ROMANIAN ACADEMY
UNIVERSITY "POLITEHNICA" OF BUCHAREST
INSTITUTE FOR COMPUTER SCIENCE, IAȘI
MARITIME UNIVERSITY OF CONSTANȚA

From Speech Processing to Spoken Language Technology

Corneliu BURILEANU
Horia-Nicolai TEODORESCU
Editors

THE PUBLISHING HOUSE OF THE ROMANIAN ACADEMY

ROMANIAN ACADEMY
Section for Science and Technology of Information,
UNIVERSITY "POLITEHNICA" OF BUCHAREST
Faculty of Electronics, Telecommunications and Information Technology
INSTITUTE FOR COMPUTER SCIENCE, IAŞI
MARITIME UNIVERSITY OF CONSTANŢA

In cooperation with

The European Association for Signal and Image Processing (EURASIP)
IEEE

From Speech Processing to Spoken Language Technology

Editors:
Corneliu BURILEANU
Horia-Nicolai TEODORESCU

The Publishing House of the Romanian Academy
Bucharest, 2009

Address: THE PUBLISHING OF THE ROMANIAN ACADEMY
(EDITURA ACADEMIEI ROMÂNE)
Calea 13 Septembrie, nr. 13, Sector 5
050711, București, România,
Tel: 4021-318 81 46, 4021-318 81 06
Fax: 4021-318 24 44
E-mail: edacad@ear.ro
Adresa web: www. ear.ro

Descrierea CIP a Bibliotecii Naționale a României
CONFERENCE SPEECH TECHNOLOGY AND HUMAN-COMPUTER-DIALOGUE "SPED 2009" (5; 2009; Constanța)

From speech processings to spoken language technology: Proceedings of the 5th Conference Speech Technology and Human-Computer-Dialogue "SpeD 2009": Constanța, România, June 18–21, 2009 / editors: Corneliu Burileanu, Horia-Nicolai Teodorescu. – București:
Editura Academiei Române, 2009
ISBN 978-973-27-1808-7

I. Burileanu, Corneliu (ed.)
II. Teodorescu, Horia Nicolai (ed.)

004.383.3:004.934(963)
81'322(063)

Proceedings of the 5th Conference
"Speech Technology and Human-Computer Dialogue SpeD 2009"

IEEE Catalog Number: CFP0955H-CDR

Editorial Assistant: Mihaela MARIAN
Computer Editing: Luiza DOBRIN
Cover: Nicoleta ZORZON

Bun de tipar: 12.05.2009. Format: 8/61 × 86
Coli de tipar: 25,75
C.Z. pentru biblioteci mari: 534.781:681.142-83(082)
C.Z. pentru biblioteci mici: 007

***Honor Committee*:**

Mihai Drăgănescu,	Honorary Chair, member of the Romanian Academy, President of the Section for Science and Technology of Information.
Ecaterina Andronescu,	Rector of the University "Politehnica" of Bucharest.
Cornel Panait,	Rector of Constanța Maritime University.
Teodor Petrescu,	Dean of the Faculty of Electronics, Telecommunications and Information Technology, University "Politehnica" of Bucharest.

***Scientific Committee*:**

Laurent Besacier,	Laboratory CLIPS, CNRS – Institut National Polytechnique de Grenoble – Université Joseph Fourier Grenoble, France.
Corneliu Burileanu,	University "Politehnica" of Bucharest, Faculty of Electronics, Telecommunications and Information Technology.
Dragoș Burileanu,	University "Politehnica" of Bucharest, Faculty of Electronics, Telecommunications and Information Technology.
Jean Caelen,	Laboratory CLIPS, CNRS – Institut National Polytechnique de Grenoble – Université Joseph Fourier Grenoble, France.
Dan Cristea,	"Al. I. Cuza" University of Iași, Computer Science Faculty.
Sorin Dusan	MCT Inc./NASA Ames Research Center, Intelligent Systems Division.
Daryle J. Gardner-Bonneau,	Speech Pathology and Audiology, Western Michigan University, Kalamazoo, MI, USA.
Inge Gavăt,	University "Politehnica" of Bucharest, Faculty of Electronics, Telecommunications and Information Technology.
Jean-Paul Haton,	Université "Henri Poincaré", Nancy, membre de l'Institut Universitaire de France.
Geza Nemeth,	Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Hungary.
Eugeniu Oancea,	Military Technical Academy Bucharest.
Beatrice Pesquet-Popescu,	Ecole Nationale Supérieure des Télécommunications Paris, France, AdCom member EURASIP, Secretary of the Executive Subcommittee of the IEEE Signal Processing Society (SPS) Conference Board.
Mihai Radu,	Military Technical Academy Bucharest.
Markus Rupp,	Institut for Nachrichtentechnik und Hochfrequenztechnik, Vienna University of Technology, Austria, AdCom member EURASIP.
Alexandru Şerbănescu,	Military Technical Academy Bucharest.
Jean-Francois Serignat,	Laboratory CLIPS, CNRS – Institut National Polytechnique de Grenoble – Université Joseph Fourier Grenoble, France.
Erchin Serpedin	Texas A&M University, Dept. of Electrical and Computer Engineering, College Station, TX , USA.
Horia-Nicolai Teodorescu,	c.m. of the Romanian Academy, Technical University of Iași.
Gavril Toderean,	Technical University of Cluj-Napoca, Faculty of Electronics and Telecommunications.
Dan Tufiș,	c.m. of the Romanian Academy, Director of the Research Institute for Artificial Intelligence.

Organizing Committee:

Corneliu Burileanu,	Chair of the Conference, Vice-Rector of the University "Politehnica" of Bucharest.
Horia-Nicolai Teodorescu,	c.m. of the Romanian Academy, Vice-Rector of the Technical University of Iaşi.
George Căruntu,	Dean of the Electromechanic Faculty, Constanţa Maritime University.
Călin Vlădeanu,	University "Politehnica" of Bucharest, Faculty of Electronics, Telecommunications and Information Technology.
Constantin Paleologu,	University "Politehnica" of Bucharest, Faculty of Electronics, Telecommunications and Information Technology.
Mihaela Hnatiuc,	Constanţa Maritime University.
Gabriel Raicu,	Constanţa Maritime University.
Alexandru Caranica,	Constanţa Maritime University, Webmaster.

CONTENTS

Foreword,
Corneliu BURILEANU 9

PART 1. Speech Recognition and Understanding. Human-Computer Dialogue

Speech Analysis for Automatic Speech Recognition: a Review,
Jean-Paul HATON 13

ProtoLOGOS, System for Romanian Language Automatic Speech Recognition and Understanding (ASRU),
Diana MILITARU, Inge GAVAT, Octavian DUMITRU, Tiberiu ZAHARIA, Svetlana SEGARCEANU 21

Comparing Various Voice Recognition Techniques,
Tudor BARBU 33

Optimizing a Discourse Structuring Component for Utterance Generation in Human-Computer Dialogue,
Vladimir POPESCU, Jean CAELEN, Corneliu BURILEANU 43

PART 2. Text-to-Speech Synthesis. Real-Time Speech-Enabled Communication Applications

Assessing the Quality of Voice Synthesizers,
Horia-Nicolai TEODORESCU, Monica FERARU, Marius ZBANCIOC 53

Time-Frequency Processing of Partials for High-Quality Speech Synthesis,
Amelia CIOBANU, Cristian NEGRESCU, Dragoş BURILEANU, Dumitru STANOMIR 67

A Prosodic Control Module for a Romanian TtS System, Based on Melodic Contour Dictionaries,
Doina JITCA, Vasile APOPEI 77

Automatic Rule-Based Syllabication for Romanian,
Ştefan-Adrian TOMA, Eugeniu OANCEA, Doru-Petru MUNTEANU 87

Real-Time Architectures for a Network-Based Text-to-Speech Service Implementation,
Mihai SURMEI, Dragoş BURILEANU, Cristian NEGRESCU, Cătălin UNGUREAN, Aurelian DERVIŞ 95

Experiments with the Prediction and Generation of Romanian Intonation,
Arpad Zsolt BODO, Ovidiu BUZA, Gavril TODEREAN 103

PART 3. Natural Language Processing

Factored Phrase-Based Statistical Machine Translation,
Dan TUFIŞ, Alexandru CEAUŞU 115

Dynamic Relationship Management for Personality Rich Character Presentations in Interactive Games,
Manish MEHTA, Kinshuk MISHRA, Andrea CORRADINI 125

Discourse Theories vs. Topic-Focus Articulation Applied to Prosodic Focus Assignment in Romanian,
Neculai CURTEANU, Diana TRANDABĂŢ, Mihai Alex MORUZ 135

General System for Normal and Phonetic Inflection,
Stefan DIACONESCU, Cristi INGINERU, Felicia CODIRLASU, Monica RIZEA, Oana BULIBASA 149

Text Conditioning and Statistical Language Modeling for Romanian Language,
József DOMOKOS, Gavril TODEREAN, Ovidiu BUZA 161

PART 4. Speech and Audio-Signal Processing. Speech Interface Applications

Speech Recognition in a Smart Home: Some Experiments for Telemonitoring,
Michel VACHER, Anthony FLEURY, Noé GUIRAND, Jean-François SERIGNAT, Norbert NOURY 169

Variable Step-Size Adaptive Algorithms for Echo Cancellation,
Constantin PALEOLOGU, Silviu CIOCHINĂ, Călin VLĂDEANU 181

Wavelet Analysis for Audio Signals with Music Classification Applications,
Anca POPESCU, Inge GAVAT, Mihai DATCU 189

Linear Interpolation of Spectrotemporal Excitation Pattern Representations
for Automatic Speech Recognition in the Presence of Noise,
Adriana STAN 199

TEXT CONDITIONING AND STATISTICAL LANGUAGE MODELING FOR ROMANIAN LANGUAGE

József DOMOKOS[*] [and] [**], Gavril TODEREAN[**], Ovidiu BUZA[**]

[*] Sapientia University, Faculty of Technical and Human Sciences, Electrical Engineering Department
[**] Technical University of Cluj-Napoca, Faculty of Electronics, Telecommunications and Information Technology, Communications Department

Corresponding author: József DOMOKOS

In this paper we present a synthesis of the theoretical fundamentals and some practical aspects of statistical (n-gram) language modeling which is a main part of a large vocabulary statistical speech recognition system. There are presented the unigram, bigram and trigram language models as well as the Good-Turing estimator based Katz back-off smoothing algorithm. There is also described the perplexity measure of a language model used for evaluation. The practical experiments were made on Romanian Constitution corpus. There are also presented the text normalization steps before the language model generation. The results are ARPA-MIT format language models for Romanian language. The models were tested and compared using perplexity measure. Finally some comparisons were made between Romanian and English language modeling and conclusions are drawn.

Key words: Romanian statistical language modeling; natural language processing; text conditioning; ARPA-MIT language model format; n-gram language modeling, smoothing, perplexity.

1. INTRODUCTION

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.

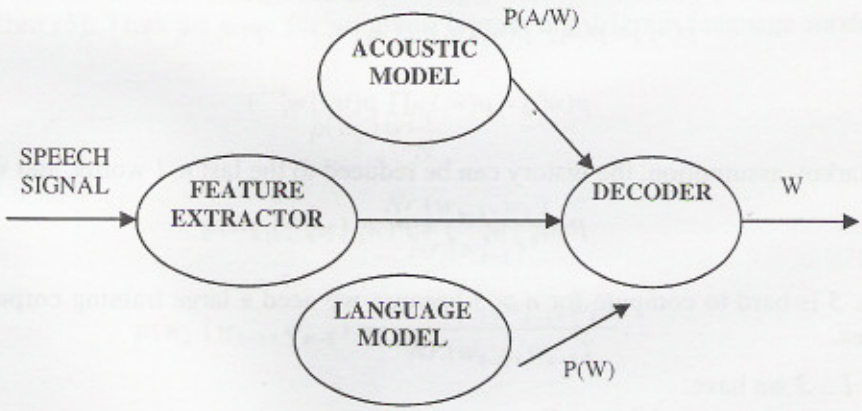

Figure 1. Statistical speech recognition system architecture.

The above system can be described mathematically as follows: we have a set of acoustic vectors $A = \{a_1, a_2, ..., a_n\}$ and we are searching the most probable word sequence: $W^* = \{w_1, w_2, ..., w_m\}$.

$$W^* = \arg\max_{W} \{P(W \mid A)\} \tag{1}$$

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

$$W^* = \arg\max_{w} \{ \frac{P(A|W) \cdot P(W)}{P(A)} \} \tag{2}$$

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

$$W^* = \arg\max_{w} \{ P(A|W) \cdot P(W) \} \tag{3}$$

In *Eq. 3* we can distinguish:

- $P(W)$ – 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:

- 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, and we want to create ARPA MIT format Romanian language models for large vocabulary continuous speech recognition systems.

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) [6] [7] [8] [10][11].

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

$$p(w_1^n) = p(w_1) \cdot p(w_2 | w_1) \cdot p(w_3 | w_1^2) \cdot ... \cdot p(w_n | w_1^{n-1}) \tag{4a}$$

$$p(w_1^n) = p(w_1) \cdot \prod_{k=2}^{n} p(w_k | w_1^{k-1}) \tag{4b}$$

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

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

Even *Eq. 5* is hard to compute for $n > 3$ because we need a large 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.

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

If we use *Eq. 4*, the probability estimation for the unigram model will be:

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

2.2. Bigram language model

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

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

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

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

2.3. Trigram language model

The trigram language model uses a two word history.

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

The probability estimation formula is given by *Eq. 11*.

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

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:

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

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

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

where:

n_k – is the number of occurrence of word w_k;
N – is the total number of words in training corpus;
Nr. (...) – is the number of occurrence of a specific word sequence;

These probabilities calculated using MLE algorithm do not provide useful results. In order to use the probabilities in language modeling experiments, they must be smoothed.

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
- Jelinek - Mercer smoothing or interpolation
- Kneser - Ney smoothing

For the practical experiments we have 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 [8]:

$$P(X)=\frac{r^*}{N}$$
$$where, r^*=(r+1)\cdot\frac{E(N_{r+1})}{E(N_r)}, \tag{15}$$

In *Eq. 15* we have the following notations:

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 [8] is presented in *Eq. 16*.

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

3.2. Back - off smoothing

Back-off smoothing was firstly introduced by Katz [9]. He showed that MLE estimation of probabilities is good enough if the number of occurrences of a word is bigger than a threshold value $K = 6$ [8][9].

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 its 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:

$$\hat{p}(w_3\mid w_1,w_2)=\begin{cases} f(w_3\mid w_1,w_2), & if \quad nr.(w_1,w_2,w_3)\geq K \\ \alpha\cdot Q_T(w_3\mid w_1,w_2) & if \quad 0<nr.(w_1,w_2,w_3)<K \\ & else \quad \beta\cdot\hat{p}(w_3\mid w_2) \end{cases} \tag{17}$$

$$\hat{p}(w_3\mid w_2)=\begin{cases} f(w_3\mid w_2) & if \quad nr.(w_2,w_3)\geq L \\ \alpha\cdot Q_T(w_3\mid w_2) & if \quad 0<nr.(w_2,w_3)<L \\ & else, \quad \beta\cdot f(w_3) \end{cases} \tag{18}$$

4. LANGUAGE MODEL EVALUATION

Language model evaluation can be done in different ways [8][10][11], using:

- random sentence generation
- words reordering in sentences
- perplexity measure
- integration in an existing speech recognition system

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

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

$$H(X) = -\sum_{x \in X} p(x) \cdot \log_2 p(x) \tag{19}$$

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

$$\frac{1}{N} H(w_1^n) = -\frac{1}{N} \sum_{w \in V} p(w_1^n) \cdot \log_2 p(w_1^n) \tag{20}$$

For a real language we should consider infinitely long word sequences ($n \to \infty$):

$$H(L) = \lim_{n \to \infty} -\frac{1}{N} \sum_{w \in V} p(w_1^n) \cdot \log_2 p(w_1^n) \tag{21}$$

Using Shanon – McMilan - Breiman theorem, if the language is stationary and ergodic (which is true for the natural languages), the above formula can be simplified:

$$H(L) = \lim_{n \to \infty} -\frac{1}{N} \log_2 p(w_1^n) \tag{22}$$

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

$$LP = -\frac{1}{N} \log_2 p^*(w_1^n) \tag{23}$$

The perplexity is defined as:

$$PP = 2^{LP} \tag{24}$$

5. TEXT CONDITIONING

Collecting sufficient language model training data for good speech recognition performance in a new domain is often difficult. Collecting text data from the Web is a real alternative. For Romanian language there are also some text corpora which can be used for language modelling, but they had to be normalized. This chapter presents the text normalization tools developed to make these data more suitable for language model training.

Text is unlike speech in a variety of ways. For example, written text also includes numbers, abbreviations, acronyms, punctuation, and other "non-standard words" (NSWs), which are not written in their spoken form. In order to effectively use this text for language modelling these items must be converted to their spoken forms. This process has been referred to as text conditioning or normalization and is often used in text-to-speech systems.

A set of text conditioning tools are available from the Linguistic Data Consortium (LDC). A more systematic approach to the NSW normalization problem referred to here as the NSW tools [12]. These tools perform text normalization using a set of ad-hoc rules, converting numerals to words and expanding abbreviations listed in a table. Also they use models trained on data from several categories. The NSW tools perform well in a variety of domains, unlike the LDC tools which were developed for business news [12].

For Romanian language the text normalization is a hard process also because of the diacritic characters. Our system performs the following basic conditionings:

- it segments text into sentences on the basis of punctuation, marking with <s> and </s> tags the beginning and the end of the sentences and puts only one sentence per line
- stripping most punctuation symbols
- convert numbers into words
- converts the whole text to uppercase
- deletes all empty lines
- eliminates redundant white spaces

6. EXPERIMENTAL RESULTS

The first language model built is based on Romanian Constitution text corpus. This little corpus contains a total number of 9936 words in the train part and in the test part. We have used for testing roughly 10% of the corpus. We count n-grams up to $n = 4$, however the corpus size does not allow us to compute valuable trigram and four-gram probabilities. In *Table 1* you can see the four-grams with the greatest frequency of appearance.

The total number of distinct words in corpus is 1928 grouped in 718 sentences. 963 of them had more than one appearance. We generated a dictionary from the most probable 963 words from the corpus (in fact these words appear more than once in the training corpus), and then we mapped all the other words into an unknown word class. We than generated based on the corpus the unigram, bigram, and trigram language models using Katz back-off smoothing technique. For probability mass reallocation we have used Good – Turing estimator.

Language model evaluation was made using perplexity measure for the three models. The perplexity results of the models created using the 963 word dictionary are presented in *Table 3*.

We have made a second experiment, with a smaller dictionary, containing just the words with appearance greater than 2 (626 words). The perplexity results of our second experiment using the 626 words dictionary are synthesized in *Table 4*.

Table 1. Most frequent four-grams in Romanian Constitution corpus

Word 1	Word 2	Word 3	Word 4	Number of appearance
</S>	<S>	DREPTUL	LA	27
DE	LEGE	</S>	<S>	16
SE	STABILESC	PRIN	LEGE	12
</S>	<S>	DREPTUL	DE	11
PRIN	LEGE	ORGANICA	</S>	10
LEGE	ORGANICA	</S>	<S>	10
</S>	<S>	CAMERA	DEPUTATILOR	9
CONDITIILE	LEGII	</S>	<S>	8
IN	CONDITIILE	LEGII	</S>	8
CAMERA	DEPUTATILOR	SI	SENATUL	8

Table 2. The most probable 15 words from Romanian Constitution

Word	Appearance	Word	Appearance	Word	Appearance
</S>	718	A	195	AL	71
<S>	718	LA	151	DREPTUL	70
DE	470	SE	128	PRIN	69
SI	405	SAU	116	PENTRU	68
IN	287	ESTE	82	CU	66

Table 3. Perplexity results for Romanian Constitution corpus using a 963-word dictionary

Model	Perplexity
Unigram	559.74
Bigram	397.37
Trigram	419.52

Table 4. Perplexity results for Romanian Constitution corpus using a 626-word dictionary

Model	Perplexity
Unigram	509.64
Bigram	332.24
Trigram	419.23

7. CONCLUSIONS AND FUTURE WORKS

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

We can see from the results that if the number of words in dictionary 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. This threshold can be experimentally settled. This technique is called in the literature n-gram pruning [4][5].

We can see from *Table 3* and *4* that, the words with single occurrence does not improve the model quality, they rise perplexity and they should be eliminated from vocabulary. We have draw the same conclusion in our previous work [3] based on the Susanne English language corpus.

The used n-gram model dimension should be selected considering the amount of training data available (i.e. for the presented Romanian Constitution corpus we have to use $n \leq 2$).

As future work, we want to improve the text-conditioning tool with diacritic restoration feature and automatic abbreviation and acronym expansion.

We also want to try to generate language models based upon a much bigger training corpus of Romanian journal articles and to implement the state of the art Kneser - Ney smoothing algorithm [2][7][8][11].

To compare our language modeling tools with others we shall try to use open source language modeling toolkits (e.g, CMU-LM, SRILM) on the same corpora.

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. DOMOKOS J., TODEREAN G., BUZA O., *Statistical Language modeling on Susanne corpus*, IEEE International Conference COMMUNICATIONS 2008, Proceedings, pp. 69–72, 2008.
4. GAO J., ZHANG M ., *Improving Language Model Size Reduction using Better Pruning Criteria*, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 2002
5. GOODMAN J., GAO J., *Language model size reduction by pruning and clustering*, ICSLP-2000, International Conference on Spoken Language Processing, Beijing, 2000.
6. HUANG X., ACERO A., HON H., *Spoken Language Processing. A Guide to Theory, Algorithm & System Development*, Prentice Hall, 2001.
7. JELINEK F., *Statistical Methods for speech recognition*, The MIT Press, 2001.
8. JURAFSKY D., MARTIN J. H., *Speech and language processing. An introduction to Natural language Processing. Computational Linguistics and Speech Recognition*, Prentice Hall, 2000.

9. KATZ S. M., *Estimation of probabilities from sparse data for other language component of a speech recognizer,* IEEE transactions on Acoustics, Speech and Signal Processing, 35(3):400–401, 1987.
10. MANNING C., HEINRICH S., Foundations of statistical language processing, The MIT Press, 1999.
11. ROSENFELD R., *Two decades of statistical language modeling: where do we go from here?* Proceedings of the IEEE, Volume 88, pp. 1270–1278, 2000.
12. SCHWARM S.; OSTENDORF M., *Text normalization with varied data sources for conversational speech language modeling,* IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '02, Volume 1, pp. 789–792, 2002.
13. 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.
14. PAUL D., B., BAKER J., M., *The design for the wall street journal-based CSR corpus,* Workshop on Speech and Natural Language, Proceedings, pp. 357–362, 1992.