International Journal of Information Technology & Decision Making
Vol. 8, No. 2 (2009) 249–265


World Scientific
www.worldscientific.com

# DISTRIBUTION OF MULTI-WORDS IN CHINESE AND ENGLISH DOCUMENTS

WEN ZHANG*,† and TAKETOSHI YOSHIDA‡

*School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai Tatsunokuchi, Ishikawa 923-1292, Japan*
*and*
**Laboratory of Internet Software Technologies Institute of Software, Chinese Academy of Sciences Beijing 100190, P. R. China*
†*zhangwen@jaist.ac.jp*
‡*yoshida@jaist.ac.jp*

XIJIN TANG

*Institute of Systems Science Academy of Mathematics and Systems Science Chinese Academy of Sciences, Beijing 100190, P. R. China*
*xjtang@amss.ac.cn*

As a hybrid of N-gram in natural language processing and collocation in statistical linguistics, multi-word is becoming a hot topic in area of text mining and information retrieval. In this paper, a study concerning distribution of multi-words is carried out to explore a theoretical basis for probabilistic term-weighting scheme. Specifically, the Poisson distribution, zero-inflated binomial distribution, and G-distribution are comparatively studied on a task of predicting probabilities of multi-words' occurrences using these distributions, for both technical multi-words and nontechnical multi-words. In addition, a rule-based multi-word extraction algorithm is proposed to extract multi-words from texts based on words' occurring patterns and syntactical structures. Our experimental results demonstrate that G-distribution has the best capability to predict probabilities of frequency of multi-words' occurrence and the Poisson distribution is comparable to zero-inflated binomial distribution in estimation of multi-word distribution. The outcome of this study validates that burstiness is a universal phenomenon in linguistic count data, which is applicable not only for individual content words but also for multi-words.

*Keywords*: Multi-word; term distribution; Poisson distribution; zero-inflated distribution; G-distribution.

## 1. Introduction

In text mining and information retrieval, study on indexing terms, such as individual words, N-gram, and multi-word, has become an unavoidable even indispensable issue confronted with researchers. Of all themes concerning this topic, distribution of terms, which is a fundamental problem and focusing on word frequency

prediction, has attracted great interest in textual information processing. Thus, various term distribution models have been proposed to capture the regularities of word occurrences and discover underlying mechanisms of terms (words' behavior) in texts.

On the one hand, a good understanding of term occurrence mechanisms is useful for information retrieval when we want to assess the likelihood of a certain number of occurrences of a specific term in a collection of texts. In this aspect, Teevan and Karger have proposed an exponential probabilistic model based on computational analysis of corpora and queries. They reported that their model can better describe text probability and consequently, a significant improvement on text retrieval was produced.[1] Madsen *et al.* used the Dirichlet distribution to model word burstiness in texts with goal to remove heuristics from naive Bayes for text classification.[2] However, their model is at document level and too complex for practice, especially in parameter estimation with corpus of a small number of documents. The Poisson distribution is utilized for term-dependent smoothing to estimate query likelihood and further document scores in information retrieval.[3] Their experiment shows that in comparison with other multinomial models for smoothing, the Poisson model has better performance for producing query results, especially with two-stage smoothing.

On the other hand, term-weighting is a crucial procedure when documents are transformed into numerical vectors. However, most term-weighting methods, such as term frequency and document frequency, are based on empirical observation and linguistic intuition, rather than theoretical analysis of term distribution and properties in documents. For this reason, term distribution is studied to shed light on distinguishing significant (content, topical, semantically focused) terms from insignificant (function, noncontent, semantically unfocused) terms in texts for practical applications such as Refs. 4, 5, and 6, according to explicit statistical characteristics of terms. For instance, in term-weighting using residual inverse document frequency[7] shown in Eq. (1), it is assumed that term occurrence in documents follows the Poisson distribution and if parameter $\lambda_i$ can be accurately estimated, then weighting of the indexing term will be precisely computed out to reflect its importance in a text collection:

$$\mathrm{RIDF} = \mathrm{IDF} - \log_2 \left( \frac{1}{1 - P(0, \lambda_i)} \right). \tag{1}$$

Generally, indexing terms for texts are individual words during the process from textual information to numerical vectors. However, for some kinds of texts, such as technical and professional papers, a group of words are often considered as a feature to describe a special concept in a specific field. Multi-word features are not found too frequently in a document collection, but when they do occur they are often highly predictive and informative in explaining discovered patterns from learning methods.[8] While "multi-word" is the fundamental notion of this paper, this notion had no satisfactory formal definition until now. It can only be intuitively characterized: it occurs only in specialized types of discourse, often specific to subsets of domains; when it occurs in general types of discourse or in a variety

of domains it often has broader or more comprehensive meaning, such as name entities, terminological noun phrases (NP), etc.

In recent studies on term distribution, two kinds of disciplines are actually followed by research with this topic. The first one is to use large-scale data to study mechanisms of word occurrence in a corpus such as Refs. 9 and 10. The concentration of this branch is on collective properties of natural utterances and population distribution for all words. The second one is trying to match underlying model assumptions to text through manual analysis of a small number of terms.[11–13] Our paper covers both sides of term distribution and the main concern is on the latter one, i.e. to carry out a study on term distribution of some words with special characteristics. Although much work has been done in term distribution and many prominent proposals have been presented, little work has been done on the comparison of different term distribution models, especially on the distribution of multi-words in documents, which is specially addressed in this paper.

The remainder of this paper is organized as follows. Three different term distribution models, the Poisson model, zero-inflated binomial, and G-model, are introduced in Sec. 2. Section 3 describes our data set and data preprocessing used to examine these models. Moreover, the method for multi-word extraction from both Chinese and English documents is proposed based on syntactical structures and characteristics of terms. Section 4 is our experiments of the three different term distributions on Chinese and English multi-words and the results are demonstrated. Section 5 is discussion of experimental results. Concluding remarks and further research are also indicated.

## 2. Term Distribution Models

Classical probabilistic models of term distribution, such as the Poisson model, zero-inflated binomial distribution, and G-model, are introduced in this section with their basic assumptions for term occurrence in texts.

### 2.1. *Poisson distribution*

The classical definition of the Poisson distribution is as follows:

$$P(k, \lambda_i) = e^{-\lambda_i} \times \frac{\lambda_i^k}{k!} \quad \text{for } \lambda_i > 0. \tag{2}$$

In most cases of using the Poisson distribution in information retrieval, parameter $\lambda_i > 0$ is the average number of occurrences of a word $w_i$ per-document, that is, $\lambda_i = \frac{cf_i}{N}$, where $cf_i$ is the collection frequency of the word and $N$ is total number of documents in the collection. With the Poisson distribution, we can estimate the probability of a word occurring at a given number of times in a document. That is, $P(k; \lambda_i)$ is the probability of $w_i$ having exactly $k$ occurrences in a document, where $\lambda_i$ is appropriately estimated for each word. The basic assumption of the Poisson distribution is that occurrences of a term are independent of each other, i.e. there is no correlation between different occurrences of a term in documents.

However, this assumption may not hold in most cases, because of different occurrence patterns between content words and noncontent words in texts. Based on this idea, two-Poisson model and further Poisson mixtures are developed to estimate probabilities of occurrences of a term, while they all have a variety of problems existing in practical applications.[12] All these models have same basic assumption regarding word occurrence with classical Poisson distribution as occurrence independence. This is the root cause for their consistent inabilities to describe term distribution in previous research.

To simplify, only classic Poisson distribution is examined in this paper for multi-word probability estimation. We conjecture that only when the basic assumption with classical Poisson distribution is validated as promising in predicting probability of multi-words, more complex model combined with classical Poisson distributions would be reasonably expected as valuable and potential in the application. This is the very motivation for us to examine classical Poisson distribution on multi-words.

### 2.2. *Zero-inflated distribution*

In the area of modeling linguistic count data-like word occurrence in documents, the problem of large counts for the zero outcomes is widely observed. To tackle this problem, zero-inflated distribution is proposed to consider the zero and nonzero word occurrences separately. That is, to construe word occurrences as two-component mixture, where one component is a degenerate distribution whose entire probability mass is assigned to the outcome zero, and the other component is a standard distribution. Although there are many types of zero-inflated distributions such as zero-inflated negative binomial distribution and zero-Poisson. In this paper, the zero-inflated binomial distribution, as a three-parameter probability distribution proposed by Martin Jansche to capture the extra-Poisson variation in linguistic count data,[4] is adopted to model multi-word occurrences shown as follows:

$$P(k; z, p, n) = z\delta_{k,0} + (1 - z)\binom{n}{k} p^k (1 - p)^{n-k}. \tag{3}$$

Here, $\delta_{k,0}$ is a Kroneker symbol whose value will be assigned as 1 if and only if the outcome is zero. Otherwise, it will be assigned as zero. $z$ and $p$ are two parameters for this model with its mean as $(1 - z)np$ and variance as $(1 - z)np(znp + 1 - p)$. $k$ represents word counts ($0 \leq k \leq n$) and $n$ is the length of a document measured as the total number of words in the document.

The basic assumption with zero-inflated distribution is that distribution of word occurrence in documents conforms to a binomial distribution except the extreme distortion at outcome zero. There is a massive probability for word's absence in documents meanwhile the occurrence of word follows a binomial distribution. That is, the occurrences of terms in documents are independent from each other.

In order to estimate the parameter $z$ and $p$, expectation maximization (EM) algorithm[15] is employed to compute them based on a data sequence because it would be very difficult for both max likelihood estimation (MLE) and method-of-moment estimation to estimate them in this case as $z$ and $p$ do not have obvious

statistical implications. For more details about zero-inflated binomial and the EM solution for estimation of $z$ and $p$, readers can refer to Refs. 14, 15, respectively.

### 2.3. *G-distribution*

The G-distribution (G means general), also known as a three-parameter probability distribution, is defined as follows:

$$P(k;\alpha,\gamma,\beta) = (1-\alpha)\delta_{k,0} + (1-\gamma)\delta_{k,1} + \frac{\alpha\gamma}{B-1}\cdot\left(1-\frac{1}{B-1}\right)^{k-2} \cdot(1-\delta_{k,0}-\delta_{k,1}). \tag{4}$$

In practical application, $\alpha = \sum_{r\geq 1} p_r = 1 - p_0$ is the sum of frequency of terms whose occurrence is more than 0, and $\gamma = \frac{\sum_{r\geq 2} p_r}{\sum_{r\geq 1} p_r} = 1 - \frac{p_1}{1-p_0}$ is the proportion of frequency of not less than 2 to frequency of not less than 1. $p_r = P(k = r)$ is probability of having exactly $r$ instances of a term in a document. $B = \frac{\sum_{r\geq 2} p_r r}{\sum_{r\geq 2} p_r}$ is a measure of topical burstiness. Also, a Kroneker symbol is employed here as $\delta_{i,j} = \{^{0,\ i\neq j}_{1,\ i=j}$.

The basic assumption with G-distribution is that there are primarily two kinds of words existing in one document: one is noncontent words, and the other is content words. Usually, multi-words are content words in a document and can be separated into topical and nontopical words. When a content word is present in a document, but the concept named by a content word is not topical (nontopical) for that document, then this word would typically occur only once in this document. However, when a concept named or expressed by a content word is topical for the document, then this content word is characterized by multiple (frequency $\geq 2$) occurrence, i.e. bursty occurrence. The notion of burstiness is fundamental for obtaining G-distribution, which means multiple occurrences of a content word or phrase turn up in some documents but in other documents, they do not occur at all. For more details about G-distribution, readers can refer to Ref. 11.

## 3. Data Preprocessing

In order to investigate the three distribution models described above comparatively, two different text collections in two languages (Chinese and English) are selected as our sample data set for experiments. In this section, first, profiles of these two text collections are specified, and then the multi-word extraction method based on syntactical structure is proposed to extract multi-words from both text collections.

### 3.1. *XSSC texts in Chinese and Reuters texts in English*

Based on our previous work as Refs. 16 and 17, 184 texts concerning details of each XiangShan science conference (XSSC) are collected from XSSC web site (http://www.xssc.ac.cn), where uploaded many academic topics of a wide scope

from basic research to advanced techniques. In this paper, Chinese multi-word extraction is conducted from these documents, and Chinese multi-word distribution characterization is compared between this real data and estimated probability with the three models. By our computing, there are 14 categories related to this document collection, and the average number of sentences per-document is 41.46. More description about this corpus can refer to Ref. 17.

Reuters-21578 data set (http://www.research.att.com/∼lewis) is used as our experiment sample texts in English. It appeared as Reuters-22173 in 1991 and was indexed with 135 categories by personnel from Reuters Ltd in 1996. In the area of text mining, it is usually adopted as a bench mark data set for text categorization. In this paper, English multi-words are extracted from this textual data set, and the real distribution of English multi-words is compared with the three distribution models in order to evaluate the performance of them. By our statistics, this data set contains totally 19,403 valid texts, with an average of 5.4 sentences in each text. For convenience, the texts from four categories, "grain," "crude," "trade," and "interest," are fetched out as our target data set, because the texts from these four categories have more difference than those from other categories, i.e. less overlapping in multi-words expression, which may ensure that the extracted multi-word is a content word in that text. With this method, 574 texts from "grain," 566 texts from "crude," 424 texts from "interest," and 514 texts from "trade" are assigned as our target data set.

It should be worth noticing that the length of documents $n$ from both XSSC and Reuters varies from each other. Usually, the fixed length of documents is required for zero-inflated binomial distribution to make an estimation. Our solution for this problem is to estimate the parameters for zero-inflated binomial distribution using varying $n$ but make the estimation using the average $n$. Thus, in this case, we estimate the probability of $k$ occurrences using $\bar{n}$, i.e. the average length of those documents, which have exactly $k$ occurrences of that multi-word, as the substitute for $n$.

### 3.2. *Multi-word extraction*

Basically, there are two types of methods to extract multi-words from documents: one is to utilize the mutual information (sometimes called association) between words, which is a statistical method,[18,19] and the other is to analyze the syntactical structure of multi-words, which is a rule-based method. For instance, mutual information method, the work in Ref. 20 proposes the association ratio for measuring word association based on the information theoretic concept of mutual information. The research in Ref. 21 proposes a regular expression to characterize lexical structure of multi-word in text in order to identify multi-words from text. Usually, the multi-word extraction method varies with different languages from linguistic perspective. To simplify the process of multi-word extraction, here, we adopt syntactical rule-based method for multi-word extraction and this method is applicable for both Chinese and English. Based on the previous study in Ref. 21 on structures

and characteristics of multi-words, a widely accepted conclusion is that a multi-word has properties as an NP ending with a noun and repetition of occurrence. This results in a simple hypothesis that an NP having a frequency of two or more can be regarded as a multi-word. With this hypothesis, we propose the following multi-word extraction method to extract the multi-words from both the Chinese and English texts. The basic idea of this method is to identify repetitive patterns (a group of consecutive words) from sentences as multi-word candidates firstly, and then determine part of speeches of these identified patterns. If a candidate's part of speech is a noun (not a pronoun), it should be accepted as a multi-word. Otherwise, it should be rejected as a multi-word. The following is our method to identify repetitions of any two sentences. For co-occurring words in texts as mentioned by one of the reviews, if they construct a consecutive pattern and the pattern is ending with a noun, then we can accept it as a multi-word.

**Algorithm 3.1.** Multi-word extraction from Chinese and English documents. The common pattern between two sentences is regarded as a multi-word candidate.
Input:
$s_1$, the first sentence
$s_2$, the second sentence
Output:
Repetitive and consecutive words extracted from $s_1$ and $s_2$.
Procedure:

1. $s_1 = w_1, w_2, \ldots, w_n$
2. $s_2 = w'_1, w'_2, \ldots, w'_m$
3. $k = 0$
4. for each word $w_i$ in $s_1$
5. for each word $w'_j$ in $s_2$
6. while$(w_i = w'_j)$
7. $k$++
8. end while
9. if $k > 1$
10. combine the words from $w_i$ to $w_{i+k}$ as the
11. output of this procedure
11. End if
12. End for
13. End for

After the repetition is extracted from sentences in a document, the ICTCAS[a] and JWNL[b,c] are employed to determine the part of speech of the last word of the repetition, for Chinese and English, respectively. Moreover, in the case that the last

[a]H. P. Zhang and Q. Liu, Chinese Lexical Analysis System 2.0, http://www.ict.ac.cn/freeware/.
[b]Java WordNet Library, http://sourceforge.net/projects/jwordnet.
[c]WordNet: A Lexical Database for the English Language, http://wordnet.princeton.edu/.

word of the repetitive pattern is not a noun, such as "prime minister agree," the last noun of this repetition is found out and the extracted pattern is segmented by the last noun. Thus, "prime minister" and "agreed" are segmented and "prime minister" is regarded as a multi-word. Nevertheless, the length and the alignment of each word are also considered to make the extraction more accurate, for example, multi-word usually has a length no more than six single words.

For the XSSC documents, 5087 multi-words are extracted, and 4024 multi-words are extracted from Reuters texts. It is very interesting to notice that although the total number of documents of Reuters texts (2074) is far larger than the XSSC texts (184), the numbers of multi-words from these two text collection are approximately equivalent. We conjecture this outcome because that the algorithm for multi-word extraction is applied in sentence level. And Reuters data set has a total of 7628 sentences, while 11,200 sentences are found in XSSC text collection. Moreover, the types of text are different, XSSC text is about academic and scientific reports, which have a long length, while Reuters texts belong to brief news reports.

## 4. The Distribution of Chinese and English Multi-Words in Text

In this section, the comparison of the Poisson distribution, zero-inflated binomial distribution, and G-distribution is conducted in characterizing the multi-word distribution in XSSC and Reuters text collection. Traditionally, the words in texts are separated into noncontent words (function words, semantically unfocused words), and content words (semantically focused words, topical words). Generally, multi-word is content word and semantically focused unit in its expression but its roles in documents alternate as topical and nontopical. We make a distinction between technical multi-words and nontechnical multi-words. Technical multi-word refers to a group of words, which are highly related to the contents of the texts, such as terminological NPs, while nontechnical multi-words are less related to the content of the texts, for example, the commonly used phrases in a field, and the names of places.

Furthermore, two measures are developed to evaluate the multi-word distribution characterizing performance of the Poisson, zero-inflated binomial, and G-distribution. They are gross error as $E_g$ and local error as $E_l$ with definitions as follows:

$$E_g = \sum_{r \geq 0} |\text{act} - \text{est}|. \tag{5}$$

$$E_l = \sum_{r \geq 2} |\text{act} - \text{est}|. \tag{6}$$

Here, act is the actual frequency of multi-word occurrence in texts, and est is the estimated frequency of multi-word occurrence in texts by the assumed distribution. $E_g$ is used to compute the overall estimation error ($0 \leq r \leq n$), and $E_l$ is used to compute the local estimation ($2 \leq r \leq n$), because there are never estimation error for G-distribution if $r \leq 1$, according to its formula.

### 4.1. *Overall distribution of the multi-words in XSSC and Reuters texts*

Figures 1 and 2 show the overall distribution of multi-words in XSSC and Reuters collection, respectively. Roughly, they have a very similar curves. The only difference

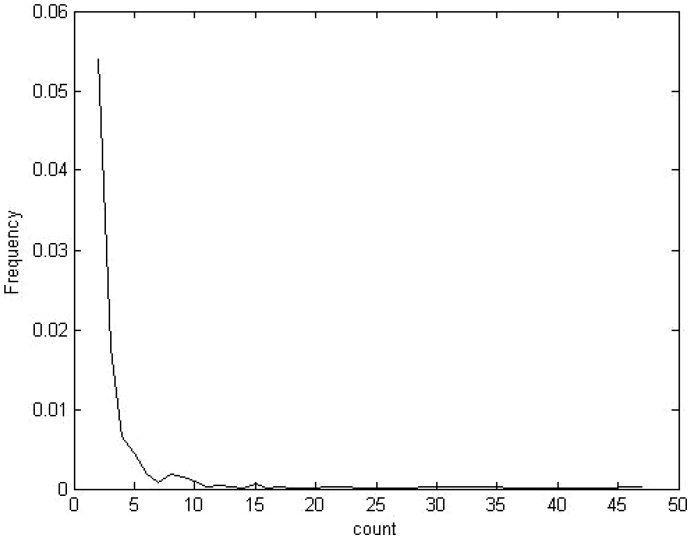


Fig. 1. The overall distribution of the numbers of multi-words and their frequency in XSSC. The product of term frequency and term rank is about 0.02–0.1 when the rank is larger than 3.

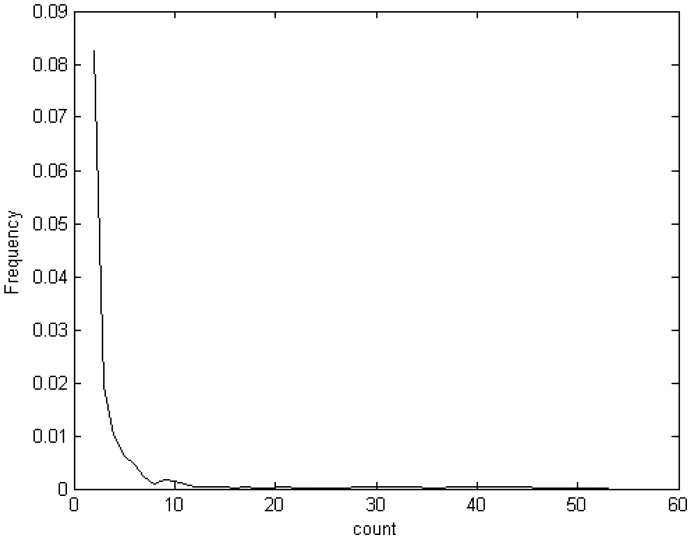


Fig. 2. The overall distribution of the numbers of multi-words and their frequency in Reuters. The product of term frequency and term rank is about 0.02–0.1 when the rank is lager than 4.

is that the latter has a larger range of number of occurrences (count) and frequency. Moreover, they conform to Zipf's law,[22] which states that product of term frequency and its rank order (order of count) is approximately constant shown as follows:

$$f = C \times \left(\frac{1}{r}\right). \tag{7}$$

Here, $f$ is the frequency of words, $r$ is the words' rank, and $C$ is a constant. This statement is usually adopted by some practical information retrieval applications such as Ref. 5 and it is also validated in our research on both XSSC and Reuters collection. For both data sets, the constant $C$ is about 0.02–0.1 for most cases but with some exceptions at the extremely smaller order.

### 4.2. *Distributions of technical multi-words*

For the technical multi-words of XSSC, "纳米材料" (nano materials) and "生态环境" (ecological environment) are assigned as the testees, because they are the hot topics in new technology areas, they have a great possibility to be the topic of the documents they do occur in XSSC collection. And for the technical multi-words of Reuters, we selected "crude oil" and "interest rates" as the testees, because they are the topics of the categories we picked out from Reuters text collection. Tables 1–4 show the estimation of the Poisson distribution, zero-inflated distribution, and G-distribution on these examined multi-words along with their global and local errors. Here, count is the frequency occurring in the text collection, act is the actual frequency ratio, P-est is the estimation given by Poisson distribution, Z-est is the estimation given by zero-inflated binomial distribution, and G-est is the estimation given by G-distribution.

From Tables 1–4, it can be seen that of these three distributions, G-distribution can most effectively characterize the multi-word's occurrence probability with its frequency, i.e. estimating the probability of exactly $r(\geq 0)$ occurrences of a multi-word in both text collections. Both the Poisson and zero-inflated binomial have a very comparable capacity of capturing the multi-word's distribution. In Tables 1 and 4, zero-inflated binomial is a little better than the Poisson distribution while Poisson distribution performs a bit better in Tables 2 and 3. Another point worth noticing is that the sum of the probability of zero-inflated binomial in all cases (for example, Table 3) is a bit more than 1. This happens because EM algorithm is an approximate algorithm based on local optimization; the length of texts in both collections for parameter estimation is not fixed. In addition, it is obvious that the estimation error of the Poisson distribution and zero-inflated binomial at outcome 0 and 1 occupies a dominant proportion of the overall estimation error for $E_g \gg E_l$. For this reason, it can be deduced that both the Poisson distribution and zero-inflated binomial are not able to capture the probability mass at 0 for frequency count data satisfyingly. Although the initiative for zero-inflated binomial is to separate the probability at outcome zero and nonzero so that the probability mass at zero could be captured by it, its performance in our study degenerate very

Table 1. The distribution of Chinese multi-word "纳米材料" and its probability estimation from the Poisson distribution, zero-inflated binomial distribution, and G-distribution.

| Count | 0 | 1 | 3 | 4 | 5 | 8 | 11 | 15 | 17 | 26 | $E_g$ | $E_l$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| act.($\times 10^{-2}$) | 89.67 | 4.89 | 1.09 | 0.54 | 0.54 | 1.09 | 0.54 | 0.54 | 0.54 | 0.54 | | |
| P-est.($\times 10^{-2}$) | 55.30 | 32.76 | 1.92 | 0.28 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 67.09 | 4.85 |
| Z-est.($\times 10^{-2}$) | 55.31 | 30.68 | 1.13 | 0.54 | 0.17 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 63.80 | 3.66 |
| G-est.($\times 10^{-2}$) | 89.67 | 4.89 | 0.60 | 0.54 | 0.42 | 0.30 | 0.21 | 0.13 | 0.10 | 0.04 | 3.08 | 3.08 |

Table 2. The distribution of Chinese multi-word "生态环境" and its probability estimation from the Poisson distribution, zero-inflated binomial distribution, and G-distribution.

| Count | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 9 | 12 | 13 | 27 | $E_g$ | $E_l$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| act.($\times 10^{-2}$) | 80.98 | 7.07 | 4.35 | 1.63 | 1.09 | 1.09 | 1.63 | 0.54 | 0.54 | 0.54 | 0.54 | | |
| P-est.($\times 10^{-2}$) | 48.01 | 35.23 | 12.92 | 3.16 | 0.58 | 0.09 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 76.45 | 15.32 |
| Z-est.($\times 10^{-2}$) | 48.70 | 35.00 | 13.52 | 5.51 | 1.00 | 0.18 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 78.03 | 17.81 |
| G-est.($\times 10^{-2}$) | 80.98 | 7.07 | 2.63 | 2.05 | 1.60 | 1.25 | 0.97 | 0.46 | 0.22 | 0.17 | 0.01 | 4.77 | 4.77 |

Table 3. The distribution of "crude oil" and its probability estimation from the Poisson distribution, zero-inflated binomial distribution, and G-distribution.

| Count | 0 | 1 | 2 | 3 | 4 | 6 | 8 | $E_g$ | $E_l$ |
|---|---|---|---|---|---|---|---|---|---|
| act.($\times 10^{-2}$) | 90.51 | 6.21 | 2.45 | 0.59 | 0.10 | 0.10 | 0.05 | | |
| P-est.($\times 10^{-2}$) | 86.73 | 12.35 | 0.88 | 0.04 | 0.00 | 0.00 | 0.00 | 12.29 | 2.37 |
| Z-est.($\times 10^{-2}$) | 86.91 | 13.38 | 1.12 | 0.07 | 0.01 | 0.00 | 0.00 | 19.81 | 3.22 |
| G-est.($\times 10^{-2}$) | 90.51 | 6.21 | 2.26 | 0.70 | 0.22 | 0.02 | 0.00 | 0.55 | 0.55 |

Table 4. The distribution of "interest rates" and its probability estimation from the Poisson distribution, zero-inflated binomial distribution, and G-distribution.

| Count | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 8 | $E_g$ | $E_l$ |
|---|---|---|---|---|---|---|---|---|---|---|
| act.($\times 10^{-2}$) | 92.22 | 4.26 | 2.01 | 0.98 | 2.04 | 0.10 | 0.15 | 0.05 | | |
| P-est.($\times 10^{-2}$) | 86.99 | 12.13 | 0.85 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 | 17.54 | 4.44 |
| Z-est.($\times 10^{-2}$) | 87.67 | 15.77 | 2.28 | 0.27 | 0.08 | 0.01 | 0.00 | 0.00 | 19.81 | 3.22 |
| G-est.($\times 10^{-2}$) | 92.22 | 4.26 | 2.01 | 0.86 | 0.37 | 0.16 | 0.07 | 0.01 | 1.97 | 1.97 |

similarly with the Poisson distribution. The theoretical reason for this degeneration is that on the one side, $z$ and $p$ have very small value ($z$ and $p$ are about $10^{-4}$ for XSSC, and for Reuters, $z$ is about $10^{-4}$, and $p$ is about $10^{-5}$) and on the other side, $n$ is a great value in our text collection (about 4000 for XSSC and 2000 for Reuters). All these make the zero-inflated binomial to become an approximate Poisson and this is the very reason they have the similar performance. Furthermore, the estimation on Reuters texts are better than the estimation on XSSC texts, as is shown that the estimation on Reuters texts always has less error using any assumed distribution. We will discuss this phenomenon in Sec. 5.

### 4.3. *Distributions of non-technical multi-words*

For the nontechnical multi-words of XSSC, "基础研究" (basic research) and "科学问题" (scientific problem) are assigned as the Chinese testees, because they are popular words in XSSC academic discussion and have a very extensive meaning other than a concrete professional concept. For the nontechnical multi-words of Reuters, we select "United States" and "Soviet Union" as our samples, as they are the names of countries and can be used anywhere related to these two countries in newswire reports. Tables 5–8 show the results of the Poisson distribution and G-distribution on these examined multi-words.

It has been shown in Tables 5–8 that the G-distribution still has the best performance in estimating the probability of exact frequency of nontechnical multi-words in text collection. Also, the Poisson distribution and zero-inflated binomial have similar capacity in capturing the multi-word distribution because the problem with parameters as $z$, $p$, and $n$ as mentioned above estimated according to the real occurrence frequency and text length makes it degenerate into an approximate Poisson distribution.

Table 5. The distribution of Chinese multi-word "基础研究" and its probability estimation from the Poisson distribution, zero-inflated binomial distribution, and G-distribution.

| Count | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 12 | 18 | $E_g$ | $E_l$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| act.($\times 10^{-2}$) | 49.46 | 20.11 | 11.41 | 6.52 | 4.35 | 2.17 | 3.26 | 0.54 | 1.09 | 0.54 | 0.54 | | |
| P-est.($\times 10^{-2}$) | 24.88 | 34.61 | 24.08 | 11.17 | 3.88 | 1.08 | 0.25 | 0.05 | 0.01 | 0.00 | 0.00 | 63.62 | 24.54 |
| Z-est.($\times 10^{-2}$) | 24.81 | 35.49 | 23.13 | 13.23 | 7.43 | 0.50 | 0.43 | 0.06 | 0.02 | 0.00 | 0.00 | 68.67 | 28.63 |
| G-est.($\times 10^{-2}$) | 49.46 | 20.11 | 10.46 | 6.86 | 4.51 | 2.96 | 1.94 | 1.27 | 0.84 | 0.16 | 0.01 | 5.45 | 5.45 |

Table 6. The distribution of Chinese multi-word "科学问题" and its probability estimation from the Poisson distribution, zero-inflated binomial distribution, and G-distribution.

| Count | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 13 | $E_g$ | $E_l$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| act.($\times 10^{-2}$) | 49.46 | 20.65 | 12.50 | 5.98 | 3.26 | 1.63 | 2.17 | 1.09 | 0.54 | 0.54 | 1.09 | 0.54 | 0.54 | | |
| P-est.($\times 10^{-2}$) | 25.01 | 34.66 | 24.02 | 11.10 | 3.84 | 1.07 | 0.25 | 0.05 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 67.30 | 28.84 |
| Z-est.($\times 10^{-2}$) | 28.63 | 33.81 | 25.03 | 14.27 | 4.14 | 1.24 | 0.68 | 0.17 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 | 61.70 | 27.70 |
| G-est.($\times 10^{-2}$) | 49.46 | 20.65 | 10.15 | 6.70 | 4.43 | 2.92 | 1.93 | 1.28 | 0.84 | 0.56 | 0.37 | 0.24 | 0.11 | 12.59 | 12.59 |

Table 7. The distribution of "United States" and its probability estimation from the Poisson distribution, zero-inflated binomial distribution, and G-distribution.

| Count | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 9 | 10 | $E_g$ | $E_l$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| act.($\times 10^{-2}$) | 87.38 | 7.97 | 2.50 | 1.27 | 0.59 | 0.15 | 0.05 | 0.05 | 0.05 | | |
| P-est.($\times 10^{-2}$) | 80.99 | 17.08 | 1.80 | 0.13 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 18.22 | 2.72 |
| Z-est.($\times 10^{-2}$) | 82.30 | 22.34 | 3.42 | 0.76 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 21.75 | 2.30 |
| G-est.($\times 10^{-2}$) | 87.38 | 7.97 | 2.55 | 1.15 | 0.52 | 0.23 | 0.11 | 0.01 | 0.00 | 0.46 | 0.46 |

Table 8. The distribution of "Soviet Union" and its probability estimation from the Poisson distribution, zero-inflated binomial distribution, and G-distribution.

| Count | 0 | 1 | 2 | 3 | 4 | 5 | $E_g$ | $E_l$ |
|---|---|---|---|---|---|---|---|---|
| act.($\times 10^{-2}$) | 94.96 | 3.13 | 1.32 | 0.39 | 0.10 | 0.10 | | |
| P-est.($\times 10^{-2}$) | 92.47 | 7.24 | 0.28 | 0.01 | 0.00 | 0.00 | 8.22 | 1.62 |
| Z-est.($\times 10^{-2}$) | 92.51 | 7.50 | 3.26 | 0.03 | 0.00 | 0.00 | 9.30 | 2.50 |
| G-est.($\times 10^{-2}$) | 94.96 | 3.13 | 1.31 | 0.41 | 0.13 | 0.04 | 0.12 | 0.12 |

The property of technical multi-word in XSSC collection takes effect since technical multi-word in XSSC has a greater maximum frequency than nontechnical multi-word. It does not work when it comes to Reuters text collection: both the technical multi-word and nontechnical multi-word has the approximate equivalent range of occurrence frequency in this collection. When comparing the estimation on technical multi-word and nontechnical multi-word, we find that the technical multi-word has less error on XSSC but greater error are presented in Reuters data in opposite to the case of technical multi-word distribution. We would like to discuss this phenomenon in Sec. 5.

## 5. Discussion and Concluding Remarks

In this paper, a comparative study on distribution of multi-words in texts is carried out. Three classical models for describing the linguistic count data, such as the Poisson distribution, zero-inflated binomial distribution, and G-distribution, are presented to describe the distribution of both technical and nontechnical multi-word in Chinese and English text collections, as XSSC text collection and Reuters data set. Moreover, a syntactical multi-word extraction method independent of language is proposed to extract the multi-words from texts, based on the syntactical structure and lexical rule of multi-words in texts.

Our experimental results have shown that G-distribution has a better capability in describing the distribution of probability on multi-word occurrence frequency than the Poisson distribution and zero-inflated binomial distribution. This result has validated the basic assumption in G-distribution about the existence of word burstiness in texts, regarding content words. The inability of the Poisson distribution to estimate the probability of outcome 0 and 1 occurrence enhanced that the

occurrences of multi-words in text may not be feasibly regarded as independent from each other. Even if the outcome 0 is separated in zero-inflated binomial distribution, the nonzero occurrences of multi-word may not be justifiably regraded as independent events. Although the zero-inflated binomial is validate as an effective way in capturing the difference between zero and nonzero outcome in linguistic count data in Ref. 14, the word examples in their experiments are not content words such as "his," "any," etc. Consequently, the effect of burstiness for content words does not exert fully on these words. Moreover, the distribution of these words has already investigated as to be conformed well to the Poisson distribution by earlier research.[12] Despite of that other researchers also argued that the two-Poisson model or negative binomial may be an out-way to solve this kind of problem,[23] and further problem with Poisson is from widely different document size,[7] the basic assumption for word's and multi-word's occurrence independence should be reconsidered.

However, some questions have turned up with our experimental results. The first one is that the estimations on Reuters texts are better than the estimations on XSSC texts. Perhaps it is because the XSSC texts are academic papers, which have more terminological NPs but fewer texts than Reuters text, then multi-word behavior is not fully exhibited on XSSC texts. The second question is that the distribution of technical multi-words follows better with G-distribution than that of nontechnical words in XSSC collection, but with opposite outcome when it comes to Reuters texts. The reason for this point is possibly because XSSC texts are academic texts, then the burstiness can more easily induced in their texts but the Reuters texts are newswire texts focusing on including information in short passages as much as possible, then burstiness of content words cannot happen in them naturally. Here, a convincing root reason for this differences has not been acquired currently and further investigations are required to disclose these phenomena.

As for our further research, term-weighting methods based on term distribution theory are a potential and valuable direction to advance, especially for multi-word features. For example, the significance of a multi-word in a document can be objectively measured by its occurrence probability and this could be used for text representation if the distribution of multi-word and term frequency of multi-word in the text is known. Nevertheless, term distributions investigated in this paper also provide a theoretical support to improve practical application of text mining such as information extraction,[24] text classification,[25] etc., on the condition that probability models are established for distributions of terms in a given text collection.

## Acknowledgments

This work is supported by Ministry of Education, Culture, Sports, Science and Technology of Japan under the "Kanazawa Region, Ishikawa High-Tech Sensing Cluster of Knowledge-Based Cluster Creation Project" and partially supported by the National Natural Science Foundation of China under Grant No. 70571078 and 70221001. The authors of this paper would like to appreciate Professor Insoo,

Kweon for his kind help to explain the mechanism of EM algorithm to estimate the parameters for zero-inflated binomial distribution. The authors would like to present their thanks to the anonymous reviewers of this paper for their help to improve the quality of the paper.

## References

1. J. Teevan and D. R. Karger, Empirical development of an exponential probabilistic model for text retrieval, in *Proc. 26th Int. ACM SIGIR Conf.* (Toronto, Canada, 2003), pp. 18–25.
2. R. E. Madsen, D. Kauchak and C. Elkan, Modeling word burstiness using the Dirichlet distribution, in *Proc. 22nd Int. Conf. Machine Learning* (Bonn, Germany, 2005), pp. 545–552.
3. Q. Mei, H. Fang and C. Zhai, A study of poisson query generation model for information retrieval, in *Proc. 30th Int. ACM SIGIR Conf.* (Amsterdam, The Netherlands, 2007), pp. 319–326.
4. Y. Peng, G. Kou, Y. Shi and Z. Chen, A descriptive framework for the field of data mining and knowledge discovery, *Int. J. Inform. Technol. Decision Making* **7**(4) (2008).
5. F. J. Ren, Automatic abstracting important sentences, *Int. J. Inform. Technol. Decision Making*, **4** (2005) 141–152.
6. R. Sakamoto, Y. Sumi and K. Kogure, Hyperlinked comic strips for sharing personal contexts, *Int. J. Inform. Technol. Decision Making* **6** (2007) 443–458.
7. C. D. Manining and S. Scheutze, *Foundations of Statistical Natural Language Processing* (MIT Press, Cambridge, Massachusetts, 1999), pp. 548–561
8. W. Zhang, T. Yoshida and X. J. Tang, Text classification based on multi-word with support vector machine, *Knowl.-Based Syst.*, in press.
9. D. Wang, M. Li and Z. Di, True reason for Zipf's law in language, *Physica A* **358** (2005) 545–550.
10. R. F. Cancho and R. V. Sole, Zipf's law and random texts, *Adv. Complex Syst.* **5**(1) (2002) 1–6.
11. S. Katz, Distribution of content words and phrases in texts and language modeling, *Nat. Language Eng.* **2**(1) (1996) 15–59.
12. K. W. Church and W. A. Gale, Poisson mixtures, *Nat. Language Eng.* **1**(2) (1995) 163–190.
13. A. D. Roeck, A. Sarkar and P. Garthwaite, Frequent term distribution measures for dataset profiling, in *Proc. 4th Int. Conf. Language Resources and Evaluation* (Lisbon, Portugal, 2004), pp. 30–37.
14. M. Jansche, Parametric models of linguistic count data, in *Proc. 41th Annual Meeting on Association for Computational Linguistics* (Sapporo, Japan, 2003), pp. 288–295.
15. C. M. Bishop, *Neural Network for Pattern Recognition* (Oxford University Press, New York, 2003), pp. 65–73.
16. W. Zhang, X. J. Tang and T. Yoshida, Web text mining on a scientific forum, *Int. J. Knowl. Syst. Sci.* **3**(4) (2006) 51–59.
17. W. Zhang, X. J. Tang and T. Yoshida, Text classification toward a scientific forum, *J. Syst. Sci. Syst. Eng.* **16**(3) (2007) 356–369.
18. M. Yamamoto and K. W. Church, Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus, in *Proc. 6th Workshop on very Large Corpora* (Montreal, Canada, 1998), pp. 285–313.

19. W. Zhang, T. Yoshida and X. J. Tang, Augmented mutual information for multi-word extraction, *Int. J. Innovative Comput. Inform. Contr.* (2009), in press.
20. K. W. Church and P. Hanks, Word association norms, mutual information, and lexicography, *Comput. Linguist.* **16**(1) (1990) 22–29.
21. F. Jueston and S. M. Katz, Technical terminology: Some linguistic properties and an algorithm for identification in text, *Nat. Language Eng.* **1**(1) (1995) 9–27.
22. G. K. Zipf, *Human Behaviour and the Principle of Least Effort* (Addison-Wesley, Cambridge, Massachusetts, 1949).
23. F. Mosteller and D. L. Wallance, *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*, 2nd edn. (Springer-Verlag, New York, 1984).
24. M. Fuketa, Y, Kadoya, E. Atlam, T. Kunikata, K. Morita, S. Kashiji and J. Aoe, A method of extracting and evaluating good and bad reputations for natural language expressions, *Int. J. Inform. Technol. Decision Making* **4** (2005) 177–196.
25. B. Boutsinas and S. Athanasiadis, On merging classification rules, *Int. J. Inform. Technol. Decision Making* **7**(3) (2008) 431–450.