12:%x[-1,0]/%x[0,0]
U13:%x[0,0]/%x[1,0]
U14:%x[1,0]/%x[2,0]
U15:%x[2,0]/%x[3,0]

# Bigram
B
```

对于命名实体识别问题:

左图是CRF的函数模板，
U(number):%x代表第(number)个特征，
%X[a, b]，a表示当前词的特征词，b表示用户自定义的特征。
其中[0, 0]代表了当前词，a代表了词的位置-3到-1是前三个词，1到3是后三个词。
后面的特征代表了，词和词之间的关系的特征。

3. CRF的学习使用-达观杯NER比赛

如下，标记的方法根据BEMOS，其中的数字代表得分，正数代表该字属于这个类别的分数高，负数代表该字属于这个类别的分数小。

```
"U06:待": [
    -0.0761171148843781,
    -0.3304252678324269,
    -0.0258093469791894,
    0.4334372103636684,
    -0.0010854806687564
],
```

通过特征模板，可以的到字和前几个字和后几个字特征，以及，不同字之间的特征。通过L-BFGS训练。最终得出模型告诉我们每个特征对于不同标签的值是多少。

解码的时候，将当前序列通过特征模板，然后去模型中乘上对应权重，最终可以得到一个得分向量，分别每个标签的得分，用维特比解码即可。

3. CRF的学习使用-达观杯NER比赛

比赛最终通过F1值，来确定通过自己的模型得到的结果的好坏。

成功提交(diindi是我的)后0.85，自己想通过在学习一下CRF的模板，看看能不能通过更好地特征模板得到更好地结果。

357	↓2	tuzi	0.85134	1
358	↓2	diindi	0.85134	3
359	↓2	lixinwuhahahaha	0.85125	1

F1值如下计算：

正确率 = 抽取出的正确字段数 / 抽取出的字段数
召回率 = 抽取出的正确字段数 / 样本的字段数
F1值 = （2 * 正确率 * 召回率）/（正确率 + 召回率）

3. CRF的学习使用-达观杯NER比赛

CRF是信息抽取的主流模型，目标函数考虑了输入的状态特征函数，同时包含了标签转移特征函数。

这样使得标注过程可以利用到内部特征和上下文特征信息。

HMM是生成模型，CRF是判别模型。