2021 IEEE International Conference on Software Engineering and Artificial Intelligence (SEAI) | 978-1-6654-1889-8/20/$31.00 ©2021 IEEE | DOI: 10.1109/SEAI52285.2021.9477550

Leveraging Zipf's Law to Analyze Statistical Distribution of Chinese Corpus

Qing Lei
University of International Business and Economics
Beijing, China
leiqing@uibe.edu.cn

Haifeng Li
KGraphX.com
Beijing, China
leeivan2008@gmail.com

Rongbin Wei
University of International Business and Economics
Beijing, China
201906058@uibe.edu.cn

***Abstract*—In term of Chinese natural language processing, it exits one particular problem that how to choose the strategy of word segmentation, which commonly includes char-based and word-based. Targeted at sentiment analysis of short text comparing with long text, the word-based segmentation faces the other problem that there are the more ambiguous or unregistered words in context of short text. The feature extraction done by the different Chinese Word Segmentation impact the statistic distribution of features, and further the accuracy of sentiment analysis. This paper evaluates five Chinese segmentation strategy effect on Sentiment Analysis of Short Text. We chose two word-based Chinese Word Segmentation (CWS), and three char-based n-gram and made usage of Zipf's law to quantify and present the result of word segmentation.**

Keywords—Zipf, Chinese Segmentation, feature extraction

I. INTRODUCTION

Currently, the more and more initiative internet gadgets continuously have been developed, which promote the rapid evolution of information spreading and communicating patterns. Billions of short texts are produced every day, in the form of search queries, ad keywords, tags, tweets, messenger conversations, social network posts, etc. [1] Short text refers to text forms that are relatively short in length and generally do not exceed 160 characters, such as Weibo, chat information, news topics, opinion comments, question texts, mobile phone text messages, and document summaries.

The short texts own some special characters instead of long document sentence. The short texts are short in length, usually the semantics expression is not complete enough, which means traditional NLP techniques, such as syntactic parsing, do not always apply to short texts. Furthermore, unregistered words in the short text greatly affect the accuracy of the word segmentation system, and its appearance can easily cause word segmentation errors in parts of the sentence. The accuracy of the word segmentation caused by unregistered words far exceeds the error caused by ambiguous segmentation. But, the recognition of unregistered words becomes more troublesome. The so-called unregistered words, also called out of vocabulary (OOV), refer to words that are not included in the word system dictionary, and also refer to words that have not appeared in the training set. The unregistered words generally can be divided into two categories: one is a proper noun such as a person's name, a place name, or an institution name; the other is a new general term or terminology. With the change of people's expression and the development of society, new nouns and new words are emerging. The proper nouns are slow to update, but they are not likely to be included in the dictionary.

On the other hand, the Indo-European languages, represented by English, have obvious spaces between words as boundaries. In general, words can be extracted relatively simply and accurately. Relatively speaking, Chinese word segmentation is much more complicated. The word in Chinese not only can be just one character, or but also they can be made up of two characters, or three or four. The characters in Chinese sentences are in a state of close connection, and there is no obvious segmentation mark and morphological change between words. In natural language understanding, the word is the basic linguistic components of meaning, and the understanding of word is a prerequisite for understanding the meaning of sentences. The process has caused a lot of trouble.

So, the different segmentation strategies produce diversities of statistic distribution for each feature. The features of short text are difficult to be extracted correctly, in addition of showing the characteristics of sparse matrix and high dimensions. Therefore, how to extract appropriate features from short texts and reflect them towards the correct sentiment events has gradually become one of the hotspots in the field of sentiment analysis.

II. RELATED WORK

Following development of Chinese NPL technology, many researchers take advantage of comparative strategy to analysis sentiment classification from different perspectives. Such as, (Meng, et al., 2019) observed that char-based models consistently outperform word-based models. Building upon these findings, they show that word-based models' inferiority is due to the sparseness of word distributions, which leads to more OOV words, overfitting and lack of domain generalization ability[2]. (Zhang et al., 2017) pointed out that word-level encoding for CJK (Chinese, Japanese and Korean) languages are competitive even without perfect segmentation, for both fastText and linear models[3].

Supported by Beijing Municipal Social Science Foundation 18GLB021.

A. Chinese Word Segmentation

Chinese Word Segmentation (CWS) technology is a fundamental task of natural language processing. Although the existence of ambiguous words has caused great difficulties for Chinese word segmentation, it is still possible to increase the correct rate of word segmentation by continuously expanding and revising the thesaurus. Relatively speaking, the recognition of unregistered words becomes more troublesome. The unregistered words, refer to words that are not only included in the word system dictionary, and but also refer to words that have not appeared in the training set. The accuracy of the Chinese word segmentation system will directly affect and pass on to the next level of tasks, which will affect the reliability and practicability of the relevant Chinese information processing system, such as sentiment analysis.

As the focus on Chinese processing increases, the Chinese Special Interests Study Group (SIGHAN [1]) under the International Computing Languages Association (ACL[2]) holds an annual International Chinese Word Segmentation Competition[4][5]. SIGHAN Bakeoff's word segmentation defines strict closed test conditions, requiring that language resources other than the training set be used, otherwise the corresponding results are considered open test categories. One of the main purposes of distinguishing between closed and open tests is to distinguish the performance improvement of machine learning from the improvement of the model itself, not the others.

The primary problem in Chinese information processing is to divide each character without a delimited mark into a reasonable sequence of words. In most intelligent natural language tasks such as machine translation, information retrieval, text classification, and speech recognition, the word segmentation system is required as its basic module and key technical point.

TABLE I. FUNDAMENTAL STATE COMPARISON AMONG THE CURRENT CWS

Python API	Latest Version	Core Function	Repository
Jieba	0.39	N/A	https://github.com/fxsjy/Jieba
SnowNLP	0.12.3	N/A	https://github.com/isnowfy/snownlp
PyNLPIR	0.5.2	NLPIR	https://github.com/tsroten/pynlpir
thulac	0.2.0	THULAC	https://github.com/thunlp/THULAC-Python
stanford-corenlp	3.9.2	CoreNLP	https://github.com/stanfordnlp/stanfordnlp
pyLTP	0.2.1	LTP	https://github.com/HIT-SCIR/pyLtp

An efficient Chinese word segmentation system with excellent performance should have several basic elements: word segmentation accuracy, word segmentation speed, system maintainability, versatility, and adaptability. During decades, several mature word segmentation systems emerged out. NLPIR Chinese word segmentation system was developed before 2004 and named ICTCLAS. It supports Chinese and English mixed word segmentation, new word recognition and adaptive word segmentation, keyword extraction and user professional dictionary function. THULAC (THU Lexical Analyzer for Chinese) is a set of Chinese lexical analysis toolkit developed by Tsinghua University's Natural Language Processing and Social Humanities Computing Laboratory. It has Chinese word segmentation and part-of-speech tagging. The Language Technology Platform (LTP) is a complete Chinese language processing system developed by the Harbin Institute of Social Computing and Information Retrieval Research Center for ten years. LTP has developed an XML-based language processing result representation, and on this basis, it provides a complete set of bottom-up rich and efficient Chinese language processing modules (including 6 Chinese processing core technologies such as lexical, syntactic and semantic), and based on The application interface and visualization tools of the Dynamic Link Library (DLL) can be used in the form of web services. Stanford CoreNLP provides a set of human language technology tools. It can give the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize dates, times, and numeric quantities, mark up the structure of sentences in terms of phrases and syntactic dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, extract particular or open-class relations between entity mentions, get the quotes people said, etc. [6]. Based on a prefix dictionary structure to achieve efficient word graph scanning. Build a directed acyclic graph (DAG) for all possible word combinations. Use dynamic programming to find the most probable combination based on the word frequency. For unregistered words, the HMM-based model is used with the Viterbi algorithm.

The benchmark data built by Boson[3] shows the results of the segmentation accuracy of the 11 segmentation engines participating in the comparison in different data. It can be seen that BosonNLP and Harbin University Language Cloud have achieved high word segmentation accuracy on the four data sets tested, especially on news data[4].

III. EXPERIMENT AND ANALYSIS

(Socher et al., 2013) propose two ways of benchmarking. First, one could consider a 5-way fine-grained classification task where the labels are (Very Negative, Negative, Neutral, Positive, Very Positive) or a 2-way coarse-grained classification task where the labels are (Negative, Positive)[7]. In our experiment, we planned to use binary classification evaluate the five CWS methods.

A. Data Preparation

In this paper, the five corpuses of short text were acquired from a public dataset community[5].

During the proceed of ETL (Extract-Transform-Load), the reviews and ratings of products, hotels or restaurants were

[1] http://sighan.cs.uchicago.edu/
[2] https://www.aclweb.org/portal/
[3] https://bosonnlp.com/
[4] https://bosonnlp.com/dev/resource, 11 open source CWS engine benchmark.
[5] https://github.com/SophonPlus/ChineseNlpCorpus

extracted and concatenate to one dataset. After duplication removal, we were able to transform the 5 scale ratings to 2 and obtain binary label in the dataset. The binary label was built such that the 1 and 2 scale belong to the negative sentiment, and the 4 and 5 scale belong to the positive sentiment, and the 3 scale is ignored. The entire dataset was split with the training

TABLE II. THE FIVE CORPUSES OF SHORT TEXT

Datasets	Paper	Resource	Data Overview
ChnSentiCorp_htl_all		Ctrip[a]	More than 7,000 hotel reviews, more than 5,000 positive reviews, more than 2,000 negative reviews
waimai_10k	N/A	N/A	The user rating collected by a takeaway platform is 4,000 positive and 8000 negative reviews
online_shopping_10_cats	N/A	N/A	10 categories, with more than 60,000 comments, about 30,000 positive and negative reviews, including books, tablets, mobile phones, fruits, shampoos, water heaters, Mengniu, clothes, computers, hotels
Dianping Review Dataset	[8]	Dianping[b]	240,000 restaurants, 540,000 users, 4.4 million reviews/rating data
JD.com E-Commerce Data	[9]	JD[c]	520,000 items, more than 1,100 categories, 1.42 million users, 7.2 million reviews / rating data

a. Ctrip.com International, Ltd. is a Chinese provider of travel services including accommodation reservation, transportation ticketing, packaged tours and corporate travel management.

b. dianping.com hosts consumer reviews of restaurants, similar to Yelp and TripAdvisor, and also offers group buying similar to Groupon.

c. JD.com, Inc., also known as Jingdong and formerly called 360buy, is a Chinese e-commerce company headquartered in Beijing.

and test datasets, there are 6,089,432 training samples with (25.31% negative, 74.69% positive), and 676,604 testing samples with (25.26% negative, 74.74% positive).

B. *Exploratory Data Analysis*

After procedure of segmentation including the Bigram, Trigram, Ltp, Jieba, the huge number of low-frequency features is produced in the training data which is sparse matrix, the models maybe capture idiosyncrasies of the training data, but not generalization for testing data. On the other hand, the Unigram produces less features than the four others and not reflect the structure of sentence, so maybe make the model impossible to fully learn context semantics. During training machine learning model, features sparsity inevitably leads to overfitting, whereas features scarcity underfitting.

In this section, we analyze features to summarize their main characteristics. The exploratory data analysis (EDA) is a crucial step to take before diving into machine learning because it provides the context needed to develop an appropriate model for the problem at hand and to correctly interpret its results. EDA is valuable to the data scientist to make certain that the results they produce are valid, correctly interpreted, and applicable to the desired business contexts.

C. *Quantitative Features Distribution*

The Zipf's law was originally formulated in terms of quantitative linguistics, stating that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Particularly, Zipf's law applies to most languages which include Chinese. The statistical results show that Zipf's laws in 50 languages all share a 3-segment structural pattern. In the classical representation of Zipf's law, $\boldsymbol{f \cdot r = C}$ is relatively simple, $\boldsymbol{f}$ is the word frequency, $\boldsymbol{r}$ is arranged in reverse order of word frequency, and $\boldsymbol{C}$ is a constant. Put $\boldsymbol{f}$ and $\boldsymbol{r}$ in a double logarithmic coordinate (log-log) system, the curve drawn is almost a straight line, and the slope is approximately $\boldsymbol{-1}$. In order to accurately solve this slope, Zipf's law can also be expressed as $\boldsymbol{f = Cr^{-a}}$, taking the logarithm of the two sides to obtain $\boldsymbol{\log f = \log C - a \log r}$, then in the double logarithmic coordinate system, a is the slope of the line, $\boldsymbol{\log C}$ is the intercept of the fitted line on the y-axis.

In experiments in different languages, empirical data indicates $\boldsymbol{a \approx 1}$, the $\boldsymbol{a}$ is very similar in different texts, but the performances of different languages are not exactly the same, as English is very consistent, but Chinese is not strictly consistent, from it has been studied that there are differences between different styles, and even the same language presents different distributions at different times. So, the $\boldsymbol{a}$ is an important parameter for distinguishing between different languages or different styles, so it is very important to calculate a accurately. This law has been verified, and foreign research has found that Zipf's law exists in large-scale text and is applicable to multiple. languages. The Zipf's curve of the five segmentation approaches are shown in Fig. 1, which plots value of rank and frequency on log-log graph.

Note that the Unigram curve is different from the curves of the four others, first with a slope less than 1 ($\boldsymbol{a < 1}$) roughly closing to the Zipf's curve of ($\boldsymbol{a = 1}$), then falling rapidly and deviating from the Zipf's curve of ($\boldsymbol{a = 1}$) after a rank of about 1,000 and below a frequency of about 50,000. The Unigram is more curved than the four others because there are more high-frequency features and less low-frequency features.

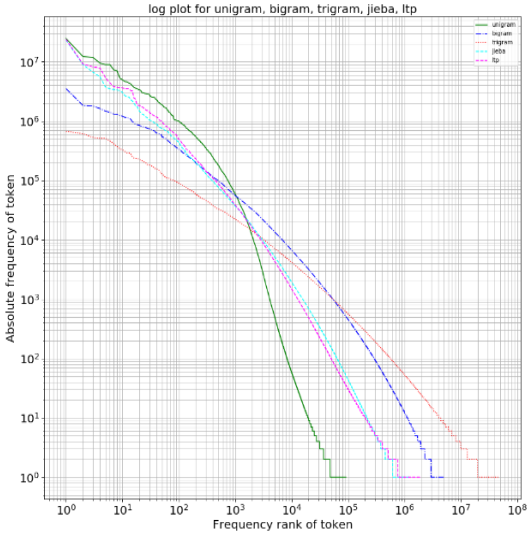
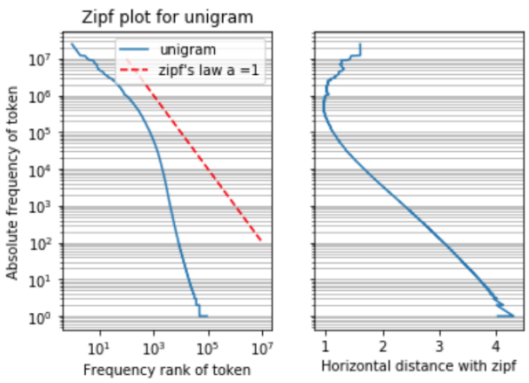
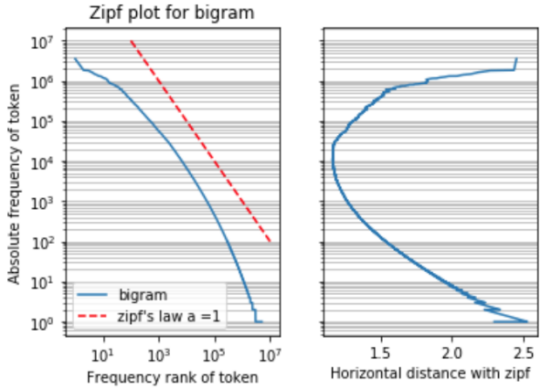


Fig. 1. log plot for Unigram, Bigram, Trigram, Jieba and Ltp

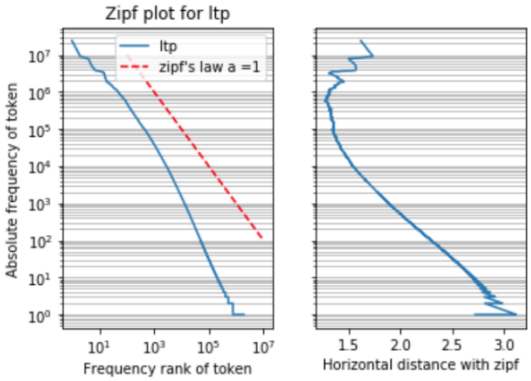


Unigrams

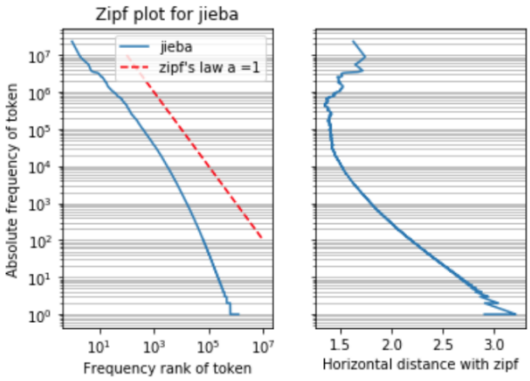


Bigrams

Fig. 2. Unigrams & Bigrams



Ltp



Jieba

Fig. 3. Ltp & Jieba

The number of high-frequency features for the five curves is arranged in descending order: $\boldsymbol{unigram > ltp > jieba > bigram > trigram}$; the number of low-frequency features is arranged in descending order: $\boldsymbol{trigram > bigram > jieba > ltp > unigram}$. The crossing-point between Unigrams and Bigrams is at frequency of about 50,000; between the Unigram and Trigram curves at frequency of about 15,000; among the Unigram, Jieba and Ltp curves at frequency of about 20,000. Furthermore, the curves of Jieba and Ltp almost fully coincide, the Ltp produce more high-frequency features and less low-frequency than the Jieba.

The Zipf's curve reflects the power law relation between the frequency and rank of feature. In the Fig. 2, Fig. 3, Fig. 4, the kind of phenomenon is invariably showed that the log-log graph (left-blue) of rand and frequency for the five segmentation exhibit a 3-segment structural pattern. The horizontal distance with Zipf's curve (right-blue) is plotted and roughly shows segment point of the pattern. The upper segment of each curve is unsmooth, with the gradient being roughly $< \mathbf{1}$; the middle segment of each curve is smooth, with the gradient being roughly $= \mathbf{1}$. The lower segment of each curve is also smooth, but bends downward to deviate from the expected line (left-red), with the gradient being roughly $> \mathbf{1}$. The inflection points of the five curves are showed in the TABLE III.

Additionally, the upper and the middle segments of curves rise in accordance with the growth of the sample size, but the lower segment rises much slower, leading to the downward bending of the curve, which suggests two quite different

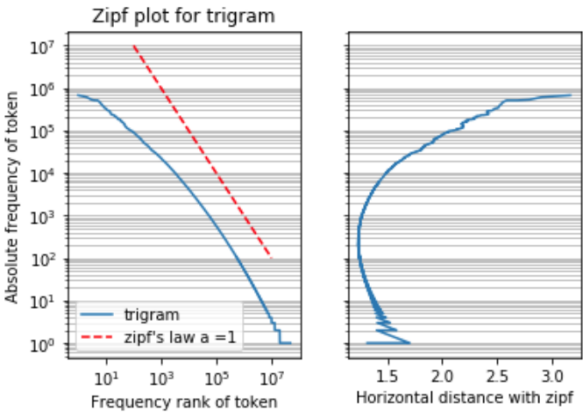


Fig. 4. Trigrams

TABLE III. THE $\boldsymbol{a}$ VALUE INTERVEL OF THE FIVE ZIPF'S CURVES

	$a < 1$	$a = 0$	$a > 1$
Unigram	$[7 * 10^5, +\infty)$	$[3 * 10^5, 7 * 10^5)$	$[0, 3 * 10^5)$
Bigram	$[3 * 10^4, +\infty)$	$[7 * 10^3, 3 * 10^4)$	$[0, 7 * 10^3)$
Trigram	$[6 * 10^2, +\infty)$	$[1 * 10^2, 6 * 10^2)$	$[0, 1 * 10^2)$
Ltp	$[4 * 10^5, +\infty)$	$[1 * 10^5, 4 * 10^5)$	$[0, 1 * 10^5)$
Jieba	$[2 * 10^5, +\infty)$	$[3 * 10^4, 2 * 10^5)$	$[0, 3 * 10^4)$

generating mechanisms for high-frequency, mid-frequency and low-frequency words.

The high-frequency words have little meaning on their own, but they do contribute a great deal to the meaning of a sentence. On the other hand, low-frequency words can be typos or rare words which cause the data sparsity issue and is likely to induce overfitting, since more words means a larger number of parameters. In addition, since it is unrealistic to maintain a huge word-vector table, many words are treated as OOVs, which may further constrain the model's learning capability. The words occurring only once in the whole corpus are called hapax legomenon (1-leg), the related terms of dis legomenon (2-leg), tris legomenon (3-leg), and tetrakis legomenon (4-leg) refer respectively to double, triple, or quadruple occurrences, but are far less commonly used. Some applications ignore most high-frequency words, usually listed in a stop-list, and pay more attention to mid-frequency or low-frequency words. But, (Saif, et al., 2014) indicated that despite the popular usage with stop-list, the precompiled (classic) stop-list has a negative impact on the classification performance, and removing singleton words is the simplest, yet most effective practice, which keeps a very good trade-off between good performance and low processing time[11]. Furthermore, (Heap, et al., 2017) A particular problem with classification of short texts is low-frequency words that do not occur at all in the text used to train models, but do occur in test data[12].

Let us delve into the value of frequency and rank. The three highest frequency of Unigrams, Bigrams, Trigrams, Ltp and Jieba are in 0 and 0. For Unigram, Ltp and Jieba, the majority of those words among the ten highest frequency are stop words, such as "的", "了", "也", "是", and so on. Words such as articles and some verbs are usually considered stop words because they don't help us to find the context or the true meaning of a sentence. These are words that can be removed without any negative consequences to the final model that you are training. But here's the catch: there's no universal stop words list because a word can be empty of meaning depending on the corpus you are using or on the problem you are analyzing. Word importance may vary depending on the dataset. But it may also change depending on the goal you are trying to achieve. Problems like sentiment analysis are much more sensitive to stop words removal than document classification.

TABLE IV. THE TEN HIGHEST TOTAL FREQUENCY OF UNIGRAMS, BIGRAMS AND TRIGRAMS

Rank	Unigrams				Bigrams				Trigrams			
	words	*positive*	*negative*	*frequency*	*words*	*positive*	*negative*	*frequency*	*words*	*positive*	*negative*	*frequency*
0	的	16429138	8258711	24687849	不 错	2823314	707982	3531296	的 时 候	423041	253171	676212
1	不	7808440	4453837	12262277	喜 欢	1423166	413588	1836754	不 错 的	486188	127073	613261
2	是	7475272	4383223	11858495	很 好	1508431	288035	1796466	很 不 错	445150	87125	532275
3	了	5786496	3710202	9496698	味 道	916025	715161	1631186	很 喜 欢	445891	67607	513498
4	一	5635078	3333005	8968083	可 以	999789	476192	1475981	还 不 错	387933	125536	513469
5	很	6795210	2160491	8955701	还 是	942625	464049	1406674	服 务 员	207321	257391	464712

TABLE V. THE TEN HIGHEST TOTAL FREQUENCY OF LTP AND JIEBA

Rank	Ltp				Jieba			
	words	positive	negative	frequency	words	positive	negative	frequency
0	的	15887899	7972763	23860662	的	15536322	7783382	23319704
1	了	5550778	3580450	9131228	了	5443481	3496153	8939634
2	很	6263990	1954167	8218157	很	5225077	1482729	6707806
3	是	4902732	2953044	7855776	是	3501188	2080385	5581573
4	不	3008282	2402189	5410471	也	2447063	1333361	3780424
5	好	3441410	965564	4406974	我	2163190	1329641	3492831

TABLE VI. THE LOW FREQUENCY WORDS PERCENTAGE OF THE FIVE SEGMENTATION

	Unigram		Bigram		Trigram		Ltp		Jieba	
NOT	539,380,924		532,616,067		525,851,291		361,889,668		332,683,312	
NOW	91,651	0.017%	5,044,482	0.947%	48,158,084	9.158%	1,890,717	0.522%	1,240,356	0.373%
1-leg	43,468	0.008%	2,081,345	0.391%	28,380,701	5.397%	1,128,192	0.312%	624,131	0.188%
2-leg	11,355	0.002%	664,847	0.125%	6,741,785	1.282%	247,078	0.068%	159,955	0.048%
3-leg	5,513	0.001%	355,347	0.067%	3083,800	0.586%	110,640	0.031%	76,840	0.023%
4-leg	3,738	0.001%	235,333	0.044%	1840,171	0.350%	66,046	0.018%	49,068	0.015%

From 0, the NOT (number of token) is total number of tokens after segmentation, the NOW (number of words) is the total number of distinct tokens. The Unigram, Bigram and Trigram are the similar level of magnitude, Ltp and Jieba are the similar level, but less than before. This situation shows that segmentation system drops the part of tokens, which belong to OOV. The percentage of the NOW to the NOT is arranged in descending order: $\boldsymbol{trigram > bigram > ltp > jieba > unigram}$. The percentage 9.158% of Trigram NOW is the first, even a lot more than the second (0.947%). This situation shows that the features of Trigram must be very huge and the vector space very sparse, which causes the issue of overfitting. On the other hand, the percentage 0.017% of Unigram is the last, a lot less than Jieba (0.373%), this situation shows that the features of Unigram are very scarce, which causes the issue of underfitting. According to the data of the 1-reg, 2-reg, 3-reg, 4-reg, they all have the same characteristics.

IV. CONCLUSION

In this paper, we completed five types of word segmentation strategy which included the Unigram, Bigram, Trigram, Ltp and Jieba. Using the Jieba, which is the most widely-used open-sourced Chinese word segmentation system, we ended up with a dataset consisting of 332,683,312 words with 1,240,356 distinct words. Among the 1,240,356 distinct words, 624,131 words appeared only once, amounting to 50.3% of the total vocabulary, yet they only took up 0.188% of the entire corpus. If we increased the frequency bar to 4, we got 909,994 words appearing less or equal to 4 times, which contribute to 73.4% of the total vocabulary but only 0.274 % of the entire corpus. Comparing the Jieba, the Ltp-segmented data is very sparse, and the Jieba tends to find more duplicate words.

In the future study, we decide to feed those CWS results into several machine learning algorithms, and evaluate influence to sentiment analysis, and could make recommendation on how to choose CWS strategy in the task of NLP.

REFERENCES

[1] Zhongyuan Wang and Haixun Wang, Understanding Short Texts, in the Association for Computational Linguistics (ACL), August 2016.

[2] Yuxian Meng, Xiaoya Li, Xiaofei Sun, Qinghong Han, Arianna Yuan, Jiwei Li, Is Word Segmentation Necessary for Deep earning of Chinese Representations?, ACL2019.

[3] Xiang Zhang, Yann LeCun, Which Encoding is the Best for Text Classification in Chinese, English, Japanese and Korean?, 8 Aug 2017

[4] Huiming Duan, Zhifang Sui, Tao Ge, The CIPS-SIGHAN CLP 2014 Chinese Word Segmentation Bake-off

[5] Lu Xiang, Xiaoqing Li, Yu Zhou, Word Segmenter for Chinese Micro-blogging Text Segmentation – Report for CIPS-SIGHAN'2014 Bakeoff

[6] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

[7] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, C. Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In Proceedings of EMNLP 2013.

[8] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu and Shaoping Ma. Explicit Factor Models for Explainable Recommendation based on Phrase-level Sentiment Analysis. In Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 2014), July 6 - 11, 2014, Gold Coast, Australia.

[9] Yongfeng Zhang, Min Zhang, Yi Zhang, Guokun Lai, Yiqun Liu, Honghui Zhang, Shaoping Ma. Daily-Aware Personalized Recommendation based on Feature-Level Time Series Analysis. In Proceedings of the 24th International World Wide Web Conference (WWW 2015), May 18 - 22, 2015, Florence, Italy.

[10] S. Yu, C. Xu, and H. Liu, "Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation," arXiv:1807.01855, 2018.

[11] Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. 2014. On stopwords, filtering and data sparsity for sentiment analysis of twitter. In Proceedings of the 9th International Language Resources and Evaluation Conference (LREC'14). European Language Resources Association (ELRA), 810–817.

[12] Bradford Heap, Michael Bain, Wayne Wobcke, Alfred Krzywicki, and Susanne Schmeidl. 2017. Word vector enrichment of low frequency words in the bag-of-words model for short text multi-class classification problems. CoRR, abs/1709.05778.