* RUSH!

Request Date: 18-JUL-2013
Expiration Date: 29-JUL-2013
ILL Number:

SHARE

Printed Date: 18-JUL-2013

TGQ or OCLC #:

ILL Number: 5906713
Rota Call Numbers: SRLF:TK7882.S65 I16 1993;SRLF:TK7882.S65 I16 1993 ;SRLF:TK7882.S65 I16 1993 ;SRLF:TK7882.S65 I16 1
Format: Article Printed
Title: ICASSP-93 : 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, April 27-30, 1993, Minneapolis Convention Center, Minneapolis, Minnesota /
Article Author: Raymond Lau, Ronald Rosenfeld, and Salim Roukos
Article Title: Trigger-based Language Models Using Maximum Likelihood Estimation of Exponential Distributions
Part Pub. Date: 1993
Pages: ? @ II-45
Pub. Place: New York, N.Y. : Institute of Electrical and Electronics Engineers ; Piscataway, N.J. Additional copies may be ordered from IEEE Order Dept., c1993.
Borrower: ULA0

TGQ or OCLC #: 5906709
ID: ULA0

ISBN/ISSN: 9780780309470

Address: Carnegie Mellon University- Sorrells E&S Library/5000 Forbes Avenue/Wean Hall Room 1001/Pittsburgh PA US 15213

Patron Name: Loring, Patricia
Patron e-mail:
Service Level: Rush - Extended Search
Delivery Method: IDS #187A01/Library Mail
Request Notes: ARIEL 128.2.21.4/128.2.114.135 (maxCost: $35.00) / PMC is a SHARES member. FAX/ARIEL:(412) 268-7385 EMAIL:es82@andrew.cmu.edu OCLC Req. Ex. Affiliations: PALCI recip. copy pilot OCLC Req. Ex. Source: ILLiad
Need By:
Verification Source: <TN:366762><ODYSSEY:128.2.20.146/ENS> OCLC
Supplier Reference:
Supplier Reference: ILLNUM:107099931
Local request number: ILLNUM:107099931
Owned By: SRLF

Service Type: Copy non returnable
Max Cost: USD35
Payment Type: IFM

Copyright Info: US:CCG

Requester Symbol: OCLC:PMC

Return To: SRLF
305 DeNeve Drive
Los Angeles, CA,
90095-1388

TRIGGER-BASED LANGUAGE MODELS: A MAXIMUM ENTROPY APPROACH

Raymond Lau
Massachusetts Institute of Technology, Cambridge, MA 02139
Ronald Rosenfeld
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213
Salim Roukos *†
IBM T. J. Watson Research Center, P.O.Box 704, Yorktown Heights, NY 10598

ABSTRACT

We describe our ongoing efforts at adaptive statistical language modeling. To extract information from the document history, we use *trigger pairs* as the basic information bearing elements. To combine statistical evidence from multiple triggers, we use the principle of Maximum Entropy (ME). To combine the trigger-based model with the static model, we absorb the latter into the ME formalism.

Given consistent statistical evidence, a unique ME solution is guaranteed to exist, and an iterative algorithm exists which is guaranteed to converge to it. Among the advantages of this approach are its simplicity, its generality, and its incremental nature. Among its disadvantages are its computational requirements. We report our current results and discuss possible improvements.

1. STATE OF THE ART

Until recently, the most successful language model (given enough training data) was the trigram [1], where the probability of a word is estimated based solely on the two words preceding it. The trigram model is simple yet powerful [2]. However, since it does not use anything but the very immediate history, it is incapable of adapting to the style or topic of the document, and is therefore considered a *static* model.

In contrast, a dynamic or *adaptive* model is one that changes its estimates as a result of "seeing" some of the text. An adaptive model may, for example, rely on the history of the current document in estimating the probability of a word. Adaptive models are superior to static ones in that they are able to improve their performance after seeing some of the data. This is particularly useful in two situations. First, when a large heterogeneous language source is composed of smaller, more homogeneous segments, such as newspaper articles. An adaptive model trained on the heterogeneous source will be able to hone in on the particular "sublanguage" used in each of the articles. Secondly, when a model trained on data from one domain is used in another domain. Again, an adaptive model will be able to adjust to the new language, thus improving its performance.

*Most of this work was done when Ray Lau and Ron Rosenfeld were summer visitors at IBM T.J. Watson Research Center.

†A fourth contributor to this work, Xuedong Huang of Carnegie-Mellon, could not be put on the authors' list due to ICASSP's technical restrictions.

The most successful adaptive LM to date is described in [3]. A cache of the last few hundred words is maintained, and is used to derive a "cache trigram". The latter is then interpolated with the static trigram. This results in a 23% reduction in perplexity, and a 5%–24% reduction in the error rate of a speech recognizer.

In what follows, we describe our efforts at improving our adaptive statistical language models by capitalizing on the information present in the document history.

2. TRIGGER-BASED MODELING

To extract information from the document history, we propose the idea of *a trigger pair as the basic information bearing element*. If a word sequence A is significantly correlated with another word sequence B, then $(A \rightarrow B)$ is considered a "trigger pair", with A being the *trigger* and B the *triggered sequence*. When A occurs in the document, it triggers B, causing its probability estimate to change.

To build a successful model based on the above principles, the following issues had to be addressed:

Filtering of the trigger pairs. Even if we restrict our attention to trigger pairs where A and B are both single words, the number of such pairs is too large. Let V be the size of the vocabulary. Note that, unlike in a bigram model, where the number of different consecutive word pairs is much less than V^2, the number of word pairs where both words occurred in the same document is a significant fraction of V^2.

Combining evidence from multiple triggers. This is a special case of the general problem of combining evidence from several knowledge sources.

Combining the trigger model with the static model. A simple-minded approach is to linearly interpolate the two. However, we sought a better solution: one which would preserve, rather than average, the advantages of both.

In the following sections, we discuss these and other issues, and our current solutions to them.

0-7803-0946-4/93 $3.00 © 1993 IEEE

3. FILTERING THE TRIGGER PAIRS

Let h denote the "history" of the document (the part of the text already seen). The goal of the language model is to estimate probabilities of the form $P(w|h)$ for any word w in the vocabulary. Let W be any word sequence. Define the events W and W_0 as follows:

W : {W occurs immediately next in the document. }
W_0 : $\{W \in h\}$

A natural measure of the information provided by A_0 on B is the average mutual information between the two:

$$I(A_0{:}B) = P(A_0, B)\log\frac{P(B|A_0)}{P(B)} + P(A_0, \overline{B})\log\frac{P(\overline{B}|A_0)}{P(\overline{B})} + P(\overline{A_0}, B)\log\frac{P(B|\overline{A_0})}{P(B)} + P(\overline{A_0}, \overline{B})\log\frac{P(\overline{B}|\overline{A_0})}{P(\overline{B})} \quad (1)$$

Should mutual information be our figure of merit in selecting the most promising trigger pairs? Consider the sentence:

> "The district attorney's office launched an investigation into loans made by several well connected banks."

A trigger based model may use "DISTRICT ATTORNEY" to trigger "INVESTIGATION", raising its probability above the default value for the rest of the document. But when "INVESTIGATION" actually occurs, it is preceded by "LAUNCHED AN", which allows a static trigram language model to predict it with a much higher probability.

Triggers are to be used as an additional component to the static model. Therefore, trigger pairs are only useful to the extent that the information they provide complements the static model. Furthermore, trigger pairs affect each others' usefulness. The utility of the trigger pair $A_1 \rightarrow B$ is diminished by the presence of the pair $A_2 \rightarrow B$. Finally, the utility of a trigger pair depends on the way it will be used in the model.

4. COMBINING KNOWLEDGE SOURCES

4.1. The Maximum Entropy Approach

Using several different probability estimates to arrive at one combined estimate is a general problem that arises in many tasks. We use the maximum entropy (ME) principle ([4, 5]), which can be summarized as follows:

1. Reformulate the different estimates as constraints on the expectation of various functions, to be satisfied by the target (combined) estimate.
2. Among all probability distributions that satisfy these constraints, choose the one that has the highest entropy.

More specifically, for estimating a probability function $P(\mathbf{x})$, each constraint i is associated with a *constraint function* $f_i(\mathbf{x})$ and a *desired expectation* c_i. The constraint is then written as:

$$E_P f_i \stackrel{\text{def}}{=} \sum_{\mathbf{x}} P(\mathbf{x}) f_i(\mathbf{x}) = c_i \,. \quad (2)$$

Given consistent constraints, a unique ME solutions is guaranteed to exist, and to be of the form:

$$P(\mathbf{x}) = \prod_i \mu_i^{f_i(\mathbf{x})} \,, \quad (3)$$

where the μ_i's are some unknown constants, to be found. Probability functions of the form (3) are called *log-linear*, and the family of functions defined by holding the f_i's fixed and varying the μ_i's is called *an exponential family*.

To search the exponential family defined by (3) for the μ_i's that will make $P(\mathbf{x})$ satisfy all the constraints, an iterative algorithm, "Generalized Iterative Scaling", exists, which is guaranteed to converge to the solution ([6]).

4.2. Formulating Triggers as Constraints

To reformulate a trigger pair $A \rightarrow B$ as a constraint, define the constraint function $f_{A\rightarrow B}$ as:

$$f_{A\rightarrow B}(h, w) = \begin{cases} 1 & \text{if } A \in h,\ w = B \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Set $c_{A\rightarrow B}$ to $\tilde{E}[f_{A\rightarrow B}]$, the *empirical expectation* of $f_{A\rightarrow B}$ (ie its expectation in the training data). Now impose on the desired probability estimate $P(h, w)$ the constraint:

$$E_P[f_{A\rightarrow B}] = \tilde{E}[f_{A\rightarrow B}] \quad (5)$$

4.3. Estimating LM Conditionals: The ML/ME Solution

Generalized Iterative Scaling can be used to find the ME estimate of a simple (non-conditional) probability distribution over some event space. But in language modeling, we often need to estimate conditional probabilities of the form $P(w|h)$. How should this be done?

One simple way is to estimate the joint, $P(h, w)$, from which the conditional, $P(w|h)$, can be readily derived. We have tried this approach without success. The likely reason is that the event space $\{(h, w)\}$ is of size $O(V^{L+1})$, where V is the vocabulary size and L is the history length. The joint distribution is over a huge space and we were subjecting it to very few (comparatively) bigram trigger constraints. Better results were obtained by estimating a conditional model $P(w|h)$ given by the exponential family corresponding to the trigger constraints, and using the method of [7] to constrain a bigram

conditional model to match the probability $P(w|h)$ derived from the trigger constraints.

A better method, which incorprates all constrains simultaneously, was proposed by colleagues at IBM [8]. Let $P(h,w)$ be the desired probability estimate, and let $\tilde{P}(h,w)$ be the empirical distribution of the training data. Let $f_i(h,w)$ be any constraint function, and let c_i be its desired expectation. Equation 5 can be rewritten as:

$$\sum_h P(h) \cdot \sum_w P(w|h) \cdot f_i(h,w) = c_i \qquad (6)$$

We now modify the constraint to be:

$$\sum_h \tilde{P}(h) \cdot \sum_w P(w|h) \cdot f_i(h,w) = c_i \qquad (7)$$

One possible interpretation of this modification is as follows. Instead of constraining the expectation of $f_i(h,w)$ with regard to $P(h,w)$, we constrain its expectation with regard to a different probability distribution, say $Q(h,w)$, whose conditional $Q(w|h)$ is the same as that of P, but whose marginal $Q(h)$ is the same as that of $\tilde{P}$. To better understand the effect of this change, define H as the set of all possible histories h, and define H_{f_i} as the partition of H induced by f_i. Then the modification is equivalent to assuming that, for every constraint f_i, $P(H_{f_i}) = \tilde{P}(H_{f_i})$. Since typically H_{f_i} is a very small set, the assumption is reasonable.

The unique ME solution that satisfies equations like (7) or (6) can be shown to also be the Maximum Likelihood (ML) solution, namely that function which, among the exponential family defined by the constraints, has the maximum likelihood of generating the data. The identity of the ML and ME solutions, apart from being aesthetically pleasing, is extremely useful when estimating the conditional $P(w|h)$. It means that hillclimbing methods can be used in conjunction with Generalized Iterative Scaling to speed up the search. Since the likelihood objective function is convex, hillclimbing will not get stuck in local minima.

4.4. Pros and Cons

The ME principle and the Generalized Iterative Scaling algorithm have several important advantages:

1. The ME principle is simple and intuitively appealing. It imposes all of the constituent constraints, but assumes nothing else. For the special case of constraints derived from marginal probabilities, it is equivalent to assuming a lack of higher-order interactions [9].
2. ME is extremely general. Any probability estimate of any subset of the event space can be used, including estimates that were not derived from the data or that are inconsistent with it. Many other knowledge sources can be incorporated, such as distance-dependent correlations and complicated higher-order effects. Note that constraints need not be independent of nor uncorrelated with each other.
3. The information captured by existing language models can be absorbed into the ME model. Later on we will show how this is done for the conventional N-gram model, and for the cache model of [3].
4. Generalized Iterative Scaling lends itself to incremental adaptation. New constraints can be added at any time. Old constraints can be maintained or else allowed to relax.
5. A unique ME solution is guaranteed to exist for consistent constraints. The Generalized Iterative Scaling algorithm is guaranteed to converge to it.

This approach also has the following weaknesses:

1. Generalized Iterative Scaling is computationally very expensive. When the complete system is trained on the entire 50M words of Wall Street Journal data, it is expected to require many thousands of MIPS-hours to run to completion.
2. While the algorithm is guaranteed to converge, we do not have a theoretical bound on its convergence rate.
3. It is sometimes useful to impose constraints that are not satisfied by the training data. For example, we may choose to use Good-Turing discounting [10], or else the constraints may be derived from other data, or be externally imposed. Under these circumstances, the constraints may no longer be consistent, and the theoretical results guaranteeing existence, uniqueness and convergence may not hold.

5. COMBINING WITH THE STATIC MODEL

We combined the trigger based model with the currently best static model, the N-Gram, by reformulating the latter to fit into the ME paradigm. The usual unigram, bigram and trigram ML estimates were replaced by unigram, bigram and trigram constraints conveying the same information. Specifically, the constraint function for the unigram w_1 is:

$$f_{w_1}(h,w) = \begin{cases} 1 & \text{if } w = w1 \\ 0 & \text{otherwise} \end{cases} \qquad (8)$$

and its associated constraint is:

$$\sum_h \tilde{P}(h) \sum_w P(w|h) f_{w_1}(h,w) = \tilde{E} f_{w_1}(h,w). \qquad (9)$$

Similarly, the constraint function for the bigram w_1, w_2 is

$$f_{w_1,w_2}(h,w) = \begin{cases} 1 & \text{if } h \text{ ends in } w_1 \text{ and } w = w2 \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

and its associated constraint is

$$\sum_h \tilde{P}(h) \sum_w P(w|h) f_{w_1,w_2}(h,w) = \tilde{E} f_{w_1,w_2}(h,w). \quad (11)$$

and similarly for higher-order ngrams.

The computational bottleneck of the Generalized Iterative Scaling algorithm is in constraints which, for typical histories h, are non-zero for a large number of w's. This means that bigram constraints are more expensive than trigram constraints. Implicit computation can be used for unigram constraints. Therefore, the time cost of bigram and trigger constraints dominates the total time cost of the algorithm.

6. INCORPORATING THE CACHE MODEL

It seems that the power of the cache model, described in section 1, comes from the "bursty" nature of language. Namely, infrequent words tend to occur in "bursts", and once a word occurred in a document, its probability of recurrence is significantly elevated.

Of course, this phenomena can be captured by a trigger pair of the form $A \rightarrow A$, which we call a "self trigger". We have done exactly that in [11]. We found that self triggers are responsible for a disproportionately large part of the reduction in perplexity. Furthermore, self triggers proved particularly robust: when tested in new domains, they maintained the correlations found in the training data better than the "regular" triggers did.

Thus self triggers are particularly important, and should be modeled separately and in more detail. For example, the trigger model as described above does not distinguish between one or more occurrences of a given word in the history, whereas the cache model does.

We plan to model self triggers in more detail. We will consider explicit modeling of frequency of occurrence, distance from last occurrence, and other factors. All of these aspects can easily be formulated as constraints and incorporated into the ME formalism.

7. RESULTS

The model described above was trained on 5 million words of Wall Street Journal text. It used some 40,000 unigram constraints, 200,000 bigram constraints, 200,000 trigram constraints, and 60,000 trigger constraints. It took about 500 MIPS-hours per iteration. After 13 iterations, it produced a language model whose perplexity was 12% lower than that of a conventional trigram, as measured on independent data.

The trigger constraints used in this run were selected very crudely, and their number was not optimized. We believe much more improvement can be achieved. Special modeling of self triggers has not been implemented yet. Similarly, we expect it to yield further improvement. We will report our latest results at the conference.

8. ACKNOWLEDGEMENTS

We are grateful to Peter Brown, Stephen Della Pietra, Vincent Della Pietra and Bob Mercer for many suggestions and discussions.

Research by Ron Rosenfeld and Xuedong Huang was sponsored in part by the Defense Advanced Research Project Agency and monitored by the Space and Naval Warfare Systems Command under contract N00039-91-C-0158, ARPA order 7239. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

1. Bahl, L., Jelinek, F., Mercer, R.L., "A Statistical Approach to Continuous Speech Recognition," *IEEE Trans. on PAMI,* 1983.
2. Jelinek, F., "Up From Trigrams!" Eurospeech 1991.
3. Jelinek, F., Merialdo, B., Roukos, S., and Strauss, M., "A Dynamic Language Model for Speech Recognition." *Proceedings of the Speech and Natural Language DARPA Workshop,* pp.293–295, Feb. 1991.
4. Jaines, E. T., "Information Theory and Statistical Mechanics." *Phys. Rev.* **106**, pp. 620–630, 1957.
5. Kullback, S., *Information Theory in Statistics.* Wiley, New York, 1959.
6. Darroch, J.N. and Ratcliff, D., "Generalized Iterative Scaling for Log-Linear Models", *The Annals of Mathematical Statistics,* Vol. 43, pp 1470–1480, 1972.
7. Della Pietra,S., Della Pietra, V., Mercer, R. L., Roukos, S., "Adaptive Language Modeling Using Minimum Discriminant Estimation," *Proceedings of ICASSP-92,* pp. I-633-636, San Francisco, March 1992.
8. Brown, P., Della Pietra, S., Della Pietra, V., Mercer, R., Nadas, A., and Roukos, S., "Maximum Entropy Methods and Their Applications to Maximum Likelihood Parameter Estimation of Conditional Exponential Models," *A forthcoming IBM technical report.*
9. Good, I. J., "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables." *Annals of Mathematical Statistics,* Vol. 34, pp. 911–934, 1963.
10. Good, I. J., "The Population Frequencies of Species and the Estimation of Population Parameters." *Biometrika,* Vol. 40, no. 3, 4, pp. 237–264, 1953.
11. Rosenfeld, R., and Huang, X. D., "Improvements in Stochastic Language Modeling." *Proceedings of the Speech and Natural Language DARPA Workshop,* Feb. 1992.

ICASSP-93

Speech Processing

Volume II of V

1993 IEEE International Conference on Acoustics, Speech, and Signal Processing

April 27-30, 1993
Minneapolis Convention Center
Minneapolis, Minnesota, USA

93CH3252-4

Sponsored by
THE INSTITUTE OF ELECTRICAL AND
ELECTRONICS ENGINEERS
SIGNAL PROCESSING SOCIETY