AN ALTERNATIVE SCHEME FOR PERPLEXITY ESTIMATION

Frédéric BIMBOT [1] *Marc EL-BEZE* [2] *Michèle JARDINO* [3]

(1) ENST - Dépt SIGNAL, CNRS - URA 820, 46 rue Barrault, 75634 PARIS cedex 13, FRANCE, European Union.
(2) LIA, BP 1228, Agroparc, 339 Chemin des Meinajariès, 84911 AVIGNON cedex 9, FRANCE, European Union.
(3) LIMSI - CNRS, BP 133, Bât. 508, Univ. Paris-Sud, 91403 ORSAY cedex, FRANCE, European Union.
bimbot@sig.enst.fr, elbeze@univ-avignon.fr, jardino@limsi.fr

ABSTRACT

Language models are usually evaluated on test texts using the perplexity derived directly from the model likelihood function. In order to use this measure in the framework of a comparative evaluation campaign, we have developped an alternative scheme for perplexity estimation. The method is derived from the Shannon game and based on a gambling approach on the next word to come in a truncated sentence. We also use entropy bounds proposed by Shannon and based on the rank of the correct answer, in order to estimate a perplexity interval for non-probabilistic language models. The relevance of the approach is assessed on an example.

1. INTRODUCTION

Comparative evaluation campaigns for speech recognition systems have been initiated a few years ago under ARPA's impulse and have proven to be a very efficient way to boost scientific progress by stimulating laboratories to compare methods on identical data and under a common protocol. A similar set of concerted research actions has been set up by AUPELF-UREF[1], an agency in charge of the promotion of the French language in post-graduate education and research. One of these actions is devoted to speech recognition systems, and includes a specific activity on the evaluation of language models.

Perplexity is a canonical figure for describing the prediction capabilities of a language model[2]. However, perplexity is usually derived from the likelihood of the model on a test corpus, but this approach is not appropriate for a comparative evaluation campaign. In practice, it would require either that each site computes itself the perplexity figure or that the software of each participant be re-implemented within the site in charge of organizing the evaluation.

In this paper, we propose to estimate the perplexity of a language model via its ability to solve a particular task. We adapt the *Shannon game* to this purpose, and we describe a *gambling scheme* that permits an alternative estimation of the perplexity for any probabilistic language model. Under this approach, a set of truncated sentences are provided to every participant. For each sentence, the participants must distribute a capital of 1.0 between each word of the vocabulary, that is they must *bet* on the word coming just after the truncation. The perplexity is then evaluated outside of the participant's site as the inverse of the geometric mean of the bets placed on the correct words.

[1] Association des Universités Partiellement ou Entièrement de Langue Française - Université des Réseaux d'Expression Française.

[2] even though other evaluations may be more meaningful to predict the behaviour of a model for speech recognition purposes.

We indicate how to choose the bets from the likelihood values yielded by the model. We describe a way to set constraints on the bets in order to remain in a tractable probabilistic framework in the realistic case of an open vocabulary. We illustrate the validity of the new scheme by a few experiments. We also report on preliminary experiments for generalizing the approach to *non*-probabilistic language models.

2. THE SHANNON GAME

In [1], Shannon describes a method for estimating the entropy of a language : a person is asked to guess the first letter of a text by proposing successively some candidates within 27 possibilities (the 26 letters of the alphabet plus the space), until success. Once the first letter has been found by the subject, he is asked to guess the second one, and so on.

Shannon relates the statistics of the number of trials required to find the right answer (namely the frequency distribution of the rank of the correct answer), to upper and lower bounds of the language entropy. Cover and King [2] show that the Shannon game can be generalized into a gambling approach, where the person is asked to place a certain fraction of a capital on each possible letter. Jelinek [3] describes variants of the Shannon game for comparing a language model and a human being. O'Boyle et al. [4] use the Shannon game for assessing a model from the number of times it predicts the correct word below a given rank.

We adapt these principles to the comparison of probabilistic (and, to a certain extent, non-probabilistic) language models in the framework of an evaluation campaign.

3. THE GAMBLING SCHEME

The estimation of the perplexity of a language model in the gambling scheme can be formalized as follows : consider the successive sentence fragments, obtained by discovering progressively each word of the n-word sentence[3] $\mathcal{W}$:

$$\mathcal{W} = < w_1\, w_2 \ldots w_k \ldots w_n >$$

1) $W_0^0 = <$
2) $W_0^1 = < w_1$
   ...
k) $W_0^{k-1} = < w_1\, w_2 \ldots w_{k-1}$
   ...
n) $W_0^{n-1} = < w_1\, w_2 \ldots w_{k-1}\, w_k \ldots w_{n-1}$

Let $\mathcal{V} = \{v_j\}_{1 \leq j \leq m}$ the vocabulary from which the words of $\mathcal{W}$ are drawn.

The evaluation of the perplexity via the gambling scheme proceeds in the following way : for each truncated sentence

[3] with < and > marking the beginning and end of the sentence.

W_0^{k-1}, the probabilistic language model puts a bet $\beta_k(v_j)$ on each possible vocabulary entry v_j.

If the bets $\beta_k(v_j)$ are derived from the language model likelihood function P as :

$$\beta_k(v_j) = p(v_j \mid W_0^{k-1}) = \frac{P(W_0^{k-1}\, v_j)}{P(W_0^{k-1})} = \frac{P(b\, w_1\, w_2 \ldots w_{k-1}\, v_j)}{\sum_{q=1}^{q=m} P(b\, w_1\, w_2 \ldots w_{k-1}\, v_q)} \quad (1)$$

they obviously add up to 1.0. Using the chain rule, the perplexity $\mathcal{PP}(\mathcal{W})$ can be computed as :

$$\mathcal{PP}(\mathcal{W}) = [P(\mathcal{W})]^{-\frac{1}{n}} = \left[\prod_{k=1}^{k=n} p(w_k \mid W_0^{k-1})\right]^{-\frac{1}{n}} \quad (2)$$

Therefore, by choosing bets $\beta_k(v_j)$ equal to the conditional probability of word v_j in context W_0^{k-1}, the perplexity $\mathcal{PP}(\mathcal{W})$ can also be obtained as :

$$\mathcal{PP}(\mathcal{W}) = \left[\prod_{k=1}^{k=n} \beta_k(w_k)\right]^{-\frac{1}{n}} \quad (3)$$

In other words, the perplexity $\mathcal{PP}$ that would be obtained via the likelihood is also the inverse of the geometric mean of the bets placed on the correct words, when the bets are computed as in equation (1).

A major feature of this scheme is that it is applicable to any kind of probabilistic model (i.e as long as the conditional probability of equation (1) can be derived from the model parameters). In the particular case when the model is an N-gram model, equation (1) simplifies into the conditional N-gram probability :

$$\beta_k(v_j) = p(v_j \mid w_{k-N+1} \ldots w_{k-1}) \quad (4)$$

4. ADAPTATION OF THE SHANNON GAME

In practice, the gambling scheme presented in the previous section must be adapted to a comparative evaluation campaign in several respects.

4.1. Tackling with out-of-vocabulary words

The computation of the perplexity estimate as in equation (3) poses no problem with a closed vocabulary. With an open vocabulary, a certain fraction of the bets must be reserved for the possibility that the word to be predicted is outside of the vocabulary (OOV), in order to avoid a null term in the geometric mean. To deal with this issue, the vocabulary $\mathcal{V}$ must include a particular *OOV-word* entry (v_{oov}) and the participants are requested to bet on this possibility as well. This implies that the participants must model the entry v_{oov}. This is the approach that was adopted for our experiments.

4.2. Randomization of the truncated sentences

For obvious reasons, it is preferable to select each truncated sentence in such a way that its linguistic content does not overlap with the one of the other truncated sentences. Therefore, the test data $\mathcal{W}$ can consist of a set of distinct sentences, each of them being truncated at a random position. The participant has to predict the word that comes next only. The approach can be generalised to truncated paragraphs instead of truncated sentences, for models using a long history. Some of the experiments reported in this paper compare perplexity scores obtained with these *randomized truncations* as opposed to those yielded by the original Shannon game (i.e *exhaustive truncations*).

4.3. Limitation of the list of candidates

In order to control the volume of data that has to be handled for the evaluation, the number of candidates for each truncated sentence can be set to a maximum $\ell \leq m$. In the case $\ell < m$, the sum A_k of the bets put on the limited list must be such that :

$$A_k = \sum_{r=1}^{r=\ell} \beta_k(u_r) < 1 \quad (5)$$

where u_r denotes the r^{th} candidate, after ordering them in decreasing bet. Then, if the correct word w_k is not in the ordered list of candidates $\mathcal{U} = (u_r)_{1 \leq r \leq \ell}$, the corresponding term in equation (3) can be set to a floor value :

$$\beta_k^* = \frac{1 - A_k}{m - \ell} \quad (6)$$

This approach assumes a uniform repartition of the unassigned fraction of the bets $(1 - A_k)$ over all non-candidate words (i.e the $m - \ell$ words ranked beyond ℓ). It can easily be shown that, if the ℓ candidates in $\mathcal{U}$ correspond indeed to the ℓ best hypotheses yielded by the model (in terms of conditional probabilities), then β_k^* must be smaller than $\beta_k(u_\ell)$. In other words, the floor value is always smaller than the smallest bet in $\mathcal{U}$. Conversely, a systematic test of the condition $0 < 1 - A_k \leq (m - \ell) \times \beta_k(u_\ell)$ signals, if not satisfied, an erroneous answer from a participant. Experimental results on the impact of this *rank limitation* on the perplexity estimation are given in the next section.

5. EXPERIMENTS AND RESULTS

In order to validate the scheme on a real case, we compare the perplexity value obtained with a particular language model when computed from the test set likelihood on the one hand, and from the gambling scheme on the other hand. We investigate on the impact of several variants of the gambling scheme : exhaustive versus randomized truncations and full list versus rank limitation.

For all the experiments, the language model is a bi-gram model, for which the conditional probability $p(v_j|v_i)$ is estimated as :

$$p(v_j|v_i) = \frac{max\{C(v_i v_j) - \alpha, 0\} + \alpha \frac{D(v_i *) D(* v_j)}{D}}{C(v_i)} \quad (7)$$

where $C(v_i)$ and $C(v_i v_j)$ denote respectively the number of occurrences of the word v_i and of the bigram $v_i v_j$, $D(v_i *)$ and $D(* v_j)$ the number of distinct bigrams respectively starting by v_i and finishing by v_j, D the total number of distinct bigrams, and α a smoothing parameter between 0 and 1. All parameters are estimated on the training set, except α which is optimized *a posteriori* so as to minimize the perplexity on the test set ($\alpha = 0.7$).

The conditional probabilities are learned on the articles from the newspaper *Le Monde* covering years 1987 and 1988 (41 257 584 words). The vocabulary $\mathcal{V}$ is composed of the 20 000 most frequent words in the training corpus, the OOV-word, plus 2 symbols indicating the beginning and the end of sentences, which yields $m = 20\,003$. The punctuation is removed from the data.

The evaluation corpus is composed of $n = 1\,707\,527$ words ($\approx 67\,000$ sentences) corresponding to excerpts of *Le Monde* sampled over the years 1989 and 1990 (the texts of the BREF speech database). Among these, 105 263 occurrences are not covered by the vocabulary $\mathcal{V}$ (i.e 6.16 % of the total test corpus). On these test data, the perplexity obtained in the conventional way, i.e using the likelihood function of the model, is equal to 159.77.

In a first series of experiments we compute estimates of the likelihood using the original Shannon game (i.e exhaustive truncations) under a gambling scheme with bets obtained as in equation (1). We consider 20 draws of n_t consecutive words in the corpus from which we compute 20 distinct estimations of the entropy ($Q = \log \mathcal{PP}$). From their average $\overline{Q}$ and their standard deviation σ, we calculate $\overline{\mathcal{PP}}^g = e^{\overline{Q}}$, and :

$$\mathcal{PP}^-_{0.95} = \overline{\mathcal{PP}}^g \, e^{-1.96\sigma} \qquad \mathcal{PP}^+_{0.95} = \overline{\mathcal{PP}}^g \, e^{+1.96\sigma} \quad (8)$$

As can be seen in Table 1, the exhaustive Shannon game yields, in the average, an estimation of the perplexity which is comparable to the likelihood-based perplexity computed on the entire test set (159.77).

In a second set of experiments, we use a random selection of truncated sentences obtained as follows : we draw one word every s words in the test corpus, we extract the corresponding sentence, and we truncate it just before the selected word. Provided that s is larger than the length of the longest sentence, this process produces $n_t = n/s$ randomized truncations on distinct sentences. If different positions for the first word are chosen, several sets of identical size n_t can be generated.

The results obtained for various numbers of truncated sentences and reported in the second half of Table 1 illustrate the accuracy of the estimation as a function of n_t. It can be seen from this table that all average values $\overline{\mathcal{PP}}$ approximate consistently the likelihood-based perplexity, but that the confidence interval for one draw is much narrower with the randomized method as the test corpus is sampled in a more uniform manner.

In a third set of experiments, we vary the parameter ℓ, i.e the rank beyond which we assume a uniform distribution of the bets that have not been allocated to the ℓ first candidates. We compute the perplexity with these approximate bets using the Shannon game on exhaustive truncations ($n_t = n$). Table 2 illustrates the effect of the rank limitation in our experiments : beyond $\ell = 5\ 000$, the perplexity $\widehat{\mathcal{PP}}$ estimated from the limited list becomes very close to the actual perplexity. On the opposite, when ℓ is too small, $\widehat{\mathcal{PP}}$ overestimates the actual perplexity (see, for an illustration, Figure 1).

6. NON-PROBABILISTIC MODELS

For non-probabilistic language models, there is no simple way to choose the bets in the gambling scheme. However, we can assume that the model is able, in the framework of the Shannon game, to rank the words in decreasing order.

It is shown in [1] that entropy bounds can be obtained from the distribution of the rank at which the correct word is predicted by the model. In other words, an ordered list of candidates for each truncated sentence is sufficient to compute an interval in which the *perplexity* of the non-probabilistic language model falls.

With ρ_k denoting the rank at which the model outputs the (correct) word w_k for the truncated sentence W_0^{k-1}, and q_r the frequency with which the model outputs the proper word at rank r ($q_{m+1} = 0$), Shannon shows that the entropy Q of the language model is bounded by :

$$\sum_{r=1}^{r=m} r\,(q_r - q_{r+1}) \log r \;\leq\; Q \;\leq\; -\sum_{r=1}^{r=m} q_r \log q_r \qquad (9)$$

In case the values of q_r are available only as far as rank ℓ (limited list), a uniform distribution for the ranks beyond ℓ can be assumed. If we denote :

$$S_\ell = \sum_{r=1}^{r=\ell} q_r \quad \text{and} \quad t_\ell = \frac{1 - S_\ell}{m - \ell} \qquad (10)$$

the bounds in equation (9) become :

$$\widehat{Q}_{inf} = \sum_{r=1}^{r=\ell-1} r(q_r - q_{r+1}) \log r \;+\; \ell(q_\ell - t_\ell) \log \ell \;+\; m t_\ell \log m \qquad (11)$$

$$\widehat{Q}_{sup} = -\sum_{r=1}^{r=\ell} q_r \log q_r \;-\; (m - \ell) t_\ell \log t_\ell \qquad (12)$$

The set $\{q_r\}_{r>\ell}$ can also be estimated from the extrapolation of a Zipf law [5], $q_r = \lambda / r$, the parameter λ being calculated in order to fit the beginning of the distribution ($\{q_r\}_{1 \leq r \leq \ell}$). In our experiement, this approach is quite consistent with the observed shape of the rank histogram (see Figure 2).

With the same bigram model as the one used in our previous experiments, we have estimated the lower and upper perplexity bounds from Shannon's equations, as well as the approximations with several values of ℓ, with both the uniform tail model and the Zipf extrapolation. The results are given in Table 2.

With no truncation of the rank histogram, the lower bound lies very far below the real perplexity, and it may happen, in a real evaluation, that the interval is so large that it does not yield any conclusion in the comparison between two models. However, if the upper perplexity bound for a non-probabilistic language model was to fall under the perplexity of a probabilistic language model, some meaningful conclusion could be drawn from this observation. With rank truncation, the Zipf extrapolation seems much more robust than the uniform extrapolation.

7. CONCLUSIONS

Our adaptation of the Shannon game to the evaluation of language models within a comparative test campaign offers a fair way to estimate perplexity in a single site, yet without requiring a software re-implementation for each tested model. Beyond this primary advantage, the use of the gambling approach provides additional diagnostic possibilities, such as evaluating the discordance between models, scoring the model for particular classes of words or contexts or being able to better predict the performance of the model within a speech recognition system (given an acoustic confusion matrix). Our work also underlines the possibility to compare probabilistic and non-probabilistic language models in terms of perplexity but additional efforts are required to increase the efficiency of the estimation.

References

[1] C. E. SHANNON : Prediction and entropy of printed English. *Bell Syst. Techn. J., pp. 50-64, Jan. 1951.*

[2] T. M. COVER, R. C. KING : A convergent gambling estimate of the entropy of English. *IEEE Trans. on Information Theory, vol. 24, n. 4, pp. 413-421, July 1978.*

[3] F. JELINEK : Self-organized language modeling for speech recognition. *In Readings in Speech Recognition, A. Waibel & K. F. Lee Ed., Morgan Kaufmann Publishers, pp. 450-506, 1990.*

[4] P. O'BOYLE, M. OWENS, F.J. SMITH : A weighted average n-gram model of natural language. *Computer Speech and Language (1994) 8, 337-349.*

[5] G. K. ZIPF : Human Behavior and the Principle of Least Effort. *Addison-Wesley Press, 1949.*

k		0	1		2		3		4		5	
w_k		<	une		société		d'		économie		mixte	
u_1	$\beta_k(u_1)$		le	0.070	v_{oov}	0.049	v_{oov}	0.117	un	0.142	et	0.113
u_2	$\beta_k(u_2)$		il	0.054	nouvelle	0.019	de	0.081	une	0.137	>	0.092
u_3	$\beta_k(u_3)$		la	0.052	fois	0.013	>	0.079	v_{oov}	0.056	de	0.071
u_4	$\beta_k(u_4)$		les	0.050	autre	0.012	française	0.040	autres	0.027	des	0.046
u_5	$\beta_k(u_5)$		l'	0.035	telle	0.011	*d'*	*0.038*	être	0.025	française	0.043
u_6	$\beta_k(u_6)$		v_{oov}	0.028	partie	0.011	qui	0.029	état	0.019	mondiale	0.032
u_7	$\beta_k(u_7)$		mais	0.028	grande	0.010	et	0.026	avoir	0.014	v_{oov}	0.029
u_8	$\beta_k(u_8)$		en	0.026	certaine	0.009	civile	0.023	autre	0.013	américaine	0.027
u_9	$\beta_k(u_9)$		à	0.022	politique	0.008	américaine	0.017	affaires	0.011	du	0.026
u_{10}	$\beta_k(u_{10})$		c'	0.022	*société*	*0.008*	des	0.015	entre	0.007	*mixte*	*0.021*
ρ_k	$10^5 \times \beta_k^*$		18	*3.1*	*10*	4.3	*5*	2.7	89	*2.7*	*10*	2.5

k		6		7		8		9		
w_k		sera		créée		fin		janvier		>
u_1	$\beta_k(u_1)$	>	0.101	pas	0.061	en	0.319	de	0.392	
u_2	$\beta_k(u_2)$	paritaire	0.100	v_{oov}	0.049	par	0.193	du	0.128	
u_3	$\beta_k(u_3)$	de	0.078	le	0.034	à	0.076	d'	0.066	
u_4	$\beta_k(u_4)$	v_{oov}	0.041	de	0.025	pour	0.041	mille	0.061	
u_5	$\beta_k(u_5)$	qui	0.037	la	0.023	il	0.040	des	0.055	
u_6	$\beta_k(u_6)$	d'	0.032	l'	0.018	>	0.039	à	0.054	
u_7	$\beta_k(u_7)$	du	0.032	plus	0.018	au	0.032	>	0.023	
u_8	$\beta_k(u_8)$	franco	0.031	t-il	0.016	le	0.026	au	0.012	
u_9	$\beta_k(u_9)$	et	0.023	en	0.016	dans	0.025	juin	0.011	
u_{10}	$\beta_k(u_{10})$	des	0.023	un	0.015	et	0.016	septembre	0.010	
ρ_k	$10^5 \times \beta_k^*$	35	*2.5*	106	*3.6*	37	*1.0*	20	*0.9*	

Figure 1. An illustration of perplexity estimation via the gambling scheme, for the french sentence : *une société d'économie mixte sera créée fin janvier.* **Exhaustive truncations, list limitation to $\ell = 10$. For this sentence, the perplexity estimate is obtained as the inverse of the geometric mean of the number in italics, namely yielding** 5094 **(see section 4.3).** *Here, as $\ell \ll m$, this figure overestimates considerably the true perplexity (174).*

n_t	1 000	2 000	5 000	10 000	20 000	50 000
$\overline{\mathcal{PP}}^g$	162.38	162.25	162.08	160.61	158.37	159.79
$\mathcal{PP}^-_{0.95} \mid \mathcal{PP}^+_{0.95}$	109.09 | 241.71	120.68 | 218.14	138.88 | 189.16	140.44 | 183.68	139.24 | 180.13	149.76 | 170.50
$\overline{\mathcal{PP}}^g$	160.90	158.57	160.06	158.27	158.80	159.99
$\mathcal{PP}^-_{0.95} \mid \mathcal{PP}^+_{0.95}$	130.52 | 198.34	144.85 | 173.59	148.64 | 172.80	152.61 | 164.13	152.73 | 165.12	156.83 | 163.20

Table 1. Geometric mean ($\overline{\mathcal{PP}}^g$) and 95 % confidence bounds ($\mathcal{PP}^-_{0.95} \mid \mathcal{PP}^+_{0.95}$) computed from 20 perplexity estimations, using the Shannon game (gambling scheme) for various set sizes (n_t). Top : exhaustive truncations. Bottom : randomized truncations. *These figures illustrate the validity of the proposed alternative scheme for perplexity estimation : both estimates converge towards the conventional perplexity value (159.77). The second estimate is more accurate as the randomized truncations are more representative of the entire corpus.*

ℓ	1	10	100	1 000	5 000	10 000	20 000
$\widehat{\mathcal{PP}}$	7349.12	1098.54	285.72	173.99	160.50	159.85	159.77
$\widehat{\mathcal{PP}}_{inf}$	5160.64	498.30	97.03	53.21	49.11	48.90	48.87
$\widehat{\mathcal{PP}}_{sup}$	7688.94	1228.87	332.11	204.31	190.05	188.95	188.05
$\widehat{\mathcal{PP}}^Z_{inf}$	49.53	37.57	39.71	46.16	48.41	48.78	48.87
$\widehat{\mathcal{PP}}^Z_{sup}$	160.51	113.92	123.40	159.07	179.65	185.11	188.05

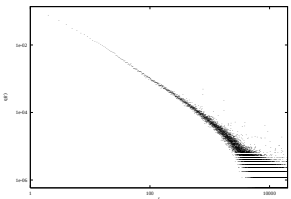

Figure 2. The log-log histogram of $\{q_r\}$, for the bigram model ($\approx$ *Zipf law*).

Table 2. Comparison various perplexity estimates with several rank truncations (ℓ). Top : Shannon game - gambling scheme. Center : Shannon game, lower and upper bounds - uniform model. Bottom : Shannon game, lower and upper bounds - Zipf model. *For this experiment, beyond a list size of 5 000, the estimated perplexity by the gambling scheme becomes very close to the exact perplexity. For the Shannon game without gambling, the lower bound falls far under the actual perplexity. The Zipf model seems to be more robust to truncation than the uniform model.*