本周报告

2020/04/09 1652792 罗吉皓

本周主要做了机器学习模型的尝试,分为以下几个部分:

- 1. 动词-地名树准确率评估
- 2. NLP 基于kashgari和BERT实现中文命名实体识别

动词-地名树准确率评估

在上次搭建动词-地名树完成以后,在这两周的工作中,我基于学弟的数据(1000个)对于整体的性能和准确率进行了评估,结果如下:

	完全准确	地点准确但标注的位置错误	完全错误
数量	346	410	244

分析得到核心的问题如下:

- 1. 在地点相同的情况下,每个人对于哪一个才是核心地点的认识可能会有不同,导致最后地点准确但标注的位置错误的情况占据了半数以上
- 2. 测试样本中存在了一些例子,其中存在地理位置和事件动词,但是无法找到核心的事件发生地点的情况,在这种例子中,动词-地名树这种类似规则制定的方式是没有办法区分的,因此会造成很多的错误
- 3. 算法本身存在一定的缺陷,在句子中存在较多干扰信息,比如报道,说,声称等情况的时候,误差特别大

通过这个实验可以得到,基于规则的准确率最多也只能达到40%~50%左右的准确率,其缺点大致有以下几点:

- 1. 其应变能力较差,跳出规则的例子基本不能识别,为了达到较好的效果,需要制定比较完善的规则。因此在规则制定上投入的人力成本会相对比较高
- 2. 抗干扰能力差,虽然现在已经使用了文本相似度匹配等方式来进行干扰词的去除,对于一些报道等词语,不同语境下表达的意思可能相差比较大,无法做到面面俱到,因此对于最后的结果影响相对也比较大

NLP 基于kashgari和BERT实现事件核心地理位置识别

在之前动词-地名树方法的基础上,我们发现基于规则的地理位置命名方法存在一定的缺陷,尤其是不 灵活,大量依靠人工的特点对于问题的解决会产生比较大的限制,因此我们也希望通过机器学习的方 法来解决这些问题

数据集的处理

在数据集处理方面,我主要碰到的是两个问题:

1. 准确的,人工标注的数据集相对比较小(2000)

2. 如何将这些文本信息以及上下文的关联传入神经网络中

在之前的调研中,我拟选择RNN网络及LSTM为实验机器学习模型。主要原因是中文核心事件地点往 往和其他句子中的一系列元素是相互关联的。选择使用RNN网络及LSTM也是希望在整体的训练过程 中,将距离,词性等特征在训练的过程中将一些信息一层一层的传下去。

在查询了相关资料后,我借鉴了目前训练ner的方式。我将原来处理得到的数据进行二次处理,为每一个元素进行标注,如下图所示:

我对每一个字符都进行了相应的标注,在实验中,我目前使用的命名实体主要有四种:

- 1. T代表目标地点, 即事件核心地理位置
- 2. P代表地点,即在句子中出现的所有的地点
- 3. V代表动词,即在句子中的核心动词
- 4. O代表其他

模型搭建

1. 使用预训练词向量

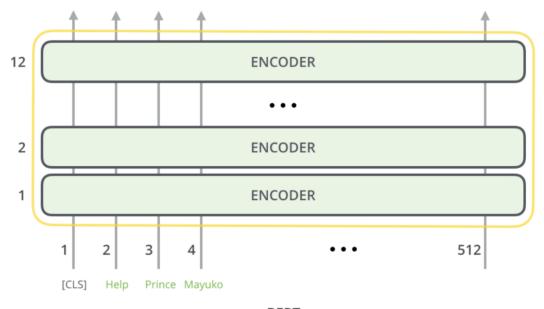
由于长新闻特征比较明显,语料量也比较大,很容易取得比较不错的结果。但是我们的语料比较少,特征不是很明显时候直接训练可能会导致模型过拟合,泛化能力很差,此时我选择使用预训练的词 Embedding 层来提高模型的泛化能力。

BERT 简介

BERT是2018年google 提出来的预训练的语言模型,并且它打破很多NLP领域的任务记录,其提出在nlp的领域具有重要意义。预训练的(pre-train)的语言模型通过无监督的学习掌握了很多自然语言的一些语法或者语义知识,之后在做下游的nlp任务时就会显得比较容易。

模型部分

首先BERT模型采用的是12层或者24层的**双向的Transformer**的Encoder作为特征提取器,如下图所示。在nlp领域,特征提取能力方面的排序大致是Transformer要比RNN和CNN的效果更好。



BERT

Transformer算法简介

Transformer引入了**self-attention**机制,同时还借鉴了CNN领域中残差机制(Residuals),由于以上原因导致transformer有如下优势:

- 模型表达能力较强,由于self-attention机制考虑到了句子之中词与词之间的关联,
- 抛弃了RNN的循环结构,同时借用了CNN中的残差结构加快了模型的训练速度。

2. 数据集再处理

3. kashgari框架搭建机器学习模型

Kashgari 是一个极简且强大的 NLP 框架,可用于文本分类和标注的学习,研究及部署上线。

- 方便易用: Kashgari 提供了简洁统一的 API 和完善的文档,使其非常方便易用。
- 内置迁移学习模块: Kashgari 通过提供 BertEmbedding , GPT2Embedding , WordEmbedding 等特征提取类,方便利用预训练语言模型实现迁移学习。
- 易扩展: Kashgari 提供简便的接口和继承关系,自行扩展新的模型结构非常方便。
- 可用于生产: 通过把 Kashgari 模型导出为 SavedModel 格式,可以使用 TensorFlow Serving 模块提供服务,直接在线上环境使用。

本次实验中我调用了BILSTM模型和CRF模型的累加,通过条件随机场对多个变量在给定观测值后的条件概率进行建模得到最后最有可能的标注。

相关模型截图在附录中显示,可以清晰的看到BERT的12层Transformer结构,以及它的参数量。最后是BILSTM和CRF层以及相关参数量。其训练完结果如下:

如这个例子中,虽然很好的找到了首尔作为整个句子的核心地理位置,但是标注的却是在新华网首尔8 月25日电其中的首尔,其中电,报道等词语的影响还是会有的

这个例子中,同其实不是核心动词,但是还是被标注了。

所以在下一次的改进中, 我有以下安排:

- 1. 训练集的修改。可能会把所有动词都标注出来,再标注核心动词,这样可以帮助他更好的发掘核 心动词
- 2. 模型的修改,可能会对相关信息进行进一步的调研来发掘相关模型
- 3. 模型的对比,对比相类似的模型,寻找最后解决方案
- 4. 数据集的进一步标注
- 5. 自己电脑训练模型耗费时间过长,可能会去寻找服务器帮助

相关网址记录

Attention is all your need 谷歌的超强特征提取网络——Transformer

基于BERT的NER模型

Kashigari文本标注

附录

Encoder-2-MultiHeadSelfAttent	io (None,	200,	768)	0	Encoder-2-MultiHeadSelfAttention[
Encoder-2-MultiHeadSelfAttent	io (None,	200,	768)	0	Encoder-1-FeedForward-Norm[0][0]
					Encoder-2-MultiHeadSelfAttention-
Encoder-2-MultiHeadSelfAttent	io (None,	200,	768)	1536	Encoder-2-MultiHeadSelfAttention-
Encoder-2-FeedForward (FeedFo	rw (None,	200,	768)	4722432	Encoder-2-MultiHeadSelfAttention-
Encoder-2-FeedForward-Dropout	((None,	200,	768)	0	Encoder-2-FeedForward[0][0]
Encoder-2-FeedForward-Add (Ad	d) (None,	200,	768)	0	Encoder-2-MultiHeadSelfAttention-
					Encoder-2-FeedForward-Dropout[0][
Encoder-2-FeedForward-Norm (La	ay (None,	200,	768)	1536	Encoder-2-FeedForward-Add[0][0]
Encoder-3-MultiHeadSelfAttent	io (None,	200,	768)	2362368	Encoder-2-FeedForward-Norm[0][0]
Encoder-3-MultiHeadSelfAttent	io (None,	200,	768)	0	Encoder-3-MultiHeadSelfAttention[
Encoder-3-MultiHeadSelfAttent	io (None,	200,	768)	0	Encoder-2-FeedForward-Norm[0][0]
					Encoder-3-MultiHeadSelfAttention-
					Elicode S Hazericaaserine celleroli
Encoder-3-MultiHeadSelfAttent	io (None.	200	768)	1536	Encoder-3-MultiHeadSelfAttention-
Encoder-3-FeedForward (FeedFo	rw (None	200	768)	4722432	Encoder-3-MultiHeadSelfAttention-
2201 21 11 11 11 1 1 1 1 1 1	,	,	1		
Encoder-3-FeedForward-Dropout	((None	200	768)	0	Encoder-3-FeedForward[0][0]
ean of mar a bropout	, ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,	,			
Encoder-3-FeedForward-Add (Ad	d) (None	200	768)	0	Encoder-3-MultiHeadSelfAttention-
Antoder 5 Cedi of Hail a Maa (Ma	, (none)		,		The state of the s
					Encoder-3-FeedForward-Dropout[0][
Encoder-3-FeedForward-Norm (La	ay (None,	200,	768)	1536	Encoder-3-FeedForward-Add[0][0]

Encoder-4-MultiHeadSelfAttentio	(None, 200	, 768)	2362368	Encoder-3-FeedForward-Norm[0][0]
Encoder-4-MultiHeadSelfAttentio	(None, 200	, 768)	0	Encoder-4-MultiHeadSelfAttention[
Encoder-4-MultiHeadSelfAttentio	(None, 200	, 768)	0	Encoder-3-FeedForward-Norm[0][0] Encoder-4-MultiHeadSelfAttention-
Encoder-4-MultiHeadSelfAttentio	(None, 200), 768)	1536	Encoder-4-MultiHeadSelfAttention-
Encoder-4-FeedForward (FeedForw	(None, 200), 768)	4722432	Encoder-4-MultiHeadSelfAttention-
Encoder-4-FeedForward-Dropout ((None, 200), 768)	0	Encoder-4-FeedForward[0][0]
Encoder-4-FeedForward-Add (Add)	(None, 200	, 768)	0	Encoder-4-MultiHeadSelfAttention- Encoder-4-FeedForward-Dropout[0][
Encoder-4-FeedForward-Norm (Lay	(None, 200), 768)	1536	Encoder-4-FeedForward-Add[0][0]
Encoder-5-MultiHeadSelfAttentio	(None, 200), 768)	2362368	Encoder-4-FeedForward-Norm[0][0]
Encoder-5-MultiHeadSelfAttentio	(None, 200	, 768)	0	Encoder-5-MultiHeadSelfAttention[
Encoder-5-MultiHeadSelfAttentio	(None, 200	, 768)	0	Encoder-4-FeedForward-Norm[0][0] Encoder-5-MultiHeadSelfAttention-
Encoder-5-MultiHeadSelfAttentio	(None, 200), 768)	1536	Encoder-5-MultiHeadSelfAttention-
Encoder-5-FeedForward (FeedForw	(None, 200), 768)	4722432	Encoder-5-MultiHeadSelfAttention-
Encoder-5-FeedForward-Dropout ((None, 200), 768)	0	Encoder-5-FeedForward[0][0]
Encoder-5-FeedForward-Add (Add)	(None, 200	, 768)	0	Encoder-5-MultiHeadSelfAttention- Encoder-5-FeedForward-Dropout[0][
Encoder-5-FeedForward-Norm (Lay	(None, 200), 768)	1536	Encoder-5-FeedForward-Add[0][0]

Encoder-6-MultiHeadSelfAttentio	(None, 200,	, 768)	2362368	Encoder-5-FeedForward-Norm[0][0]
Encoder-6-MultiHeadSelfAttentio	(None, 200)	, 768)	0	Encoder-6-MultiHeadSelfAttention[
Encoder-6-MultiHeadSelfAttentio	(None, 200)	, 768)	ø	Encoder-5-FeedForward-Norm[0][0] Encoder-6-MultiHeadSelfAttention-
Encoder-6-MultiHeadSelfAttentio	(None, 200)	, 768)	1536	Encoder-6-MultiHeadSelfAttention-
Encoder-6-FeedForward (FeedForw	(None, 200,	, 768)	4722432	Encoder-6-MultiHeadSelfAttention-
Encoder-6-FeedForward-Dropout ((None, 200,	, 768)	0	Encoder-6-FeedForward[0][0]
Encoder-6-FeedForward-Add (Add)	(None, 200,	, 768)	ø	Encoder-6-MultiHeadSelfAttention- Encoder-6-FeedForward-Dropout[0][
Encoder-6-FeedForward-Norm (Lay	(None, 200)	, 768)	1536	Encoder-6-FeedForward-Add[0][0]
Encoder-7-MultiHeadSelfAttentio	(None, 200,	, 768)	2362368	Encoder-6-FeedForward-Norm[0][0]
Encoder-7-MultiHeadSelfAttentio	(None, 200,	, 768)	0	Encoder-7-MultiHeadSelfAttention[
Encoder-7-MultiHeadSelfAttentio	(None, 200,	, 768)	ø	Encoder-6-FeedForward-Norm[0][0] Encoder-7-MultiHeadSelfAttention-
Encoder-7-MultiHeadSelfAttentio	(None, 200,	, 768)	1536	Encoder-7-MultiHeadSelfAttention-
Encoder-7-FeedForward (FeedForw	(None, 200,	, 768)	4722432	Encoder-7-MultiHeadSelfAttention-
Encoder-7-FeedForward-Dropout ((None, 200)	, 768)	0	Encoder-7-FeedForward[0][0]
Encoder-7-FeedForward-Add (Add)	(None, 200,	, 768)	0	Encoder-7-MultiHeadSelfAttention- Encoder-7-FeedForward-Dropout[0][
				1 27

ncoder-7-FeedFo	rward-Norm (Lay	(None, 200,	768)	1536	Encoder-7-FeedForward-Add[0][0]
ncoder-8-MultiH	eadSelfAttentio ((None, 200,	768)	2362368	Encoder-7-FeedForward-Norm[0][0]
ncoder-8-MultiH	leadSelfAttentio	(None, 200,	768)	0	Encoder-8-MultiHeadSelfAttention[
ncoder-8-MultiH	WeadSelfAttentio	(None, 200,	768)	0	Encoder-7-FeedForward-Norm[0][0]
					Encoder-8-MultiHeadSelfAttention-
Encoder-8-MultiH	HeadSelfAttentio	(None, 200,	768)	1536	Encoder-8-MultiHeadSelfAttention-
Encoder-8-FeedFo	orward (FeedForw	(None, 200,	768)	4722432	Encoder-8-MultiHeadSelfAttention-
Encoder-8-FeedFo	orward-Dropout ((None, 200,	768)	0	Encoder-8-FeedForward[0][0]
Encoder-8-FeedFo	orward-Add (Add)	(None, 200,	768)	0	Encoder-8-MultiHeadSelfAttention-
					Encoder-8-FeedForward-Dropout[0][
incoder-8-FeedFo	orward-Norm (Lay	(None, 200,	768)	1536	Encoder-8-FeedForward-Add[0][0]
Encoder-9-MultiH	eadSelfAttentio	(None, 200,	768)	2362368	Encoder-8-FeedForward-Norm[0][0]
Encoder-9-MultiH	HeadSelfAttentio	(None, 200,	768)	0	Encoder-9-MultiHeadSelfAttention[
Encoder-9-MultiH	HeadSelfAttentio	(None, 200,	768)	0	Encoder-8-FeedForward-Norm[0][0]
					Encoder-9-MultiHeadSelfAttention-
Encoder-9-MultiH	leadSelfAttentio	(None, 200,	768)	1536	Encoder-9-MultiHeadSelfAttention-
Encoder-9-FeedFo	orward (FeedForw	(None, 200,	768)	4722432	Encoder-9-MultiHeadSelfAttention-
Encoder-9-FeedFo	orward-Dropout ((None, 200,	768)	0	Encoder-9-FeedForward[0][0]
			100000000000000000000000000000000000000		

Encoder-9-FeedForward-Add (Add)	(None, 200, 7	768) 0		Encoder-9-MultiHeadSelfAttention- Encoder-9-FeedForward-Dropout[0][
				enesser 3 recard ward proposetojt
Encoder-9-FeedForward-Norm (La	(None, 200, 7	'68) 1	.536	Encoder-9-FeedForward-Add[0][0]
Encoder- <mark>10</mark> -MultiHeadSelfAttenti	(None, 200, 7	768) 2	362368	Encoder-9-FeedForward-Norm[0][0]
Encoder-10-MultiHeadSelfAttent	(None, 200, 7	768) 0	,	Encoder-10-MultiHeadSelfAttention
Encoder-10-MultiHeadSelfAttenti	(None, 200, 7	68) 0		Encoder-9-FeedForward-Norm[0][0]
				Encoder-10-MultiHeadSelfAttention
Encoder-10-MultiHeadSelfAttenti	(None, 200, 7	68) 1	536	Encoder-10-MultiHeadSelfAttention
Encoder-10-FeedForward (FeedFor	(None, 200, 7	(68) 4	722432	Encoder-10-MultiHeadSelfAttention
Encoder-10-FeedForward-Dropout	(None, 200, 7	68) 0		Encoder-10-FeedForward[0][0]
Encoder-10-FeedForward-Add (Add	(None, 200, 7	68) 0	,	Encoder-10-MultiHeadSelfAttention
				Encoder-10-FeedForward-Dropout[0]
Encoder-10-FeedForward-Norm (La	(None, 200, 7	68) 1	536	Encoder-10-FeedForward-Add[0][0]
Encoder-11-MultiHeadSelfAttenti	(None, 200, 7	68) 2	362368	Encoder-10-FeedForward-Norm[0][0]
Encoder-11-MultiHeadSelfAttenti	(None, 200, 7	68) 0		Encoder-11-MultiHeadSelfAttention
Encoder-11-MultiHeadSelfAttenti	(None, 200, 7	68) 0		Encoder-10-FeedForward-Norm[0][0]
				Encoder-11-MultiHeadSelfAttention
Encoder-11-MultiHeadSelfAttenti	(None, 200, 7	68) 1	536	Encoder-11-MultiHeadSelfAttention
Encoder-11-FeedForward (FeedFor	(None, 200, 7	(68) 4	722432	Encoder-11-MultiHeadSelfAttention

ncoder-11-FeedForward-Dropout	(None, 200,	768)	0	Encoder-11-FeedForward[0][0]
ncoder-11-FeedForward-Add (Add	(None, 200,	768)	0	Encoder-11-MultiHeadSelfAttention Encoder-11-FeedForward-Dropout[0]
ncoder-11-FeedForward-Norm (La	(None, 200,	768)	1536	Encoder-11-FeedForward-Add[0][0]
ncoder- <mark>12</mark> -MultiHeadSelfAttenti	(None, 200,	768)	2362368	Encoder-11-FeedForward-Norm[0][0]
ncoder- <mark>12</mark> -MultiHeadSelfAttenti	(None, 200,	768)	0	Encoder-12-MultiHeadSelfAttention
ncoder- <mark>12</mark> -MultiHeadSelfAttenti	(None, 200,	, 768)	0	Encoder-11-FeedForward-Norm[0][0] Encoder-12-MultiHeadSelfAttention
ncoder- <mark>12</mark> -MultiHeadSelfAttenti	(None, 200,	, 768)	1536	Encoder-12-MultiHeadSelfAttention
ncoder- <mark>12</mark> -FeedForward (FeedFor	(None, 200,	, 768)	4722432	Encoder-12-MultiHeadSelfAttention
ncoder- <mark>12</mark> -FeedForward-Dropout	(None, 200,	, 768)	0	Encoder-12-FeedForward[0][0]
ncoder-12-FeedForward-Add (Add	(None, 200,	, 768)	0	Encoder-12-MultiHeadSelfAttention Encoder-12-FeedForward-Dropout[0]
ncoder- <mark>12</mark> -FeedForward-Norm (La	(None, 200)	, 768)	1536	Encoder-12-FeedForward-Add[0][0]
on_masking_layer_4 (NonMasking	(None, 200,	, 768)	0	Encoder-12-FeedForward-Norm[0][0]
idirectional_3 (Bidirectional)	(None, 200,	, 512)	2099200	non_masking_layer_4[0][0]
ense_3 (Dense)	(None, 200,	, 128)	65664	bidirectional_3[0][0]
rf_3 (CRF)	(None, 200,	, 10)	1410	dense_3[0][0]