STATISTICAL MODELS OF LANGUAGE AND ZIPF'S LAW

Victoria Bobicev, Anatol Popescu, Tatiana Zidraşco

Technical University of Moldova

Abstract: Statistical models based on text words became very widespread for the last years. Estimation of words never met in corpus is one of word probability estimation subtasks. Attempts to find the number of never met words, using Zipf's formula give rather big values for the words never met in corpus. Making several experiments we observed that the number of words never met in corpus is proportional to the number of words met only once and depends on the text vocabulary. If the following texts are of the same type with corpus, estimation of never met words is rather adequate. But if the following texts differ from the corpus, the number of never met words can either increase or decrease considerably.

Key words: Zipf's law, statistical language modelling, zero frequency problem.

INTRODUCTION

Statistical language modeling (SLM) is the attempt to capture regularities of natural language for the purpose of improving the performance of various natural language applications [Rosenfeld 2000]. Most often the statistical natural language model is based on word sequences and their probabilities in text. SLM employs statistical estimation techniques using language training data, that is, corpus of texts. The problem is that a lot of words do not appear in corpus and their probability is equal to 0.

As it was mentioned in [Борщевич 1997] in 1916 a Frenchman J. Estu arranged the words according to their use frequency and introduced the term **"rang"** of the word meaning its number in frequency dictionary. But this law attracted scientists' attention only later when published by G. Zipf [Zipf 1949]. For the frequency dictionary of words the law looks like follows:

$$r * n \approx constant \quad \text{or} \quad n = K / r \text{ , where } K \text{ – constant} \qquad (1)$$

or in the logarithmic representation

$$\log (n) \approx - \log (r) \qquad (2)$$

where

r – word rang, that is its ordinal number in the frequency dictionary
n – word frequency in the text

SOME STATISTICS FOR MY CORPORA

Corpora of texts used for the investigations in this work and their abbreviation:

- 885 documents, represent The Appeal Court decisions (http://moldova.wjin.net) - **Hot**;
- 6339 texts from the site ***România literară*** archive (www.romlit.ro) – **RL;**
- 3160 ***Adevărul*** newspaper online articles (http://www.adevarulonline.ro) - **Ad**;
- 2464 ***Evenimentul zilei*** newspaper online articles (http://www.expres.ro) - **EZ**.

In base of Zipf's law [Salton 1988] estimated word proportions with frequency ***n*** in frequency dictionary:

$$t = 1/ n(n+1) \qquad (3)$$

In the table is presented the proportion of words with frequencies from 10 to 1 calculated according to the formula and as well as numbers, found on the corpora bases.

Table 1. Word proportions with frequencies from 10 to 1.

n	***t***	**RL**	**EZ**	**Ad**	**Ho**	**Brown**	**LOB**
10	0,9	1,2	1,3	1,4	1,2	1,4	1,3
9	1,1	1,5	1,6	1,5	1,7	1,5	1,7
8	1,4	1,7	1,8	1,9	1,6	1,9	1,9
7	1,8	2,1	2,1	2,3	2,2	2,6	2,3
6	2,4	2,6	2,8	2,8	3,4	2,9	2,9
5	3,3	3,4	3,6	3,6	3,5	4,0	3,9
4	5,0	4,8	4,9	5,2	5,0	5,4	5,6
3	8,3	7,1	7,4	7,9	11,	8,5	8,5
2	16,	13,	13,3	13,	17,	14,6	14,6
1	50,	43,	38,2	39,	33,	38,3	38,4

According to the table it seems that real proportion distributions have various coefficients for different corpora. Zipf's formula diagram asymptotically approaches to the axe, without reaching it. Real curve has a definite ending having frequency **1**. Besides, all the word frequency estimation methods assume some non-zero probability for words that were never met in the corpus. So, a real diagram must end on the frequency axe, showing the frequency of words never met in the corpus. We modified the formula in order to approach it to the real distribution. The following result was obtained:

$$t \approx 1/(n2+1,5*n+0,3) + 0,05*n \qquad (4)$$

If we equal ***n*** to **zero**, we get the proportion of words never met in the text. In our case it is ***t=1/0,3=10/3*** that is three times more than the whole frequency dictionary size.

It is interesting that dependence of word number having the same frequencies on these frequencies has the same form of Zipf's law:

$$N * n \approx \text{constant} \quad \text{or} \quad n = K / N \quad , \quad \text{where } K \text{ - constant} \qquad (5)$$

or in the logarithmic way **log (n) ≈ - log (N)** (6)

where **n** – word frequency, **N** – number of words having this frequency

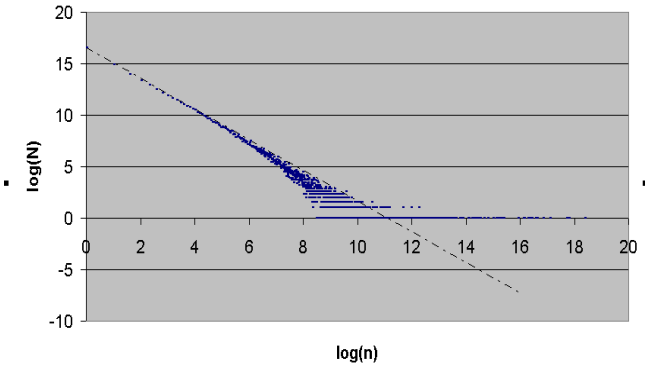

Fig. 1. Graph showing the dependence of the logarithm of the number of words with a definite frequency on this frequency logarithm.

Dash line from figure corresponds to the function:

$$y = K - m*x, \quad (7)$$

where x and y – corresponding axes values, K and m – some coefficients

In our case the coefficients are K = 16,5 and m =1,49. On the axe Y diagram and graph values almost coincide and equal 16,5, that corresponds to the logarithm of word number met once (96926). If we extend this graph on the other side of Y axe we will get the logarithm never met word number. In our case extending the graph we obtain the logarithm value 18,05 corresponding to 272255 words never met in corpus.

To estimate the number of words never met in training corpus we made some experiments.

Table 2. Percentage of words met in one part and never met in another.

	Words in part 1	Words in part 1 f.d.	Met in part 1 never met in part 2	Words in part 2	Words in part 2 f.d.	Met in part 2 never met in part 1
RL	4336900	153512	75114(44%)	3272128	172125	56501(37%)
Ad	1128185	56152	20622(36%)	1175105	57006	21476(38%)
EZ	3502180	108911	34857(28,5%)	4530138	122281	48227(44%)
Hot	306951	20860	7795(43%)	344667	18236	5170(25%)

From the table 2 we can see that number of words appeared in the second part varies and is about 30-45%. We can conclude that the number of words supposed to appear in further texts is proportional to total number of words. Number of words met only once is also proportional to the total number of words. So we can equal the number of words never met and those met only once. But it is necessary to mention that this assumption is true if and only if the texts used further will be of the same type with training texts.

Table 3. Number and percentage of words never met in one corpus and appeared in another.

	RL	Ad	EZ	Hot
RL		185442(239%)	165030(105%)	205455(789%)
Ad	38854(17%)		19242(12%)	66928(257%)
EZ	98050(44%)	98850(127%)		145600(559%)
Hot	7382(3%)	15445(20%)	14510(9%)	

All four corpora are compared in the next table 3. In each line there is a number of words met in corpus marked in the row and never met in corpus marked in the column. The maximal and the minimal results were obtained on comparing RL and Hot, literary and law texts. To compare with **Hot** frequency dictionary 789% of new words were met in RL. However 3% of words never met in RL appeared in Hot. So, while training the statistical model on the small corpus, one should expect many never met words, more that those ever met.

CONCLUSION

Estimation of words never met in corpus is one of word probability estimation subtasks. Attempts to find the number of never met words, using Zipf's formula give rather big values for the words never met in corpus. According to one estimation there are three times more words met in corpus, according to another – similar to the number of words met in text. Making several experiments we observed that the number of words never met in the corpus is proportional to the number of words met only once and depends on the text vocabulary; it also depends on the text type. If the following texts are of the same type with corpus, estimation of never met words is rather adequate. But if the following texts differ from the corpus, the number of never met words can either increase or decrease considerably.

REFERENCES

[Борщевич 1997] В. Борщевич. Информационный феномен закона Ципфа. ACta Academia, Chişinău: Evrica, 1997.

[Rosenfeld 1996] Ronald Rosenfeld. A maximum entropy approach to adaptive statistical language modeling. Computer Speech and Language, 10:187–228, 1996.

[Salton 1988] Salton G. 1988 Automatic text processing. Addison Wesley Pub. Co.

[Zipf 1949] Zipf, George K. 1949. Human Behaviour and the Principle of Least Effort. Addison-Wesley. Reading, Massachusetts.