COMET

作者：李一飞 时间：2019-07-22

本文提出了一种能够用于自动构建常识知识图谱的Commonsense Transformer。利用一些pre-trained的深度学习语言模型，通过训练它们，使其能够生成常识知识。文章在ATOMIC和ConceptNet两个数据集上做了实验，分别达到了77.5%和91.7%的准确率，基本达到了人类的水平，并且提出利用这种生成commonsense的模型来自动构建常识知识图谱可能很快就会成为extractive method（*这里是指关系抽取的方法吗？*）的合理替代方案。（*这块没太懂，KG constructing和relation extraction有什么联系？*）

1 Introduction

本文提出了COMmonsEnse Transformer (COMET)，利用已有的知识三元组作为seed set进行训练，使得pre-trained的语言模型将其学到的表示向知识生成迁移，并且生成高质量的新tuple。

主要的Contribution：

1. 提出了一个生成的方法（generative method）来构建知识库。模型通过学习产生新节点
2. 开发了一套框架，利用大规模的transformer来学习生成常识知识三元组(commonsense knowledge tuple)
3. 对利用我们的方法在两个知识库（ATOMIC和ConceptNet）上生成的常识知识的质量、新颖程度、多样性进行了实证研究，以及训练处一个有效的模型所需的seed tuples的数量问题。

2 Learning to Generate Commonsense

按照文中所说，COMET是一套adaption framework，在由knowledge tuple构成的seed set上进行训练。这些seed tuples为COMET提供了KB的结构信息和关系信息。COMET学习去将pre-trained的语言模型学到的representation迁移到生成新的知识、关系上，作为节点和边，添加到seed KG中去。

任务：对于形如$\{s, r, o\}$ (s-subject, r-relation, o-object) 的三元组，给定s和r，模型需要生成o。

文章中说COMET和语言模型是无关的，本文选用了GPT模型（这个还没看，需要之后补一下）。下文基本也是介绍了一下用到的Transformer和Multi-headed Attention的具体结构，看起来和GPT原模型没太大区别。

Encoder部分，模型将X^s, X^r, X^o拼接起来作为输入：$X = \{X^s, X^r, X^o\}$。其中X^s, X^r, X^o均是一个序列，由若干个词构成。对于X中的每个词，还加了一个位置信息：$h_t^0 = e_t + p_t$，其中e_t是序列中每个词的embedding，p_t将当前词的位置信息编码到和embedding同维度的向量中，将二者相加作为输入。

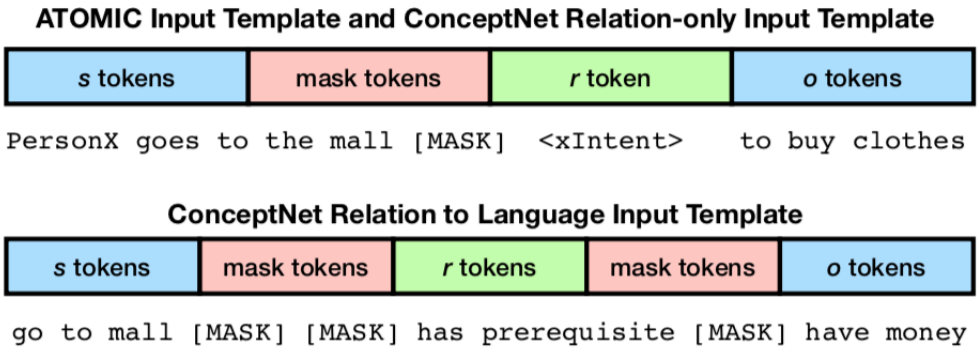

Figure 3: Input token setup for training configurations. For the ATOMIC dataset, the tokens of the subject, X^s (e.g., PersonX goes to the mall) are followed by masking tokens, which is followed by a single relation token X^r (e.g., xIntent), and then the object tokens X^o (e.g., to buy clothes). The model receives the same input for ConceptNet, except that a second set of masking tokens separate X^r and X^o because X^r can have a variable number of tokens for ConceptNet (§5.2)

（*这个图片里的mask没懂是干什么的。。貌似是和数据集有关系？*）

3 Training COMET

目的是给定s和r，使得模型生成o。因此，模型将X^s与X^r的拼接作为输入，然后输出序列X^o。

Loss Function: $\mathcal{L} = -\sum_{t=|s|+|r|}^{|s|+|r|+|o|} \log P\left(x_t | x_{<t}\right)$。

Datasets：ATOMIC和ConceptNet作为seed set。文章说也可以用其他数据集，因为COMET是domain-agnostic的。

Initialization：基本和GPT训练好的保持一致。一些新添加进去的比如oReact等是由标准正态分布sample出来的。

Hyperparameters：和GPT一样，12层，hidden维度是768，12个attention heads。用了0.1的dropout，GeLU作为激活函数。训练时的batch size是64。

4 Experiments

本部分介绍在ATOMIC和ConceptNet上做实验的一些细节。

自动化评价指标使用了BLEU-2指标、全新的sro tuple比例 (% N/T sro)、新生成的object比例 (% N/T o)、新生成的unique的object比例 (% N/U o)。手工评价在Amazon Mechanical Turk上建立了一些任务，志愿者们对其进行评价。COMET用的是Transformer，文章与只使用seq2seq的模型进行了对比（其实就是ATOMIC里边最初用到的那个），以及自己模型不用pre-trained好的参数做对比。结果是COMET相比于seq2seq、以及未pre-trained的模型，效果均有较大的提升。

结果：较ATOMIC那些基础的模型，整体上都有很大提升。自动评价指标有51%的提升。人工评价的指标也有18%的提升。COMET生成的tuple (object) 在质量和数量上均有较大提升。使用pre-trained的参数和随机初始化的参数相比，效果有14%的提升，说明通过GPT模型学到的自然语言表示的信息可以被迁移到生成自然语言的常识之水中去。（***什么是beam search和gold ATOMIC (distribution)?***）另外，模型只用10%的数据训练，效果也说得过去，并且使用pre-trained语言模型的COMET效果确实要比未使用pre-trained的要好。

ATOMIC：

Model	PPL[5]	BLEU-2	N/T sro[6]	N/T o	N/U o
9ENC9DEC (Sap et al., 2019)	-	10.01	100.00	8.61	40.77
NearestNeighbor (Sap et al., 2019)	-	6.61	-	-	-
Event2(IN)VOLUN (Sap et al., 2019)	-	9.67	100.00	9.52	45.06
Event2PERSONX/Y (Sap et al., 2019)	-	9.24	100.00	8.22	41.66
Event2PRE/POST (Sap et al., 2019)	-	9.93	100.00	7.38	41.99
COMET (- pretrain)	15.42	13.88	100.00	7.25	45.71
COMET	**11.14**	**15.10**	100.00	**9.71**	**51.20**

Table 1: Automatic evaluations of quality and novelty for generations of ATOMIC commonsense. No novelty scores are reported for the NearestNeighbor baseline because all retrieved sequences are in the training set.

Model	oEffect	oReact	oWant	xAttr	xEffect	xIntent	xNeed	xReact	xWant	Avg
9Enc9Dec (Sap et al., 2019)	22.92	32.92	35.50	52.20	47.52	51.70	48.74	63.57	51.56	45.32
Event2(In)voluntary (Sap et al., 2019)	26.46	36.04	34.70	52.58	46.76	61.32	49.82	71.22	52.44	47.93
Event2PersonX/Y (Sap et al., 2019)	24.72	33.80	35.08	52.98	48.86	53.93	54.05	66.42	54.04	46.41
Event2Pre/Post (Sap et al., 2019)	26.26	34.48	35.78	52.20	46.78	57.77	47.94	72.22	47.94	46.76
COMET (- pretrain)	25.90	35.40	40.76	48.04	47.20	58.88	59.16	64.52	65.66	49.50
COMET	**29.02**	**37.68**	**44.48**	**57.48**	**55.50**	**68.32**	**64.24**	**76.18**	**75.16**	**56.45**

Table 2: Human score of generations of ATOMIC commonsense. We present comparisons to the baselines from Sap et al. (2019). Underlined results are those where COMET is not significantly better at $p < 0.05$

ConceptNet：

Seed	Relation	Completion	Plausible
piece	PartOf	machine	✓
bread	IsA	food	✓
oldsmobile	IsA	car	✓
happiness	IsA	feel	✓
math	IsA	subject	✓
mango	IsA	fruit	✓
maine	IsA	state	✓
planet	AtLocation	space	✓
dust	AtLocation	fridge	
puzzle	AtLocation	your mind	🤔
college	AtLocation	town	✓
dental chair	AtLocation	dentist	✓
finger	AtLocation	your finger	
sing	Causes	you feel good	✓
doctor	CapableOf	save life	✓
post office	CapableOf	receive letter	✓
dove	SymbolOf	purity	✓
sun	HasProperty	big	✓
bird bone	HasProperty	fragile	✓
earth	HasA	many plant	✓
yard	UsedFor	play game	✓
get pay	HasPrerequisite	work	✓
print on printer	HasPrerequisite	get printer	✓
play game	HasPrerequisite	have game	✓
live	HasLastSubevent	die	✓
swim	HasSubevent	get wet	✓
sit down	MotivatedByGoal	you be tire	✓
all paper	ReceivesAction	recycle	✓
chair	MadeOf	wood	✓
earth	DefinedAs	planet	✓

Table 7: **Randomly selected and novel** generations from the ConceptNet development set. Novel generations are *sro* tuples not found in the training set. Manual evaluation of each tuple indicates whether the tuple is considered plausible by a human annotator

Conclusion

本文提出了一种Commonsense Transformer，用于自动构建常识知识三元组。模型个在ATOMIC和ConceptNet上的自动、人为评价指标都达到了很好的效果。