2018年3月26日 星期一

基于断言的开放域信息抽取式问答技术

- 摘要:

我们提出了基于断言的问答任务(ABQA)。ABQA是一项开放域自动问答任务，其输入为一个问题和一段文字，输出一个包含主谓宾三元组的半结构化断言。在阅读理解的任务中，断言可以比仅回答词组(answer span)传递更多的信息，但比返回整段乏味的文字来得精炼。以上特点使得ABQA更适合诸如语音交互的人机对话场景。如果想要进一步提升ABQA的准确率，我们需要为文本理解构建更丰富的监督数据集和强有力的模型。因此，我们构造了一个全新的数据集WebAssertions，它包含了经人工标注问答标签的55960个<问题，段落>对及从中产生的358427个<问题，断言>对。

同时，为实现ABQA，我们提出了生成式和抽取式两种方法。我们的生成式方法是基于端到端的学习策略：为捕捉输出断言的结构，我们使用了一种先产生断言结构，再向内填充词的层级解码器。而抽取式则建立在排序学习的基础上：我们构造了不同粒度的特征来表征问题与断言之间的语义相关性。实验结果显示，我们的方法可以实现从段落中抽取与问题相关的断言。在进一步评估时，我们将ABQA的结果作为特征整合到了基于段落的自动问答任务(PBQA)中，结果证明在两个数据集上，后者的准确率都有了明显的提高。

- 简介:

在NLP领域，实现计算机在开放域上的自动问答是一个长期任务。在这里，我们提出了基于断言的问答任务(ABQA)，其目的是用半结构化断言实现自动问答，而非类似机器阅读式的词组抽取(MRC)或答案抽取任务下的句子/段落匹配(PBQA)。

断言，指的是从问题所指的段落中推理得到的一组具有主谓宾结构的词。我们认为ABQA有很多优势。从工业界角度看，ABQA可以优化智能语音助手，诸如Amazon Echo，Google Home和Microsoft Invoke，因为它们的工作场景都是通过给出一个简洁但语义充分的表达来回答用户的问题。在这样的场景下，简单的几个词无法传递论据，而回答一整段话则太过冗余。从学界角度看，ABQA是一个研究解释性自动问答技术的潜在方向，因为它明确地揭示了段落中用以回答问题的知识。此外，ABQA还可以帮助提升其它类型自动问答任务的准确率，比如句子抽取式。基于对一系列断言节点的整合，我们还可以构造能实现详尽推理的断言网。

2018年3月26日 星期一

ABQA与PBQA、MRC任务都相关。尽管这些任务的输入都是<问题，段落>对，ABQA与其它任务也显著不同于其它任务。如表1所示，断言是将完整而简洁的信息用一定的结构组织起来。ABQA也与知识问答(KBQA)不同，后者通常是从大规模互联网文档中抽取知识。ABQA的目标是对文档的深度理解，并基于此作出自动问答。在知识问答中，表述形式的多样性使得知识到文档的直接关联变得富有挑战性。ABQA与开放信息(OIE)抽取有关，因为OIE的目标正是从文档中抽取全部的断言，只不过ABQA的最终目标不只是从问题和文档中抽取断言，更是正确地回答问题。

为深入研究ABQA，我们构建了一个人工标注集WebAssertions。该数据集中的问题和相应段落是从一个商业搜索引擎的查询日志中获取的，这样的做法关注了用户对信息的真实需求。对于每个<问题，段落>对，我们采用一种最先进的OIE算法抽取候选断言。标注者被要求标注出每个断言是否能正确、精简地回答相应问题。这个数据集总共包括经标注的55960篇互联网段落及其中的358427个断言。

我们解决ABQA有两种方法：生成式和抽取式。我们的生成式算法Seq2Ast是基于Seq2Seq learning的。它的扩展在于融入了一个分层解码器：先通过一个元组型解码器产生断言结构，然后再用另一个词级解码器为每个槽位产生填充词。抽取式算法则是基于排序学习的：通过精心设计的不同粒度的匹配特征对候选断言进行排序。

我们进行了两个实验。我们先测试了我们的两种方法在ABQA任务上的表现。结果显示，Seq2Ast的BLEU-4分数为35.76，超过了单纯的Seq2Seq模型。我们还测试了将ABQA的结果作为特征整合到PBQA时的效果。在两个数据集上都证明了ABQA特征可以显著提升PBQA的准确率。

总而言之，我们的成果总结如下：

1.我们提出了基于段落内容产生断言的ABQA任务，并为之构造了一个人工标注的数据集。这个数据集将会开源。

2.我们提出了融合层级解码器的生成式算法，也提出了另一种抽取式的方案。

3.我们进行了详尽的实验，并证明了我们的算法在ABQA和PBQA上都具备有效性。

- 任务定义及数据集构造：

在这一节，我们将给出ABQA的定义，并描述WebAssertions的构造。

- 任务定义：

2018年3月26日 星期一

给定一个自然语言问题q和一个段落p，ABQA的目的是输出一个能基于p来回答q的半结构化断言ast。每个断言表示为一个大小为3及以上的元组，包括一个主语、一个谓语及一个或多个宾语（这些语言成分都由1个或多个词构成）。

- 数据集构造：

由于目前没有可用的开源ABQA数据集，我们构造了人工标注集WebAssertions。表2展示了这个数据集的构造步骤。

在这儿我们想深入描述一些构造数据集时的重要细节。目前有若干种开源的OIE算法：TextRunner、Reverb和OLLIE等，而OIE算法的计算结果与我们的断言有着一致的结构。我们从语料中随机抽取了部分样本段，并将OIE算法通过实现工具应用到这些样本上。我们发现：ClausIE算法能产生较多针对这些问题的回答。ClausIE是一种无需训练数据的、基于规则的开源OIE算法。它的基石是一系列通过依存语法树解构得到的句子结构而预设的规则。更多关于ClausIE的细节，可以参考这篇论文。

我们使用一条简单的规则来提升断言的推理效力。我们相信ABQA是一个能够从文档中抽取可解释的问答与推理成分的好方法：与使用无法解释的深度神经网络策略的问题-文档匹配任务不同，结构化的断言揭示了部分文档中包含的用来回答问题的知识。请记住以上假设，由此，本文中我们初步尝试对文档中提取到的初级断言进行二次升级：考虑“is-a”结构，并对它进行拓展——假设已经抽取到两个断言<A, is, B>和<A, pre, C>，我们就会构造一个新断言<B, pre, C>。你可以在表3中看到一个例子：我们利用第一和第二个断言构造了第四个新断言。同时，表3也描述了人工标注的结果。表4给出了对这个数据集一些成分的统计描述。

- 基于断言的问答任务ABQA:

在这一节，我们将给出分别用生成法和抽取法来完成ABQA的过程。

- 生成法Seq2Ast：

我们提出了一种序列-断言的生成法来完成ABQA任务。这种方法是基于已经在众多自然语言生成任务上取得优异表现的Seq2Seq技术的。Seq2Seq框架中包含一个编码器和一个解码器。编码器负责将输入的序列编码成一系列隐式向量，而解码器则通过在每一个时间步上输出一个词来有序生成另一组序列。

2018年3月26日 星期一

ABQA的特点在于其输出是一个由一系列槽位及填充到槽位中的一组词构成的断言。为此，我们设计了先通过一个元组型解码器产生断言结构中的各个槽位，然后再用另一个词级解码器为每个槽位产生填充词的层级解码器。从算法整体看，一层的元组解码器负责记忆断言结构，二层词解码器负责学习槽位中的短信息。

技术上，元组解码器利用GRU-RNN模型来产生槽位表示。在元组解码器之上，我们用另一个GRU-RNN模型来搭建词解码器。图2描述了Seq2Ast的结构。这其中，元组解码器利用词解码器第k-1位信息和自身的第k-1位信息，通过GRU生成其第k位信息的隐藏状态（详见论文公式1）。

我们认为槽位是一种能够帮助预测序列中每个词的全局信息，因此我们也把元组解码器的第k位信息添加给词解码器，作为后者预测当前槽位中词的特征之一。我们可以这样表述第k个槽中第j位词的产生：该词通过第k个槽中第j-1位词的信息和k槽的信息被预测。同时，词解码器中还添加了注意力机制，以便从输入信息中抽取最重要的内容。对于含有罕见词的情况，我们设计了简单而有效的拷贝机制：使用输入信息中注意力得分最高的词直接代替产生的罕见词。

除了解码器，在编码器端，我们使用了双向GRU-RNN。整个模型的输入是一串用特殊字符将段落与问题连接起来形成的序列。

整体来说，整个模型是通过后向传播算法进行端到端训练的，旨在最大化为给定<问题，段落>对选取到正确断言的概率。在进行实验时，Seq2Ast的参数是随机初始化的，并通过AdeDelta算法进行更新。

- 抽取法ExtAst：

抽取法是通过对不同粒度的特征进行排序学习，从而使模型从一系列候选断言中选出排序最高的那些断言。我们的抽取法包含以下3步：I.与上文所述数据集的构建过程一致，获取全部候选断言；II.抽取基于问题的匹配特征；III.对候选断言进行排序。

- 抽取基于问题的匹配特征：

我们从3个不同粒度构建了衡量问题q与断言ast语义相关性的特征：

1.词级别：我们使用了词匹配特征Fwn和词级翻译特征Fw2w。Fwn的依据是如果一个断言和一个问题相关，那么它们一定有很高的词重合度。因此Fwn就是问题和断言间公共词的个数。Fw2w表示基于翻译的词到词特征，这一问题与断言的相似性特征通过IBM model 1计算所得。其中，词对齐的概率通过GIZA++提供的11.6M相似句对训练得到。

2.词组级别：我们设计了一个基于涵义的特征Fpp和词组到词组翻译特征Fp2p来解决断言与问题可能对同一对象使用不同表述方式的情况。这两种特征的相同点是都基于现有统计机器翻译算法抽取到的词组列表，不同点则在于Fpp的词组列表是从0.5M中英双语对中抽取的，而Fp2p的词组列表是从4M问答对中抽取的。

3.句级别：我们构造了基于CNN的特征Fcnn和基于RNN的特征Frnn来对问题和断言进行匹配。Fcnn特征是基于CDSSM模型得到的，这一模型曾在句子匹配任务上表现优良。该模型通过两个独立的CNN模型生成问题向量和断言向量，并用cosine函数计算它们的相似度。至于Frnn，我们先使用两个RNN来分别定长向量化问题与断言，然后用相同的双向GRU来提取它们的向量表示。以问题表示为例，当前词的向量表示Eq,t使用上一个时间步Hq,t-1递归转化得到；在向量表示层，我们将最后四个隐藏状态向量与双向逐一生成的向量拼接作为最终的向量表示。之后，我们将<问题，断言>对向量传入到一个全连接神经网络中。

Fcnn与Frnn的参数通过随机梯度下降算法在4M问答数据上训练得到。每一组问答对的排序损失(margin ranking loss)通过公式4计算所得，其中f+(q, ast)和f-(q, ast)分别是模型计算的问答对相似/不相似得分，m是margin。

- 候选断言排序：

我们使用一种解决真实世界排名问题的算法LambdaMART来计算每个<问题，断言>对的排序得分。LambdaMART的基本思想是：构建一组决策树，输出它们结果的线性组合。决策树上的每个分支表征了应用某一特征的阈值，而每个叶子结点都是一个真实值。特别地，对于一个包含N棵决策树的组，公式5计算了它赋予一个<问题，断言>对的得分。在公式5中，Wi是第i棵回归树的权重，tr(·)是通过使用特征[f_1(q, ast), ..., f_K (q, ast)]对第i棵树进行估计后叶子结点获得的值。Wi的值和tr(·)的参数通过梯度下降算法训练得到。

- 实验：

接着，我们将汇报分别针对ABQA和A-PBQA任务实验的环境设置及经验结果。

- ABQA任务：

我们先测试了生成法和抽取法在ABQA任务上的表现。在本次试验中，我们将WebAssertions数据集按80:10:10随机划分为训练、开发和测试集。开发集用于参数调

试，测试集用于结果生成和汇报。测试集包含了来自5575个<问题，段落>对的36165个<问题，段落，断言>三元组。

首先是生成法。我们使用BLEU-4分数作为评价指标：该指标计算的是生成断言与参考断言之间的ngram匹配程度。我们的模型与标准Seq2Seq、Seq2Seq+attention模型都进行了比较。表6展示了这些结果。可以看到，Seq2Ast超过了标准Seq2Seq，证明了层级解码器的有效性。顺便一提，尽管抽取法不适合用来做对比，但是我们也计算了抽取法的BLEU-4分数：72.27——即便对于文本生成任务来说也是相当高的一个分数了。不过从道理上，这也是可以解释的，因为抽取法是从包含正确结果在内的候选断言中去选出最可能的。因此，对于正确的排序结果而言，BLEU-4分数应是100。在下文，我们会汇报在进一步篇章级问答任务中抽取法和生成法取得的结果。

我们将抽取法视作一个排序问题：为给定<问题，段落>对的候选断言排序，并选出能够以最大概率来正确回答该问题的断言。因此，我们选用Top1准确率、平均准确率和MRR(Mean Reciprocal Rank)来评价我们模型。

我们使用控制变量法来研究以上特征在抽取法中的效果：不出意外，句级特征能够对<问题，断言>对的全局语义相关性进行建模，因此效果远高于词/词组级特征。最终，我们的ExtAst综合了所有特征，以达到最佳效果。

表5展示了我们的生成法和抽取法在样例上的表现。可以看到，生成模型拥有生成槽位、槽填充和在一定程度上形成完整断言句子的能力。在这个例子中，生成法的结果甚至比抽取法还要精准。不过，生成法还可以进一步优化：1.经研究，我们发现槽填充时短语的流畅与否不是特别重要；2.生成法当前最大的问题是会产生重复的内容及和问题无关的断言。对于第二个问题，重复性可以通过能明确记忆是否来源词已经被复制的复写机制来解决，无关性则可能需要涉及深度问题理解和设计一个由问题深度驱动的解码器。

我们也对抽取法进行了错误分析，并将主要错误来源总结成以下三类。第一类是关键信息缺失，比如说为问题“When were the Mongols defeated by the Tran?”抽取得到的断言是一个逻辑通顺但却没有包含任何时间信息的断言。第二类问题是问题中的实体与给定段落的同一实体采用了不同的表述方式。指代消解类问题都应当归入这类问题中。第三类问题就是模型逻辑推理能力不足，比如无法处理“not”问题：“Which is the largest city not connected to an interstate highway?”。

- A-PBQA任务：

通过将ABQA的结果应用到PBQA任务上，我们进一步评估了上述的ABQA算法，并使用在PBQA上端到端的表现来衡量我们算法的效果。要指出的是，这里PBQA的任务要求是将<问题，段落>对作为输入，然后从段落中抽取一个句子作为答案。

当给定<问题，文档>后，我们先使用ABQA算法(抽取法和生成法)来输出最佳断言。之后，将<问题，断言>对作为额外的特征添加到原有的PBQA算法的特征向量中。此处特征集与我们在抽取法中使用的完全一致。PBQA的基础特征包括：一个基于问题和段落中共现词的词级特征，以及一个通过CNN将问题和段落编码成连续向量的句级特征。对于PBQA，我们也同样使用LambdaMART算法进行排序训练。排序模型中的特征权重是通过SGD分类器在训练集上获得的——这个训练集由标注好的<问题，句子，标签>三元组组成，其中标签表示是否这个句子是该问题的正确回答。

下文总结了我们的ABQA算法在WikiQA和MARCO这两个数据集上的表现。这两个数据集与我们的WebAssertion数据集一脉相承，因为它们都来源于搜索引擎中真实的用户查询数据。WikiQA是一个句子选择型问答任务方面的基准数据集，它是基于自然语言问题与Wikipedia文档而精准构建的。该数据集也由训练集、开发集和测试集构成，分别包含20360、2733和6165条数据。而MARCO数据集一开始是为阅读理解任务构造的，不过它也可用于段落排序。在MARCO数据集中，问题和候选段落都来自Bing的搜索日志。如果一个段落包含支持回答给定问题的证据，标注者就会给这个段落标1。由于MARCO的测试集答案不对公众开放，我们就将原有的验证集随机分割为开发和测试集。在本文中，我们仅使用段落排序的信息来测试我们的模型。MARCO数据集本身由训练集、开发集和测试集构成，分别包含676193、39510和42850条数据。在这个实验中，我们沿用平均准确率和MRR作为评价指标。遵照其它发表作品的惯例，在计算以上指标时，我们排除了那些所有候选答案全正和全负的数据。

图8展示了我们将PBQA的不同算法一起进行比较的结果。这两个数据集的基准模型结果可以在之前的论文中查到。CNN+Cnt模型是将bigramCNN与基于逻辑回归的共现词模型相结合。LSTM+Att+Cnt则组合了基于注意力的LSTM和基于逻辑回归的共现词模型。ABCNN直接使用了基于注意力的CNN模型，这一模型已经被证明在多个句子匹配任务中有出色表现。Dual-QA把问答与句子生成作为双重任务进行处理，并给出了MARCO数据集上ABCNN模型的运行结果。IARNN-Occam模型是一个基于内部注意力机制的RNN模型，而conv-RNN则是融合了CNN与RNN的混合模型。CNN+CH模型则混合了CNN与一些卷积树核特征。另外，正如之前所述，我们的基准模型中包含了一个基于重叠词的词特征和一个基于CDSSM的句特征。

最后，要补充说明我们在PBQA上的其它实现策略。除了抽取法和生成法，我们可以使用OIE工具从段落中抽取全部断言，然后把这些断言整合成特征传递给PBQA。这一思路有两种实现方法。RndAst是指我们随机挑选一个断言，并使用它来计算额外的断言特征。MaxAst则近似于CNN中的最大池化法：我们先产生所有候选断言的特征向量，然后在每个维度选择最大值组成最终的特征向量。最后，可以看到，我们的A-PBQA策略显著提升了基准模型的效率，特别是基于抽取法的A-PBQA模型。

- 相关工作介绍:

总的来说，我们的工作涉及了开放信息抽取OIE、开放知识问答、PBQA以及机器阅读理解等领域的内容。

ABQA任务与机器阅读理解MRC的相关性在于它们的输入都是<问题，段落>对，不同点则在于ABQA的输出是一个包含完整、精准信息的半结构化断言，而MRC的输出则是短词组形式的答案。ABQA与答案为长段落形式的PBQA也不相同。上文提出的抽取法与PBQA的现有工作相关：LCLR模型中使用了一些从包括WordNet、PILSA模型和不同向量空间模型在内的工作中收集到的富含词义、语义的特征；ABCNN是基于注意力机制的CNN模型，它首次计算了相似度矩阵并将之作为了CNN的新通道。此外，最近的研究还探索了使用问题生成技术来提升整个问答系统的效果。

开放信息抽取OIE负责从自然语言文本中无监督地抽取<主语，谓语，宾语>三元组式的断言。TextRunner是一项OIE方面的先进研究，旨在构建一个通用模型来表述基于词性和组块(Chunking)特征的关系。ReVerb能够把谓语限定在动词短语的范围内，并基于语法结构再进行抽取。ClausIE使用基于依存树的人工语法模版来检测和抽取基于结构的断言。本文的工作与OIE有所不同，因为我们的最终目标并不仅仅是从问题和文档中抽取断言，更是要能正确回答问题。值得一提的是，我们的生成法还能产生源文档中没有的词。

知识问答KBQA有两条研究思路。一是使用本地知识库来回答自然语言问题。这条思路的核心是将自然语言问题与预搭建的本地知识库中结构化的对应知识联系起来。另一种思路是使用开源OIE技术从互联网语料中获取大规模的开放知识。为了实现KBQA，Fader研制了第一个能够从大规模语料中学习到问答知识的开源KBQA系统。OQA系统则是一种融合了本地知识与开放知识的问题解决系统。TAQA是一个通过处理n-tuple断言来回答具有复杂语义限制问题的开源KBQA系统。TUPLEINF系统能够回答一些复杂的问题并搜寻断言中的最佳子集，因为它内置了一个能够基于OIE知识进

行推理的线性程序优化模型。ABQA和KBQA的区别在于断言/知识是从文档中提取的，且ABQA的重点是基于这些断言/知识进行文档理解与问题回答。KBQA一般都依赖于预先完成的本地知识库或大规模互联网文本知识抽取，但我们的ABQA仅从当前给定的问题和文档中去推理知识。

- 总结:

在本文中，我们介绍了ABQA，一项基于从文档中推理出的半结构化断言来完成的开放域QA任务。我们为此构建了WebAssertions数据集并开发了生成法和抽取法两种模型。我们还完成了一些拓展性实验，结果显示我们的ABQA方法确实能够从文档中抽取问题相关的断言。我们也进一步将ABQA的结果作为额外的特征整合到PBQA任务上，这一做法大大提升了PBQA基准模型的准确率。在未来工作中，我们计划提升现有模型的理解和推理能力，以便跨句形成的断言能用于最终答案的生成。