INTERACTIVE FEATURE INDUCTION AND LOGISTIC REGRESSION FOR WHOLE SENTENCE EXPONENTIAL LANGUAGE MODELS

Ronald Rosenfeld, Larry Wasserman, Can Cai, Xiaojin Zhu

Carnegie Mellon University
Pittsburgh, PA 15213
USA

ABSTRACT

Whole sentence exponential language models directly model the probability of an entire sentence using arbitrary computable properties of that sentence. We present an interactive methodology for feature induction, and demonstrate it in the simple but common case of a trigram baseline, focusing on features that capture the linguistic notion of semantic coherence. We then show how parametric regression can be used in this setup to efficiently estimate the model's parameters, whereas non-parametric regression can be used to construct more powerful exponential models from the raw features.

1. INTRODUCTION

Conventional language models such as n-grams take little advantage of the nature of human language. There have been several attempts to improve on these models by exploiting hypothesized linguistic structure. These attempts have not succeeded for the most part: the difficulty of parameter estimation made it impossible to find a good solution, or else improvements over the conventional model were too small to justify the added complexity. We believe that these attempts were hampered by at least the following problems:

- **Conditional framework:** Language modeling to date has been done by using the chain rule to decompose the probability of a sentence or utterance into a product of conditional word probabilities. But this conditional formulation is not conducive to thinking about, let alone incorporating, global (sentence level) linguistic information.
- **Maximum entropy training:** When integrating diverse and overlapping knowledge sources into a language model, the most successful method to date makes use of the exponential family and the maximum entropy principle. However, training a conditional exponential model is computationally prohibitive. The most burdensome step in the training is the computation of the normalizing term, which depends on the conditioning event (the history).

We have recently been working on a language modeling framework that addresses these problems. Our solution is a whole-sentence exponential model:

$$p(s) = \frac{1}{Z} \cdot p_0(s) \cdot \exp\left[\sum_i \lambda_i f_i(s)\right] \qquad (1)$$

where the λ_i's are the parameters of the model, Z is a universal normalization constant which depends only on the λ_i's, and the $f_i(s)$'s are arbitrary computable properties, or *features*, of the sentence s. $p_0(s)$ is an arbitrary probability distribution, which can be thought of as the starting point, or baseline, for further modeling improvements. Often, $p_0(s)$ will be simply derived from a trigram.

The features $\{f_i(s)\}$ are selected by the modeler to capture those aspects of the data they consider appropriate or profitable. These can vary from class n-grams, longer-distance dependencies, or simple global sentence properties, to more complex functions based on part-of-speech tagging, parsing, or other types of linguistic analysis (person and number agreement, semantic coherence, etc.). For each feature $f_i(s)$, its expectation under $P(s)$ is constrained to a specific value K_i:

$$E_p f_i = K_i \quad . \qquad (2)$$

These target values are typically set to the expectation of that feature under the empirical distribution $\tilde{p}$ of the training corpus $T = \{s_1, \ldots, s_N\}$ (For binary features, this is simply the prevalence of that feature in the corpus.) Then, the constraint becomes:

$$\sum_s p(s) \cdot f_i(s) = E_{\tilde{p}} f_i \equiv \frac{1}{N} \sum_{j=1}^{N} f_i(s_j) \quad . \qquad (3)$$

If the constraints (2) are consistent, there exists a unique solution $\{\lambda_i\}$ within the exponential family (1) which satisfies them. Among all (not necessarily exponential) solutions to equations (2), the exponential solution is the one closest

to the baseline $p_0(s)$ (in the Kullback-Liebler sense), and is thus called the Minimum Divergence or Minimum Discrimination Information (MDI) solution. If the baseline $p(s)$ is flat (uniform), this becomes the Maximum Entropy (ME) solution. Furthermore, if the feature target values K_i are the empirical expectations over some training corpus (as in equations (3)), the MDI or ME solution is also the Maximum Likelihood solution of the exponential family. For more about the ME principle, see [7]. For the application of ME to conditional language models, see [2, 8].

The whole sentence exponential model of equation (1) was first proposed in [9], where we discussed training via Monte Carlo Markov Chain (MCMC) and other sampling methods. In [3] we studied efficient sampling, smoothing and automatic feature selection for such models. In [10], we used these techniques to add parse-based features into a baseline trigram, and showed how to accurately estimate the (universal) normalizing constant.

Once the above framework has been proven practical, there remains the important challenge of *feature induction*. Namely, we need a methodology for searching for and selecting profitable features.

Joint (i.e. non-conditional) exponential modeling was first applied to a natural language processing problem by [4]. They used a joint exponential form to model the spelling of individual words. Because of the relatively small space of this problem, feature induction could be done by iteratively considering a small set of atomic features and their combination with existing features. When modeling whole sentences, however, the space of possible features is considerably larger, and such completely automatic methods may no longer suffice.

In what follows, section 2 presents an interactive methodology for feature induction. Section 3 demonstrates the methodology in the simple but common case of a trigram baseline. This leads to a focus on features that capture the linguistic notion of semantic coherence, which is taken up in section 4. Finally, section 5 shows how parametric regression can be used in this setup to efficiently estimate the model's parameters, whereas non-parametric regression can be used to construct more powerful exponential models from the raw features.

2. INTERACTIVE METHODOLOGY FOR FEATURE INDUCTION

Our goal is to choose features $f_i(s)$ that capture aspects of language which are not captured (or inadequately captured) by the current baseline modeling technique. To this end, we have developed the following interactive methodology.

Given a corpus T of natural language sentences with empirical distribution $\tilde{p}$, presumably representative of the unknown target distribution p, we use it to train our best baseline model p_0. Next, we use p_0 to generate a corpus T_0 of 'pseudo sentences' . We then manually compare T_0 with T (or some other dataset from the same distribution p). We ask human subjects to look for systematic differences between the two corpora. Any such difference points to a deficiency in the way p_0 models the unknown target distribution p. Any such deficiency can now be readily fixed, by defining an appropriate set of features $f_1(s), \ldots, f_k(s)$ which have different expectations under p and p_0 (as evidenced by their respective samples T and T_0). The new features are then added, resulting in a new model:

$$p_1(s) = \frac{1}{Z} \cdot p_0(s) \cdot \exp\left[\sum_i \lambda_i f_i(s) \right] \qquad (4)$$

Once p_1 is trained, the appropriate constraint (equation 3) guarantees that it consistently captures the new feature, and the previously observed difference between our model and the target distribution has been eliminated.

The process can now be repeated by generating a corpus T_1 of 'pseudo sentences' from the improved model p_1, and comparing it to the original corpus T, looking for new differences. The latter will be captured with new features, and so on. Note that, in contrast with [4], the emphasis in our methodology is on manual inspection of two corpora and the linguistic analysis and 'detective work' of searching for and evaluating families of linguistically motivated features.

3. WHAT'S WRONG WITH A TRIGRAM?

To demonstrate the methodology presented above, we applied it to the simple but common case of a trigram baseline. Let T be the 1992–1996 Broadcast News corpus [5]. Example sentences from T are given in table 1. Let $p_0(s)$ be derived by chain rule from a well-smoothed trigram model trained on T, and let T_0 be a corpus of "pseudo sentences" generated according to p_0. Examples from T_0 are given in table 2.

How inherently different are these two sentence sources? Even though some of the true sentences are by no means grammatical or complete, there is something about them which seems to "make sense". In contrast, the pseudo sentences (except the very short ones) do not generally "make sense".

How well can the two sources be told apart? In an informal experiment, we presented a blind mixture of 40 average length sentences from T and T_0 to 17 members of the Sphinx research group at Carnegie Mellon University. An example of such a mixture is given in table 3.

The on-the-spot individual classification accuracies achieved by this group were 90%±5%. It is likely that better performance can be achieved given more deliberation time and/or experience. But in any case, human performance is neither an upper bound nor a lower bound on automatic perfor-

Table 1: Example sentences from the Broadcast News Corpus

```
THAT'S YOUR NEWS ON THE DAY BEFORE CHRISTMAS THIRTY FIVE PAST THE HOUR </s>
BUT WHAT ABOUT THE FLAWLESS SYMMETRY OF THE IMAGE ON THOSE WINDOWS </s>
STEVE GREEN WITH THE U. S. POSTAL SERVICE A SPOKESMAN </s>
RELATIONSHIPS AND ALLIANCES QUOTE WE ALSO HAVE A NUMBER OF DEVELOPING RELATIONSHIPS AND ALLIANCES </s>
TOYNX HAS THE CHRISTMAS SPIRIT ALL YEAR ROUND QUOTE </s>
THERE ARE PEOPLE ALL THESE CHRISTMAS MOVIES ARE IN A SENSE ABOUT THE SECULAR CHRISTMAS </s>
HE'S GOING TO ARGUE THAT THE JURY'S VERDICT SUGGESTS THAT NICHOLS' PARTICIPATION WAS SO MINOR IN THIS CONSPIRACY
            THAT IT WOULD BE UNCONSTITUTIONAL TO EVEN ALLOW THE JURY TO IMPOSE THE DEATH PENALTY </s>
THE SINGLE EUROPEAN CURRENCY </s>
```

Table 2: Example "pseudo sentences" generated by a trigram

```
IT WAS A HUMAN RIGHTS AND RESPONSIBILITIES WASN'T SAFE FOR MY CAPITAL GAINS ARE JOINING US FROM GETTING GUNS
            BECAUSE THERE WAS NO CRIMINAL WRONGDOING THIS TREATY </s>
SO OF COURSE UNLESS THEY'VE DIVIDED MUNCHAUSEN PAYNE </s>
THAT'S AWFULLY INFLAMMATORY ATTENDING U. TWO TO ONE IN FIVE YEARS BACK YOU KNOW ALL THE OTHER THING ROYCE HAS
            MORE ON THIS SUIT </s>
YEAH SURE </s>
CLARK CIRCUS COPS FEET AND PREPARE TO PUT BIBLE STORIES FROM THE GAME </s>
WITH THE CHANGES WOULD COST </s>
AND FRANKLY I FIND TO PEOPLE ALL OVER THE UNITED STATES FROM THE FEDERAL INTIMIDATION CAN BE SEEN AS A NON
            MARRIED CHOOSES TO FIGHT INFLATION </s>
```

mance. These numbers should therefore be taken as merely a rough indication of the potential of automatic discrimination methods (such methods can in turn be automatically converted into features, as will be shown in section 5).

What "features" did the human subjects use in discriminating T from T_0? We mentioned that the pseudo sentences did not "make sense". In fact, they violate just about all of our linguistic notions (with the not surprising exception of short-term word correlations). These include notions of lexical relations, syntax, semantics, topic coherence and pragmatics. Can we capture any of these glaring differences computationally? We decided to focus on a single aspect: the semantic coherence of the T sentences, as opposed to the apparent semantic *in*coherence of T_0 sentences. To do so, we repeated the classification experiment after removing all non-content words[1]. The test set then looked as in table 4[2].

The on-the-spot individual classification accuracies for this condition was down to 66%±9%. Again, it is likely that better performance can be achieved given more deliberation time and/or experience. It is clear though that, although a great deal of information has been lost, much information still remains in the content-bearing words. Thus, the linguistic notion of "semantic coherence" could prove quite useful. In the next section we derive raw features based on that notion.

[1]These were heuristically defined as the 200 most common words in T

[2]In practice, this latter test set was of course presented first.

4. MODELING SEMANTIC COHERENCE

There are tens of thousands of "content words" in natural language. Each sentence contains a very small subset of them (from zero to, say, twenty). Clearly, some subsets are much more likely than others. Modeling the distribution of such subsets in natural language is a non-trivial challenge. One might consider hidden-variable models, or dimension reduction techniques such as Singular Value Decomposition, recently applied to language modeling by [1].

As a first and admittedly crude attempt, we chose to collect a set of raw features, one for each possible pair of content words. Our hope was that significant differences between T and T_0 will be observable even with relatively simple features. Given a pair of content words, (w_A, w_B), a 2x2 contingency table can be constructed with the following counts:

- C_{11}: the number of sentences in T in which w_A and w_B co-occurred[3].
- C_{12}: the number of all other sentences in T in which w_A occurred.
- C_{21}: the number of all other sentences in T in which w_B occurred.

[3]To exclude "trigram effects", which can be overwhelming, we only considered w_A, w_B to have co-occurred if they were separated by at least 5 words.

Table 3: Example mixture of real- and pseudo-sentences, presented to human subjects for classification

IT'S DIFFICULT REALLY I THINK FOR ANY OF US TO YET ENTIRELY COMPREHEND WHAT THIS MEANS </s>
YOU WERE GOING TO TAKE THEIR CUE FROM ANCHORAGE LIFTED OFF EVERYTHING WILL WORK SITE VERDI </s>
HE URGES RESTAURANTS SYNAGOGUES AND SCHOOLS TO SET UP THE CELLULAR EQUIVALENT OF NO SMOKING SECTIONS </s>
I'D LIKE TO BE IDENTIFIED WITH A DAY THEY'RE FULL OF FLAMMABLE HYDROGEN IN REAL LIFE </s>
ANN IF I COULD ASK YOU TO JUST HOLD IT THERE FOR A SECOND JONATHAN ALSO </s>
LOOK I UNDERSTAND THE WAY SUCH AS DAVID SAID LITTLE INCUMBENCY FOUR TELEVISION YOU LIKE IT </s>
IN NEW YORK CITY TEAMS OF PLOWS PUSHED AWAY UP TO HALF A FOOT OF SNOW </s>
IT'S A REALLY SURE VOTE THAT POP UP A SMALL BUT IMPORTANT CORN AND SOYBEAN FIELDS </s>

Table 4: Same mixture of real- and pseudo-sentences, with non-content words removed

- DIFFICULT - - - - - - - - YET ENTIRELY COMPREHEND - - MEANS
- - - - - - CUE - ANCHORAGE LIFTED - EVERYTHING - - SITE VERDI
- URGES RESTAURANTS SYNAGOGUES - SCHOOLS - SET - - CELLULAR EQUIVALENT - - SMOKING SECTIONS
I'D - - - IDENTIFIED - - - - FULL - FLAMMABLE HYDROGEN - REAL LIFE
ANN - - - ASK - - - HOLD - - - - SECOND JONATHAN -
- - UNDERSTAND - - SUCH - DAVID - - INCUMBENCY - TELEVISION - - -
- - YORK CITY TEAMS - PLOWS PUSHED AWAY - - HALF - FOOT - SNOW
- - - SURE VOTE - POP - - SMALL - IMPORTANT CORN - SOYBEAN FIELDS

- C_{22}: the number of sentences in T in which neither w_A nor w_B occurred.

Let the marginals of this table be designated by C_{1+}, C_{2+}, C_{+1} and C_{+2}, and let $C_{++} = N$ be the size of the corpus. We have considered the following measures of association:

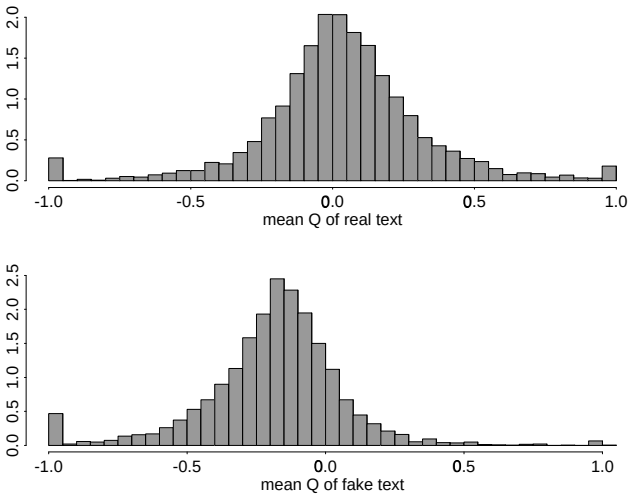
- correlation coefficient: $\hat{\rho} = \frac{C_{11}C_{22} - C_{12}C_{21}}{\sqrt{C_{1+}C_{2+}C_{+1}C_{+2}}}$
- Yule's measure of association: $\hat{Q} = \frac{C_{11}C_{22} - C_{12}C_{21}}{C_{11}C_{22} + C_{12}C_{21}}$
- mutual information: $\hat{I} = \sum_{i,j=1,2} \frac{C_{ij}}{N} \log N \left(\frac{C_{ij}}{C_{i+}C_{j+}} \right)$

All these measure are of course strongly correlated. But some are more appropriate than others in this setup. For example, ρ has a very narrow dynamic range for this data, whereas the estimate of I is undefined for zero counts. For these reasons we chose to focus on Q for now.

We estimated Q values for all content-word pairs in a substantial fraction of the vocabulary. Then for each sentence s, the list of content words pairs co-occuring in it can be derived, resulting in a variable length list of Q values. The distribution of these lists in natural language can then be studied and contrasted with that in P_0-generated sentences.

For our preliminary analysis, we have further simplified the modeling task by extracting a small set of statistics from each sentence-based Q list: its maximum, minimum, mean, and median. In figure 1, the distribution of the sentence-mean of Q is compared between "real" sentences and "fake" (trigram generated) sentences. Although the shape of the two distributions is similar, their modes and their variances are quite distinct. Significant differences were also found with the other three statistics.

Figure 1: The distribution of sentence-mean Q values in "real" (natural language) and "fake" (trigram-generated) sentences.



Given these clear distributional differences, how can we

best exploit them? In the next section we try to provide a general answer, by making use of existing statistical techniques.

5. EXPONENTIAL MODELING AND LOGISTIC REGRESSION

In this section we show that the discrimination setup of section 2 leads to an interesting relationship between exponential modeling and logistic regression. We first show that, given our setup, simple parametric logistic regression can be used to estimate the model's parameters (instead of the more computationally expensive iterative scaling algorithm). Next, we show how non-parametric logistic regression can be used to construct more powerful exponential models from the raw features.

5.1. Parameter Estimation via Parametric Regression

Given a fixed set of features $f_1(s), \ldots, f_k(s)$, let us assume that the distribution of "real" (natural language) sentences belongs to the exponential family:

$$p(s;\lambda) = \frac{1}{Z(\lambda)}\, p_0(s) \cdot \exp\,[\, \sum_i \lambda_i f_i(s)\,] \qquad (5)$$

Our goal is then to find the Maximum Likelihood estimate for the λ's, based on some corpus T. Let $s_1, \ldots, s_n$ be a sample of "real" sentences drawn from T, and let $s_{n+1}, \ldots, s_{2n}$ be a sample of "fake" sentences drawn from p_0. Define $Y = (Y_1, \ldots, Y_{2n})$ as follows: $Y_i = 1$ for $i = 1, \ldots, n$ and $Y_i = 0$ for $i = n+1, \ldots, 2n$. Note that $\Pr(s|Y=1) = p(s)$ and $\Pr(s|Y=0) = p_0(s)$. Also, $\Pr(Y=0) = \Pr(Y=1) = 1/2$ by construction.

Define $h(s) = Pr(Y=1|S=s)$. By Bayes' theorem,

$$\begin{aligned} h(s) &= \Pr(Y=1|S=s) \\ &= \frac{p(s|Y=1)Pr(Y=1)}{p(s|Y=1)Pr(Y=1) + p(s|Y=0)Pr(Y=0)} \\ &= \frac{p(s)}{p(s)+p_0(s)} \\ &= \frac{\frac{1}{Z}\exp\,[\,\sum_j \lambda_j f_j(s)\,]}{\frac{1}{Z}\exp\,[\,\sum_j \lambda_j f_j(s)\,]\, + 1}. \end{aligned}$$

Hence,

$$\begin{aligned} \mathrm{logit}(s) &\equiv \log\left(\frac{h(s)}{1-h(s)}\right) \\ &= \sum_j \lambda_j f_j(s) - \log Z \\ &= \beta_0 + \sum_j \beta_j f_j(s) \end{aligned}$$

where $\beta_j = \lambda_j$ and $\beta_0 = -\log Z$.

Now suppose we perform a logistic regression of the $Y_i's$ on the $f_j's$. This means we fit a regression model of the form $\mathrm{logit}(s) = \beta_0 + \sum_j \beta_j f_j(s)$ yielding estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. The above algebra suggests that we can use $\hat{\beta}_j$ as an estimate of λ_j and we can use $\hat{\beta}_0$ as an estimate of $-\log Z$.

In addition, given a large set of candidate features, any of the myriad existing statistical tools for variable selection and significance testing can be employed to search for an optimal set of features. This approximation has a significant computational advantage over the step-by-step application of Iterative Scaling[4].

5.2. Powerful Features via Non Parametric Regression

Given the same set of features $f_1(s), \ldots, f_k(s)$, we can combine them non-linearly by using a more general exponential model:

$$p(s) = \frac{1}{Z}\, p_0(s) \exp\,[\, \sum_{j=1}^{k} g_j(f_j(s))\,] \qquad (6)$$

where g_j is an arbitrary smooth function of f_j.

Maximizing the likelihood over all smooth functions g_1, $\ldots$, g_k is an ill-posed problem. One could use regularization techniques, or maximum penalized likelihood, to favor simple functions over more complicated ones. Alternatively, sticking with the regression framework, we can fit a generalized additive logistic regression [6] of the form

$$\mathrm{logit}(s) = \beta_0 + \sum_j g_j(f_j(s)) \qquad (7)$$

using any standard non-parametric regression software. If $\hat{\beta}_0$ and $\hat{g}_j$ are the resulting estimates, the estimate of the exponential model is then given by:

$$\hat{p}(s) = e^{\hat{\beta}_0} \cdot p_0(s) \cdot \exp\,[\, \sum_{j=1}^{k} \hat{g}_j(s_j)\,]\,. \qquad (8)$$

6. REFERENCES

[1] Bellegarda, J. R. 1998 A Multi-Span Language Modeling Framework for Large Vocabulary Speech Recognition. *IEEE Trans. SAP*, **6**, 456–467.

[2] Berger, A., DellaPietra, S. & DellaPietra, V. 1996 A maximum entropy approach to natural language processing. *Comput. Linguistics* **22**, 39–71.

[4] The loss of information due to using regression instead of the true MLE is the difference between the Fisher information of the MLE and the Fisher information of the regression model.

[3] Chen, S. F. & Rosenfeld, R. 1999 Efficient Sampling and Feature Selection in Whole Sentence Maximum Entropy Language Models. In *Proc. ICASSP, Phoenix, Arizona, March 1999*.

[4] DellaPietra, S., DellaPietra, V. & Lafferty, J. 1997 Inducing features of random fields. *IEEE Trans. PAMI*, **19**.

[5] Graff, D. 1997 The 1996 Broadcast news Speech and Language Model Corpus. In *Proc. DARPA Workshop on Spoken Language technology*.

[6] Hastie, T. J. Tibshirani, R. J. (1990). Generalized additive models. Chapman and Hall.

[7] Jaynes, E. T., 1957 Information Theory and Statistical Mechanics. *Physics Reviews*, **106**, 620–630.

[8] Rosenfeld, R. 1996 A Maximum Entropy Approach to Adaptive Statistical Language Modeling. *Computer, Speech and Language*, **10**, 187–228. Longer version: *Carnegie Mellon Tech. Rep. CMU-CS-94-138*.

[9] Rosenfeld, R. 1997 A Whole Sentence Maximum Entropy Language Model. In *Proc. IEEE workshop on Speech Recognition and Understanding, Santa Barbara, California, December 1997*.

[10] Zhu, X. J., Chen, S. F. & Rosenfeld, R. 1999 Linguistic Features for Whole Sentence Maximum Entropy Language Models. In *Proc. Eurospeech, September 1999, Hungary*.