面向上下文语境分析的语境向量法

李星辰[a]，程逸航[a]，冯礼群[a]，刘泽楠[a]，刘嘉煜[a]

[a]天津大学，天津 300072，中国

文章类别 研究

研究 观点

摘要

本文主要介绍了面向上下文语境分析的语境向量法。利用统计学和概率论方法，通过对问答环境中不同词性词汇的特征的考察，提出了提取问答环节中语境的语境向量法。文章对特定语料库中问答对出现词汇的词性进行了统计测量，分析了可能作为问句语境的词性，并针对该语料库对这些词性进行分布的统计、参数估计与假设检验，得出了构造语境向量空间的一般方法。同时参考其他语料库，得出了针对不同语料对语境向量的简单修正方法。最后，文章介绍了一个“基于语境的简单对话系统”的搭建，为如何将语境向量加入对话系统提供了一个范例。

关键词：词性标注，参数估计，假设检验，自然语言处理

Context Analysis Based on Context Vector

Xingchen Li[a], Yihang Cheng[a],Liqun Feng[a], Zenan Liu[a], Jiayu Liu[a]

[a]Nankai District, Tianjin, 300072, China

Column of article: Research

Type of article: Perspective

ABSTRACT

The article introduces the method of Context Vector used to analyse context situation. We checked the features of different part-of-speeches in questioning and answering environment with the method of Statistics and Possibility, and raised the Context Vector which is used to extract situations in chat circumstances. The article shows the statistic feature of each part-of-speech appearing in the corpus, analysts the possibility of appearing in questions and studies the distribution, Parameter Estimation and hypothetical test of them. By referencing to another corpus, we also raised a brief way of establishing and correcting the vector. Finally, the article also introduced a simple example of chat system based on context vector.

KEYWORDS: part-of-speech tagging, parameter estimation, hypothetical test, nature language processing

1 引言

目前在自然语言处理方面，上下文语境常常被用来做机器翻译、词义消歧、情感分析等，而很少有人考虑实际对话的语境。传统的语境依旧是语言学的研究范畴。目前的主流聊天机器人在应对某个具体问题时已经可以达到一个很不错的回答效果，但是在多次对话中，问答系统就显得力不从心。例如在对话中，将之前出现的人名用人称替代，对话系统往往应答就会很困难，这说明对话系统对语境的了解依旧有限。我们就是基于此，对问句营造的语境进行了统计学研究，总结出一个简单的帮助系统识别语境的方法。

2 基于词性的语境统计分析

2.1 词性标注的n-gram统计建模

要想分析词性的影响，首先我们要对语料库进行词性标注，这里我们利用n-gram统计建模的方法[1-3]。由于本文重点不在考察词性标注的具体方法，因此只在这里对具体实现的方法进行介绍。在python的 NLTK 处理包中，有事先集成好的多元语法标注器，可以利用这个package进行语料的训练与验证[2-3]。我们利用已经完成标注的语料库 treebank 进行标注器的训练与验证。利用回退标注的方法，我们选取前7000句作为训练集，2000句之后的作为测试集进行标注器的训练。训练过程包括如下三步：

1.尝试使用bigram标注器标注标识符

2.如果bigram标注器无法找到标记，尝试unigram标注器。

3.如果unigram标注器也无法找到标记，使用默认标注器。

最后的标注器在测试过程中，得到的准确度为：0.9875608103947288。说明这个标注器可以比较准确地标注词性。但是有一点值得注意，就是在标注器标注此语料库以外的语料时，标注准确率会有明显的下降，这也是之后的一个主要误差来源之一。

2.2 基于语料库的词性筛选

2.2.1 初步的词性筛选

要想合理正确的判断语境，就要了解语境的构成，以及问句与其他语句的区别，我们的思路是通过语句中不同词性出现的频率来判断。

首先我们试图区分问句群和普通句群中哪些词性的词汇出现频率会有明显变化，为此我们对已有的问答语料库（语料库：overheard）中的问答语句进行统计分析：

第一步，我们对所有语句中的 NN、JJ、PRP、VB、WP、IN、DT、CD、WDT 这 9 种词性进行统计。由于我们目的是为了区分问句中表现特殊的那些词性，所以我们忽略了句子的界限概念而将同类句子中的所有词汇进行了整体频率分析，即：

$$\mathrm{P}(x，\ i)=\frac{n(x)}{N(i)} \tag{2-1}$$

式中x代表某词性，i代表考察点，n代表该词性出现的词语数，N代表到考察点为止所有的词语数，P代表该词的出现频率。这个式子意味着，从被考察群体的第一个词开始，到考察点为止目标词性的词出现的次数。我们针对两种句群，对0<i<15000的所有情况进行了遍历，得到如下结果：

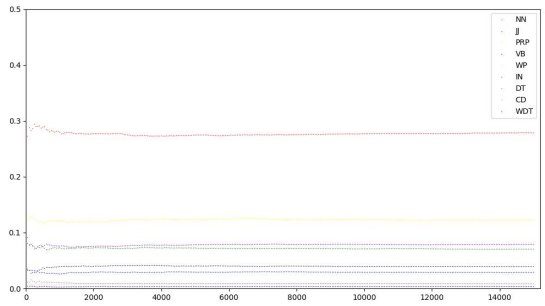

图2-1.不同词性词汇在所有句中出现概率

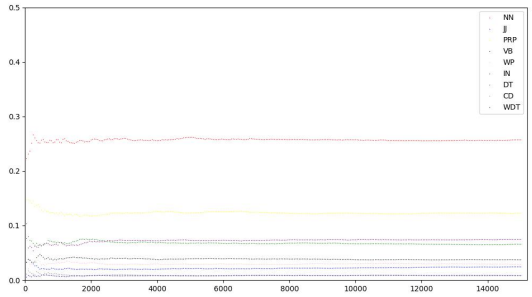

图2-2.不同词性词汇在问句中出现概率

图2-1描述了在问句群中各个词性词汇出现的频率随采样数目增加趋于稳定的现象，图2-2描述了所有句群中各个词性词汇出现频率随采样数目增加而趋于稳定的现象。对这两个图的解释为：1.在数据抽样足够的情况下，词性出现频率会趋于一个稳定的值，说明词性在语言中出现是收敛的，可以进行统计分析。这也符合伯努利大数定律[5]；2.在不同的语句群中存在词性的出现频率发生了明显变化，说明词性可以用于区分不同类型的语句群。这两点，也是之后我们可以进行深入分析的基本假设。

为了进一步看出这两种采样中的差别，我们将各种情况下频率的收敛值展示在表2-1中：

表2-1　各词性的出现频率

词性	所有句群中频率	问句中频率
NN:	0.278476128	0.255956956
JJ:	0.028991852	0.024232372
PRP	0.122586486	0.122375501
VB:	0.039276651	0.037541972
WP:	0.008051231	0.032687406
IN:	0.078968849	0.07702577
DT:	0.070361081	0.065981634
CD:	0.008422256	0.008010033
WDT	0.003502471	0.004692746

2.2.2 细化的词性筛选与选择

事实上，在2.2.1中考察的词性只是一个词性的大体分类，每个词性下都有更为具体的子词性。例如，NN下就分为NNP，NNS等。该部分来讨论这些具体词性的频率。我们将词性细分，如将NN分为NNS与NNP，VB分为VBN、VBG、VBD、VBZ，并对这六种词性分别在所有语句和问句中进行统计，做出散点图如2-3，2-4。

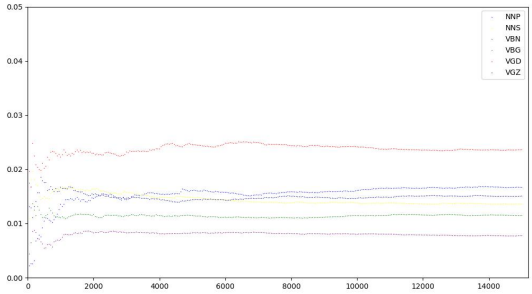

图2-3.细分的词性词汇在所有句中出现概率

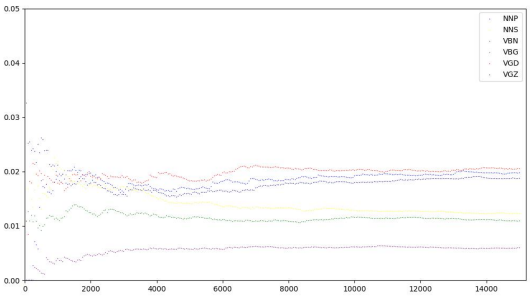

图2-4.细分的词性词汇在问句中出现概率

可以看出，这些词性词汇依旧满足前面所述的两条基本假设。我们将粗分与细分的结果合并进行列表显示如下：

表2-2　各词性与细分词性的出现频率

词性	所有句群中频率	问句中频率
NN:	0.278476128	0.255956956
NNP:	0.016636737	0.019782354
NNS:	0.013572075	0.012217323
JJ:	0.028991852	0.024232372
PRP	0.122586486	0.122375501
VB:	0.039276651	0.037541972
VBN:	0.007769252	0.005987297
VBG:	0.011494338	0.010922772
VBD:	0.023619418	0.020510538
VBZ:	0.015056173	0.018770986
WP:	0.008051231	0.032687406
IN:	0.078968849	0.07702577
DT:	0.070361081	0.065981634
CD:	0.008422256	0.008010033
WDT	0.003502471	0.004692746

现在我们要从这些数据中挑选出词性频率明显发生变化，即可能作为区分问句和全部句子判据的词性。我们使用了绝对误差和相对误差结合的方法，即：1.计算各词性在不同句群的相对误差，对误差进行标准化（该词性误差/所有误差总和），取相对误差序列极差下四分位以上的所有结果为样本空间A；2.计算各词性在不同句群的绝对误差，，对误差进行标准化（方法同上），取绝对误差序列极差下四分位以上的所有结果为样本空间B。那么我们最后选取的词性样本空间C可以表达为：

$$C = A \cap B \tag{2.2}$$

下面的表格显示了筛选的过程：

表2-3　各词性与细分词性的误差，标准化与筛选

词性	相对误差	标准化的相对误差	是否选取（A）	绝对误差	标准化的绝对误差	是否选取（B）
NN:	0.080865718	0.016942992	1	0.022519172	0.298413031	1
NNP:	0.189076513	0.039615328	1	0.003145616	0.041684165	1
NNS:	0.099819113	0.020914109	1	0.001354753	0.017952516	0
JJ:	0.164166139	0.034396104	1	0.00475948	0.063070302	1
PRP	0.001721113	0.000360608	0	0.000210985	0.002795872	0
VB:	0.044165653	0.009253592	0	0.001734679	0.022987115	0
VBN:	0.229359935	0.048055514	1	0.001781955	0.023613597	0
VBG:	0.049725891	0.010418573	0	0.000571566	0.007574115	0
VBD:	0.131623882	0.027577848	1	0.003108879	0.041197347	1
VBZ:	0.246730212	0.051694936	1	0.003714813	0.049226878	1
WP:	3.059926514	0.641116075	1	0.024636175	0.326466521	1
IN:	0.024605641	0.005155376	0	0.001943079	0.025748732	0
DT:	0.06224247	0.013041048	0	0.004379447	0.058034291	1
CD:	0.048944462	0.010254848	0	0.000412223	0.005462574	0
WDT	0.339838771	0.071203049	1	0.001190275	0.015772947	0

可以看出，最后我们选取的词性样本空间为：

$$C = \{\text{NN}，\ \text{NNP}，\ \text{JJ}，\ \text{VBD}，\ \text{VBZ}，\ \text{WP}\}$$

2.3 参数估计与假设检验

2.3.1 参数估计

在获得初步选择后，我们考察这些词性在单独的英语问句中出现的频率的分布情况。图2-5至2-10直观反映了各个词性词汇出现频率样本在我们抽样空间（overheard的前1500个问句）的分布：

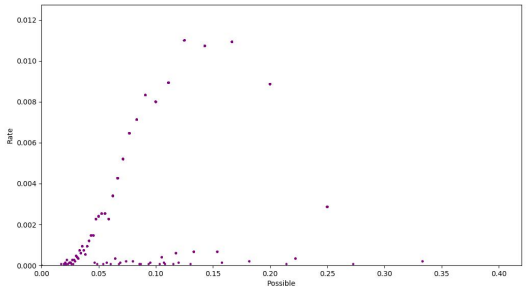

图2-5 VBZ类词性分布情况

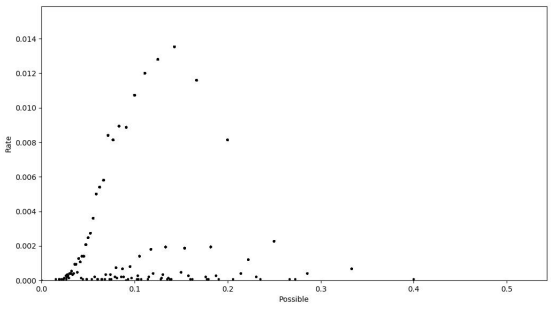

图2-6 VBD类词性分布情况图

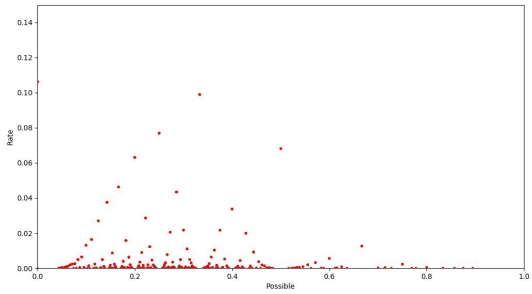

图2-7 NN类词性分布情况

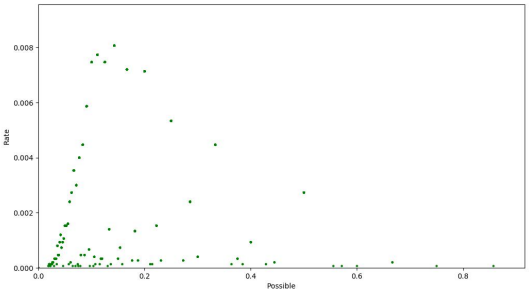

图2-8NNP 类词性分布情况图

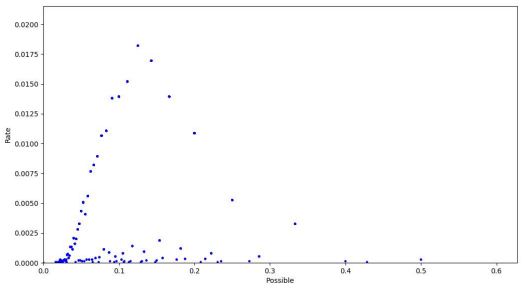

图2-9 JJ类词性分布情况

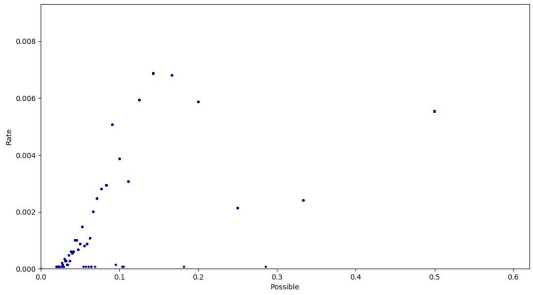

图2-10 WP类词性分布情况图

对于这些抽样数据，我们希望通过样本均值和方差获得在一定置信区间下的总体均值和方差。由中心极限定理，各个词性的频率分布近似于正态总体。由区间估计和抽样检验定理的相关概念，在总体方差未知的情况下，有枢轴量[4-5]：

$$Z=\frac{\overline{X}-\mu}{S/\sqrt{n}}\sim t_{n-1} \tag{2.3}$$

因此在1-α的置信水平下参数 μ 的置信区间为：

$$\left[\overline{X}-t_{n-1}(\frac{\alpha}{2})\frac{S}{\sqrt{n}},\overline{X}+t_{n-1}(\frac{\alpha}{2})\frac{S}{\sqrt{n}}\right] \tag{2.4}$$

在总体期望未知的情况下，有枢轴量：

$$X^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \quad (2.5)$$

因此在1-α的置信水平下参数σ²的置信区间为：

$$\left[\frac{(n-1)\ S^2}{\chi^2_{n-1}(\frac{\alpha}{2})}, \frac{(n-1)\ S^2}{\chi^2_{n-1}(1-\frac{\alpha}{2})} \right] \quad (2.6)$$

由式（2.3）至（2.6）可以得到（置信概率95%）：

表2-4 区间估计

词性	样本均值	样本方差	均值置信区间	方差置信区间
NN:	0.2094382385	0.0292305667	[0.20679987,0.21207660]	[0.02860386,0.029972035]
NNP:	0.0125098626	0.0025678244	[0.01172788,0.01329185]	[0.00251277,0.00263296]
JJ:	0.0191493457	0.0023432374	[0.01840234 ,0.01989635]	[0.00229300 ,0.002402676]
WP:	0.0020205686	0.0002466787	[0.00177820,0.00226294]	[0.00024139,0.000252936]
VBD:	0.0143269703	0.0016465985	[0.01370077,0.01495317]	[0.00161130 ,0.001688366]
VBZ:	0.0097626955	0.0011409091	[0.00924145,0.01028394]	[0.00111645,0.00116985]

2.3.2 假设检验

我们需要对所得到的均值和方差做假设检验。我们从语料库 overheard 中另外随机抽样了 15000 个样本。由于数据数量在 15000 个左右，所得样本近似为正态总体，因此我们可以用正态总体的假设检验对所得样本进行处理。

对于均值 μ 的假设检验，在总体方差 σ 2 未知的情况下，我们选取检验统计量：

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \quad (2.7)$$

其拒绝条件为：

$$\{|\ T\ | \geq\ t_{n-1}(\frac{\alpha}{2})\} \quad (2.8)$$

带入相关数值，可以观察其值是否落入拒绝域内。

同理对于方差 σ 2 的假设检验，在均值 μ 未知的情况下，我们选择检验统计量：

$$X^2 = \frac{(n-1)S^2}{\sigma^2} \quad (2.9)$$

其拒绝条件为：

$$\{\chi^2 \leq \chi^2_n(1-\frac{\alpha}{2})\ 或\chi^2 \geq \chi^2_n(\frac{\alpha}{2})\} \quad (2.10)$$

带入相关数值，观察其值是否落入拒绝域内。

在选择α=90%的条件下，由上述公式可以得到下表：

表2-5 假设检验

词性	原假设	样本均值	样本方差	拒绝域	是否拒绝
NN:	μ =0.209438	0.206693724		|T|≥1.644955225	是
	σ 2=0.029231		0.028755056	c2≤14715 或 c2≥ 15285	否
NNP:	μ =0.01251	0.013087245		|T|≥1.644955225	否
	σ 2=0.002567		0.002706519	c2≤14715 或 c2≥ 15285	是
JJ:	μ =0.019149	0.019136055		|T|≥1.644955225	否
	σ 2=0.002343		0.002332654	c2≤14715 或 c2≥	否

				15285	
WP:	μ=0.002021	0.001961551		|T|≥1.644955225	否
	σ2=0.000246		0.000238607	c2≤14715 或 c2≥15285	是
VBD:	μ=0.01432697	0.014593306		|T|≥1.644955225	否
	σ2=0.001646598		0.001677898	c2≤14715 或 c2≥15285	否
VBZ:	μ=0.009763	0.009722346		|T|≥1.644955225	否
	σ2=0.00114		0.001129685	c2≤14715 或 c2≥15285	否

在拒绝均值的要求下，我们最后选出的词性有：NNP（专有名词）,JJ（形容词）,WP（wh 开头的词语，主要是疑问词）,VBD（动词过去式）,VBZ（动词第三人称单数）。这说明，最终这五种词性可以作为语境的生成来源。

需要注意的是，在另外一个语料库中做同样的统计建模，得到的结果发生了一定的变化，这说明对于不同语料而言生成语境的词性有不同的种类。所以要想获得更为一般的结果，需要综合多个语料进行统计建模。

3 语境向量法与语境空间

3.1 基本假设

语境向量法的提出，是基于之上对于词性的分析得出的。在 2.2 的统计分析中，我们发现可以代表问句的词性有 NN,NNP,JJ,VBD,VBZ,WP 六种，而在 2.3 的统计分析中，发现这六种里面可以作为有明确指代性的词性有NNP,JJ,WP,VBD,VBZ。据此我们提出三条基本假设，只要承认这些假设（利用实际交际中的经验，这种假设是合理的），这些词性就是与语境有联系的，可以作为判断语境的标准：

1. 实词有效性：认为句子中任何一个实词都有具体明确要指代的意义。
2. 对话连续性：人们在对话过程中存在对话回合的概念，每回合都是是在同一语境环境下的。
3. 动词弱化性：在对话中，尤其是英语对话中，动作通常不是想要传递的主要信息。

在得到的五种词性中，基于这三条假设，语境将主要从NNP和WP中产生。

3.2 具体内容

语境向量，即在某个特定的大环境（即超出语境的整体环境，如对话发生的场所之类）下构造向量，使它的每一个维度都可以表示当前对话回合的语境的内容，从而可以标记对话回合所在的语境。多个这样的线性无关向量就构成了该大环境下的语境空间。它的基本表示是：

$$[\text{语境1，语境2,…，语境n}]$$

语境向量中的每个维度，都是从一个特定的词性中分离出来的。但是要注意的是一个词性可能创造多个语境向量中的维度，只要该词性是利用2中的方法筛选出的。但也要注意的是在不同的大环境里从词性里分离出的语境可能是不同的，这也是为什么在2中我们一再强调词性的获得只是为了得到语境存在的范围，而绝不是直接获得语境。语境还需要通过将该词性中的具体类别分离出来，组成语境向量来获得。

对于语境向量间的比较，类似于欧氏空间向量的比较，即两向量若n个维度中有i个维度相同，我们就认为它们的距离为 i/n。同时当i/n小于某个值时，我们就可以认为这两个语境已经无关。

有了语境向量的概念，就可以利用语境向量配合其他等技术构建可以有“时间记忆”功能的对话系统。3.3中将给出一个例子。

3.3 适用范围

在我们考察的语料库中，抽象出的词性有，对应的具体内容就是：姓名，位置，时间和W开头的疑问词。因此抽象出的语境向量是：

$$[\text{姓名，地点，时间，疑问词}]$$

获得语境向量后，由于每个维度的内容需要利用已有的数据进行筛选，我们可以进行一定扩展方便在数据库中筛选，比如该向量可以扩展为：

$$[(\text{男性名字，女性名字}),(\text{城市，国家}),(\text{年，月，日，时间}),\ \text{疑问词}]$$

当然扩展的方式是不唯一的。这样，在具体搭建系统时，我们就可以通过与数据库比较，更方便地从句子中提取出表示语境的词汇。我们搭建了一个最简单的对话系统框架来反应这个过程，它的输出结果如下图：

图3-1 一个对话系统框架

可以看出系统成功返回了语境向量，值得注意的是，在Obama用he指代后，语境向量依旧可以判断出这个he指的是Obama。这就解决了1中提出的问题。利用这些返回结果，就可以结合其他的自然语言处理工具构造可以识别上下文语境的对话系统。

4 结果和讨论

4.1 项目创新点

本项目主要的创新点如下：

1.关注对话过程中的实际语境，更加贴近人们对问答系统的真实需求

2.利用词性来进行语境提取，关注问句与普通句子在词性方面的差异

3.进行了假设检验，论证了方法的可行性

4.对不同情况进行了分类，实现了总体意义上的全面性

4.2 主要误差来源

在之前的过程中，可能的误差来源有：词性标注器在脱离训练语料库后准确度下降；不同类型的对话库中各个词性出现的频率可能不同；不同词性之间的搭配可能导致分布的区别；构造语境向量时来自不同词性的语境选择可能不同。

4.3 可改进的部分

基于主要误差，可以改进的部分有：设计更加合理的词性标注器；对多类型对话库进行统一的频率分析；深入研究从筛选好的词性中提取语境的准确方法；将语境向量与问答系统的结合。

5 结论

该研究立足于对词性在不同类型句子中出现的频率不同，对问答语料进行统计建模，提取出可以区分不同类型句子的特征词性。通过词性的筛选，最终构造了描述对话时营造的语境的语境向量，并进行了简单的应用。文中所有程序的代码都可以在 github: https://github.com/MullerLee/ChatBot2.0 上找到。

致谢

在此非常感谢计算机学院张鹏老师一直以来的的帮助和支持，给我们提供了很多知识上和实践上的宝贵资源与建议，解决了我们研究过程中的很多难题；感谢全体参与研究的同学们。同时也十分感谢微电子学院一直以来对本项目的支持，没有这些支持我们难以走到现在。最后还要感谢马立群，王本友，苏展等学长们的帮助，为我们提供了不少有价值的思路。

参考文献

[1] 宗成庆，《统计自然语言处理（第二版）》北京：清华大学出版社，2013.

[2] 何敏煌，《Python程序设计入门到实战》 北京：清华大学出版社，2017.

[3] Deepti Chopra, Nisheeth Joshi, Iti Mathur, eds. Mastering Natural Language Processing with Python. 北京：人民邮电出版社，2017.

[4] 贾俊平，何小群，金勇进等. 《统计学（第五版）》北京：中国人民大学出版社, 2012.

[5] 天津大学数学系，《概率论与数理统计讲义》北京：人民教育出版社，2012.