Literature review on Zipf's law, its effect and applications to various topics in Natural Language Processing

Morteza Abdolrahim Kashi
Concordia University, Department of Computer Science
Montreal, Quebec, Canada
m_abdolr@cs.concordia.ca
April, 2004

1-Abstract:

Zipf's law is a very useful law in Natural Language Processing. It has many useful aspects such as the principle of least effort, relationship between the word frequency and the rank of the word if words are sorted in decreasing order of their frequency in a text, relationship between the number of senses for a word and its frequency in a text, and the relationship between the frequency of a particular distance between same words and that particular distance. Each of above mentioned aspects of Zipf's law could be applied in different topics of Natural Language Processing (NLP). Here, I will go over the literature of Zipf's law and some of its applications in NLP.

2-Introduction:

One of Zipf's Laws is a tool to predict the frequency of words in a text. This law states that if we have a very large text and we sort the words in that text in the decreasing order of their frequency, the rank of any word in the made list multiplied by its frequency is almost equal to a constant number. In other words we have the following equation for the mentioned relationship: $r \times f \approx k$. Where r is the rank of the word in the mentioned sorted list and f is the frequency of the same word, and k is a constant. Therefore, when *log (f)* is drawn against *log(r)* in a graph, we should see a line whose slope is -1. Manning and Schütze[5] argue about one aspect of Zipf's law which is about the Principle of Least Effort which says that people do their work some how they do the least work with the

lowest pace. They also argue about the other two aspects of Zif's law. According to them, one aspect of Zipf's law relates the number of senses for a word to the frequency of the same word in a text. They also argue that one aspect of Zipf's law relates the frequency of space size between same words in a text to the size of the space in that text. All of above aspects could be useful in NLP.

In the remainder of this paper, I will bring a literature review on different aspects of Zipf's law. Then I will bring the correlation of Zipf's law to some topics in NLP such as n-gram models, citations, random texts, word sense, co-reference resolution, and summarization. At the end, I will bring my conclusion and then the references I used to prepare this article.

3-Literature Review:

First Zipf's law has been created by looking at a text and experimenting it by hand. Le Quan Ha, E.I. Sicilia-Garcia, Ji Ming, and F.J. Smith [1] argue that Zipf found out about the law by looking at a text and the frequencies of its words to see if there is any relationship between the frequencies of words and their rank if they are sorted by decreasing order of their frequencies. That is: $: r \times f \approx k$. They argue that if we draw *log(f)* against *log(r)* we will have an almost straight line with the slope of (-1). They also argue that further research showed that, in some cases, we do not have the slope of (-1) if we draw the log(f) against log(r) which necessitated some correction in Zipf's formula by Mandelbrot as following:

$$f = \frac{k}{(r+\alpha)^{\beta}} \quad (0)$$

where α and β are real numbers and they are different from corpus to corpus.

As you can see, the type of the corpus matters in having a correct formula for Zipf's law and we have a slightly different formula as the type and probably the subject of our text changes. Therefore, we conclude that the Zipf's law does not work exactly the same way for different texts.

Manning and Schütze [5] argue about another aspect of Zipf's law which says that the number of senses for a word in a corpus is proportional to the square root of the frequency of that word in that corpus. That is: $m\alpha\sqrt{f}$ where *m* is the number of senses for a certain word in a corpus and *f* is the frequency of that word in the corpus.

They argue about the other aspect of Zipf's law about the size of the intervals between occurrences of same words. They mention that same words occur in different distance from each other and, therefore, we have different size of distances between the occurrences of a word in a text. .Then, they argue that the frequency of a certain interval size is proportional to the converse of that size powered to a constant. That is: $f\alpha\frac{1}{I^p}$ where *f* is the frequency of a certain interval and *I* is the size of that interval and *p* is a constant number between 1 and 1.3.

4-Effects and applications of Zipf's Law :

In the above, I briefly explained about the different aspects of Zipf's law. These different aspects could have different effects and applications in Natural Language Processing. In the remainder of this paper, I will bring some applications of Zipf's law in NLP.

4-1 – Zipf's law in n-gram model:

Just like we have individual words in a language, we have phrases of 2 or more words (n-grams) in a language. Le QuanHa, E.I. Sicilia-Garcia, Ji Ming and F.J. Smith [1], have shown that the frequency of an n-gram in an English text has a Zipf's law like relationship with its rank when n-grams are sorted in decreasing order of their frequency. They have computed the frequencies of all n-grams with n between 2 and 5 in three English and one Mandarin corpus and put them in a rank order. Then, they drew the Zipf's law like diagram for 2-grams to 5-grams. They concluded that the curves look like the Zipf's law curve saying that the frequency of a word in a corpus is proportional to the inverse of its rank when all words in the corpus are sorted in the decreasing order of their frequency. Le QuanHa, E.I. Sicilia-Garcia, Ji Ming and F.J. Smith [1] experimented that all lines are almost straight lines when they draw *log(f)* against *log(r)* and the frequency-rank relationship follows the formula: $f = \frac{k}{r^{\beta}}$ with different β in each case as following:

	WSJ87	WSJ88	WSJ89	Mandarin
2-gram	0.67	0.66	0.65	0.75
3-gram	0.51	0.50	0.46	0.59
4-gram	0.42	0.42	0.39	0.53
5-gram	0.42	0.41	0.34	0.48

Table1: Slopes for best-fit straight line approximate to the Zipf curves

They have taken all English texts from Wall Street Journal articles. Table 1 is taken from [1] and it shows that the amount of slope is almost the same for each n-gram from one text to another text.

This fact clearly shows that n-grams almost obey the Zipf's law. Therefore, we can conclude the following points about n-grams in a text: a)- There are a few very frequent

n-grams b)- There are a medium number of medium frequent n-grams c)- There are a large number of infrequent n-grams . Therefore, we conclude that only for a few n-grams we have a lot of occurrences. But what are those occurrences? These occurrences are probably the ones having more grammatical words inside themselves. Therefore, the structure of these n-grams is less important compared with that of other n-grams in the corpus as far as it is concerned to the meaning of the text. But other n-grams with the a frequency have less grammatical words and meaningful ones. These n-grams are important for the meaning of the text. Therefore, we might conclude that we will almost have most important n-grams in our corpus if we drop those n-grams with high frequency from our corpus. We know that the number of these n-grams is considerably high since they are high in their frequency. Therefore, we might shrink our corpus considerably by dropping those n-grams without hurting the meaningful n-grams in our corpus. This way, we can have a smaller corpus with the almost same meaning as our original one which might leads us to a summarization procedure. That is: keeping the less frequent n-grams and dropping the most frequent ones believing that the most frequent ones are less important since they occur a lot according to the Zipf's law and they might carry more grammatical words.

We might also conclude the second Zipf's law which relates the number of meaning of an occurrence to the square root of its frequency in the corpus. Since the first Zipf's law happened to be almost true for n-grams, we might think that the second one might be true. If so, we could conclude that less frequent n-grams have less meaning and in the extreme case they have only one meaning and they can be used only for conveying a special meaning. This description looks like the meaning of a collocation: a group of

words that have a special meaning and its meaning does not depend on the meaning of its component. Therefore a less frequent n-gram has a better chance to be a good candidate for a collocation than the more frequent one. This fact might be used as criteria for qualifying candidates for collocation. However, we should not forget that some consider the most frequent n-grams as candidates for collocations too. But also we should not forget that the most frequent n-grams contain many grammatical words as I explained in this part of the paper earlier. Besides, there are some very good collocations that occurred only once in the corpus and they make perfect candidates for collocations.

4-2 Zipf's Law in Random Texts:

Wentian Li[2] argues that a random text follow one of Zipf's laws saying that the frequency of a word is proportional to the inverse of the rank of that word when all words are sorted according to the decreasing order of their frequencies. He (she) explains a random text as following: If we have *M* alphabets in a language we will have *M+1* possible characters (including space) to print or type randomly. Therefore, we will have a random text after a while like the following small text: erpgj bnd bh a sphsjcd cg ap. Each random string "erpgj" , "bnd", and etc. is considered to be a word in this random text. Li [2] concludes that if we sort the words in the decreasing order of their frequency in a random text ,the relationship between the frequency of a word in a random text and the rank of the word is as following(the same Zipf's-Mandelbrot formula for normal text)

$$P(r) = \frac{C}{(r+B)^{\alpha}}$$

where $\alpha = 1.01158$, $C = 0.04$, and $B=1.04$. He (she) also points this fact out that α is very close to α in an English text.

The important observation that Li[2] makes in here is the fact that the random text is created by typing characters randomly and , therefore, each character has the same probability to be typed which leading us to believe that words with the same length have the same frequency in a random text. Ferrer and Sole [3] also point to this fact by saying that in Monkey language(random text) , words with the same size have the same frequency. They also give the formula for the probability (*P (L)*) of a word to have a size of *L* as following: $P(L)\alpha(1-q)^L$ where *q* is the probability of having a space.

About a real text, I believe that it is too premature if we assume that all characters have the same chance to be printed since, for example, some characters like "x" and "z" have a lower chance to be typed than other characters. But we might assume that many characters have almost the same chance to be typed more or less. This assumption can lead us to believe that, in a real text, words with the same length might have the same frequency. On the other hand longer words have lower frequencies.Considering the second Zipf's law mentioning that the number of senses for a word is proportional to the square root of the frequency of that word in a text, we might conclude that longer words might have fewer senses since they might have a lower frequency.

The other conclusion we might get in this concern is that the words with same length might probably have almost the same number of senses. The above two points might help us a lot in the field of word sense disambiguation in Natural Language Processing. That is: we might have the approximate number of senses of a word having the length of that word. And the other point is that we might be able to calculate the number of senses for a word X if we have its length, number of senses for word Y, and the length of word Y.

4-3- Zipf's law and Citations:

Silagadze[4] argues that we can have the Zipf's law like relationship in scientific citations if we have a group of citations from different articles and we count the number of occurrences of a particular citation and we rank all different citations in terms of their frequencies. He (she) argues that we can have the following formula in the above mentioned context:

$$f(r) = \frac{p_1}{(p_2 + r)^{p_3}}$$

Where f is the frequency of a citation , r is the rank of that citation when all citations are ranked in the decreasing order of their frequencies, and p_1, p_2 and p_3 are some constant numbers depending on citations.

Silagadze[4] looked at the citations of a particular author and found out that p_1, p_2 and p_3 are 0.1360×10^{19}, 50.18 , and 8.892 respectively in the above formula.

As you can see, this equation is similar to equation (0) which is the Zipf-Mandelbrot law for the word frequency in a corpus. As I understand from this result, we can probably conclude the following points:

a) - There are a few very frequent citations b)- There are a medium number medium frequent citations c)- There are a large number of low frequent citations

Therefore, we might conclude that a large number of citations for a particular author have not been cited and used frequently. On the other hand, a few references have been cited many times. This result clearly shows that only a few references have been referred over and over again by an author in a particular field of science. For example if we consider the science of Chemistry and a Chemical scientist who has many publications, we can

conclude that there are only a few references that have been referred by that scientist frequently. Those references are probably the ones that have applications in many parts of Chemistry which are probably references about the fundamental Chemistry because they are referred many times and they are used in any aspect of Chemistry. Also they are few because each science has a few fundamental books or references. On the other hand, chemistry has many other references that have not referred frequently since we have many special research subjects and only a small group of people are interested in a special field but there are a lot of such a group of people in different fields. These people might be Chemical Engineers, chemists, physical chemists, doctors, animal scientists, and etc. and each individual in each group could be interested in a different field in the same group.

What I want to conclude in here is that if we want to look for a fundamental and basic reference in a particular field of science, the Zipf's law is a helping tool to let us do so. That is : Find those reference among the references that have been referred the most. Also Zipf's law might help us find more advanced and special references in particular fields of science. That is: if we want to find a reference about a special part of a particular science, we should not waste our time looking among the references that have been referred very frequently.

4-4 Zipf's law and word senses:

As it is mentioned in introduction and literature review, Manning and Schütze[5] argue about one important aspect of Zipf's law which is about the word senses. The law says that the number of senses that any word might have is proportional to the square root of its frequency in the corpus. What we understand from this law is that the more frequent a

word is in a large text in a language, the more senses it might have in that language. It sounds that this fact is very interesting when we are dealing with senses of words and sense disambiguation in a corpus.

For example, this law tells us one word in a corpus with frequency 81 might have 3 times as many sense as one word with frequency 9 can have in the corpus. This fact could give us a useful tool in word sense disambiguation topic. If we know the frequency and number of senses of a simple word in a corpus by looking at a dictionary or by our mind because of its simplicity, we can guess and estimate the number of senses that another word might have using its frequency in the corpus. Finding out about the frequency of words in texts is an easy task using a computer program but this easy task together with the mentioned aspect of Zipf's law gives us a strong tool in guessing the number of senses for a word in NLP.

Once we have the approximate number of senses for a word, we can develop and enlarge our training data to include most senses of a considered word in our training corpus and we could have a more realistic result about the accuracy of our procedure in word sense disambiguation.

Knowing about the approximate number of senses of a word could be useful in other way too. If this approximate number is low, we know that we are dealing with a word which does not have many senses in that language. Therefore we know that it is easier to find the most popular sense and it is safer to assign this sense to our word. But if that approximate number is large, we know that we have to look at a wider range of senses to find out about the correct sense and if we simply assign the most popular sense, we are looking at a higher risk of being mistaken.

4-5: Zipf's law's relationship with co-reference resolution:

As I mentioned earlier in the literature review part of this paper, Christopher D. Manning and H. Schütze[5] , argued about one aspect of Zipf's law saying that the frequency of a special interval size between the occurrences of same words in a corpus is proportional to the converse of that size powered to a constant.

This law says that shorter distances between same words are more frequent or, in other words, same words like to be close to each other. I believe that this fact could be very interesting when we are talking about the co-referencing of words and particularly co-referencing nouns in a corpus. Co-referencing of noun phrases is a basic tool in text summarization in Natural language processing. The above mentioned aspect of Zipf's law tells us that if we have a word as a noun and we are looking for the same noun in the corpus, same nouns should be somewhere near each other in the corpus. Besides, we might also assume that all other noun phrases that co-refer to our first word (noun), might occur close to each other as a result of this aspect of Zipf's law, since other noun phrases that containing more than one word have the same meaning of the first noun that has only one word and they can be used instead of that single noun word since they are referring to the same noun.

Therefore, we can make the use of this fact in making an algorithm to make co-reference chains for noun phrases in a corpus. In other words, our algorithm might take this fact into its account that all co-referent noun phrases probably happen to be close to each other.

5-Conclusion:

Zipf's law has been created by Zipf as a result of experimenting a corpus manually and it has been completed by Mandelbrot later on. Zipf's law has many aspects in Natural Language Processing and each aspect could have many uses in NLP. Not only can Zipf's law be used for individual word tokens, it can be used for n-grams too. This fact might lead us to achieve some kind of text summarization technique and the technique for finding collocations as it is discussed in section 4-1.

We have a law like Zipf's law in random texts. This fact might give us some useful ideas about random texts and also the number of senses for a word. It also might be useful in word sense disambiguation topic in normal texts as it is explained in 4-2.

Zipf's law also works in scientific citations and it might help us to get the reference we need about a particular field of science as it is discussed in 4-3.

We have a useful aspect of Zipf's law in word senses and it again could help us finding the number of senses for a word and it could be useful in word sense disambiguation technique as it is discussed in 4-4.

Finally, there is a useful aspect of Zipf's law about the frequency of a special interval size between the occurrences of same words in a corpus. This aspect might be useful in co-reference resolution of noun phrases and, therefore, it might be useful in text summarization as it is explained in 4-5.

6-Bibliography:

1- Le Quan Ha, E.I. Sicilia-Garcia, Ji Ming,F.J. Smith, *Extension of Zipf's Law to Words and phrases,* School of Computer Science , Queen's University of Belfast

2- Wentain Li (1992), *Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution* ,Santa Fe Institute

3-Ramon Ferrer Cancho & Ricard V. Sole (2001) *"Zipf's Law and Random Texts"* Complex Systems Research Group, FEN , Universitat Politecnica de Catalunya

4-Silagadze, Z. K. (1997) *"Citations and Zip-Mandelbort Law"* Budker Institute of Nuclear Physics, Novosibirsk, Russia

5- Christopher D. Manning and H. Schütze (1999*), Foundations of Statistical Natural Language Processing*