Probabilistic logic with minimum perplexity: Application to language modeling

Djamel Bouchaffra

Department of Computer Science, Oakland University, 131 Dodge Hall, Rochester, MI 48309, USA

Abstract

Any statistical model based on training encounters sparse configurations. These data are those that have not been encountered (or seen) during the training phase. This inherent problem is a big challenge to many scientific communities. The statistical estimation of rare events is usually performed through the maximum likelihood (ML) criterion. However, it is well-known that the ML estimator is sensitive to extreme values that is therefore non-reliable. To answer this challenge, we propose a novel approach based on probabilistic logic (PL) and the minimal perplexity criterion. In our approach, configurations are considered as probabilistic events such as predicates related through logical connectors. Our method was applied to estimate word trigram probability values from a corpus. Experimental results conducted on several test sets show that the PL method with minimal perplexity has outperformed both the "Absolute Discounting", and the "Good-Turing Discounting" techniques.
© 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

Keywords: Word trigrams; Probabilistic logic; Statistical language model; Maximum likelihood estimation; Sparseness problem; Minimum perplexity; Entropy maximization

1. Introduction

Techniques for statistical processing of natural language are present in many areas such as information retrieval, language modeling, parts-of-speech tagging and word sense disambiguation. All these techniques face the same challenging problem known as: *data sparseness*. In the case of language modeling, this inherent problem arises while collecting frequency statistics on observations from a database of finite size. Statistical methods compute in the first phase relative frequencies derived from a maximum likelihood (ML) estimation of configurations (e.g., *word*$_1$ followed by *word*$_2$) from a training corpus. In the second phase, the statistical methods attempt to evaluate alternative interpretations of new textual or speech samples. *The problem of data sparseness arises during this second phase*, *when we are*

dealing with configurations that have never (*or rarely*) *been encountered in the training corpus.* Therefore, the estimation of probabilities assigned to these configurations based on observed frequencies becomes unreliable. From an estimation theory standpoint, the variances of these observation frequencies are large. Therefore the frequencies computed are inaccurate estimators of the probabilities assigned to these configurations.

To overcome this challenging problem due to the traditional ML estimation, a number of different approaches have been proposed in the literature. The various discounting methods [1–4], word clustering [5], link grammars [6], sentence mixtures, decision trees, caching [7], maximum entropy models [8], latent semantic analysis [9], neural network models [10] and web-based improved trigrams [11] are examples of these techniques that deal with the sparse data problem in language modeling. Traditionally, the dominant motivation for language modeling emerged from

E-mail address: dbouchaffra@ieee.org.

speech recognition [5,12–16]. However, statistical language models have recently become more widely used in many other application areas, such as machine translation, information retrieval, handwriting recognition [17–19], spelling correction, information extraction and bioinformatics.

We propose in this paper a novel technique based on probabilistic logic (PL) that solves the sparseness problem in an efficient manner. Words are viewed as predicates that interact according to rules of logic within a corpus. The motivation of our approach is threefold: (i) it is original since it is based on a "logical" relationship between *n*-grams, (ii) the information about a trigram probability value is constructed from the information provided by all consecutive unigrams and bigrams involved in this trigram, and (iii) it is general, it can easily be applied to other types of configurations. We show that through the context of PL, it is possible *to directly minimize the perplexity* assigned to a testing corpus and derive an optimal statistical language model.

This paper is organized as follows: Section 2 introduces the perplexity measure for models comparison. Section 3 underscores the background by emphasizing two discounting techniques: The "Absolute Discounting" and the "Good-Turing Discounting". We introduce the PL paradigm for sparse data modeling in Section 4. Experiments are laid out in Section 5 and the conclusion and future work are the objects of Section 6.

2. Model evaluation

In order to compare our approach with the discounting methods, we need to define a model evaluator measure. This measure quantifies the "quality" of the language model. The cross-entropy function enables us to measure in some scale the power of prediction of the statistical language model proposed. The third-order cross-entropy of a set of *n* random variables $(w, t)_{1,n} = (w_1, w_2, \ldots, w_n)$, (a word path representing the test set) where the correct model is $Pr(w_{1,n})$ but the probabilities are estimated using the model $Pr_M(w_{1,n})$, is given by

$$\mathcal{H}(w_{1,n}, Pr_M) = -\sum_{w_{1,n}} Pr(w_{1,n}) \times \log Pr_M(w_{1,n}),$$
(1)

where "log" represents the base 2 logarithm.

Using some assumptions and some classical transformations [8], we can estimate the cross-entropy as

$$\mathcal{H}(w_{1,n}, Pr_M)$$
$$\approx -\frac{1}{n} \left[\sum_{i=1}^{i=n} \log Pr_M(w_i \mid w_{i-1} \wedge w_{i-2}) \right].$$
(2)

The perplexity defined as $PP_p(T) = 2^{\mathcal{H}}$ (where $p = Pr_M(w_{1,n})$ and $T = (w, t)_{1,n}$) is the function used to validate statistical language models. Although this measure have been receiving criticism and some new measures are

being proposed [20], *it is still the standard measure used by the language modeling community*. This expression of the perplexity is used in the experiments section to compare the performance of our approach with two discounting techniques.

3. Background in statistical language modeling

Many statistical techniques have been proposed by the language modeling community in order to solve the sparse data problem. Brown et al. [5] compute words co-occurrence based on their distributions. They have clustered words with respect to their co-occurrence distributions. Words with similar co-occurrence distributions were assigned a same word class. The probability of any pair of words is transformed into the probability of the pair of classes they belong to. Jelinek proposed the "linear interpolation" by smoothing the specific probabilities with less sparse probabilities [15,16]. Katz proposed the "discounting technique" using the Good-Turing formula [4,21]. Besides, in case of estimation of lexical associations on the basis of word forms, a fairly large amount of training data is required. The reason is that all inflected forms of a word are to be considered as different words. A better result using the same amount of data can be achieved by analyzing *lemmata* (inflectional differences are ignored) or even semantic classes. However, even with this refinement, the sparseness problem persists. The models used in the literature circumvent this problem by assuming that there exists a functional relationship between the probability estimate for a previously unseen co-occurrence and the probability estimate for the word contained in the co-occurrence [14], while other techniques use class-based and similarity-based models to overcome the problem [13]. The modified Kneser–Ney algorithm [22] which is an extension of Kneser and Ney's algorithm introduced in 1995 [23] which is itself an extension of absolute discounting has shown promising results. Like absolute discounting, the Kneser–Ney approach computes the probability of a word following a particular context by calculating the raw probability of the word following the context and subtracting a discounting value. Jaakola [7] presented a general framework for a discriminative estimation based on the maximum entropy principle and its extensions when the labels in the training set are uncertain or incomplete. Bellegarda [9] used latent semantic analysis that uncovers the salient semantic relationships between words and documents in a given corpus. Bengio et al. [10] propose a technique that "fights" the curse of dimensionality. It consists of learning a distributed representation for words which allow each training sentence to inform the model about an exponential number of semantically neighboring sentences. Finally, Xu et al. [24] use random forests that incorporate syntactical information to predict the next word based on words already seen before.

For performance comparison purposes, we describe in this paper three discounting techniques which are: the

"Good-Turing Discounting", and the "Absolute Discounting". The perplexity computed on several test sets using these two techniques will be compared with the perplexity of the "PL" approach. But first, we need to briefly introduce these three techniques in this section.

3.1. Back-off with discounting

Traditional approaches to smoothing probability estimates that cope with sparse data problem consist of using some sort of back-off or interpolated estimator [25]. The classical model mostly used is the "Discounted Back-Off n-grams". This model is defined as

$$Pr(w_i \mid w_{i-n+1}, \ldots, w_{i-1}) \approx \begin{cases} \hat{P}r(w_i \mid w_{i-n+1}, \ldots, w_{i-1}) & \text{if } \#(w_{i-n+1}, \ldots, w_i) > 0, \\ \beta(w_{i-n+1}, \ldots, w_{i-1}) \times Pr(w_i \mid w_{i-n+2}, \ldots, w_{i-1}) & \text{otherwise,} \end{cases} \quad (3)$$

where

$$\hat{P}r(w_i \mid w_{i-n+1}, \ldots, w_{i-1})$$
$$= \frac{discount \; \#(w_{i-n+1}, \ldots, w_i)}{\#(w_{i-n+1}, \ldots, w_{i-1})} \quad (4)$$

and $\beta(w_{i-n+1}, \ldots, w_{i-1})$ is a normalizing term expressed as

$$\beta(w_{i-n+1}, \ldots, w_{i-1})$$
$$= \frac{1 - \sum_{x \in (w_{i-n+1}, \ldots, w_{i-1}x)} \hat{P}r(x \mid w_{i-n+1, \ldots, w_{i-1}})}{1 - \sum_{x \in (w_{i-n+1}, \ldots, w_{i-1}x)} \hat{P}r(x \mid w_{i-n+2, \ldots, w_{i-1}})}. \quad (5)$$

Therefore, if n_r represents the number of words occurring r times in the corpus, then the "Good-Turing" (GT) discounting can be estimated as

$$\hat{P}r(w_i \mid w_{i-n+1}, \ldots, w_{i-1}) = \frac{GT_{\#(w_{i-n+1}, \ldots, w_i)}}{\#(w_{i-n+1}, \ldots, w_{i-1})}, \quad (6)$$

where $GT_r = (r+1)[(n_{r+1})/n_r]$.

However, in the "Absolute Discounting", the frequency of a word is subtracted by a constant K. This constant is equal $n_1/(n_1 + 2n_2)$. Finally, the "Absolute Discounting" estimation is expressed as

$$\hat{P}r(w_i \mid w_{i-n+1}, \ldots, w_{i-1})$$
$$= \frac{\#(w_{i-n+1}, \ldots, w_i) - K}{\#(w_{i-n+1}, \ldots, w_{i-1})}. \quad (7)$$

In the case of trigrams ($n = 3$) and $r = 1$, the "Good-Turing" estimation becomes

$$\hat{P}r(w_3 \mid w_1, w_2) = \frac{\#(w_1, w_2, w_3)}{\#(w_1, w_2)} \times \frac{2n_2}{n_1}, \quad (8)$$

and the "Absolute Discounting" becomes

$$\hat{P}r(w_3 \mid w_1, w_2)$$
$$= \frac{\#(w_1, w_2, w_3) - (n_1/(n_1 + 2n_2))}{\#(w_1, w_2)}. \quad (9)$$

4. Solving the sparseness problem through probabilistic logic

Traditional methods dealing with the sparseness problem assumes that the probability estimate for a previously unseen co-occurrence is a function of the probability estimates for the word in the co-occurrence. For example, in the trigram models that we study in this paper, the probability $Pr(w_i \mid w_{i-1} \wedge w_{i-2})$ of a conditioned word w_i that has never been encountered during training following the conditioning words (w_{i-1}, w_{i-2}) is computed from the probability of the word w_i. This technique is based on an independence assumption on the co-occurrence of the pairs (w_{i-1}, w_{i-2}) and w_i.

To circumvent this independence assumption, we propose a novel approach that explores the PL paradigm [26,27]. We make an analogy between word configurations w_i and logical predicates. *The predicate W_i is true if the word w_i is present in the training corpus and false otherwise.* Our approach uses information captured by all n-grams that *logically impact* the estimation of $Pr(w_i \mid w_{i-1} \wedge w_{i-2})$. Our method involves a set $\mathcal{M}$ constituted of three unigrams, two bigrams and one trigram. This set is

$$\mathcal{M} = \{w_i, w_{i-1}, w_{i-2}, (w_i \mid w_{i-1}), (w_{i-1} \mid w_{i-2}),$$
$$(w_i \mid w_{i-1} \wedge w_{i-2})\}. \quad (10)$$

This set $\mathcal{M}$ is chosen because all elements of this set are consecutive and "logically" reconstruct $Pr(w_i \mid w_{i-1} \wedge w_{i-2})$. In fact, PL paradigm enables us to express a linear relationship between probabilities of all elements contained in the set $\mathcal{M}$.

Since the estimated quantity $\hat{P}r(w_i \mid w_{i-1} \wedge w_{i-2})$ is not reliable, then it should be treated separately. The truth value of this quantity is built from the truth values of all unigrams and bigrams in $\mathcal{M}$ that logically depend on it. Therefore the logical interaction we are looking for is the following:

$$Pr(w_i \mid w_{i-1} \wedge w_{i-2})$$
$$\simeq \log(\lambda_1^h) \times \hat{P}r(w_i)$$
$$+ \log(\lambda_2^h) \times \hat{P}r(w_{i-1}) + \log(\lambda_3^h)$$
$$\times \hat{P}r(w_{i-2}) + \log(\lambda_4^h) \times \hat{P}r(w_i \mid w_{i-1}) + \log(\lambda_5^h)$$
$$\times \hat{P}r(w_{i-1} \mid w_{i-2}) + K^h. \quad (11)$$

For obtaining a more general model, we added a constant K^h (that depends on the history h) in this linear combination. Our goal is to compute an "optimal" solution for the coefficients λ_i^h $(i = 1, \ldots, 5)$ and K^h. The motivation behind Eq. (11) is that the trigram $(w_i \mid w_{i-1} \wedge w_{i-2})$ might be (probabilistically) reconstructed from all unigrams and bigrams that are involved in it. The logarithm assigned to the coefficients are introduced only to put these coefficients into the interval [0..1]. Each probability involved in this decomposition is computed from a training corpus using the ML estimation.

We will show in the next section how we use PL tools to determine an estimation of: $Pr(w_i \mid w_{i-1} \wedge w_{i-2})$ by computing uniquely the parameters λ_i^h $(i = 1, \ldots, 5)$ and the constant K^h.

4.1. A bounded set of solutions

In order to compute this logical "reconstruction", we need to pass from the conditional events to the joint events. We first define a linear relationship that involves joint probabilities, and then we express this relationship into conditional probability terms. Table 1 depicts the logical truth-table assigned to this set of predicates assigned to words. This table has $8 = 2^3$ columns, where 3 corresponds to the number of base predicates. As outlined previously, a predicate W_i is either true or false. It is true if the word w_i (at position i) is observed (or encountered) in a corpus and false whenever it is absent. Each time, we obtain two worlds (or interpretations) $\mathcal{V}_i$, $(i = 1, 2)$ that are assigned to one predicate, and the values of the predicate are opposite in these two worlds [28]. We define a probability distribution over the sets of possible worlds (the columns of Table 1) associated with the different sets of possible truth values of predicates. This probability specifies for each set of possible worlds what is the probability $Pr(\mathcal{V}_i)$ that the actual world $\mathcal{V}_a$ is contained in $\mathcal{V}_i$. If the actual world $\mathcal{V}_a$ is for example the fourth column $[0, 1, 1, 0, 1, 0]^T$ of Table 1, and if the trigram of interest is $\langle w_{i-2} \, w_{i-1} \, w_i \rangle = \langle$ate an apple$\rangle$, then this means that $\langle$apple$\rangle$ is absent from the corpus at position i, $\langle$an$\rangle$ is present at position $(i-1)$, $\langle$ate$\rangle$ is present in the corpus at position $(i-2)$, $\langle$an apple$\rangle$ is absent, $\langle$ate an$\rangle$ is present and $\langle$ate an apple$\rangle$ is absent. Since we do not know the actual world which depends on the corpus contents, therefore there is an uncertainty about it expressed by a probability measure. The probabilities $Pr(\mathcal{V}_i) = p_i$ are subject to the constraint $\sum_i Pr(\mathcal{V}_i) = 1$ since the set of possible worlds are mutually exhaustive and exclusive (the actual world is one of the column of Table 1).

A mapping from the space of possible worlds to the space of predicates is therefore built. *A probability of any predicate (π_j) is the sum of the probabilities of the worlds $(\mathcal{V}_i)$ where the predicate is true.* The linear

Table 1
The consistent matrix C assigned to the set of predicates

$$C = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{12}$$

equation mapping possible world probabilities and predicate probabilities is

$$C \times P = \Pi, \tag{13}$$

where the consistent matrix C whose columns are the possible worlds $(\mathcal{V}_i)$ represents this linear mapping between the space of possible worlds and the space of predicates. The vectors P and Π are defined as

$$P = [Pr(\mathcal{V}_1), Pr(\mathcal{V}_2), \ldots, Pr(\mathcal{V}_8)]^T$$
$$= [p_1, p_2, \ldots, p_8]^T,$$
$$\Pi = [Pr(w_i), Pr(w_{i-1}), Pr(w_{i-2}), Pr(w_i \wedge w_{i-1}),$$

$$Pr(w_{i-1} \wedge w_{i-2}), Pr(w_i \wedge w_{i-1} \wedge w_{i-2})]^T$$
$$= [\pi_1, \pi_2, \ldots, \pi_6]^T, \tag{14}$$

where "T" stands for the transpose form vector. Using the linear system of equations expressed by Eq. (13), we can easily derive bounds for the set of solutions:

$$\max\{0, \pi_5 + \pi_4 - \pi_1\} \leqslant \pi_6 \leqslant \min\{\pi_i, i = 1, \ldots, 5\}. \tag{15}$$

Therefore, there is an infinity of solutions for $Pr(w_i \wedge w_{i-1} \wedge w_{i-2})$, which means an infinity of solutions for $Pr(w_i \mid w_{i-1} \wedge w_{i-2})$.

4.2. The entropy maximization criterion

As expressed by Eq. (15), the solution obtained so far is not unique. In this section, we use the maximum entropy criterion assigned to the world distribution in order to determine a unique estimate for $Pr(w_i \mid w_{i-1} \wedge w_{i-2})$. This is explained as follows:

Because π_6 is the unknown variable, we first disregard the last row of the consistent matrix C (C becomes C_d) (and the corresponding entry of Π which becomes Π_d) and add a tautology (all components $=1$) as the first constraint. The tautology is justified by the fact that $\sum_1^8 p_i = 1$. We finally obtain a constraint

$$C_d \times P = \Pi_d, \tag{16}$$

or, spelled out,

$$
\begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \\ p_7 \\ p_8
\end{bmatrix}
$$

$$
=
\begin{bmatrix}
1 \\
\widehat{Pr}(w_i) \\
\widehat{Pr}(w_{i-1}) \\
\widehat{Pr}(w_{i-2}) \\
\widehat{Pr}(w_i \wedge w_{i-1}) \\
\widehat{Pr}(w_{i-1} \wedge w_{i-2})
\end{bmatrix}
=
\begin{bmatrix}
1 \\ \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \\ \pi_5
\end{bmatrix}.
\tag{17}
$$

Since the solution to (16) is not unique (6 equations and 8 variables), we select among the possible solutions for P that maximizes the Shannon entropy $\sum -p_i \log p_i$ of the distribution P (or minimally discriminate among solutions). The resulting optimization program (as a minimization, reversing the sign of the entropy) is

$$
\min \left\{ S(P) = \sum_{i=1}^{i=8} p_i \log p_i \mid C_d \times P = \Pi_d, 0 \leqslant P \right\}. \tag{18}
$$

Note that we actually need a box constraint $0 \leqslant P \leqslant 1$ since we are looking for probability, but non-negativity of P, in addition to the first constraint $\sum_{i=1}^{i=8} p_i = 1$ ensures the upper bound of 1. The tautology acts as a normalizing condition.

Note also that (18) is a convex program since the gradient is

$$
\nabla S(P) = [\log p_i + 1]_{i=1}^{i=8},
$$

and therefore the Hessian

$$
\nabla^2 S(P) = \mathrm{Diag} \left[\frac{1}{p_i} \right]_{i=1}^{i=8},
$$

is a diagonal non-negative matrix.

Moreover, since (18) has a Slater point, the usual constraint qualification holds and a necessary and sufficient condition for optimality is given by the *Karush–Kuhn–Tucker* condition. But, before differentiating, we notice that we can reduce the number of variables.

Since the rank of C is 6, we can express (16), after some simple algebra, as

$$
P =
\begin{bmatrix}
1 - \pi_1 - \pi_2 - \pi_3 + \pi_4 + \pi_5 \\
\pi_3 - \pi_5 \\
\pi_2 - \pi_4 - \pi_5 \\
\pi_5 \\
\pi_1 - \pi_4 \\
0 \\
\pi_4 \\
0
\end{bmatrix}
+ p_6
\begin{bmatrix}
1 \\ -1 \\ 0 \\ 0 \\ -1 \\ 1 \\ 0 \\ 0
\end{bmatrix}
+ p_8
\begin{bmatrix}
0 \\ 0 \\ 1 \\ -1 \\ 0 \\ 0 \\ -1 \\ 1
\end{bmatrix},
\tag{19}
$$

a two-parameter family. Using this reduction, the objective function (the Shannon entropy) can also be simplified to a two variable function,

$$
\begin{aligned}
\overline{S}(p_6, p_8) = {}& p_6 \log(p_6) + p_8 \log(p_8) \\
& + (\pi_5 - p_8) \log(\pi_5 - p_8) \\
& + (\pi_5 - p_8) \log(\pi_5 - p_8) \\
& + (\pi_1 - p_6 - \pi_4) \log(\pi_1 - p_6 - \pi_4) \\
& + (\pi_3 - p_6 - \pi_5) \log(\pi_3 - p_6 - \pi_5) \\
& + (p_8 + \pi_2 - \pi_4 - \pi_5) \\
& \times \log(p_8 + \pi_2 - \pi_4 - \pi_5) \\
& + (p_6 + 1 - \pi_1 - \pi_2 - \pi_3 + \pi_4 + \pi_5) \\
& \times \log(p_6 + 1 - \pi_1 - \pi_2 \\
& - \pi_3 + \pi_4 + \pi_5).
\end{aligned}
\tag{20}
$$

With these simplifications, program (18) can be expressed as

$$
\min \{ \overline{S}(p_6, p_8) \mid 0 \leqslant p_6, 0 \leqslant p_8 \}, \tag{21}
$$

which is an unconstrained program since $\overline{S}$ is defined only in the positive orthant. To solve this latter problem, we finally differentiate and obtain $\nabla \overline{S}(p_6, p_8) = 0$, or

$$
\begin{aligned}
& \log(p_8) - \log(\pi_4 - p_8) - \log(\pi_5 - p_8) \\
& \quad + \log(p_8 + \pi_2 - \pi_4 - \pi_5) = 0, \\
& \log(p_6) - \log(\pi_1 - p_6 - \pi_4) - \log(\pi_3 - p_6 - \pi_5) \\
& \quad + \log(p_6 + 1 - \pi_1 - \pi_2 - \pi_3 + \pi_4 + \pi_5) = 0,
\end{aligned}
\tag{22}
$$

which simplifies to

$$
\frac{p_8(p_8 + \pi_2 - \pi_4 - \pi_5)}{(\pi_4 - p_8)(\pi_5 - p_8)} = 1, \tag{23a}
$$

$$
\frac{p_6(p_6 + 1 - \pi_1 - \pi_2 - \pi_3 + \pi_4 + \pi_5)}{(\pi_1 - p_6 - \pi_4)(\pi_3 - p_6 - \pi_5)} = 1. \tag{23b}
$$

We note that Eq. (23) are separated. From a convex, nonlinear optimization program in 8 variables, we have obtained 2 simple linear equations with solutions:

$$
p_8 = \frac{\pi_4 \pi_5}{\pi_2}, \tag{24a}
$$

$$
p_6 = \frac{(\pi_1 - \pi_4)(\pi_3 - \pi_5)}{1 - \pi_2}. \tag{24b}
$$

The final solution to the optimization problem (18) is now obtained from substituting (24) into (19). This yields the

optimal solution

$$P^* = \begin{bmatrix} 1 - \pi_1 - \pi_2 - \pi_3 + \pi_4 + \pi_5 + \frac{(\pi_1 - \pi_4)(\pi_3 - \pi_5)}{1 - \pi_2} \\ \pi_3 - \pi_5 - \frac{(\pi_1 - \pi_4)(\pi_3 - \pi_5)}{1 - \pi_2} \\ \pi_2 - \pi_4 - \pi_5 + \frac{\pi_4 \pi_5}{\pi_2} \\ \pi_5 - \frac{\pi_4 \pi_5}{\pi_2} \\ \pi_1 - \pi_4 - \frac{(\pi_1 - \pi_4)(\pi_3 - \pi_5)}{1 - \pi_2} \\ \frac{(\pi_1 - \pi_4)(\pi_3 - \pi_5)}{1 - \pi_2} \\ \pi_4 - \frac{\pi_4 \pi_5}{\pi_2} \\ \frac{\pi_4 \pi_5}{\pi_2} \end{bmatrix}. \tag{25}$$

Theorem 4.1. *The estimation using the maximum entropy criterion for the possible worlds distribution P implies that the word trigram distribution is a Markov chain of order* 1.

Proof. Using Bayes' rule, we can write the following:

$$Pr(w_i \mid w_{i-1} \wedge w_{i-2}) = \frac{Pr(w_i \wedge w_{i-1} \wedge w_{i-2})}{Pr(w_{i-1} \wedge w_{i-2})}$$

$$\simeq \frac{p_8^*}{p_4^* + p_8^*}. \tag{26}$$

Using the optimal values of p_8^* and p_4^* from the vector P^* of Eq. (25) and computing the ratio $p_8^*/(p_4^* + p_8^*)$, we obtain: $Pr(w_i \mid w_{i-1} \wedge w_{i-2}) = Pr(w_i \mid w_{i-1})$.

Using Eq. (11), we can conclude that the maximum-entropy estimation technique provides $\lambda_i^h = 1$ *for* $h = 1, 2, 3, 5$, $\lambda_4^h = e$, and $K^h = 0$. *Finally, we can conclude that the entropy maximization method enables to obtain a unique solution to our problem but provides a poor estimator of the trigram since it backs-off to the bigram.* □

4.3. The minimum perplexity criterion

Our ultimate goal in this approach is to determine a general expression for λ_i^h and K^h of the interpolation model of Eq. (11). We show how the minimum perplexity criterion enables to determine a unique solution for λ_i^h and K^h.

4.3.1. The determination of the interpolation coefficients

To determine a closed form for the coefficients λ_i^h and K^h, we use the following theorem [29]:

Theorem 4.2. *Let* $\mathscr{P} = \{P = [p_1, p_2, \ldots, p_n]^T$ *such that* $\|P\|_1 = \sum_{i=1}^{i=n} p_i = 1\}$, *let* U *be a linear operator in* $\Re^n$ *(Euclidean space), and* $D = (d_{ij})$ *its associated matrix relative to the classical base of* $\Re^n$. *If* $\sum_i d_{ij} = n$, $\forall j$, *and*

$d_{ij} \geqslant 0$, *then we can write the following assertions*:

(1) $U(\mathscr{P}) \subset n \times \mathscr{P}$,
(2) $\|\Pi\|_1 = n \times \|P\|_1 = n$,

where $\|.\|_1$ *is the* L_1 *norm* (*sum of components*).

Proof. $\|D \times P\|_1 = \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} d_{ij} \times p_j = \sum_{j=1}^{j=n} \sum_{i=1}^{i=n} d_{ij} = n \times \sum_{j=1}^{j=n} p_j$, and both of the relations (1) and (2) are proved. □

In order to use Theorem 4.2, we need to have a linear operator of $\Re^n$, which means a square matrix. Therefore, we add an *artificial row* $\mathscr{A}$ in the consistent matrix C (C becomes D) such that the sum of the components for each column is equal to the dimension of the space which is 8 in this problem (see Table 2). The vector Π is therefore extended with two components $\mu^h(\mathscr{A}_1)$ which is a tautology and $\mu^h(\mathscr{A}_2)$ corresponding to the row vectors $\mathscr{A}_1 = [1, 1, 1, 1, 1, 1, 1, 1]^T$ and $\mathscr{A}_2 = [7, 6, 6, 4, 6, 5, 4, 1]^T$, respectively. By adding a tautology which expresses exclusivity and exhaustivity, we made this two vector extension unique. The predicate vector Π is now extended and called Π_e. For a sake of brevity, $\mu^h(\mathscr{A}_2)$ is now called $\mu^h(\mathscr{A})$. From equation $D \times P = \Pi_e$, we can write

Table 2
The matrix D results from adding the two vectors $\mathscr{A}_1$ and $\mathscr{A}_2$ as two last rows

$$D = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 7 & 6 & 6 & 4 & 6 & 5 & 4 & 1 \end{pmatrix}. \tag{27}$$

$$\mu^h(\mathscr{A}) = 7p_1 + 6(p_2 + p_3 + p_5) + 4(p_4 + p_7) + 5p_6 + p_8. \tag{28}$$

Using Theorem 4.2, we obtain the following closed form

$$Pr(w_i \wedge w_{i-1} \wedge w_{i-2})$$
$$= 7 - \mu^h(\mathscr{A}) - \hat{P}r(w_i) - \hat{P}r(w_{i-1})$$
$$- \hat{P}r(w_{i-2}) - \hat{P}r(w_i \wedge w_{i-1})$$
$$- \hat{P}r(w_{i-1} \wedge w_{i-2}). \tag{29}$$

Using Bayes' rule, we finally express the conditional probability as

$$Pr(w_i \mid w_{i-1} \wedge w_{i-2}) \simeq \frac{7 - \mu^h(\mathscr{A}) - \hat{P}r(w_i) - \hat{P}r(w_{i-1}) - \hat{P}r(w_{i-2}) - \hat{P}r(w_i \wedge w_{i-1})}{Q} - 1, \tag{30}$$

where the quantity $Q = \hat{P}r(w_{i-1} \wedge w_{i-2})$ is an estimation of the probability assigned to the history of the configuration. This latter quantity has to be different from 0.

Given this latter equation, we can write the constant K^h introduced in Eq. (11) as

$$K^h = \frac{(7 - \mu^h(\mathscr{A}))}{Q} - 1. \tag{31}$$

In order to emphasize the coefficients λ_i^h $(i = 1, \ldots, 5)$, we write Eq. (29) into the following form

$$
\begin{aligned}
Pr&(w_i \mid w_{i-1} \wedge w_{i-2}) \\
&\simeq \frac{-1}{Q}[\hat{P}r(w_i) + \hat{P}r(w_{i-1}) + \hat{P}r(w_{i-2}) \\
&\quad + \hat{P}r(w_i \mid w_{i-1}) \times \hat{P}r(w_{i-1})] + 0 \\
&\quad \times \hat{P}r(w_{i-1} \mid w_{i-2}) + \frac{7 - \mu^h(\mathscr{A}) - Q}{Q \times \hat{P}r(w_i \mid w_{i-1} \wedge w_{i-2})} \\
&\quad \times \hat{P}r(w_i \mid w_{i-1} \wedge w_{i-2}).
\end{aligned}
\tag{32}
$$

We finally equate all coefficients of this equation with the interpolation model of Eq. (11) and derive the following:

$$
\begin{aligned}
\log(\lambda_i^*) &= \frac{1}{-Q} \iff \lambda_i^* = e^{\frac{1}{-Q}}, \quad \forall i = 1, \ldots, 3, \\
\log(\lambda_4^*) &= \frac{\hat{P}r(w_{i-1})}{-Q} \iff \lambda_4^* = e^{\frac{\hat{P}r(w_{i-1})}{-Q}}, \\
\log(\lambda_5^*) &= 0 \iff \lambda_5^* = 1, \\
K^h &= \frac{7 - \mu^h(\mathscr{A}) - Q}{Q}.
\end{aligned}
\tag{33}
$$

Because of the logarithmic scale, all coefficients λ_i $(i = 1, \ldots, 5)$ belong to the interval [0..1].

4.3.2. Determination of $\mu^h(\mathscr{A})$: minimum perplexity criterion

In order to determine a unique solution for $\mu^h(\mathscr{A})$ and compute uniquely the value for K^h, we use the minimum perplexity criterion. However, minimizing the perplexity $PP_p(T)$ is equivalent to maximizing $Pr(w_i \mid w_{i-2}, w_{i-1})$. Using the matrix D and some algebraic transformations, we derive

$$\mu^h(\mathscr{A}) = -\pi_1 - \pi_2 - \pi_3 - 2\pi_4 - \pi_5 + p_7 + 7. \tag{34}$$

Replacing this value of $\mu^h(\mathscr{A})$ in Eq. (30), we obtain

$$Pr(w_i \mid w_{i-2} \wedge w_{i-1}) = \frac{\pi_4 - p_7}{\pi_5}. \tag{35}$$

Therefore, we have the following equivalent problems:

$$
\begin{aligned}
\min\{PP_p(T)\} &\iff \max\{Pr(w_i \mid w_{i-2} \wedge w_{i-1})\} \\
&\iff \max_{p_7} \left\{ \frac{\pi_4 - p_7}{\pi_5} \right\}.
\end{aligned}
\tag{36}
$$

However, the trigram probability value and p_7 have to be in the interval [0..1]. This is expressed by the following constraints:

$$
\begin{aligned}
0 &\leqslant \frac{\pi_4 - p_7}{\pi_5} \leqslant 1, \\
0 &\leqslant p_7 \leqslant 1.
\end{aligned}
\tag{37}
$$

This constrained optimization problem provides $p_7 = \max\{0, \pi_4 - \pi_5\}$ as a solution. Finally, the trigram probability can be written as

$$Pr(w_i \mid w_{i-2} \wedge w_{i-1}) = \frac{\pi_4 - \max\{0, \pi_4 - \pi_5\}}{\pi_5}, \tag{38}$$

which is equivalent to

$$
\begin{aligned}
Pr&(w_i \mid w_{i-2} \wedge w_{i-1}) \\
&= \frac{Pr(w_{i-1} \wedge w_i) - \max\{0, Pr(w_{i-1} \wedge w_i) - Pr(w_{i-2} \wedge w_{i-1})\}}{Pr(w_{i-2} \wedge w_{i-1})}.
\end{aligned}
\tag{39}
$$

Finally, we can write the estimation result using the minimum perplexity criterion as

$$
\begin{aligned}
Pr&(w_i \mid w_{i-2} \wedge w_{i-1}) \\
&\approx \begin{cases}
\dfrac{\#(w_{i-2}, w_{i-1}, w_i)}{\#(w_{i-2}, w_{i-1})} & \text{if } \#(w_{i-2}, w_{i-1}, w_i) > 0, \\[2mm]
\dfrac{\#(w_{i-1}, w_i)}{\#(w_{i-2}, w_{i-1})} & \text{if } \#(w_{i-2}, w_{i-1}, w_i) = 0 \\
& \quad \wedge Pr(w_{i-1} \wedge w_i) \\
& \quad \leqslant Pr(w_{i-2} \wedge w_{i-1}), \\[2mm]
1 & \text{if } \#(w_{i-2}, w_{i-1}, w_i) = 0 \\
& \quad \wedge Pr(w_{i-1} \wedge w_i) \\
& \quad > Pr(w_{i-2} \wedge w_{i-1}).
\end{cases}
\end{aligned}
\tag{40}
$$

5. Experiments

We have conducted several experiments in order to compare the "PL" approach using the minimum perplexity criterion with the state-of-the-arts techniques described in this paper. Several samples with different sizes were extracted from *the Brown Corpus available at the Linguistic Data Consortium hosted by the University of Pennsylvania server* (*www.ldc.upenn.edu*). In each experiment, the corpus has been divided into 2/3 for training and 1/3 for testing. All models have been trained and tested on the same corpora. We have conducted four experiments with 19,422 words, 154,251 words, 466,514 words and 1,224,336 words, respectively. Table 3 depicts the perplexity values obtained for each model on these different test corpora. The results show that the "PL" language model using the minimum perplexity criterion outperforms the "Good-Turing Discounting" and the "Absolute Discounting" language models with a significant margin.

Table 3
Perplexity of each language model as a function of the corpus size (number of words) C_i

Perplexity				
Models	C_1: 19,422	C_2: 154,251	C_3: 466,514	C_4: 1,224336
Abs.Discount	80.736343	106.988535	132.733558	139.246739
Good-Turing	141.225636	150.514389	185.708934	180.390850
PL	**47.257053**	**26.359698**	**30.184792**	**21.629367**

6. Conclusion and future work

Our approach to language modeling incorporates the rules of logic within a probabilistic framework. This novel approach provides an insight into the sparseness problem encountered in statistical language modeling. Our methodology views word n-grams as predicates interrelated through logical connectors. We have proven that the solution obtained using the maximum entropy criterion is a first-order Markov chain. Within the same PL context, we used the minimum perplexity criterion and obtained an "optimal" solution to the sparseness problem. We conducted several experiments for comparison purposes. The results obtained show that "PL using minimum perplexity" is a promising method since it has outperformed both the "Good-Turing Discounting" and the "Absolute Discounting" techniques.

We will focus in our future work on (i) performing comparisons with other techniques such as the modified Kneser–Key method, (ii) extending the history on the n-grams, and (ii) incorporating semantical relationships between words using the WordNet lexical database [30]. Synonymy relation provided by WordNet will be expressed as an equivalence logical connector. Through this logical framework, we will be capable to fall-back on synonyms of a word in a trigram in case this word is not seen in the training corpus.

7. Summary

We have proposed a novel language modeling technique that is founded on the rules of logic and probability theory to solve the data sparseness problem inherent to many applications. This probabilistic logic paradigm views word n-grams as predicates connected by logical operators and assigned some probability values. Our approach expresses the unseen trigram as a linear combination of probabilities of all possible unigrams, and bigrams that can be extracted from it. The probability distribution was defined over the sets of all possible worlds associated with the different sets of possible truth values of predicates. This probability specifies for each set of possible worlds what is the chance that the actual world V_a is contained in V_i. If the actual world is for example the column $[0, 1, 1, 0, 1, 0]^T$, and if the trigram of interest is $\langle w_{i-2}\ w_{i-1}\ w_i \rangle = \langle \text{ate an apple} \rangle$, then this means that $\langle \text{apple} \rangle$ is absent from the corpus at position i, $\langle \text{an} \rangle$ is

present at position $(i - 1)$, $\langle \text{ate} \rangle$ is present in the corpus at position $(i - 2)$, $\langle \text{an apple} \rangle$ is absent, $\langle \text{ate an} \rangle$ is present and $\langle \text{ate an apple} \rangle$ is absent. Since we do not know the actual world which depends on the corpus contents, therefore there is an uncertainty about it expressed by a probability measure. The probability values $Pr(V_i)$ are subject to the constraint $\sum Pr(V_i) = 1$ since the set of possible worlds are mutually exhaustive and exclusive. We have proven that the maximum entropy-based model provides a Markov chain of order 1 and therefore is a poor estimate of the unseen trigrams. Using some topological knowledge (theorem), we were capable to minimize the perplexity of a language model assigned to a test corpus. We have thus derived the optimal model (the one with the lowest perplexity) and compared our approach with two traditional language models which are the "Good Turing" and the "Absolute Discounting". The results obtained show that the probabilistic logic language model has significantly outperformed these two traditional language models when tested on three different test sets of the Brown corpus containing 19,422 words, 154,251 words, 466,514 words and 1,224336 words, respectively.

References

[1] C. Manning, H. Schutze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA, 1999.

[2] H. Ney, U. Essen, R. Kneser, On structuring probabilistic dependencies in stochastic language modeling, Comput. Speech Lang. 8 (1) (1994) 1–28.

[3] I. Witten, T. Bell, The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression, IEEE Trans. Inf. Theory 37 (4) (1991).

[4] S.M. Katz, Estimation of probabilities from sparse data for the language model component of a speech recognizer, IEEE Transt. Acoust. Speech Signal Process. 35 (3) (1981) 400–401.

[5] P.F. Brown, V.J. Della Pietra, P.V. deSouza, J.C. Lai, R.L. Mercer, Class-based n-grams models of natural language, Comput. Linguist. 1814 (1992) 467–479.

[6] J. Lafferty, D. Sleator, D. Temperley, Grammatical trigrams: a probabilistic model of link grammar, Proceedings of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language, Cambridge, MA, 1992.

[7] T. Jaakola, M. Meila, T. Jebara, Maximum entropy discrimination, Advances in Neural Information Processing Systems, vol. 12 (NIPS1999), MIT Press, Cambridge, MA, 1999.

[8] R. Rosenfeld, Adaptive statistical language modeling: a maximum entropy approach, Ph.D. Thesis, School of Computer Science, Carnegie Mellon University, 1994, p. 24.

[9] J. Bellegarda, Exploiting latent semantic information in statistical language modeling, Proc. IEEE 88 (8) (2000) 1279–1296.

[10] Y. Bengio, R. Ducharme, P. Vincent, A neural probabilistic language model, Advances in Neural Information Processing Systems, vol. 13, 2001.

[11] X. Zhu, R. Rosenfeld, Improving trigram language modeling with the world wide web, Proceedings of International Conference on Acoustics, Speech, and Signal Processing, 2001.

[12] K. Church, W. Gale, Enhanced good turing and cat-cal: two new methods for estimating probabilities of English bigrams, Comput. Speech Lang. (1990).

[13] I. Dagan, F. Pereira, L. Lee, Similarity-based estimation of word cooccurrence probabilities, Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, New Mexico State University, June 1994.

[14] F. Jelinek, R. Mercer, S. Roukos, Principles of lexical language modeling for speech recognition, in: S. Furui, M. Moham Sondhi (Eds.), Advances in Speech Signal Processing, Marcel Dekker, Inc., New York, 1992, pp. 651–699.

[15] F. Jelinek, J. Lafferty, R. Mercer, Interpolated estimation of Markov source parameters from sparse data, Pattern Recognition in Practice, North-Holland, Amsterdam, 1981, pp. 381–397.

[16] A. Nadas, Estimation of probabilities in the language model of the IBM speech recognition system, IEEE Trans. Acoust. Speech Signal Process. 32 (1981) 819–861.

[17] D. Bouchaffra, V. Govindaraju, S.N. Srihari, Postprocessing of recognized strings using nonstationary Markovian models, IEEE Transactions Pattern Analysis and Machine Intelligence, PAMI, vol. 21(10), October 1999.

[18] D. Bouchaffra, E. Koontz, V Kripasundar, R.K. Srihari, Integrating signal and language context to improve handwritten phrase recognition: alternative approaches, Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics, Fort Lauderdale, Florida, January, 1997, p. 47.

[19] D. Bouchaffra, E. Koontz, V. Kripasundar, R.K. Srihari, Incorporating diverse information sources in handwriting recognition postprocessing, Int. J. Imaging Syst. Technol. 7 (1996) 320–329.

[20] P. Clarkson, Adaptation of statistical language models for automatic speech recognition, Ph.D. Thesis, Cambridge University, Engineering Department, 1999.

[21] I.J. Good, The population frequencies of species and the estimation of population parameters, Biometrika 40 (1953) 237–264.

[22] F. James, Modified Knesser–Ney smoothing of n-gram models, RIACS Technical Report 00.07, October 2000.

[23] R. Kneser, H. Kney, Improved backing-off for m-gram language modeling, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, 1995, pp. 181–184.

[24] P. Xu, I. Jelinek, Random forests in language modeling, Technical Report, Center for Language and Speech Processing, John Hopkins University, Baltimore, Maryland, 2004.

[25] S.F. Chen, R. Rosenfeld, A survey of smoothing techniques for ME models, IEEE Trans. Speech Audio Process. 8 (1) (2000) 3750.

[26] N.J. Nilsson, Probabilistic logic, Artificial Intelligence, vol. 281, Elsevier, North-Holland, Amsterdam, 1981, pp. 71–81.

[27] D. Bouchaffra, Theory and algorithms for analysing the consistent region in probabilistic logic, Int. J. Comput. Math. 25 (3) (1993) 13,25.

[28] S. Russel, P. Norvig, Artificial Intelligence: A Modern Approach, Prentice-Hall, Englewood Cliffs, NJ, 1995.

[29] J. Dieudonné, Foundations of Modern Analysis, Academic Press, New York, 1960.

[30] G. Miller, R. Beckwith, C. Fellbaum, Gross, K.J. Miller, Introduction to WordNet: an online lexical database, J. Lexicog. (4) (1990).

About the Author—DJAMEL BOUCHAFFRA holds a Master of Science in Mathematics and a Ph.D. in Computer Science from Grenoble University, France. After his Ph.D., he was awarded a grant of excellence for a 1 year postdoctoral position at the University of Quebec at Montreal, Canada. His work was on Markov random fields and Galois lattices for classification of textual documents. Dr. Djamel Bouchaffra joined The Center of Excellence for Document Analysis and Recognition (CEDAR) located at the State University of New York at Buffalo. He was a senior research scientist for a period of 5 years. His work was focused on recognition of zipcodes in handwritten mailpieces. In August of the year 2000, Dr. Bouchaffra joined Oakland University as an assistant professor of Computer Science and Engineering. He is teaching Pattern Recognition (CSE 616), Soft Computing (CSE 513), Discrete Mathematics (CSE 504), Artificial Intelligence (CSE 516), and Operating Systems (CSE 450/550). His field of research is in Structural Pattern Recognition and Artificial Intelligence. Professor Bouchaffra is the director of the pattern recognition and machine intelligence laboratory (PARMIL). He has written several papers in peer-reviewed conferences and premier journals. He has chaired several sessions in conferences. He is one of the general chairs of the Computer Science and Information technology Conference CSIT'2005. He is a regular reviewer for many journals such as IEEE TPAMI, TNN, TKDE, and Image Processing. He was nominated for the 2004 Oakland University Teaching Excellence Award. Professor Bouchaffra is a senior member of the IEEE, a member of the Computer Society and the vice chair of the IEEE/SEM chapter V.