



Cognitive Workload and Vocabulary Sparseness: Theory and Practice

Ron M. Hecht^{1,2}, Aharon Bar-Hillel³, Stas Tiomkin¹, Hadar Levi¹, Omer Tsimhoni², Naftali Tishby¹

¹ The Hebrew University of Jerusalem, Israel

² Advance Technical Center Israel, General Motors, Israel

³ Microsoft Research, Israel

hadaron@gmail.com, aharonb@microsoft.com, stasti@gmail.com, hadarl@gmail.com
omer.tsimhoni@gm.com, tishby@cs.huji.ac.il

Abstract

In this work we present a Language Model (LM) that accounts for the effects of speaker workload by drawing on recent findings in cognitive psychology research. Under workload, speakers tend to shorten their utterances, but still aim to convey their message; hence they use more informative words. Inspired by the Perception and Action Cycle Method (PACM), the LM is used as a baseline dictionary that is constrained to have higher entropy. We show that the resulting LM has a power law relation to the baseline dictionary; i.e., there is a linear relation between the word log-probability under workload and its baseline log-probability. We then test for the existence of this relation in transcriptions of audio text messages (SMS) dictated while driving under different workload conditions. Significance tests were conducted using Monte Carlo simulations, with the data modeled by principal component analysis (PCA) and linear regression (LR). Based on this power law, we suggest a simple algorithm for LM adaptation under workload. Experiments show encouraging results in perplexity improvement of the LM under workload, thus providing empirical support for our model.

Index Terms: language model, speaker workload.

1. Introduction

Studies in cognitive psychology have made important connections between mental workload and human attention patterns (Lavie et al. in a series of papers [1,2]). At low workload, human attention is spread over many stimuli, some of which are not task related. This attention spread constitutes the natural comfort zone behavior of a human agent. Under high workload conditions, however, individuals tend to focus most of their attentional resources on the stimuli that are most relevant for the task. Most studies in this area have focused on human visual perception, but a clear analogy can be drawn with speech behavior. In the latter case, we are dealing with choice of words rather than perceptual actions, but a similar phenomenon can be observed: workload moves the speaker from a baseline word distribution to a more task-oriented distribution. In this short manuscript we quantify this effect in theory and practice.

An analogous phenomenon has been addressed in the field of Reinforcement Learning (RL)[3] in studies on reward maximization under complexity constraints [4,5,6,7]. Specifically, the Perception and Action Cycle Method (PACM) [6,7] explores this tradeoff by introducing a baseline ‘comfort zone’ action distribution. A policy (behavior pattern) is considered ‘simple’ if it is similar to the comfort zone policy. The agent’s behavior is determined as a tradeoff between reward maximization and policy simplicity; i.e., its distance from the agent’s comfort zone. In this analogy, high workload

corresponds to high reward demands, leading the agent away from his/her comfort-zone policy into goal-oriented behavior.

We consider a Language Model (LM) shift under speaker workload. Recent work [8] has shown that under workload conditions the speaker LM changes significantly, and specifically that the usage probability of common words is reduced [9]. We address these empirical findings as a specific case of the cognitive phenomena observed in [1,2], and use the PACM computational model introduced in [4,5] to quantify it. Following [1,2,4,5], we conjecture that the human agent has a ‘comfort-zone’ LM, and that s/he diverges from it under the pressure of cognitive load, formulated as a reward-rate constraint. Specifically, high cognitive load constrains the number of utterances the agent uses, and thus each word needs to carry more reward. We conjecture that the reward, in the case of a language model, is related to word information; i.e., its negative log probability. Hence high cognitive load requires usage of a LM with higher entropy: the less informative words, which are the frequent ones, decline in frequency and the more informative words, which are rare, increase.

We suggest a simple model for this human LM shift under workload, and show empirical evidence with respect to the prediction of word selection. We formalize the LM shift as a constrained optimization problem, where the agent wishes to stay close to the baseline LM, but increase its entropy. The problem is convex, with the solution indicating a power law relation between LM with and without the workload constraint. We show empirically that this power-law behavior is exhibited by drivers subject to a driving workload. Finally, we show that language model correction according to the workload condition improves the perplexity of a unigram LM.

2. Model

The formalism we suggest to analyze the effect of workload on LM is inspired by PACM, but can be simply stated without this background. The organism generates a sequence of actions from a discrete set A , and we denote by π the organism’s policy, where π is a probability function over $a \in A$. It is assumed that a reward is associated with each action executed, denoted by R_a . The average value of a policy π is:

$$V^\pi = \sum_{a \in A} \pi(a) R_a \quad (1)$$

The organism strives to behave i.e. selects a policy that is as close as possible to a baseline distribution $\hat{P}(a)$. The similarity between the selected policy π and the baseline policy \hat{P} is estimated by the Kullback-Leibler Divergence [10].

$$D_{kl}[\pi || \hat{P}] = \sum_{a \in A} \pi(a) \log \frac{\pi(a)}{\hat{P}(a)} \quad (2)$$

The organism selects its behavior according to two contradictory forces: maximization of reward and similarity to baseline distribution. This tradeoff can be presented as a minimization of the $D_{kl}[\cdot || \cdot]$ subject to a reward constraint:

$$\min_{\pi} \sum_{a \in A} \pi(a) \log \frac{\pi(a)}{\hat{P}(a)} \quad (3)$$

s.t.
 $\sum_{a \in A} \pi(a) R_a \geq \theta$
 $\sum_{a \in A} \pi(a) = 1$

where θ is the required reward level.

A strategy under this model has one degree of freedom – the change in θ . Intuitively, the organism selects a behavior that is as similar as possible to $\hat{P}(a)$, while maintaining at least a reward level of θ .

2.1. Log Likelihood Based Reward

If we select the reward function to be minus the log likelihood, the constraint becomes entropy- based.

$$\min_{\pi} \sum_{a \in A} \pi(a) \log \frac{\pi(a)}{\hat{P}(a)} \quad (4)$$

s.t.
 $\sum_{a \in A} -\pi(a) \log \pi(a) \geq \theta$
 $\sum_{a \in A} \pi(a) = 1$

In this formulation, the $D_{kl}[\cdot || \cdot]$ is convex in π , whereas the entropy constraint creates a convex set. Hence in these conditions there is a single global optimum which can be found by equating the Lagrangian to zero and using the Kuhn-Tucker conditions. The solution to this equation takes the following form:

$$\log \pi(a) = \begin{cases} \frac{1}{1+\beta} \log \hat{P}(a) - \frac{1+\beta+\lambda}{1+\beta} & \theta \geq H(\hat{P}) \\ \log \hat{P}(a) & otherwise \end{cases} \quad (5)$$

where β and λ are the Lagrange multipliers of the first and second constraints, $H(\hat{P}) = \sum_a \hat{P}(a) \log \hat{P}(a)$. We see that the solution retains $\pi(a) = \hat{P}(a)$ if possible, and if the constraint does not allow it, a power law emerges.

$$\pi(a) \propto \hat{P}(a)^{\frac{1}{1+\beta}} \quad (6)$$

An important property of this solution is the linear relation between the log likelihood of both policies expressed in Eq. 5. This phenomenon can be easily tested on datasets.

2.2. Unknown Reward

In Eq. 5 there is a linear relation between the log probabilities of π and \hat{P} . Hence if we solve for the policy π using two different thresholds θ_1, θ_2 , we will get a linear relation between the log probabilities of the two corresponding solutions π_1, π_2 , i.e., $\log \pi_1 = a \log \pi_2 + b$. As shown below, this relation can also be obtained even without limiting ourselves to an entropy-based reward, if we assume a uniform baseline distribution \hat{P} .

Solving the problem in Eq. 3 for a general reward, the solution for the case of $\theta \geq \sum_a \pi(a) R_a$ takes the form [5]:

$$\pi(a) = \frac{\hat{P}(a) e^{\beta R_a}}{z(\beta)} \quad (7)$$

where β is a Lagrange multiplier and $z(\beta)$ is the normalization factor (partition function):

$$z(\beta) \equiv \sum_{a \in A} \hat{P}(a) e^{\beta R_a} \quad (8)$$

This solution requires explicit knowledge of the reward function. Unfortunately in some real world tasks, the reward is unknown. Nevertheless, a relation between the solutions obtained for two θ values can be established. Rewriting Eq. 7 we get

$$\log \frac{\pi(a)}{\hat{P}(a)} = \beta R_a - \log z(\beta) \quad (9)$$

For two θ values θ_1, θ_2 , the corresponding policies π_1, π_2 preserve the relation:

$$\frac{\log \frac{\pi_1(a)}{\hat{P}(a)} + \log z(\beta_1)}{\beta_1} = R_a = \frac{\log \frac{\pi_2(a)}{\hat{P}(a)} + \log z(\beta_2)}{\beta_2} \quad (10)$$

And by rearranging the equation, the linearity emerges:

$$\log \frac{\pi_2(a)}{\hat{P}(a)} = \frac{\beta_2}{\beta_1} \log \frac{\pi_1(a)}{\hat{P}(a)} + \frac{\beta_2}{\beta_1} \log z(\beta_1) - \log z(\beta_2) \quad (11)$$

If we further assume that $\hat{P}(a)$ is a uniform distribution, we can expect a linear relation between $\log \pi_2(a)$ and $\log \pi_1(a)$.

$$\log \pi_2(a) = \frac{\beta_2}{\beta_1} \log \pi_1(a) + \frac{\beta_2}{\beta_1} \log z(\beta_1) - \log z(\beta_2) \quad (12)$$

We suggest testing for linearity between the log probability of π_1, π_2 as a way to validate PACM-like behavior.

3. Workload

We work along the lines presented by Lavie, and suggest the PACM as a specific mechanism that captures the tradeoff between two behavior patterns: the pure intentional goal oriented pattern π and the unintentional exploring pattern \hat{P} .

We assume two workload conditions: low and high, and their corresponding strategies π_{low} and π_{high} . We assume that a higher workload requires a higher average reward: $\theta_{low} < \theta_{high}$. Intuitively, in cases where there is no rush, an organism can perform a large set of actions, of which only a few will yield a significant reward. However, under high workload conditions, the organism performs fewer actions, and hence a higher reward level per action is required to maintain performance. This logic also applies to secondary actions performed during high workload on a primary task. In our experiments the primary task was driving, and the secondary task was SMS dictation. A high driving workload leaves fewer cognitive resources for SMS dictation, causing a Language Model shift.

4. Approach Validation

We applied the machinery and methods presented in the previous sections to analyze the speech interaction of drivers under different workload conditions. Specifically, we tested the linearity properties suggested by the formalism on text messages (Short Message Service - SMS) dictation while driving. Two linearity tests were carried out on the data: Principal Component Analysis (PCA) and Linear Regression (LR). We conducted significance tests using Monte Carlo simulations. The two workload conditions were dictation while driving along a straight road, and dictation while in parked

position, idling. The actions were the words uttered by the driver. We assumed that the words (actions) were generated i.i.d. (Unigram) and that for each word there was an associated reward. Since our goal was to validate a mechanism and not to boost performance, we selected the simplest model; namely Unigram. We tested the linearity in terms of the significance of the phenomenon and as a method to improve recognition performance; namely, – perplexity.

4.1. Corpus and data collection

The corpus was collected by Green et al [11] at the University of Michigan Transportation Research Institute Library (UMTRI) driving simulator. The UMTRI simulator is based on a real cab system. The subjects were twenty four pairs of friends, for purposes of familiarity between subjects. Subjects were recruited in a balanced demographic. All pairs participated in all the workload conditions (repeated measures). The order of the conditions was counter-balanced. During the experiment, each pair of subjects was separated. One was seated in the cab in the simulator room and the other in a different room. The subjects could not hear or see one another. Communication between the subjects was not direct, but rather by proxy, as shown in the following figure.

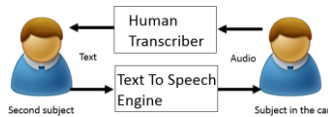


Figure 1: Experiment setup.

The first message was dictated by the subject in the cab. This message was a pre-selected one, that was intended to reduce topic and vocabulary diversity. (This message, for obvious reasons, was eliminated from the final corpus). The message was transcribed by a human transcriber and passed on as a text message to the second subject. Next, the second subject sent a text message response that was converted to speech by a Text To Speech (TTS) system and played to the subject in the cab. Then the subject in the cab dictated an answer and the cycle repeated itself several times. After several messages, the cycle was interrupted and a new pre-selected message was presented. The process was repeated both for the parked and driving scenarios. Overall, a set of more than thousand messages was collected:

Table 1. UMTRI corpus size

Workload scenario	Number of messages
Driving	580
Parked	570

4.2. Results

Our model predicts a linear relation between $\log \pi_{parked}(w)$ and $\log \pi_{driving}(w)$ for most words w . We limited ourselves to words that appeared at least ten times in both scenarios (yielding a set of 81 words). The log probabilities of these words are presented in Figure 2. The X and Y axis are the log probability according to the parked and driving models. The solid black line is the H_0 hypothesis in which the workload does not change the probability the word will appear. Under the H_0 hypothesis the slope is 1. The red dashed line is a Linear Regression (LR) regressing Y from X.

Each word is represented by a dot. The center of the dot corresponds to the word probability under the ‘parked’ and ‘driving’ conditions. It is clear that the slope is different from 1. The slope was estimated to be 0.83 and the R^2 is about 0.8.

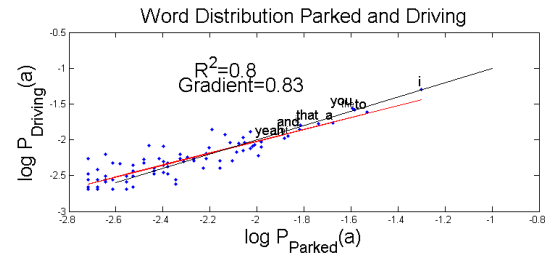


Figure 2: Log-likelihood of parked to log-likelihood of driving plot. Blue represents the observed data. Black and red lines represent H_0 and the linear regression.

4.3. Linear Regression - Monte Carlo Significance Test

In this and the following sections we tested whether the observed slope was significantly different from the null hypothesis H_0 . We used a Monte Carlo method. First, the data from both scenarios were pool and randomly divided into two fictitious scenarios. Then, the slope between the scenarios was estimated using linear regression, as shown in Figure 2. This process was repeated 1000 times and the histogram of the slope values is presented in Fig 3.

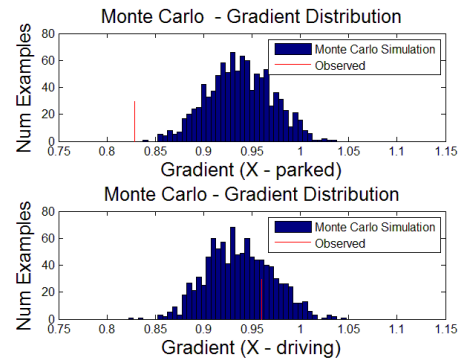


Figure 3: Observed linear regression slope, and the histogram of slope values estimated by the Monte Carlo method.

This figure presents the histogram of slopes obtained in two linear regressions estimations. In one, the parked model was on the X axis and in the other it was on the Y axis. The upper figure indicates high significance; and the lower is not significant. Despite the non-symmetric nature of the LR and the fact that the data are not fully linear, in both cases the slope was less than 1. To better understand these findings, we ran additional tests.

4.4. Monte Carlo PCA – Significance Test

To confirm these findings, the slope was estimated by Principal Component Analysis (PCA) as the direction of the first eigenvector. The results are shown in Figure 4. As expected, this time the average of the MC distribution was around 1. In Figure 3, the red line represents the observed slope. The advantage of this approach is its symmetry as compared to LR.

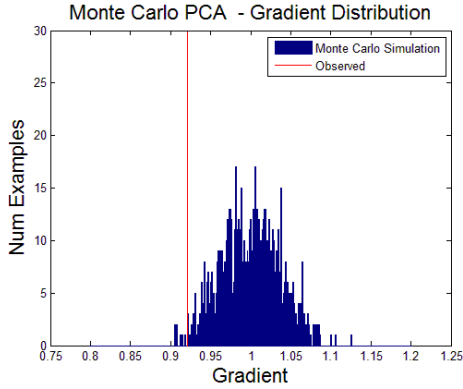


Figure 4: Observed PCA slope and histogram of the slope values estimated by the Monte Carlo method

Thus, in the last two sections two very different approaches supported our finding with high level of significance. The test that failed to reach significance is the one where the probability of a word while driving was used to predict its probability at rest.

5. Language Model Experiment

In this section, we examine whether the observed workload behavior can be measured by the known metric of Language Model (LM) perplexity.

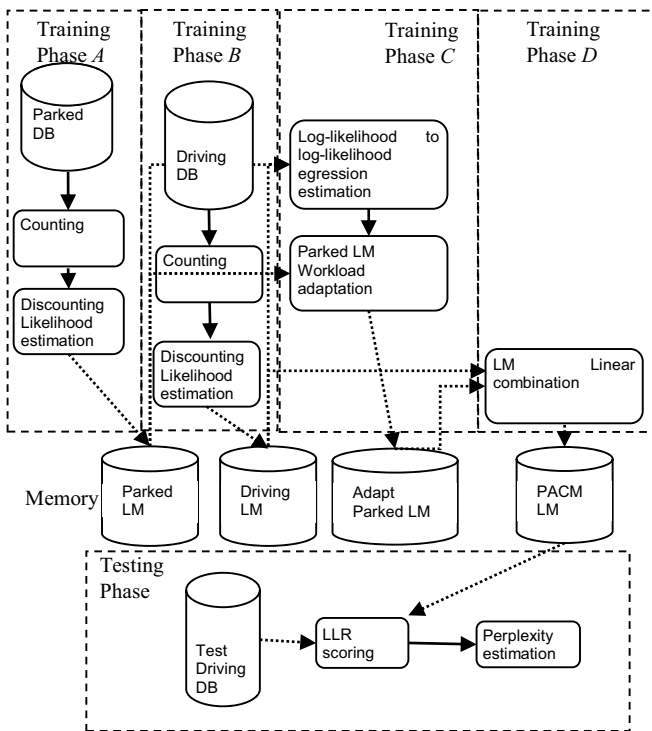


Figure 5: PACM based LM adaptation system.

5.1. Experimental setup

The experimental setup is depicted in Figure 5. It is composed of four training stages and a single test stage. We conducted a k-fold (k=20) jackknife procedure to exploit all of the data and estimate the standard deviation of our LM prediction

performance. The data for the experiment were divided into three sets: the parked data – used only for training, most of the driving data which was used for training as well, and the remaining driving data that was reserved solely for testing.

During stage A, a unigram LM was estimated based on the transcribed text collected from participants while seated in the parked car. We refer to this as the parked LM. To evaluate statistical significance, we only examined words that appeared at least ten times in both the parked and driving corpora, yielding a set of 81 words. Other words were omitted. This stage was made up of two parts: counting and likelihood estimation (discounting). In stage B, an LM was estimated from the main part of the driving data. Training was similar to stage A. A model workload adaptation based on the PACM approach was run in stage C. First the parameters of a linear regression between the log likelihoods of words according to the parked and driving LMs were estimated. The parked LM was the X axis and the driving LM was on the Y axis. Based on the parameters the likelihood of each word to appear was re-estimated:

$$\log P_{adapted\ parked}(a) = \frac{1}{z} (m \log P_{parked}(a) + n) \quad (13)$$

where z is a normalization factor so that the probabilities sum to one.

In Stage D a linear combination of the driving LM and adapted PACM parked LM were combined. To avoid overfitting we restricted ourselves to an equal weight combination. In the testing stage, the PACM LM was tested on the reserved driving data.

5.2. Results

The PACM LM was compared to the other LMs: the parked LM (estimated in stage A), the driving LM (estimated in stage B) and the linear combination of the parked and driving LMs (without the adapted PACM). In this combined linear model, we tested for the optimal weighting of the two to make sure that the linear combined model used its full potential. The perplexity performance of all the tested systems are shown in Table 2. The standard deviation of the perplexity differences between the linear combined and the PACM was estimated over the different jackknife experiments and is presented in the table as well.

Table 2. Perplexity performance of the tested models

Training Method	Perplexity
(1) Driving	68.8
(2) Parked	70.5
(3) linear combined	68.6
(4) PACM PCA	68.3
(5) PACM LR	67.9
Std of Δ (3)-(5)	0.03

Both PACM based approaches outperformed the three other approaches significantly.

6. Conclusion and Future Work

The PACM approach proved itself to be an effective tool to model the change in language models under workload conditions. Two types of significance tests support the findings, as well as the perplexity reduction in the estimate of a unigram LM. In further work, we plan to extend the finding to larger corpora and more complex language models.

7. References

- [1] Lavie, N., "Attention, distraction, and cognitive control under load" *Current Directions in Psychological Science*, 19(3), 143-148, 2010.
- [2] Cartwright-Finch, U., & Lavie, N. "The role of perceptual load in inattention blindness". *Cognition*, 102(3), 321-340, 2007.
- [3] Sutton, R. S., & Barto, A. G. "Introduction to reinforcement learning". MIT Press, 1998.
- [4] Tishby, N., & Polani, D., "Information theory of decisions and actions". In *Perception-Action Cycle* (pp. 601-636). Springer New York, 2011.
- [5] Rubin, J., Shamir, O., & Tishby, N., "Trading value and information in MDPs". In *Decision Making with Imperfect Decision Makers* (pp. 57-74). Springer Berlin Heidelberg, 2012.
- [6] Tishby, N., Pereira, F. and Bialek, W., "The Information Bottleneck Method", The 37th annual Allerton conference on communication, control and computing, 1999.
- [7] Chechik, G., Globerson, A., Tishby, N., & Weiss, Y., "Information Bottleneck for Gaussian Variables", *Journal of Machine Learning Research*, 6(1), 165-188, 2005.
- [8] Hecht, R. M., Tzirke E., and Tsimhoni, O., "Adjusting Language Models for Text Messaging based on Driving Workload", *Proceedings of Applied Human Factors and Ergonomics Conference*, 2012.
- [9] Gasic M., Tsiakoulis, P., Henderson, M., Thomson, B., Yu, K., Tzirke E. and Young, S. "The effect of cognitive load on a statistical dialogue system." *SigDial* 2012,
- [10] Cover, T. M., & Thomas, J. A., "Elements of information theory" John Wiley & Sons, 2012.
- [11] Green, P. A., Lin, B., Kang, T., and Best, A., "Manual and speech entry of text messages while driving", Technical Report UMTRI-2011-47, University of Michigan Transportation Research Institute, 2011.