ATOMIC

作者：李一飞 时间：2019-07-22

本文主要提出了一个名为ATOMIC的自动生成的常识知识图谱，含有877k个文本描述的推理得来的常识tuple。将if-then关系划分为9个维度来生成后件（暂且这么称呼）。文章还使用神经网络模型在ATOMIC上进行了训练，使模型可以对unseen event给出reason和effect。通过自动评价（BLEU指标）和人工评价验证了模型的效果，并验证了多任务模型比单任务效果要好。

Introduction

本文主要聚焦于if-then形式的知识，目标是创建一个scale、coverage、quality三方面均可靠的知识库。文章提出了一种新颖的方式对if-then关系的类型进行分类。一种是基于将要发生的内容，可以分为：(1) *If-Event-Then- Mental-State*, (2) *If-Event-Then-Event*, (3) *If-Event-Then-Persona*。另一种是基于关系，可以分为：(1) “causes”, (2) “effects”, (3) “stative”。与传统的从已有的知识库中抽取知识的做法不同，作者采用了众包分配任务的方式，按类型来获取知识。接下来，作者尝试了用神经网络模型在ATOMIC数据集上跑，模型可以学得对没见过的event进行原因推理的能力。

If-Then Relaton Types

如上文提到的，文章介绍了两种分类方式，一种是按mental-state、event、persona分，另一种是按causes、effects、stative分，并且其中也细分为了九种情况，文章称为9个维度。具体如下图。

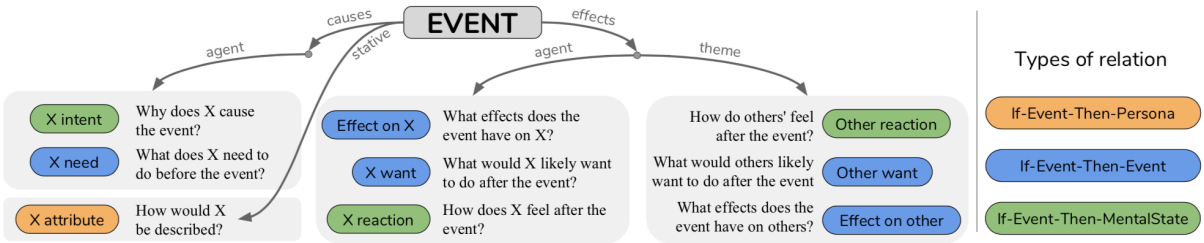

Figure 2: The taxonomy of *if-then* reasoning types. We consider nine *if-then* relations that have overlapping hierarchical structures as visualized above. One way to categorize the types is based on the type of content being predicted: (1) **If-Event-Then-Mental-State**, (2) **If-Event-Then-Event**, and (3) **If-Event-Then-Persona**. Another way is to categorize the types based on their causal relations: (1) **“causes”**, (2) **“effects”**, and (3) **“stative”**. Some of these categories can further divide depending on whether the reasoning focuses on the “agent” (X) or the “theme” (Other) of the event.

Data

作者从故事、书籍、Google Ngrams和Wikitionary等来源中抽取了24K个常见的动词短语作为base events。对于不常出现的名词改用一个抽象名词来替代。并且对动作的参与者进行了消歧处理。作者创建了一个众包任务（挂在亚马逊平台上，分发给许多人做）。每个phrase由三个人共同判断，且至少由两人判定为Valid才算可以。接下来就是介绍了一些在亚马逊上分配任务的细节。下图为所得到数据的情况：

	Count	#words
# triples: If-Event-Then-*	877,108	-
- Mental-State	212,598	-
- Event	521,334	-
- Persona	143,176	-
# nodes: If-Event-Then-*	309,515	2.7
- Mental-State	51,928	2.1
- Event	245,905	3.3
- Persona	11,495	1.0
Base events	24,313	4.6
# nodes appearing > 1	47,356	–

Table 2: Statistics of ATOMIC. Triples represent distinct <event, relation, event>. #words represents the average number of words per node.

Methods

这里文章将其看作是一个常规的序列生成问题，使用encoder-decoder模型对输入的序列进行处理。（*回去补一下Seq2Seq的论文，详细了解一下encoder-decoder模型*）

Encoder：输入 $e = \{e_0, e_1, \ldots, e_{n-1}\} \in R^{n \times i_{enc}}$。将输入序列e压缩至hidden空间：$f_{enc} : R^{i \times h_{enc}} \rightarrow R_h$。

这里采用了300维的静态GloVe预训练的Embedding（*这个回去需要详细看一下*），并用1024维的ELMo预训练的Embedding与其拼接，增强一下输入数据。

Decoder：用的是维度为h_{dec}的单向GRU，目的是生成序列$t = \{t_0, t_1, \ldots\}$，对于每个ti，相当于最大化$p(t_{i+1}|h(i), t_0, \ldots, t_i) = softmax(W_o \times GRU(h(i), t_i) + b_o)$。

单任务 vs 多任务：这里作者用单任务和多任务的模型分别做了几组对比，如下：

- EVENT2(IN)VOLUNTARY： 相当于按照是否是Voluntary的关系，将9个维度中属于同一类的进行合并归类。此模型包含1个encoder和4个decoder负责处理4种Voluntary的情况，以及另一个encoder和5个decoder负责处理Involuntary的情况。这里的数量和Figure 2中列举的9个维度可以对应上（其实就是九种关系）。
- EVENT2PERSONX/Y： 和上一条类似，按照对象是"agent" (自己) 和 "theme" (旁人) 为划分依据。模型包含1个encoder和6个decoder负责 "agent"，以及另一个encoder和3个decoder负责 "theme"。
- EVENT2PRE/POST：这个是按照关系的影响划分的。省略了xAttr这个维度，使用了2个encoder和8个decoder来表示8种关系。
- 9ENC9DEC：为9个维度单独训练9个encoder-decoders。

(* Mark: Nearest Neighbor算法也要补一下 *)

训练细节：将数据集按训练集，验证集，测试集80%，10%，10%的比例划分。将300维的GloVe词向量和1000维的ELMo词向量拼接组成1324维的向量作为encoder的输入。h_{enc}和h_{dec}均为100。

Results

作者对模型生成之前没见过的event的解释能力进行评估。用了Automatic的评分（比如BLEU），结果表明多任务的模型效果比单任务更好。还用了人为的评估，让志愿者对模型给出的top-10结果进行评估，将他们认为合理的设为valid，得到了86.2%的准确率。最后还与ConceptNet的工作进行了对比（***ConceptNet可能也需要补一下***）

Conclusion

本文主要是将if-then划分为了几种类型，并且建立众包任务分发给志愿者们完成了知识库的构建工作，建立了目前最大的推断知识库。另外，文章还使用神经网络模型，从KG中学习，获得对之前未出现过的event给出原因解释和影响的能力。