PROCEEDINGS

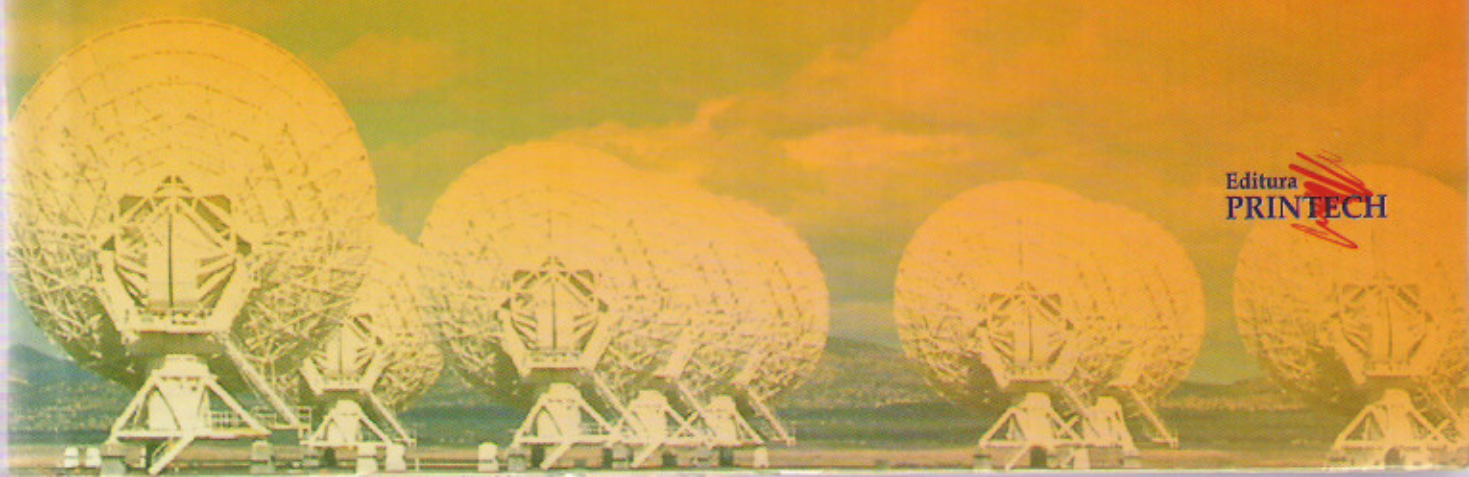
of the 7th International Conference

COMMUNICATIONS 2008

Edited by:

Adelaida MATEESCU
Alexandru ŞERBĂNESCU
Mihai RADU
Victor CROITORU

Bucharest, June 5-7
ROMANIA



Editura PRINTECH

INTERNATIONAL CONFERENCE

COMMUNICATIONS 2008

June 05 – 07, 2008
BUCHAREST, ROMANIA

PROCEEDINGS

Organized by:
Military Technical Academy
"Politehnica" University of Bucharest
"Electronica 2000" Foundation
and
IEEE Romanian Section

Editura PRINTECH
2008

Honorary Chairperson:

Adelaida MATEESCU,
"Politehnica" University of Bucharest

Conference Chairpersons:

Mihai RADU,
Military Technical Academy
Alexandru ŞERBĂNESCU,
Military Technical Academy
Victor CROITORU,
"Politehnica" University of Bucharest

Honorific Committee:

Doru SAFTA,
Rector of Military Technical Academy, Bucharest
Ecaterina ANDRONESCU,
Rector of "Politehnica" University of Bucharest

Organizing Committee

"Politehnica" University of Bucharest	*Military Technical Academy*
Theodor PETRESCU	Ioan NICOLAESCU
Silviu CIOCHINĂ	Adrian STOICA
Ion BĂNICĂ	Victor PATRICIU
Ion MARGHESCU	Doru MUNTEANU
Eugen BORCOCI	Cristian MOLDER
Constantin PALEOLOGU	Cristian RÎNCU
Călin VLĂDEANU	

Editura acreditată de Consiliul Național al Cercetării Științifice din Învățământul Superior - Cod CNCSIS 54

TIPAR:
Editura PRINTECH (S.C. ANDOR TIPO S.R.L.)
Str.TUNARI nr.11, sector 2, BUCURESTI
Tel/Fax : 211.37.12

Certificat nr.1303

Descrierea CIP a Bibliotecii Naționale a României

COLECTIV AUTORI
COMMUNICATIONS 2008

Bucureşti, Printech, 2008
Bibliogr., Index
ISBN 978-606-521-008-0

NATURAL LANGUAGE AND SPEECH PROCESSING

STATISTICAL LANGUAGE MODELING ON SUSANNE CORPUS

Domokos József[1,2], *Toderean Gavril*[2], *Ovidiu Buza*[2]

[1]Faculty of Technical and Human Sciences/Electrical engineering department, Sapientia University
Sighişoarei street, 1C, 540485, Tîrgu-Mureş/Corunca, România
phone: + 40265-208.170, fax: +40265-206.211, email: domi@ms.sapientia.ro
web: www.ms.sapientia.ro

[2]Faculty of Electronics, Telecommunications and Information Technology/Communications department,
Technical University of Cluj-Napoca
Gheorghe Barițiu street, 26-28, 400027, Cluj-Napoca, România
phone: + 40264-596.285, fax: +40264-591.689
email: toderean@pro3soft.ro
email: Ovidiu.Buza@com.utcluj.ro
web: www.etti.utcluj.ro

ABSTRACT

In this paper are presented the theoretical background and some experiments for statistical language modeling which is a main part of a statistical speech recognition system. There are presented the unigram, bigram and trigram language model as well as the Good-Turing estimator based back-off smoothing algorithm. The experiments were made on Susanne Corpus, a grammatically tagged speech corpus, based on Brown corpus. The built language models are tested and compared using perplexity measure.

1. STATISTICAL SPEECH RECOGNITION

Statistical speech recognition is based on Hidden Markov Models (HMM's). Such a system, depicted in *Fig. 1*, is built using multiple chained HMM's for acoustic modeling and language modeling.

The above system can be described mathematically as fallows [3] [4] [5] [8]: we have a set of acoustic vectors $A = \{a_1, a_2, ..., a_n\}$ and we are searching the most prob-

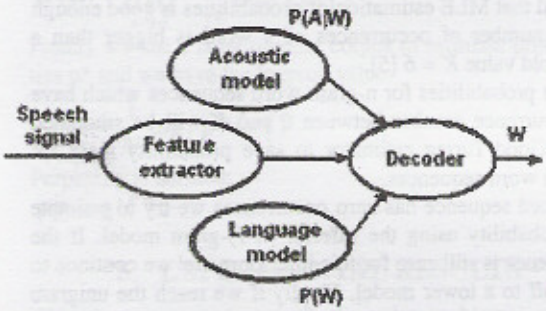


Fig. 1: Statistical speech recognition system architecture

able word sequence $W^* = \{w_1, w_2, ..., w_m\}$.

$$W^* = \arg\max_w \{P(W \mid A)\} \quad (1)$$

Using Bayes theorem, we can transcribe *Eq. 1*:

$$W^* = \arg\max_w \{\frac{P(A \mid W) \cdot P(W)}{P(A)}\} \quad (2)$$

We know, that probability of acoustic vector *P(A)* is constant, and we have:

$$W^* = \arg\max_w \{P(A \mid W) \cdot P(W)\} \quad (3)$$

In *Eq. 3* we have:

- *P(W)* – the language model
- *P(A|W)* – the acoustic model

The acoustic modeling part of the speech recognition system can be developed using HMMs, Gaussian Mixture Models (GMMs) or Artificial Neural Networks (ANNs).

The language modeling part of the system can be: [1] [3] [4] [5]:

- statistical language model
- context free grammar (CFG)
- probabilistic context free grammar (PCFG).

In this paper we want to present the statistical n-gram type language model which is the most powerful and the most widely used one.

2. STATISTICAL LANGUAGE MODELING

The speech can be considered a stochastic process and every linguistic unit (phoneme, syllabus, or word) can be considered a random variable with a random probability distribution. The n-gram language models try to estimate the probability of the next word based on the history (the last *n-1* preceding words) [4] [5] [7] [8].

The language model must estimate the probability of word sequence $w_1^n = (w_1, w_2, ..., w_n)$, which is:

$$p(w_1^n) = p(w_1) \cdot p(w_2 \mid w_1) \cdot p(w_3 \mid w_1^2) \cdot ... \cdot p(w_n \mid w_1^{n-1})$$

$$p(w_1^n) = p(w_1) \cdot \prod_{k=2}^{n} p(w_k \mid w_1^{k-1}) \quad (4)$$

Using Markov assumption, the history can be reduced to the last n-1 words, and we have:

$$P(w_k \mid w_1^{k-1}) \approx P(w_k \mid w_{k-n+1}^{k-1}) \quad (5)$$

Even *Eq. 5* is hard to compute for $n > 3$ because we need a huge training corpus to properly evaluate the probabilities. For $n = 1 .. 3$ we have:

- Unigram language model ($n=1$)
- Bigram language model ($n=2$)
- Trigram language model ($n=3$)

2.1 Unigram language model

The unigram language model considers all the words independent. This means that no history information is involved.

$$P(w_k \mid w_1^{k-1}) \approx P(w_k) \quad (6)$$

If we use *Eq. 4*, we have the probability estimation for the unigram model.

$$p(w_1^n) = \prod_{k=1}^{n} p(w_k) \quad (7)$$

2.2 Bigram language model

The bigram language model takes in consideration one word for history.

$$P(w_k \mid w_1^{k-1}) \approx P(w_k \mid w_{k-1}) \quad (8)$$

If we put *Eq. 8* in *Eq. 4*, we have the formula for probability estimation for bigram language model:

$$p(w_1^n) = p(w_1) \cdot \prod_{k=2}^{n} p(w_k \mid w_{k-1}) \quad (9)$$

2.3 Trigram language model

The trigram language model uses a two word history.

$$P(w_k \mid w_1^{k-1}) \approx P(w_k \mid w_{k-1}, w_{k-2}) = P(w_k \mid w_{k-2}^{k-1}) \quad (10)$$

The formula for probability estimation is given in *Eq. 11*.

$$p(w_1^n) = p(w_1) \cdot p(w_2 \mid w_1) \cdot \prod_{k=3}^{n} p(w_k \mid w_{k-1}, w_{k-2}) \quad (11)$$

3. PROBABILITY ESTIMATION AND SMOOTHING

The probabilities for *Eq. 7, 9,* and *11* can be simply calculated using MLE (Maximum Likelihood Expectation) algorithm [5]. Thus we have for unigram, bigram and trigram language models the fallowing MLE estimators.

$$p(w_k) = \frac{n_k}{N} \quad (12)$$

$$p(w_k \mid w_{k-1}) \approx \frac{Nr.(w_{k-1}, w_k)}{Nr.(w_{k-1})} \quad (13)$$

$$p(w_k \mid w_{k-1}, w_{k-2}) \approx \frac{Nr.(w_{k-2}, w_{k-1}, w_k)}{Nr.(w_{k-2}, w_{k-1})} \quad (14)$$

n_k – the number of occurrence of word w_k

N – the total number of words in training corpus

$Nr.(...)$ – the number of occurrence of a specific word sequence

These probabilities calculated using MLE algorithm do not provide useful results. In order to can use the probabilities in language modelling experiments, they must be smoothed [1] [2] [4] [5].

Smoothing means that a probability mass is retained from high probabilities to be reallocated to zero or small probability values. There are a lot of useful smoothing techniques:

- Add one or Laplace smoothing
- Good-Turing estimator
- Back-off or Katz smoothing
- Kneser - Ney smoothing
- Jelinek - Mercer smoothing or interpolation

For the experiments we used Good - Turing estimator and back-off smoothing.

3.1. Good-Turing estimator

The Good-Turing estimator comes from biology where it was used for species estimation.

The general form of the estimator is [6]:

$$P(X) = \frac{r^*}{N} \quad (15)$$

$$where, r^* = (r+1) \cdot \frac{E(N_{r+1})}{E(N_r)}$$

where:

r is the number of occurrence of word X;

N_r is the number of words which occurs exactly r times in the training corpus;

N is the total number of words from the training corpus;

E is an estimation function for N_r;

r^* is the adjusted number of occurrence;

The total value of probability calculated using Good-Turing estimator is always smaller than *1*. The remaining probability mass is reallocated to the unseen words from the vocabulary. The simplest way to choose the estimation function E [6] is presented in *Eq. 16*.

$$\frac{E(n+1)}{E(n)} = \frac{n}{n+1} \cdot (1 - \frac{E(1)}{N}) \quad (16)$$

3.2. Back-off smoothing

Back-off smoothing was firstly introduced by Katz. He showed that MLE estimation of probabilities is good enough if the number of occurrences of a word is bigger than a threshold value $K = 6$ [5].

All the probabilities for n-gram word sequences which have an occurrence number between *0* and *K* will be smoothed using Good-Turing estimator to save probability mass for unseen word sequences.

If a word sequence has zero occurrences we try to estimate his probability using the inferior *(n-1)*-gram model. If the occurrence is still zero for this inferior model we continue to back-off to a lower model. Finally if we reach the unigram model, we have the relative frequency of a word bigger than zero.

For a trigram back-off model we have the fallowing relations:

$$\hat{p}(w_3|w_1,w_2)=\begin{cases} f(w_3|w_1,w_2), & if \quad nr.(w_1,w_2,w_3)\geq K \\ \alpha\cdot Q_r(w_3|w_1,w_2) & if \quad 0<nr.(w_1,w_2,w_3)<K \\ & else \quad \beta\cdot \hat{p}(w3|w2) \end{cases}$$

$$\hat{p}(w_3|w_2)=\begin{cases} f(w_3|w_2) & if \quad nr.(w_2,w_3)\geq L \\ \alpha\cdot Q_r(w_3|w_2) & if \quad 0<nr.(w_2,w_3)<L \\ & else \quad \beta\cdot f(w_3) \end{cases}$$

4. LANGUAGE MODEL EVALUATION

Language model evaluation can be done in different ways [1] [2] [4] [5]:

- random sentence generation
- words reordering in sentences
- perplexity

For the experiments we used perplexity to measure the quality of language models. Perplexity is the most used measure for language model evaluation.

Perplexity can be defined using entropy from information theory. For a random variable $X=\{x_1, x_2, ...,x_N\}$, the entropy can be defined:

$$H(X)=-\sum_{x\in X} p(x)\cdot \log_2 p(x) \tag{17}$$

Instead of entropy, we use the entropy rate calculated as fallows:

$$\frac{1}{N}H(w_1^n)=-\frac{1}{N}\sum_{w\in V} p(w_1^n)\cdot \log_2 p(w_1^n) \tag{18}$$

For a real language we should consider infinitely long word sequences:

$$\frac{1}{N}H(w_1^n)=-\frac{1}{N}\sum_{w\in V} p(w_1^n)\cdot \log_2 p(w_1^n) \tag{19}$$

Using Shanon-McMilan-Breiman theorem, if the language is stationary and ergodic, the above formula can be simplified:

$$H(L)=\lim_{n\to\infty} -\frac{1}{N}\log_2 p(w_1^n) \tag{20}$$

Finally we use a large training corpus to estimate probabilities p* and we have the logprob value:

$$LP=-\frac{1}{N}\log_2 p^*(w_1^n) \tag{21}$$

Perplexity is defined:

$$PP=2^{LP} \tag{22}$$

5. EXPERIMENTAL RESULTS

We try to create and compare the unigram, bigram and trigram Katz back-off language models on Susanne Corpus, developed at University of Sussex, England [8]. This corpus contains 64 grammatically tagged files from Brown Corpus, each of them with more than 2000 words. The 64 files are grouped in 4 categories (each with 16 files) noted as fallows [8]:

- A – press reportage
- G - belles lettres, biography, memoirs
- J – learned (mainly scientific and technical) writing
- N – adventure and Western fiction

We divide the corpus in two sets: one training set and one test set. The test set consists of one file from each category. The rest of files were used for training the language models.

In our first experiment we have built unigram, bigram and trigram language models for each category using the most probable 5000 words for dictionary and test them using the corresponding test file.

We calculate the perplexity of language models to evaluate and compare them.

The results are presented in *Tables 1* to *4*.

For *A* type model we have used 33175 words in the training set and 2226 words from 85 sentences in the test set. The number of out-of-vocabulary words was 505, that means 23.59%.

Table 1. Perplexity results for language models built using *A* category files from Susanne Corpus

Model	Perplexity	Words predicted
Unigram	525.85	1615
Bigram	368.07	1227
Trigram	376.24	959

For *G* type texts model we used 33866 words for training and 78 sentences with a number of 2232 words for testing. The number of out of vocabulary words was 394 (18,29%).

Table 2. Perplexity results for language models built using *G* category files from Susanne Corpus

Model	Perplexity	Words predicted
Unigram	525.61	1760
Bigram	403.48	1427
Trigram	409.81	1180

For *J* type texts model we used 33373 words for training and 62 sentences with a number of 2164 words for testing. The number of out of vocabulary words was 441 (20.98%).

Table 3. Perplexity results for language models built using *J* category files from Susanne Corpus

Model	Perplexity	Words predicted
Unigram	559.74	1637
Bigram	397.37	1277
Trigram	419.52	1011

For N type texts model we used 34845 words for training and 103 sentences with a number of 2230 words for testing. The number of out of vocabulary words was 267 (12.55%)

Table 4. Perplexity results for language models built using *N* category files from Susanne Corpus

Model	Perplexity	Words predicted
Unigram	410.7	1836
Bigram	260.52	1612
Trigram	261.48	1448

In the second experiment we have built unigram, bigram and trigram models for the whole corpus using three different dictionaries (5000, 10000, 15000 words) and test them with all test set. The 5000 word dictionary is useful to compare this model with those adapted to text genre. The model built based on 10000 word dictionary is relevant because appreciatively 10000 words occur more tan once in training set. The last model built on a 15000 word dictionary includes almost all of the vocabulary of the training set of Susanne corpus.

Thus we trained the models using 135259 words (15444 distinct words) from 6259 sentences. The test set contains 8855 words from 328 sentences.

The results are resumed in *Table 5*.

Table 5. Perplexity results for language models built using the entire Susanne Corpus

Vocabulary	Model	Perplexity	Words predicted
5000	Unigram	625.03	7126
	Bigram	355.64	5957
	Trigram	352.7	5078
10000	Unigram	815.53	7542
	Bigram	476.64	6673
	Trigram	470.62	5984
15000	Unigram	936.22	7753
	Bigram	550.26	7065
	Trigram	545.62	6502

5. CONCLUSIONS AND FURTHER WORKS

In the first experiment the best results are achieved using bigram model. The trigram model can't improve the results because there are insufficient data for training the models.

We can see from the results that if the number of words increase, the perplexity of the model increase too, and the model has weaker quality. The models built by eliminating the words with occurrences smaller than a threshold are simpler and performs better.

The best perplexity achieved for N type files from Susanne corpus outperforms those reported in the literature (i.e. [9]).

The words which occurs once does not improve the model quality, they rise perplexity and they should be eliminated from vocabulary.

In the second experiment the trigram models performs better than bigram models. This performance can be explained by the large amount of available training data.

In conclusion the used n-gram model should be choosed considering the amount of training data.

REFERENCES

[1] Becchetti C., Ricotti L. P., *"Speech recognition. Theory and C++ implementations"*, John Wiley & sons, 1999.

[2] Chen S., Goodman J., *"An Empirical Study of Smoothing Techniques for Language Modeling"*, Harvard Computer Science Technical report TR-10-98, 1998.

[3] Huang X., Acero A., Hon H., *"Spoken Language Processing. A Guide to Theory, Algorithm & System Development"*, Prentice Hall, 2001.

[4] Jelinek F., *"Statistical Methods for speech recognition"*, The MIT Press, 2001.

[5] Jurafsky D., Martin J. H., *„Speech and language processing. An introduction to Natural language Processing. Computational Linguistics and Speech Recognition"*, Prentice Hall, 2000.

[6] Manning C., Heinrich S., "Foundations of statistical language processing", The MIT Press, 1999.

[7] Rosenfeld, R., *„Two decades of statistical language modeling: where do we go from here?"*, Proceedings of the IEEE, Volume 88, Issue 8, Aug 2000, Page(s):1270 – 1278.

[8] Nelleke O., P. de Haan, eds., *„Corpus-Based Research into Language - SUSANNE: a Domesday Book of English grammar" by Geoffrey Sampson*, Rodopi (Amsterdam), 1994.

[9] Quiniou S., Anquetil E., Carbonnel S., *"Statistical Language Models for On-line Handwritten Sentence Recognition"*, Proceedings of Eighth International Conference on Document Analysis and Recognition. Volume I, 29 Aug. - 1 Sept. 2005, Page(s): 516 – 520.

[10] Young S., Evermann G., Gales M., Hain T., Kershaw D., Moore G., Odell J., Ollason D., Povey D., Valtchev V., Woodland P., *"The HTK Book"*, Cambridge University Engineering Department, 2005.

CONTENTS

SESSION OF INVITED PAPERS

SIGNAL PROCESSING 1

SIGNAL PROCESSING 2

NATURAL LANGUAGE AND SPEECH PROCESSING

IMAGE PROCESSING 1