

2019.8.16汇报-王鹏

1. 论文阅读

1. Is word segmentation necessary for DL of Chinese representations?
2. COMET : Commonsense Transformers for Automatic Knowledge Graph Construction (ACL2019)
3. Fine-Grained Entity Typing in Hyperbolic Space

2. Web Data Mining

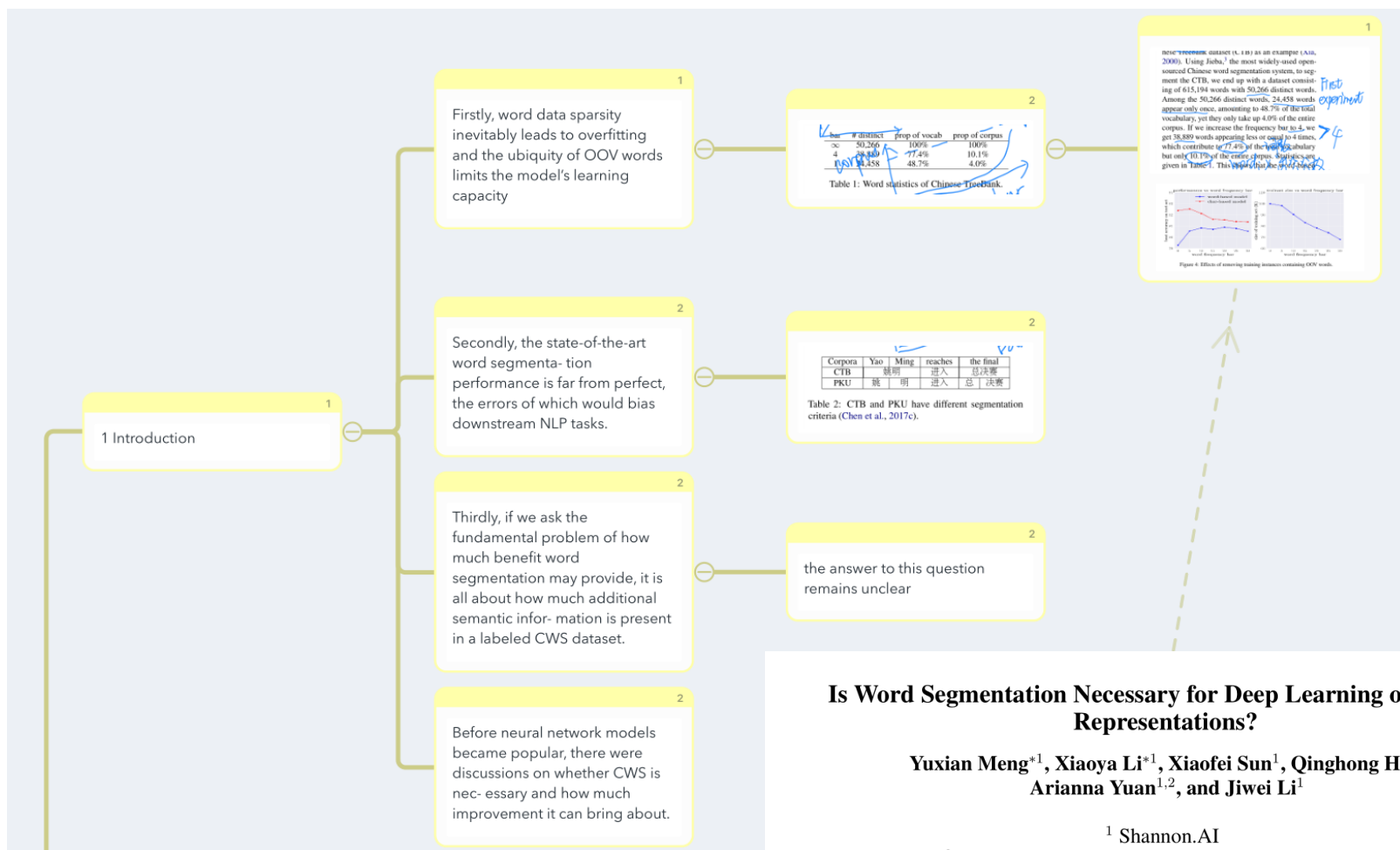
3. NLP 基础知识

4. 参加达观杯NER比赛的体会

5. Github: <https://github.com/RelativeWang/word2vec-study>

1.1 Is word segmentation necessary for DL of Chinese representations?

Introduction中说明词库稀疏性会导致过拟合并且OOV会限制模型的学习能力；分词标准不同会产生不同分词结果；而且分词后多少语义信息留在词中也并不明确。



Is Word Segmentation Necessary for Deep Learning of Chinese Representations?

Yuxian Meng^{*1}, Xiaoya Li^{*1}, Xiaofei Sun¹, Qinghong Han¹
Arianna Yuan^{1,2}, and Jiwei Li¹

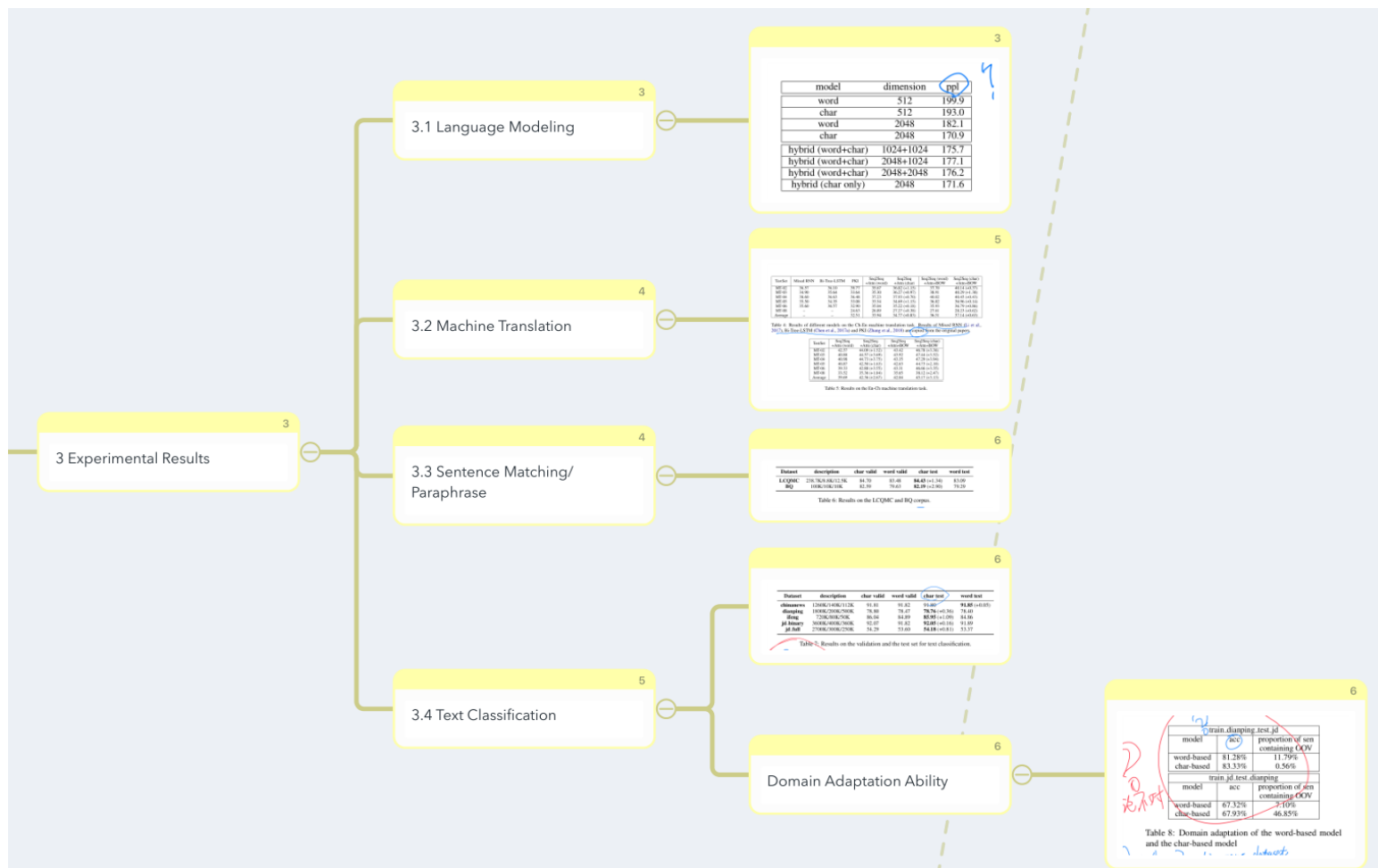
¹ Shannon.AI

² Computer Science Department, Stanford University

{ yuxian_meng, xiaoya_li, xiaofei_sun, qinghong_han
arianna_yuan, jiwei_li }@shannonai.com

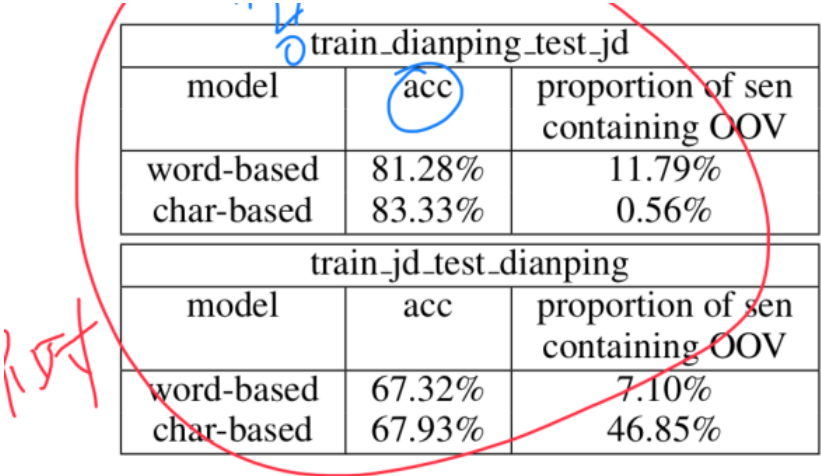
1.1 Is word segmentation necessary for DL of Chinese representations?

比较char-based 模型和 word-based 模型在四种应用上的表现，得出char-based模型均优于word-based 模型。



1.1 Is word segmentation necessary for DL of Chinese representations?

在domain adaptation ability中，有一个问题，在京东评价训练的模型拿到点评上测试时，char-based模型包括OOV的句子远远多于word-based模型，因为只有两组对比试验，也无法说明，char-based模型词库适应能力一定好于word-based模型。



1.54

train_dianping_test_jd		
model	acc	proportion of sen containing OOV
word-based	81.28%	11.79%
char-based	83.33%	0.56%

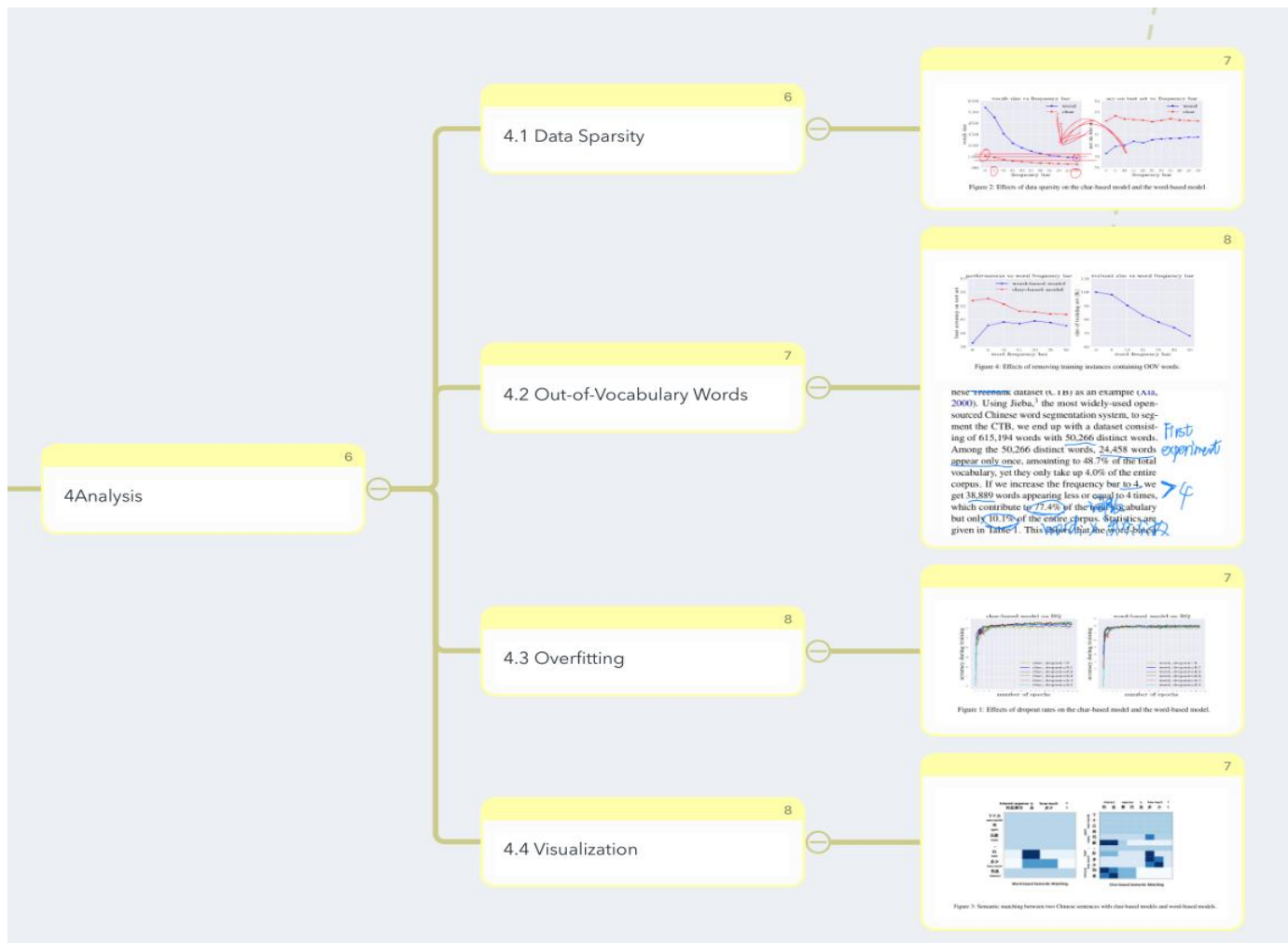
train_jd_test_dianping		
model	acc	proportion of sen containing OOV
word-based	67.32%	7.10%
char-based	67.93%	46.85%

Table 8: Domain adaptation of the word-based model and the char-based model

去了香侬慧语科技的知乎问了一下，我认为可能是有专有名词的原因，因为点评数据中会出现很多的词是少用到的地名或者菜名，这就远远超出了京东评价的范围了。反过来，京东上的评价词这种专有词的含量就没有那么大了。

1.1 Is word segmentation necessary for DL of Chinese representations?

然后，文章说明对于词向量模型不如字向量的原因，有稀疏性，库外词，过拟合三个方面。最后通过一个图形象对比词和字对于语义匹配的不同，在这例子上，字向量模型更容易准确的揣测文本的意思。



1.1 Is word segmentation necessary for DL of Chinese representations?

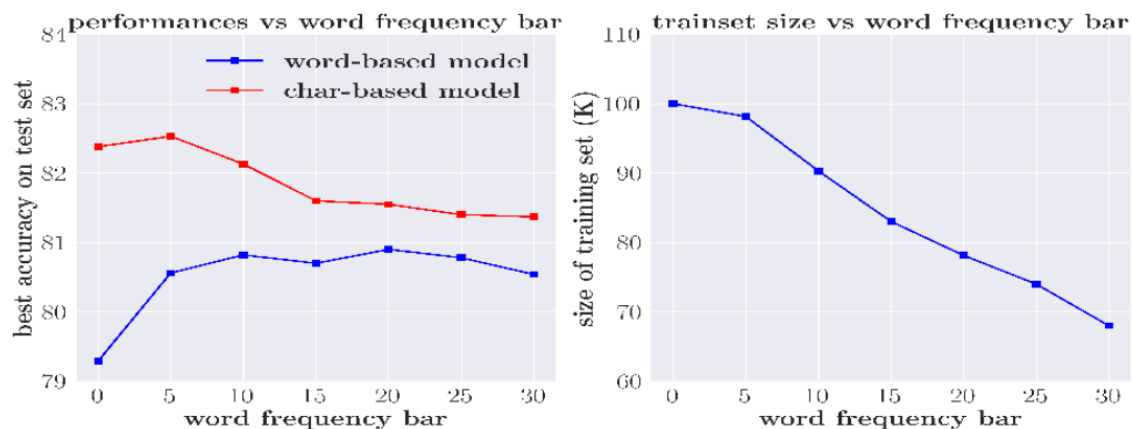


Figure 4: Effects of removing training instances containing OOV words.

左图word-base模型表现先有很大提升，然后下降：因为当frequency bar比较小的时候，对于那些infrequency 但是比较接近frequency bar的词会被归入词库中，从而不被判定为OOV。随着frequency bar的增大，这些词的进入会越来越地导致data sparsity，使得很多具有词库代表性的词，即【频率低，有特征】的词，被划分到词库外，使参数拟合效果变差。

根据作者的表述，frequency bar = 4时，有38889个词的词频小于4，这些词占了词表（vocab）的77.4%，但是却仅仅是数据集（corpus）的10.1%

1.1 Is word segmentation necessary for DL of Chinese representations?

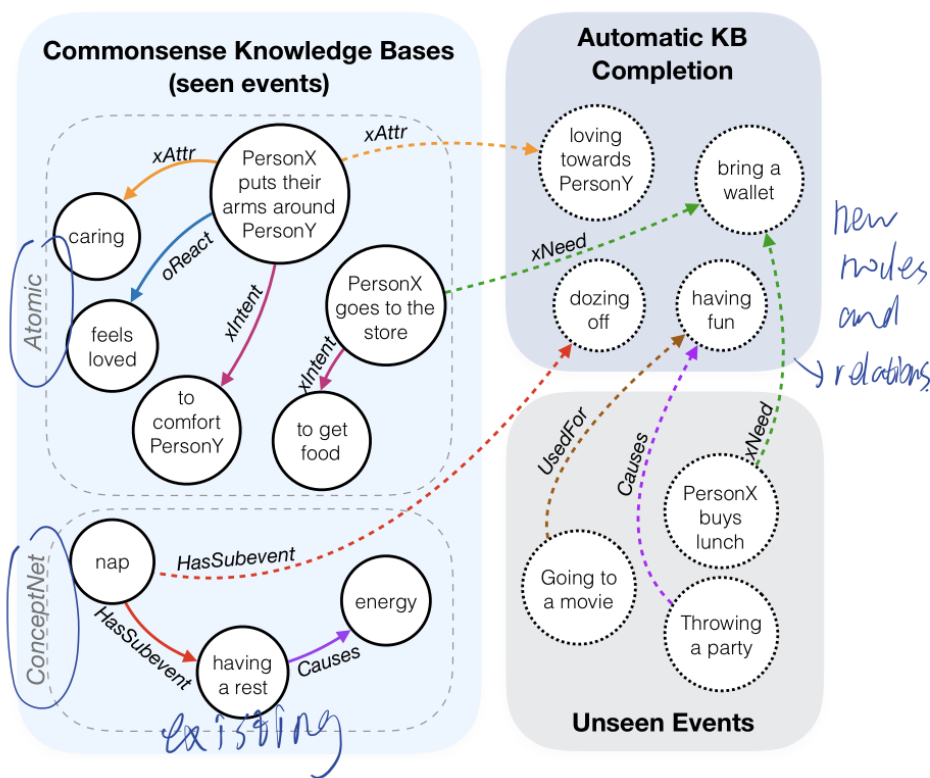
从这篇论文：

1. 前段时间看fastText时，提到了n-gram的方法，是一种将句子标记为n个单词组合的方法。这篇论文中char-based的在Text classification 上对于不同数据集的表现还可以提升。或许在char和word之间还会存在一个平衡点，这个sub-word应该可能会参与到前后不相等，或者其他的关系上。
2. 另外没看到考虑标点，句意的影响，自己这段时间的论文中也没有涉及到，或许通过识别标点，分句子判定会提升词库的表现？这有点像命名实体识别，例如主语，宾语，排名高一些。连词，冠词可以排名低一些。
3. 自己在做复现实验的时候，找到的分字工具时transformer。
4. CTB-6是，Chinese Tree Bank，来自新闻文章等的用于分词，位置标记等中文词库。
5. 各个实验的评价标准还需要了解并且分类。

1.2 COMET

论文: <https://www.aclweb.org/anthology/P19-1470>

- 这篇文章介绍了一个用来自动生成常识知识库的Commonsense transformers, 它能够调整语言模型的权重来学习产生新的知识库。
- 并且文章通过在两个知识库 (ATOMIC和ConceptNet) 上的实验, 展示了通过COMET产生的新知识是可以得到人类的认可 (准确率分别77.5%和99.1%)。
- 未来可以通过COMET扩展其他类型的数据库, 为构建知识图谱提供了另一种方案。



COMET: Commonsense Transformers
for Automatic Knowledge Graph Construction

Antoine Bosselut ♦ ♦ Hannah Rashkin ♦ ♦ Maarten Sap ♦ ♦ Chaitanya Malaviya ♦

Asli Celikyilmaz ♦ Yejin Choi ♦ ♦

♦ Allen Institute for Artificial Intelligence, Seattle, WA, USA

♦ Paul G. Allen School of Computer Science & Engineering, Seattle, WA, USA

♦ Microsoft Research, Redmond, WA, USA

左图刻画Atomic和ConceptNet的内部关系和属性; 以及如何从已有的数据集推测出新的关系的模型思路, 即COMET的思路。实线代表已有, 虚线代表生成。

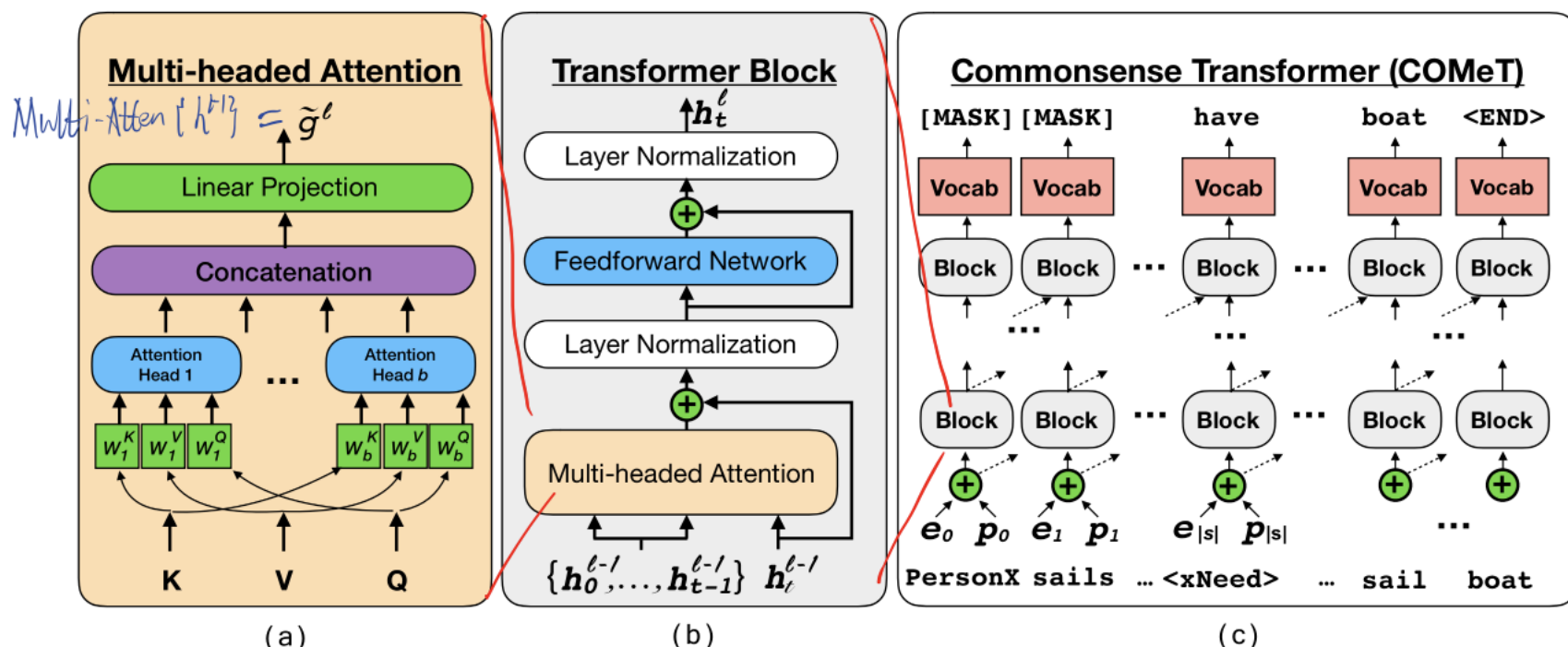
1.2 COMET

论文: <https://www.aclweb.org/anthology/P19-1470>

GPT: [Language Models are Unsupervised Multitask Learners](#)

GPT2.0: <https://zhuanlan.zhihu.com/p/56865533>

下图c是Commonsense Transformers的模型, a是b的细节, b是c的细节。
模型的输入涉及到了用GPT生成Transformer语言模型的内容, 因此还不太明白关于多头Attention的具体好处, 只能从感觉上发现其加入了前面的内容。
最后生成的是尾实体。



1.2 COMET

论文: <https://www.aclweb.org/anthology/P19-1470>

ConceptNet: <http://conceptnet.io/>

ATOMIC: <https://arxiv.org/pdf/1811.00146.pdf>

下图说明了同一个句子在两个不同的数据集中，标注方法不同。
ConceptNet 在subject和relation之间多了一个mask。

ATOMIC Input Template and ConceptNet Relation-only Input Template



PersonX goes to the mall [MASK] <xIntent> to buy clothes

ConceptNet Relation to Language Input Template



go to mall [MASK] [MASK] has prerequisite [MASK] have money

1.2 COMET

Score评价原理: <https://www.aclweb.org/anthology/P16-1137>

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	PPL	Score	N/T _{sro}	N/T _o	Human
LSTM - s	-	60.83	86.25	7.83	63.86
CKBG (Saito et al., 2018)	-	57.17	86.25	8.67	53.95
COMET (- pretrain)	8.05	89.25	36.17	6.00	83.49
COMET - RELTOK	4.39	95.17	56.42	2.62	92.11
COMET	4.32	95.25	59.25	3.75	91.69

Table 6: ConceptNet generation Results

左图和右图COMET分别在ATOMIC和ConceptNet上的表现。

左图：表明，在BLEU-2评估上，COMET超过了所有baseline的表现；从 N/T_{sro}^6 ， N/T_o ， N/U_o 三个指标中，出COMET产生的新的元组对象同样超过baseline。

右图：perplexity低，表明COMET对于预测结果的准确性比较高。同时Score高表明其产生的新的语言元组得到了认可。

指标 (%)	PPL	BLEU-2	N/T_{sro}^6	N/T_o	N/U_o	Human
含义	perplexity	自动评估 测量法 (双语互 译质量评 测)	产生的语 言元组为 新元组	产生的语 言元组含 有新 object	新对象是 生成唯一 的对象集 合的占比	来自AMT 的员工人 工标注

1.2 COMET

Seed Concept	Relation	Generated	Plausible
X holds out X's hand to Y	xAttr	helpful	✓
X meets Y eyes	xAttr	intense	✓
X watches Y every ____	xAttr	observant	✓
X eats red meat	xEffect	gets fat	✓
X makes crafts	xEffect	gets dirty	✓
X turns X's phone	xEffect	gets a text	✓
X pours ____ over Y's head	oEffect	gets hurt	✓
X takes Y's head off	oEffect	bleeds	✓
X pisses on Y's bonfire	oEffect	gets burned	✓
X spoils somebody rotten	xIntent	to be mean	✓
X gives Y some pills	xIntent	to help	✓
X provides for Y's needs	xIntent	to be helpful	✓
X explains Y's reasons	xNeed	to know Y	✓
X fulfils X's needs	xNeed	to have a plan	✓
X gives Y everything	xNeed	to buy something	✓
X eats pancakes	xReact	satisfied	✓
X makes ____ at work	xReact	proud	✓
X moves house	xReact	happy	✓
X gives birth to the Y	oReact	happy	✓
X gives Y's friend ____	oReact	grateful	✓
X goes ____ with friends	oReact	happy	✓
X gets all the supplies	xWant	to make a list	✓
X murders Y's wife	xWant	to hide the body	✓
X starts shopping	xWant	to go home	✓
X develops Y theory	oWant	to thank X	✓
X offer Y a position	oWant	to accept the job	✓
X takes ____ out for dinner	oWant	to eat	✓

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	✓

左图和右图的第三列分别是COMET在ATOMIC和ConceptNet上生成的新知识的随机取样，最后一列都是人工识别的结果，可以看到这几个例子的意思还是非常接近真实情况的意思。

之所以看这篇文章，是因为自己最近参加的比赛中需要NER的相关知识，正好ACL2019出现了一些这些方面的相关研究，从这个文章中，只能看到了扩展知识库，扩展知识图谱的作用，对于实体关系识别没有看到相关的方法。

今天发现COMET的实现：<https://github.com/atcbosselut/comet-commonsense>

1.3 Fine-Grained Entity Typing in Hyperbolic Space

这篇文章，主要介绍了在双曲空间上的属性分类问题。由于投影在双曲空间上的距离被放大，相对于欧几里得空间投影来讲，导致在训练集合中频繁共现的属性距离更近，提升了基于最近邻策略的属性的预测效果。

Fine-Grained Entity Typing in Hyperbolic Space

Federico López*

Benjamin Heinzerling

Michael Strube

*Research Training Group AIPHES

Heidelberg Institute for Theoretical Studies

`firstname.lastname@h-its.org`

[1] 原文:

<https://arxiv.org/abs/1906.02505>

[2] Ultra-Fine Entity Typing:

[定义 论文链接](#)

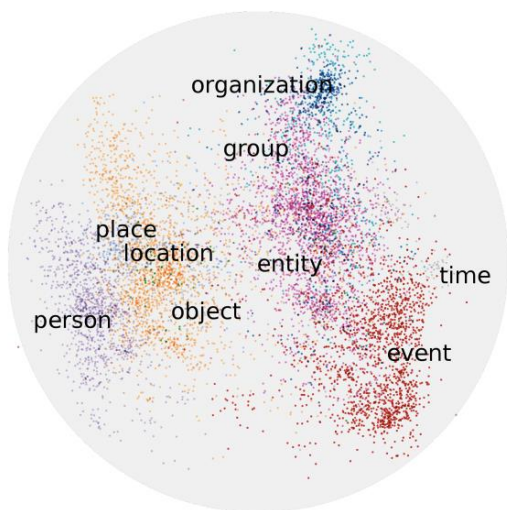
[3] 庞加莱圆盘:

[论文链接](#)

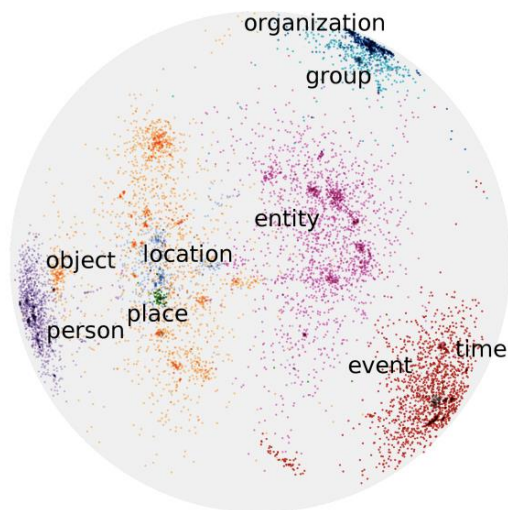
1.3 Fine-Grained Entity Typing in Hyperbolic Space

从下图可以看出，经过微调的WordNet具有层级结构的名词，通过细粒度的属性分类，投影到两个空间上的效果，其中Hyperbolic Space是采用的庞加莱圆盘模型。

在图（b）中，属性更加靠近边界，使得不同的属性区分更加明显。这意味着，从层级结构上讲，在图中离得越近的元素越会共享同一个父节点，并且会越靠近其中心位置，反之，在层级上更低的元素会远离中心位置。



(a) Euclidean Space.



(b) Hyperbolic Space.

[1] 原文:

<https://arxiv.org/abs/1906.02505>

[2] Ultra-Fine Entity Typing:

[定义 论文链接](#)

[3] 庞加莱圆盘:

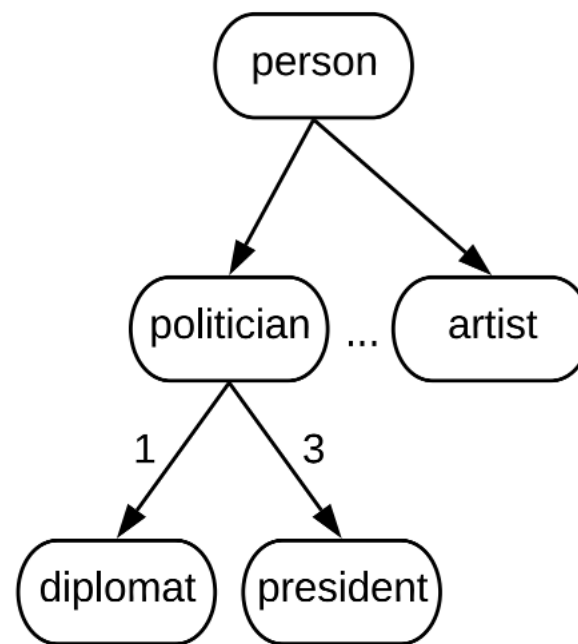
[论文链接](#)

1.3 Fine-Grained Entity Typing in Hyperbolic Space

层级结构表示如下，对于person属性来讲，如果有politician属性，那么还会存在其子节点的属性，diplomat。

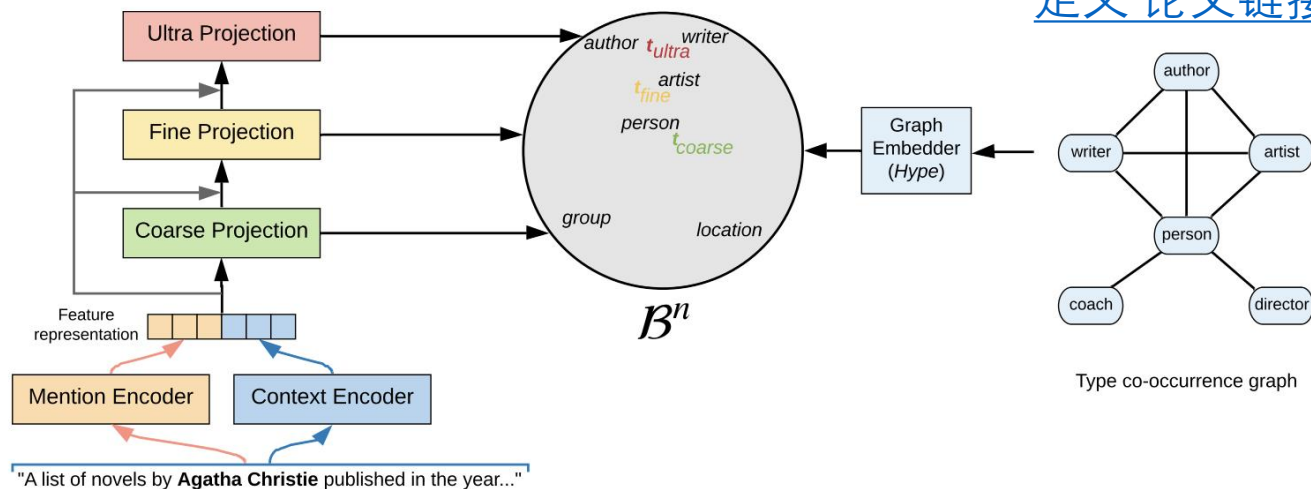
从层级结构的上往下，粒度由粗到细，即coarse, fine, ultra-fine。一个模型学习到了diplomat便可以将这个知识联系到politician。

Sentence	Annotation
...when the president said...	politician, president
...during the negotiation, he ...	politician, diplomat
...after the last meeting, she ...	politician, president
...the president argued...	politician, president



1.3 Fine-Grained Entity Typing in Hyperbolic Space

[1] Ultra-Fine Entity Typing:
[定义 论文链接](#)



(a) Projection layers.

(b) Incorporation of hierarchical information.

模型结构如上。

在Mention Encoder，通过将char-based CNN和Glove两种方法的得到的特征用similar self-attention encoder合并。

在Context Encoder上，改进了[1]中的方法，即用一个词位置嵌入来反映低i个词和entity mention得距离，可以减少attention层得偏差，从而使得特征更多得关心mention而不是context。最后用Bi-LSTM和self-attentive encoder得到context representation。

这篇文章提出了三层模型，就是用三种不同粒度的layer预测三种不同的标签。模型损失函数是最小化节点间距离的方差，因为cosin作为距离在庞加莱圆盘中也即是双曲空间中同样适用。

1.3 Fine-Grained Entity Typing in Hyperbolic Space

Model	Space	Coarse		Fine		Ultra-fine		Coarse + Ultra		Variation	
		MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1
MULTITASK	-	60.6	58.0	37.8	34.7	13.6	11.7	-	-	-	-
WORDNET	Hyper	45.9	44.3	22.5	21.5	7.0	6.7	41.8	37.2	-4.1	-7.1
	Euclid	56.1	54.2	26.6	25.3	7.2	6.5	56.6	48.5	0.6	-5.7
WORDNET + FREQ	Hyper	54.6	52.8	18.4	18.0	11.3	10.8	46.5	40.6	-8.0	-12.2
	Euclid	56.7	54.9	27.3	26.0	12.1	11.5	55.8	49.1	-0.9	-5.8
FREQ	Hyper	56.5	54.6	26.8	25.7	16.0	15.2	59.7	53.5	3.2	-1.1
	Euclid	56.1	54.2	25.8	24.4	12.1	11.4	60.0	53.0	3.9	-1.3
PMI	Hyper	54.7	53.0	26.9	25.8	16.0	15.4	57.5	51.8	2.8	-1.2
	Euclid	56.5	54.6	26.9	25.6	12.2	11.5	59.7	53.0	3.2	-1.5

(a) Results on the same three granularities analyzed by Choi et al. (2018).

(b) Comparison to previous *coarse* results.

模型测试了四个具有层级结构的数据集上的最细粒度表现，在双曲空间上的模型，经过特征上的调整，网络结构的调整，在MaF1和MiF1两个指标，FREQ（type co-occurrence frequency）和PMI（pointwise mutual information）都好于原来的工作MULTITASK以及欧几里得空间上的测试。

但是在另外两个粒度上，表现并没有明显的提升，作者认为是模型的回归有问题，1998个实例中仅有1318个fine的标签，Ultra-fine有7511个，Coarse也有1904个

Split	Coarse	Fine	Ultra-fine
Train	2,416,593	4,146,143	3,997,318
Dev	1,918	1,289	7,594
Test	1,904	1,318	7,511

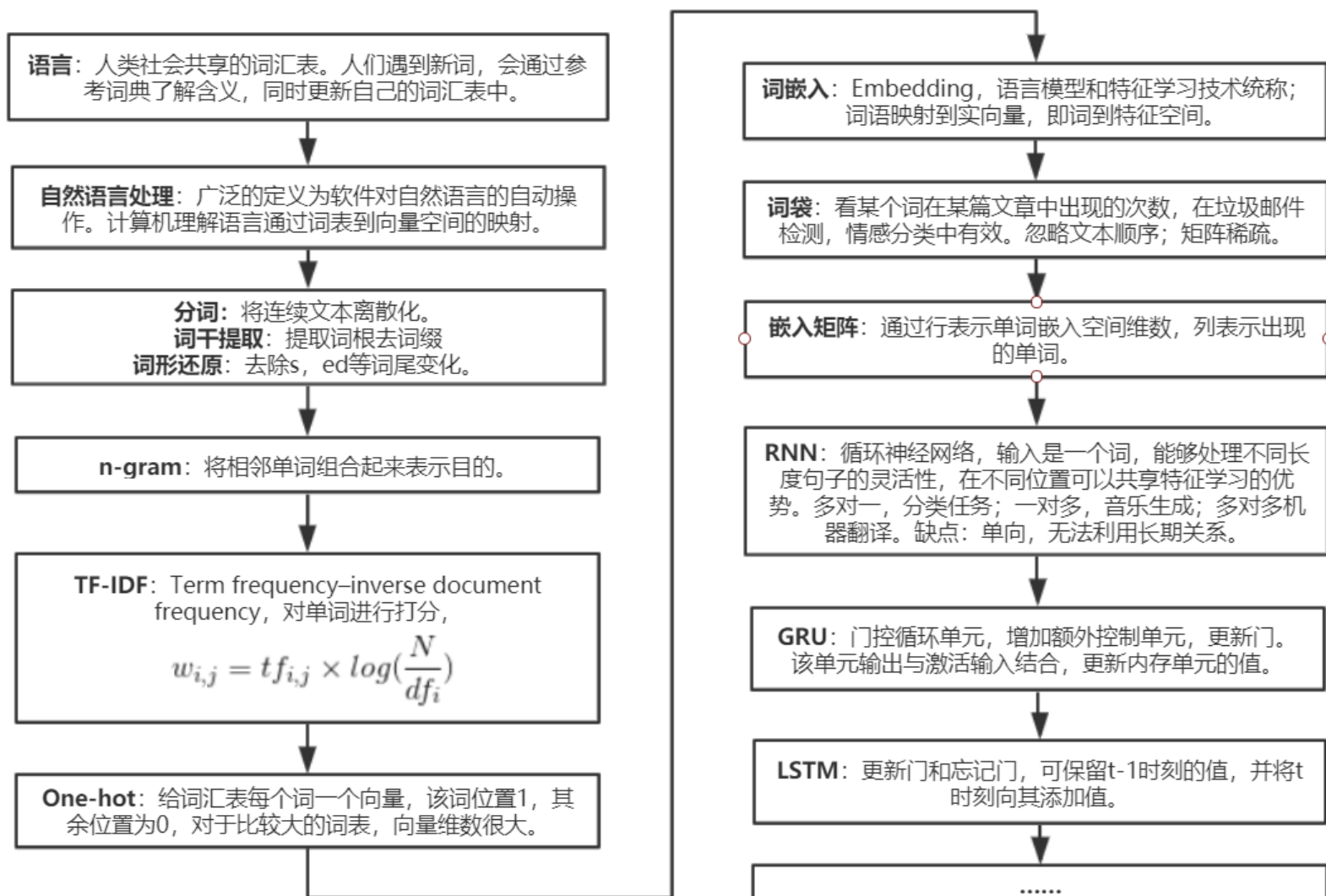
模型在表(b)合并了粗粒度和最细粒度，发现双曲空间MaF1改善，MiF1反而下降。

2 Web Data Mining

这本书是刘兵写的，其个人主页[个人主页](#)和[书籍主页](#)（第二版）。
该书两个版本差别主要在11.12章，体现在观点挖掘上。

现在正在看前五章基础内容，目前大概以每天五到六页，大概两到三个算法的速度进行学习，目前进度到了第三章用类联系规则进行分类。在看的过程中，自己逐渐从看理论，到尽可能实现代码或者参考相关代码，来加深各个算法的理解，通过看效果也更容易明白。

3 NLP基础知识



4 达观杯

前段时间搜索了一个比赛，想要在比赛场景下，快速成长一下，但是.....

这个比赛是关于命名实体分类的比赛，官方给的CRF++，我测试官方模型效果不是很好。CRF主要关注整体的影响。

不过，看群里同学们有用BiLSTM + CRF这个方案，我稍微学习了一下，还没有进行测试。

还有同学分享了一个BERT的使用，本来自己想尝试用会议上新出的RoBert，即具有鲁棒特性的Bert，但是训练时间感觉要很久，20hours可能还不够。我打算先把一些参数的利弊分清，再去做测试。

这个比赛也是我之所以想学习一下前面两篇COMET以及实体分类的原因，想试试能不能用新模型方法做一下这个实验。

4 Github: <https://github.com/RelativeWang/word2vec-study>

目录

- 语料预处理
 - 环境
 - 下载语料包
 - 解压
 - 繁体转简体
 - 删除非中文字符
 - 分词
 - 测试截图
- 训练模型
 - Word2Vec
 - 环境
 - 过程
 - Glove
 - FastText
- 测试
 - 环境
 - 相似度测试
 - 类比测试
 - 通过类比测试看一些变化
 - 可视化
 - PCA二维
 - PCA三维
- 参考&感谢

整理了Github，更加系统清晰，主要是将一些词向量可视化的内容加入进去，给词向量降维，观察词之间的关系。左侧是整理完后的目录。