PREDICTING THE OUT-OF-VOCABULARY RATE AND THE REQUIRED VOCABULARY SIZE FOR SPEECH PROCESSING APPLICATIONS

Johannes Müller, Holger Stahl, Manfred Lang

Institute for Human-Machine-Communication
Munich University of Technology
Arcisstrasse 21, D-80290 Munich, Germany
email: {mue,sta,lg}@mmk.e-technik.tu-muenchen.de

ABSTRACT

This paper describes an approach for predicting both the vocabulary size and the resulting out-of-vocabulary rate (OOV-rate) for a hypothetical extension of an existing text corpus. By splitting the original corpus into two different sub-corpora, vocabulary and OOV-rate can be determined for that special constellation. Average values are calculated for all combinations of sub-corpora and can be approximated by analytic function terms. These functions enable the easy prediction of the vocabulary size and the OOV-rate. The prediction accuracy results in a relative error below 4.6%.

Keywords: out-of-vocabulary rate, OOV-rate, vocabulary size, text corpus, test corpus, training corpus

1. INTRODUCTION

The occurrence of out-of-vocabulary words (OOV-words) within a speech processing application is very harmful for the user's acceptance, since each OOV-word causes usually at least one recognition error. Actual investigations report an average of 1.2 [1][3] to 1.6 [9] recognition errors, which would not occur, if the respective word would have been within the vocabulary. Hence, avoiding unknown words must have great priority during the collection of the vocabulary! In contrast to other investigations for minimizing the OOV-rate (i.e. the probability for the occurrence of an OOV-word) of very large corpora [2][6], this paper regards a limited-domain corpus with a relatively small vocabulary of a few hundred words. In this context, the following questions are of great interest:

- How much is the probability (in the following denoted as OOV-rate) that at least one word of an expected word chain is not within the current vocabulary?
- Is a prediction possible at all, how far a hypothetical corpus extension (i.e. hypothetical addition of word chains) influences the vocabulary size and the OOV-rate?
- How many word chains have to be added to the corpus to drop the OOV-rate below a certain value ε?
- How many words are in the resulting vocabulary?

2. CORPUS PARAMETER

A text corpus K consisting of L word chains W_l with $1 \le l \le L$ forms the basis for the following explanations:

$$K = \{ W_1, W_2, \ldots, W_l, \ldots, W_L \} \tag{1}$$

Each word chain W_l consists of one or more words and represents an expected user utterance within the regarded domain. The corpus size $|K|$ is described by the number of word chains:

$$|K| = L \tag{2}$$

The vocabulary $V(K)$ is the set of those T words w_t with $1 \le t \le T$, which can be found at least once within the corpus K. Multiply occurring words are included only once into this set.

$$V(K) = \{ w_1, w_2, \ldots, w_t, \ldots, w_T \} \tag{3}$$

The vocabulary size $|V(K)|$ is the number of all at least once occurring words within the corpus K:

$$|V(K)| = T \tag{4}$$

3. DETERMINING THE OOV-RATE OF ANY SUB-CORPUS

The given corpus K is split into two sub-corpora, into a training corpus K_{train} and into a test corpus K_{test}.

$$K = K_{\text{train}} \cup K_{\text{test}} \tag{5}$$

The test corpus K_{test} embodies any combination of M word chains $W_{l_1}, \ldots, W_{l_M}$ of the original corpus K.

$$K_{\text{train}} = K \setminus \{ W_{l_1}, \ldots, W_{l_M} \} \quad \text{and} \quad \left| K_{\text{train}} \right| = L - M \tag{6}$$

$$K_{\text{test}} = \{ W_{l_1}, \ldots, W_{l_M} \} \quad \text{and} \quad \left| K_{\text{test}} \right| = M \tag{7}$$

It is required that $l_1 \neq l_2 \neq \ldots \neq l_M$ with $1 \le l_m \le L$ as well as $1 \le m \le M$ and $1 \le M \le L - 1$.

The vocabulary $V(K_{\text{train}})$ is a subset of the original one $V(K)$:

$$V(K_{\text{train}}) \subseteq V(K) \tag{8}$$

Following statement can be made for the vocabulary size $\left|V(K_{\text{train}})\right|$:

$$\left|V(K_{\text{train}})\right| \leq |V(K)| \tag{9}$$

The vocabulary is reduced, if one or more words of $V(K)$ exclusively occur within K_{test} and not in K_{train}. Due to the division into the corpora K_{train} and K_{test}, it is possible to estimate the probability that at least one word of any word chain out of K_{test} is not included within the vocabulary $V(K_{\text{train}})$. This probability is denoted as OOV-rate $OOV(K_{\text{train}})$. It represents the ratio between the frequency of OOV-word chains and the overall test corpus size:

$$OOV(K_{\text{train}}) = \frac{\sum_{m=1}^{M} \text{oov}\left(W_{l_m} \middle| V(K_{\text{train}})\right)}{M} \tag{10}$$

An examined word chain W_{l_m} is considered as OOV-word chain, if at least one word of W_{l_m} is not included in the regarded vocabulary $V(K_{\text{train}})$:

$$\text{oov}\left(W_{l_m} \middle| V(K_{\text{train}})\right) = \begin{cases} 1, & \text{if one or more words of } W_{l_m} \text{ are not within } V(K_{\text{train}}) \\ 0, & \text{if all words of } W_{l_m} \text{ are within } V(K_{\text{train}}) \end{cases} \tag{11}$$

4. DETERMINING AVERAGE VOCABULARY SIZE AND AVERAGE OOV-RATE

Now, in contrast to the previous chapter, vocabulary sizes and OOV-rates are determined and averaged for all $\binom{L}{M}$ constellations of different training and test corpora as well as for constant test corpus size $\left|K_{\text{test}}\right| = M = const.$ and $1 \leq M \leq L-1$.

Depending on the respective test corpus size M, the average vocabulary size $\varphi_K(M)$ results in:

$$\varphi_K(M) = \overline{V(K_{\text{train}})} = \frac{\sum_{\substack{\text{all combinations of} \\ K_{\text{train}} \text{ and } K_{\text{test}}}} V(K_{\text{train}})}{\binom{L}{M}} \tag{12}$$

Depending on the respective test corpus size M, the average OOV-rate $\omega_K(M)$ results in:

$$\omega_K(M) = \overline{OOV(K_{\text{train}})} = \frac{\sum_{\substack{\text{all combinations of} \\ K_{\text{train}} \text{ and } K_{\text{test}}}} OOV(K_{\text{train}})}{\binom{L}{M}} \tag{13}$$

With these determined values of eq. (12) and (13), the average vocabulary size and the average OOV-rate of the corpus K, reduced by M word chains, can be calculated.

Therefore, the computation effort could be enormous, since $\binom{L}{M}$ combinations with all together $M \cdot \binom{L}{M}$ word chains have to be examined.[1] For reducing the computational load, a coincidental sample can be created from the $\binom{L}{M}$ imaginable combinations. For such a sample, the average vocabulary size $\varphi_{K,S}(M)$ and the average OOV-rate $\omega_{K,S}(M)$ are calculated. For a sufficient sample size[2], it can be assumed:

$$\varphi_K(M) \approx \varphi_{K,S}(M) \tag{14}$$

$$\omega_K(M) \approx \omega_{K,S}(M) \tag{15}$$

If the training corpus is equal to the original corpus K (i.e. M=0), the whole vocabulary is available:

$$\varphi_K(0) = |V(K)| \tag{16}$$

If the training corpus does not contain any word chain (i.e. M=L), the vocabulary degenerates to an empty set $\varnothing$ with the vocabulary size $|\varnothing| = 0$. Since any word of any word chain cannot be included within an empty vocabulary, all words are OOV-words and the resulting OOV-rate rises to one.

$$\varphi_K(L) = 0 \tag{17}$$

$$\omega_K(L) = 1 \tag{18}$$

By the help of equations (12)-(18), $\varphi_K(M)$ and $\omega_K(M)$ can be calculated for all M with 0 or $1 \leq M \leq L$. It turns out that:

- $\varphi_K(M)$ steadily decreases for $0 \leq M \leq L$, i.e. the bigger the test corpus, the smaller the training corpus, the smaller is the resulting vocabulary.
- $\omega_K(M)$ steadily increases for $1 \leq M \leq L$, i.e. the bigger the test corpus, the smaller the training corpus, the bigger is the resulting OOV-rate. This seems to be logical, since the probability inevitably decreases that words of an increasing test corpus are included in a decreasing training corpus. The reverse is an increasing probability for words being not included.

5. ANALYTIC FUNCTION MODEL

The prediction of vocabulary size and OOV-rate for a given corpus K can be done by quasi-continuous analytic function terms $\Phi_K(M)$ and $\Omega_K(M)$ dependent on M. At least for $M < \frac{L}{2}$, these terms should be a good approximation of the previously calculated values $\varphi_K(M)$ und $\omega_K(M)$. Furthermore, they must be defined for $M \leq 0$.

$$\Phi_K(M) \approx \varphi_K(M) \tag{19}$$

$$\Omega_K(M) \approx \omega_K(M) \tag{20}$$

The OOV-rate of the corpus K is equal to that probability for at least one word of an expected word chain within the regarded domain be-

---

1) For $L = 1843$ (training corpus size of our 'graphic editor' domain) and $M = 20$, there would be $\binom{1843}{20} \approx 7,576 \cdot 10^{47}$ combinations and twenty times as much word chains.

2) If the result should be precise on three fractional digits, 10^5 coincidental combinations of K_{train} and K_{test} should be enough.

ing not included in the vocabulary $V(K)$. This can be effectively approximated by the value $\Omega_K(0)$. (If there is no analytic function, the similar value $\omega_K(1)$ can be used.)

If these functions are defined for negative and integer M, the expected vocabulary size and OOV-rate can be estimated, if the corpus is enlarged by $(-M)$ additional word chains. Of course, these new word chains should be collected within the same domain and under the same conditions as those within the original corpus K.

To fall below a certain OOV-rate ε, that M has to be found, which fulfils the condition $\Omega_K(M) < \varepsilon$.

5.1. Function Term Structure

On the basis of the available values of $\varphi_K(M)$ and $\omega_K(M)$, it can be assumed that:

- The functions $\Phi_K(M)$ and $\Omega_K(M)$ must be defined for $M \le L$ (also for $M \le 0$!).
- $\Phi_K(M)$ is positive or zero within the definition range and steadily decreasing:

$$\Phi_K(M) \ge 0 \tag{21}$$

$$\frac{\partial \Phi_K(M)}{\partial M} < 0 \tag{22}$$

- In "negative direction", $\Phi_K(M)$ grows more and more slowly:

$$\lim_{M \to -\infty} \frac{\partial \Phi_K(M)}{\partial M} = 0 \tag{23}$$

- $\Omega_K(M)$ is positive within the definition range and steadily increasing:

$$\Omega_K(M) > 0 \tag{24}$$

$$\frac{\partial \Omega_K(M)}{\partial M} > 0 \tag{25}$$

- In "negative direction", $\Omega_K(M)$ approaches more and more zero and decreases more and more slowly:

$$\lim_{M \to -\infty} \Omega_K(M) = 0 \tag{26}$$

$$\lim_{M \to -\infty} \frac{\partial \Omega_K(M)}{\partial M} = 0 \tag{27}$$

The desired function terms are extracted under these preconditions and with the calculated values $\varphi_K(M)$ and $\omega_K(M)$. The structures of following functions comply with the affected requirements for suitable values a_1, a_2, a_3 und a_4. Better than other function terms, they fit very well to the values of $\varphi_K(M)$ and $\omega_K(M)$.

$$\boxed{\Phi_K(M) = a_1 \cdot \sqrt{a_2 - M} + a_3} \tag{28}$$

$$\boxed{\Omega_K(M) = \frac{a_4}{\sqrt{a_2 - M}}} \tag{29}$$

It is necessary that $a_1 > 0$, $a_2 > L$, $a_3 > -a_1\sqrt{a_2 - L}$, $a_4 > 0$ and $M \le L$ to meet all the above requirements. Note, that the OOV-rate becomes proportional to the first derivation of the vocabulary size:

$$\Omega_K(M) = -\frac{2a_4}{a_1} \cdot \frac{\partial \Phi_K(M)}{\partial M} \tag{30}$$

This seems to be a logical consequence, since the lower the growth of the vocabulary (in "negative direction"), the closer to zero the OOV-rate moves. If the vocabulary would not grow, all expected words are already included and the OOV-rate must become zero. This is theoretically the case, if an infinite number of word chains $(M \to -\infty)$ is added to the corpus.

To fall below of a certain OOV-rate ε, eq. (29) delivers the number $(-M)$ of additional word chains with $M < 0$:

$$\frac{a_4}{\sqrt{a_2 - M}} < \varepsilon \quad \Rightarrow \quad -M > \left(\frac{a_4}{\varepsilon}\right)^2 - a_2 \tag{31}$$

The resulting vocabulary size $\Phi(\varepsilon)$ depending on ε results to:

$$\Phi(\varepsilon) = a_1 \cdot \sqrt{a_2 - M} + a_3 = \frac{a_1 \cdot a_4}{\varepsilon} + a_3 \tag{32}$$

Please note in eq. (32), that $\Phi(\varepsilon)$ does not depend on a_2.

5.2. Example

Now we look at a real existing text corpus K_{wiz}[1] with $L = 1843$ word chains and a vocabulary size of $|V(K_{\text{wiz}})| = 853$. The discrete function values $\varphi_{K\text{wiz}}(M)$ and $\omega_{K\text{wiz}}(M)$ have been calculated according to eq. (12) and (13) and are shown in figure 1.

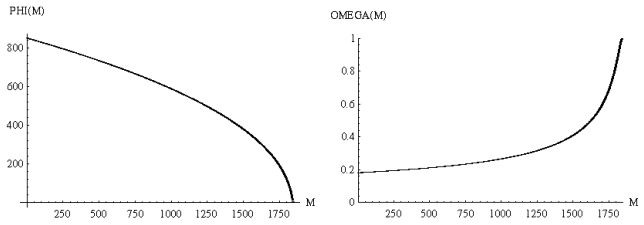

Figure 1: Average vocabulary size $\varphi_{K\text{wiz}}(M)$ and average OOV-rate $\omega_{K\text{wiz}}(M)$ depending on M

By polynomial curve fitting (function *Fit* of the software program *Mathematica* [8]), following function terms, which are derived from the previously explained structure, can be found for the calculated values $\varphi_{K\text{wiz}}(M)$ and $\omega_{K\text{wiz}}(M)$:

$$\Phi_{K_{\text{wiz}}}(M) = 19,5216 \cdot \sqrt{1893 - M} + 6,23525 \tag{33}$$

$$\Omega_{K_{\text{wiz}}}(M) = \frac{7,92816}{\sqrt{1893 - M}} \tag{34}$$

[1] This corpus was collected during a Wizard-of-Oz simulation [4] within the 'graphic editor' domain for a speech understanding task [7] and a speech translation task [5].

For a desired OOV-rate of 5%,

$$-M = \left(\frac{7,92816}{0,05}\right)^2 - 1893 \approx 23250 \quad (35)$$

must be added to the corpus K_{wiz}. Then, the vocabulary size becomes

$$\Phi_{K_{wiz}}(\varepsilon = 0,05) = \frac{19,5216 \cdot 7,92816}{0,05} + 6,23525 \approx 3102 \; . \quad (36)$$

6. INSPECTION OF THE MODEL

The model developed in the previous chapter can be checked by creating a sub-corpus $\tilde{K}$ from the original corpus K.

$$\tilde{K} \subset K \quad (37)$$

This corpus $\tilde{K}$ contains $\tilde{L}$ word chains, i.e. the corpus size is:

$$|\tilde{K}| = \tilde{L} \quad \text{with} \quad \tilde{L} < L \quad (38)$$

The corresponding functions $\Phi_{\tilde{K}}(M)$ and $\Omega_{\tilde{K}}(M)$ are determined. However, the corpus $\tilde{K}$ should be large enough to create meaningful functions[1]. Now, one can check, if the predictions made with these functions match to the available values $\Phi_K(M)$ and $\Omega_K(M)$ of the complete corpus K.

$$\Phi_{\tilde{K}}(M - L + \tilde{L}) \stackrel{?}{\approx} \Phi_K(M) \approx \varphi_K(M) \quad (39)$$

$$\Omega_{\tilde{K}}(M - L + \tilde{L}) \stackrel{?}{\approx} \Omega_K(M) \approx \omega_K(M) \quad (40)$$

For that, an example is given: The corpus K_{wiz} of the previous chapter is reduced. Every second word chain is extracted for the new corpus $\tilde{K}_{wiz}$. It contains $|\tilde{K}_{wiz}| = \tilde{L} = 922$ word chains with a resulting vocabulary size of $|\tilde{V}(K_{wiz})| = \tilde{T} = 625$ words. We received following analytic function terms:

$$\Phi_{\tilde{K}_{wiz}}(M) = 20,8573 \cdot \sqrt{955 - M} - 17,9162 \quad (41)$$

$$\Omega_{\tilde{K}_{wiz}}(M) = \frac{7,93053}{\sqrt{955 - M}} \quad (42)$$

According to eq. (39) and (40), for $M = 0$ or $M = 1$ should be valid:

$$\Phi_{\tilde{K}_{wiz}}(-921) = 885,47 \approx \Phi_{K_{wiz}}(0) = 855,59 \approx \approx \varphi_{K_{wiz}}(0) = 853 \quad (43)$$

$$\Omega_{\tilde{K}_{wiz}}(-920) = 0,1831 \approx \Omega_{K_{wiz}}(1) = 0,1822 \approx \approx \omega_{K_{wiz}}(1) = 0,1812 \quad (44)$$

The relative error of φ_{Kwiz} and ω_{Kwiz} is 3,8% for the modelled vocabulary size $\Phi_{\tilde{K}_{wiz}}$ or 1,0% for the modelled OOV-rate $\Omega_{\tilde{K}_{wiz}}$.

[1] As a rule of thumb, it can be assumed that a corpus with a corpus size greater than its vocabulary size, i.e. $|\tilde{K}| >> |V(\tilde{K})|$, is large enough for estimating the analytic functions.

An addition of 2921 word chains to the reduced corpus $\tilde{K}$ corresponds to an addition of 2000 word chains to the original one K:

$$\Phi_{\tilde{K}_{wiz}}(-2921) = 1280,61 \approx \Phi_{K_{wiz}}(-2000) = 1224,26 \quad (45)$$

$$\Omega_{\tilde{K}_{wiz}}(-2921) = 0,1274 \approx \Omega_{K_{wiz}}(-2000) = 0,1271 \quad (46)$$

The relative error of Φ_{Kwiz} and Ω_{Kwiz} is 4,6% for the modelled vocabulary size $\Phi_{\tilde{K}_{wiz}}$ or 0,2% for the modelled OOV-rate $\Omega_{\tilde{K}_{wiz}}$.

7. CONCLUSION

The small relative errors of the previous example show that the prediction of the expected vocabulary size and the expected OOV-rate deliver useful values. The introduced approach cannot and does not want to make any exact prediction, but it is able to estimate the tendency and the order of magnitude of both the vocabulary size and the OOV-rate. The values can be hypothetically estimated without real corpus extension by a relatively simple procedure. As soon as training data are collected for a speech processing application, a statement can be given about usability and completeness of the vocabulary set up by the training data.

8. REFERENCES

1. L. Chase, R. Rosenfeld, A. Hauptmann, M. Ravishankar, et al.: *Improvements in Language, Lexical, and Phonetic Modeling in Spinx-II*, Proc. „ARPA Spoken Language Systems Technology Workshop“ (Austin, USA), 1995, pp. 60-65
2. P. Fetter, F. Class, U. Haiber, A. Kaltenmeier, U. Kilian, P. Regel-Brietzmann: *Detection of Unknown Words in Spontaneous Speech*, Proc. Eurospeech 1995 (Madrid, Spain), pp. 1637-1640
3. J. Gauvin, L. Lamel, M. Adda-Decker: *Developments in Continuous Speech Dictation Using the ARPA WSJ Task*, Proc. „ARPA Spoken Lang. Syst. Techn. Workshop“, 1994
4. J. Müller, H. Stahl: *Collecting and Analyzing Spoken Utterances for a Speech Controlled Application*, Proc. Eurospeech 1995 (Madrid, Spain), pp. 1437-1440
5. J. Müller, H. Stahl, M. Lang: *Automatic Speech Translation Based on the Semantic Structure*, Proc. ICSLP 1996 (Philadelphia, USA), to be published in these proceedings
6. R. Rosenfeld: *Optimizing Lexical and N-gram Coverage via Judicious Use of Linguistic Data*, Proc. Eurospeech 1995 (Madrid, Spain), pp. 1763-1766
7. H. Stahl, J. Müller, M. Lang: *An Efficient Top-Down Parsing Algorithm for Understanding Speech by Using Stochastic Syntactic and Semantic Models*, Proc. ICASSP 1996 (Atlanta, USA), pp. I.397-I.400
8. S. Wolfram: *Mathematica - a System for Doing Mathematics by Computer*, Addison-Wesley, Redwood City, 1991
9. P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev, S.J. Young: *The Development of the 1994 HTK Large Vocabulary Speech Recognition System*, Proc. „ARPA Spoken Language Techn. Workshop“ (Austin, USA), 1995, pp. 104-109