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

ABSTRACT

The article introduces the method of Context Vector used to analyses 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 Introduction

At present, in natural language processing, certain contexts are often used as machine translation, word sense disambiguation, sentiment analysis, etc., but few people consider the context of actual dialogue. The traditional context is still the research scale of linguistics. The mainstream chat bots can already achieve a very good answering effect when responding to a specific question, but in many conversations, the question and answer system causes powerlessness. For example, in a dialogue, it is difficult for the dialogue system to answer frequently if the name of the person before it appears is replaced, which shows that the dialogue system's understanding of the context is still limited. Based on this, we conducted statistical research on the context of question creation, and summarized a simple method to help the system recognize the context.

2 Statistical analysis of context based on part of speech

2.1 N-gram statistical modeling of part-of-speech tagging

To analyze the impact of part of speech, we must first tag the corpus. Here we use the method of n-gram statistical modeling [1-3]. Since the focus of this article is not on the specific methods of part-of-speech tagging, only the specific implementation methods are introduced here. In the python NLTK processing package, there is a pre-integrated multi-grammar tagger, which can be used to train and verify the corpus [2-3]. We use the already annotated corpus treebank to train and verify the tagger. Using the method of back labeling, we select the first 7000 sentences as the training set, and the 2000 sentences later as the test set for tagger training. The training process includes the following three steps:

1. Try to use the bigram tagger to tag the identifier

2. If the bigram tagger cannot find the tag, try the unigram tagger.

3. If the unigram tagger cannot find the tag, use the default tagger.

The accuracy of the final tagger in the testing process is 0.9875608103947288. It shows that this tagger can tag parts of speech more accurately. However, it is worth noting that when the tagger is annotating the corpus outside this corpus, the accuracy of the tagging will be significantly reduced, which is also one of the main sources of error afterwards.

2.2 Corpus-based part-of-speech screening

2.2.1 Preliminary part-of-speech screening

To judge the context reasonably and correctly, we must understand the composition of the context and the difference between the question sentence and other sentences. Our idea is to judge by the frequency of different parts of speech in the sentence.

First, we try to distinguish which part of speech words in question group and common sentence group will have obvious changes. For this reason, we perform statistical analysis on the question and answer sentences in the existing question and answer corpus (corpus: overheard):

In the first step, we count the 9 parts of speech in all sentences: NN, JJ, PRP, VB, WP, IN, DT, CD, and WDT. Since our purpose is to distinguish those special parts of speech in question sentences, we ignored the concept of sentence boundaries and performed an overall frequency analysis of all words in similar sentences, namely:

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

In the formula, x represents a certain part of speech, i represents the survey point, n represents the number of words that appear in the part of speech, N represents the number of words up to the survey point, and P represents the frequency of occurrence of the word. This formula means the number of occurrences of the target part-of-speech word from the first word of the surveyed group to the survey point. We traverse all the cases of 0<i<15000 for the two sentence groups, and get the following results:

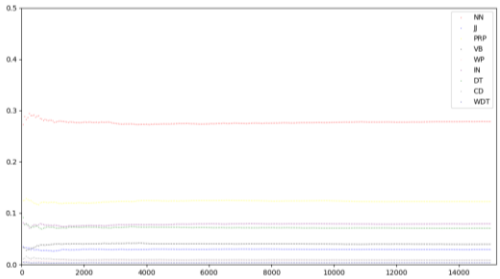

Figure 2-1. Probability of different part-of-speech words in all sentences

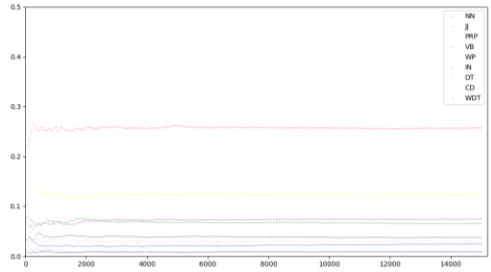

Figure 2-2. Probability of different part-of-speech words in question sentences

Figure 2-1 describes the phenomenon that the frequency of each part-of-speech vocabulary in the question group tends to stabilize with the increase in the number of samples. Figure 2-2 describes the frequency of each part-of-speech vocabulary in all sentence groups tends to stabilize with the increase in the number of samples. The phenomenon. The interpretation of these two figures is as follows: 1. In the case of sufficient data sampling, the frequency of part of speech will tend to a stable value, indicating that the appearance of part of speech in the language is convergent, and statistical analysis can be carried out. This also conforms to Bernoulli's Law of Large Numbers [5];

Table 2-1 The frequency of each part of speech

Part of speech	Frequency in all sentence groups	Frequency in question
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. The frequency of occurrence of part of speech in different sentence groups has changed significantly, indicating that part of speech can be used to distinguish different types of sentence groups. These two points are also the basic assumptions that we can conduct in-depth analysis later.

In order to further see the difference between the two kinds of sampling, we show the frequency convergence value under various conditions in Table 2-1.

2.2.2 Refined part-of-speech screening and selection

In fact, the part of speech examined in 2.2.1 is only a general classification of part of speech, and each part of speech has more specific subparts. For example, NN is divided into NNP, NNS, etc. This section discusses the frequency of these specific parts of speech. We subdivide the part of speech, such as dividing NN into NNS and NNP, and dividing VB into VBN, VBG, VBD, and VBZ, and count these six parts of speech in all sentences and questions respectively, and make a scatter diagram like 2 -3, 2-4.

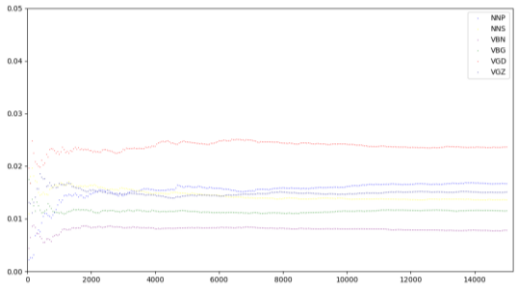

Figure 2-3. The occurrence probability of subdivided part-of-speech words in all sentences

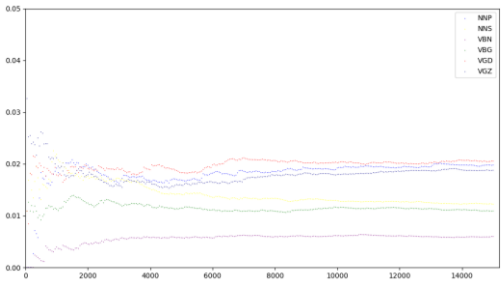

Figure 2-4. Probability of segmented part-of-speech words in question sentences

Table 2-2 Frequency of appearance of each part of speech and subdivided part of speech

Part of speech	Frequency in all sentence groups	Frequency in question
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

These part-of-speech words still satisfy the two basic assumptions mentioned above. We will combine the results of the coarse score and the subdivision and display the list in Table 2-2:

Now we must select from these data the part of speech frequency that has obviously changed, that is, the part of speech that may be used as a criterion for distinguishing question sentences from all sentences. We use the method of combining absolute error and relative error, that is: 1. Calculate the relative error of each part of speech in different sentence groups, standardize the error (the part of speech error/the sum of all errors), and take the relative error sequence range lower four points All results above digits are sample space A; 2. Calculate the absolute error of each part of speech in different sentence groups, standardize the errors (the method is the same as above), and take all results above the quartile of the absolute error sequence range as the

sample space B. Then the part of speech sample space C we finally selected can be expressed as:

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

Table 2-3 table shows the screening process:

Table 2-3 Errors, standardization, and screening of each part of speech and segmented part of speech

Part of speech	Relative error	Standardized relative error	Whether selected (A)	Absolute error	Standardized absolute error	Whether selected (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

The part of speech sample space we selected at the end is:

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

2.3 Parameter estimation and hypothesis testing

2.3.1 Parameter Estimation

After obtaining a preliminary selection, we examine the frequency distribution of these parts of speech in individual English questions. Figures 2-5 to 2-10 intuitively reflect the distribution of frequency samples of each part-of-speech vocabulary in our sampling space (overheard's first 1500 questions):

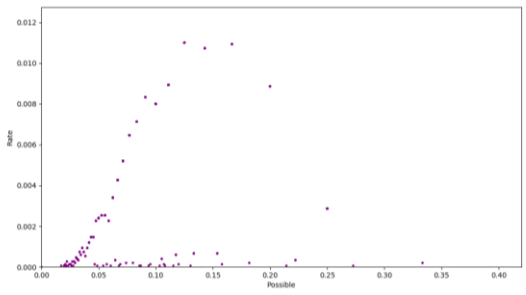

Figure 2-5 Distribution of VBZ

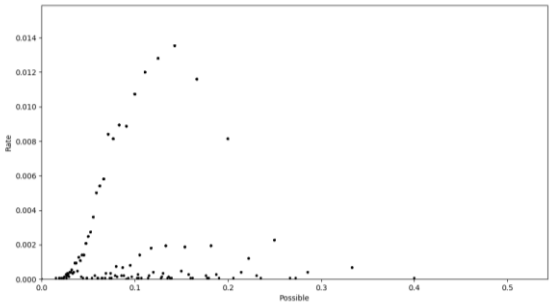

Figure 2-6 Distribution of VBD

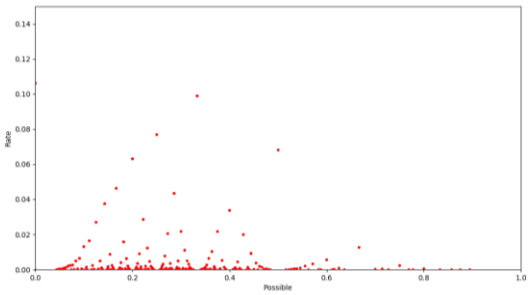

Figure 2-7 Distribution of NN

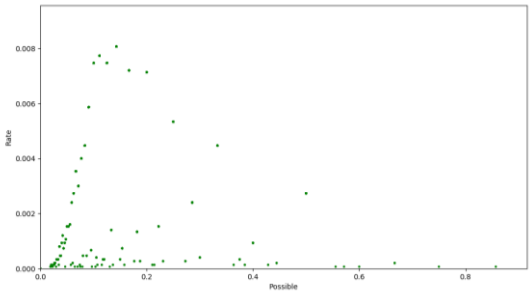

Figure 2-8 Distribution of NNP

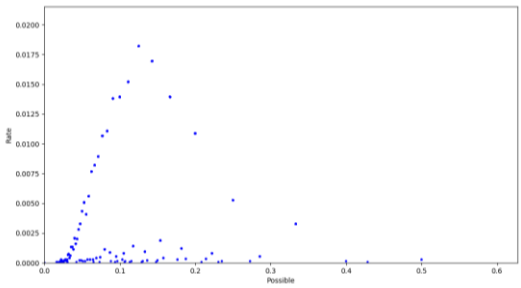

Figure 2-9 Distribution of JJ

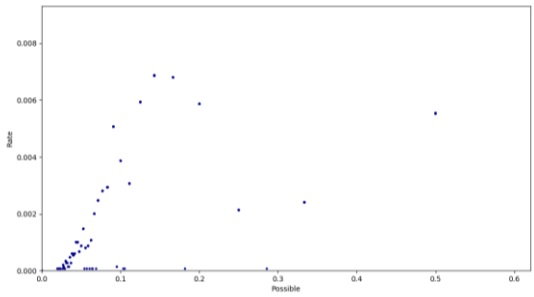

Figure 2-10 Distribution of WP

For these sampled data, we hope to obtain the population mean and variance under a certain confidence interval through the sample mean and variance. According to the central limit theorem, the frequency distribution of each part of speech is similar to the normal population. From the related concepts of interval estimation and sampling inspection theorem, when the overall variance is unknown, there is a pivot quantity [4-5]:

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

Therefore, the confidence interval of the parameter μ under the confidence level of 1-α is:

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

When the overall expectation is unknown, there are pivot quantities:

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

Therefore, the confidence interval of parameter σ2 at the confidence level of 1-α is:

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

From equations (2.3) to (2.6), we can get (confidence probability 95%):

Table 2-4 Interval estimation

Part of speech	Sample mean	Sample variance	Mean confidence interval	Variance confidence interval
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 hypothetical test

We need to do a hypothesis test on the mean and variance obtained. We randomly sampled another 15,000 samples from the overheard corpus. Since the number of data is about 15,000, the sample obtained is approximately a normal population, so we can use the hypothesis test of the normal population to process the sample.

For the hypothesis test of the mean μ, when the population variance σ2 is unknown, we select the test statistic:

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

The rejection conditions are:

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

Bring in the relevant value, you can observe whether its value falls within the rejection range.

Similarly, for the hypothesis test of variance σ2, when the mean μ is unknown, we choose the test statistic:

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

The rejection conditions are:

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

Bring in the relevant value and observe whether the value falls within the rejection range.

Under the condition of choosing α=90%, the following table can be obtained from the above formula:

Table 2-5 Hypothesis test

Part of speech	Null hypothesis	Sample mean	Sample variance	Rejection domain	Whether to reject
NN:	μ=0.209438	0.206693724		|T|≥1.644955225	Yes
	σ2=0.029231		0.028755056	c2≤14715orc2≥15285	No
NNP:	μ=0.01251	0.013087245		|T|≥1.644955225	No
	σ2=0.002567		0.002706519	c2≤14715orc2≥15285	Yes
JJ:	μ=0.019149	0.019136055		|T|≥1.644955225	No
	σ2=0.002343		0.002332654	c2≤14715orc2≥15285	No
WP:	μ=0.002021	0.001961551		|T|≥1.644955225	No
	σ2=0.000246		0.000238607	c2≤14715orc2≥15285	Yes
VBD:	μ=0.01432697	0.014593306		|T|≥1.644955225	No
	σ2=0.001646598		0.001677898	c2≤14715orc2≥15285	No
VBZ:	μ=0.009763	0.009722346		|T|≥1.644955225	No
	σ2=0.00114		0.001129685	c2≤14715orc2≥15285	No

Under the requirement of rejecting the mean value, the part of speech we finally selected were: NNP (proper noun), JJ (adjective), WP (words beginning with wh-, mainly Yes interrogative words), VBD (vertical past tense), VBZ (verb Third person singular). This shows that these five parts of speech can be used as the source of context.

It should be noted that Yes, doing the same statistical modeling in another corpus, the results obtained have undergone certain changes, which shows that there are different types of parts of speech for generating context for different corpora. Therefore, in order to obtain more general results, it is necessary to synthesize multiple corpora for statistical modeling.

3 Context Vector Method and Context Space

3.1 Basic assumption

The context vector method is proposed, and Yes is based on the above analysis of parts of speech. In the statistical analysis of 2.2, we found that the part of speech that can represent the question sentence has six types: NN,

NNP, JJ, VBD, VBZ, and WP. In the statistical analysis of 2.3, we found that these six types can be used as clear referents. The part of speech is NNP, JJ, WP, VBD, VBZ. Based on this, we put forward three basic hypotheses. If these hypotheses are recognized (using the experience in actual communication, this hypothesis Yes is reasonable), these parts of speech are related to the context and can be used as the criterion for judging the context.

1. The validity of the content word: It is believed that any content word in the sentence has a specific meaning to be referred to.

2. Dialogue continuity: People have the concept of dialogue rounds in the process of dialogue, and each round is in the same context.

3. Verb weakening: In dialogues, especially in Yes English dialogues, actions are usually not the main message to be conveyed.

Among the five parts of speech obtained, based on these three hypotheses, the context will be mainly generated from NNP and WP.

3.2 Details

Context vector, that is, construct a vector under a certain general environment (that is, the overall environment beyond the context, such as the place where the dialogue occurs), so that each dimension of it can represent the content of the context of the current dialogue round , So that you can mark the context of the dialogue round. Several such linearly independent vectors constitute the context space in this general environment. Its basic representation is:

[Context 1, Context 2, Context 3, …]

Each dimension in the context vector is separated from a specific part of speech. But it should be noted that one part of speech may create multiple dimensions in the context vector, if the part of speech is filtered by the method in 2. But it should also be noted that the context separated from the part of speech may be different in different contexts. Therefore in 2 we repeatedly emphasized that the acquisition of part of speech is only to obtain the scope of context, not directly. Get context. The context also needs to be obtained by separating the specific categories of the part of speech and forming the context vector.

The comparison between context vectors is like the comparison of Euclidean space vectors, that is, if two vectors have the same dimension in n dimensions, we consider their distance to be i/n. At the same time, when i/n is less than a certain value, we can think that these two contexts are irrelevant.

With the concept of context vector, context vector can be used in conjunction with other technologies to construct a dialogue system with the function of "time memory". An example will be given in 3.3.

3.3 Scope of application

In the corpus we examined, the abstract part of speech has, and the corresponding specific content is: name, location, time, and question words starting with W. Therefore, the abstracted context vector is:

$$\begin{bmatrix} Name, place, time, question \\ \\ \end{bmatrix}$$

After obtaining the context vector, since the content of each dimension needs to be filtered using the existing data, we can make certain extensions to facilitate screening in the database. For example, the vector can be expanded to:

$$[(Male\ name,\ Female\ name),\ (City,\ Country),\ (Year,\ Month,\ Day,\ Time),\ Question]$$

Of course, the way to expand is not unique. In this way, when building a specific system, we can more easily extract contextual vocabulary from sentences by comparing with the database. We built a simplest dialogue system framework to reflect this process, and its output results are as follows:

Figure 3-1 A dialog system framework

The system successfully returned the context vector. It is worth noting that after Obama uses he to refer to it, the context vector can still determine that he refers to Obama. This solves the problem raised in 1. Using these returned results, you can combine other natural language processing tools to construct a dialogue system that can recognize context.

4 Results and discussion

4.1 Project innovation

The main innovations of this project are as follows:

1. Pay attention to the actual context in the dialogue process and get closer to people's real needs for the question answering system

2. Use part of speech to extract context, focusing on the difference between question sentences and ordinary sentences in terms of part of speech

3. Hypothesis testing was conducted to demonstrate the feasibility of the method

4. Different situations are classified to achieve comprehensiveness in the overall sense

4.2 Main source of error

In the previous process, the possible sources of error include: the accuracy of the part-of-speech tagger decreases after being separated from the training corpus; the frequency of each part-of-speech in different types of dialogue databases may be different; the collocation between different parts of speech may lead to differences in distribution; When constructing the context vector, the context choices from different parts of speech may be different.

4.3 To be improved

Based on the main errors, the parts that can be improved include: designing a more reasonable part-of-speech tagger; performing a unified frequency analysis of multiple types of dialogue libraries; in-depth research on the accurate method of extracting context from the filtered part of speech; combining context vectors with question and answer System integration.

5 Conclusion

The research is based on the different frequency of part of speech in different types of sentences, statistical modeling of question and answer corpus, and extracting the characteristic parts of speech that can distinguish different types of sentences. Through the screening of parts of speech, the context vector describing the context created in the dialogue is finally constructed, and simple application is carried out. The code of all programs in the article can be found on github: https://github.com/MullerLee/ChatBot2.0

Acknowledgement

I am very grateful to Teacher Zhang Peng from the School of Computer Science for his continuous help and support, who provided us with a lot of valuable knowledge and practical resources and suggestions, and solved many problems in our research process; thanks to all the students who participated in the research. At the same time, I am very grateful to the School of Microelectronics for its continuous support for this project. Without this support, we would not be able to go to the present. Finally, I would like to thank Ma Liqun, Wang Benyou, Su Zhan and other seniors for their help, who provided us with many valuable ideas.

Reference

[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.