Learning from Human-Generated Lists

Kwang-Sung Jun, Xiaojin Zhu {DELTAKAM,JERRYZHU}@CS.WISC.EDU
Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI 53706 USA

Burr Settles BSETTLES@CS.CMU.EDU
Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Timothy T. Rogers TTROGERS@WISC.EDU
Department of Psychology, University of Wisconsin-Madison, Madison, WI 53706 USA

Abstract

Human-generated lists are a form of non-*iid* data with important applications in machine learning and cognitive psychology. We propose a generative model — *sampling with reduced replacement* (SWIRL) — for such lists. We discuss SWIRL's relation to standard sampling paradigms, provide the maximum likelihood estimate for learning, and demonstrate its value with two real-world applications: (*i*) In a "feature volunteering" task where non-experts spontaneously generate feature⇒label pairs for text classification, SWIRL improves the accuracy of state-of-the-art feature-learning frameworks. (*ii*) In a "verbal fluency" task where brain-damaged patients generate word lists when prompted with a category, SWIRL parameters align well with existing psychological theories, and our model can classify healthy people vs. patients from the lists they generate.

1. Introduction

We present a probabilistic model describing the human process of ordered list generation. For machine learning, such a model can enhance the way computers learn from people. Consider a user who is training a system to classify sports articles. She might begin by generating a list of relevant phrase⇒label rules (e.g., "touchdown⇒football, home-run⇒baseball"). Incorporating such "feature volunteering" as training data in a machine learning algorithm has been an area of considerable research interest (Druck et al., 2008; Liu et al., 2004; Settles, 2011). Less effort has gone into modeling the *human* process of generating these lists, which (as we show) can be combined with these algorithms for further improvement. For cognitive psychology, such list-generation tasks are used to probe the structure of human memory and to diagnose forms of cognitive impairment. For instance, the "verbal fluency" task requires a subject to generate as many words of a given category (e.g., "vehicle") as possible within 60 seconds (e.g., "car, plane, boat, ..."). Performance on this simple task is highly sensitive to even mild mental dysfunction.

Learning from human-generated lists differs from more familiar machine learning settings. Unlike ranking, the human teacher does not have the collection of items in front of her to arrange in order. Unlike active learning, she is not prompted with item queries, either. Instead, she must *generate* the list internally via memory search. Such lists have two key characteristics:

1. **Order matters.** Items (e.g., vehicle names or phrase⇒label rules) that appear earlier in a list tend to be more "important." This suggests that we can estimate the "value" of an item according to its position in the list. This relation is not deterministic, however, but stochastic. Very important items can appear later or be missing from any single list altogether.

2. **Repeats happen.** Humans tend to repeat items in their list even though they know they should not (see Table 1). Indeed, we will see that people tend to repeat items even when their lists are displayed right in front of them (e.g., Figure 1). Though this is a nuisance for some applications, the propensity to repeat can provide useful information in others.

We propose a new sampling paradigm, *sampling with reduced replacement* (SWIRL), to model human list production. Informally, SWIRL is "in-between" sampling with and without replacement, since a drawn

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

Order	Item
1	baseball bat⇒BS
...	...
7	quarterback⇒F
8	**football field⇒F**
9	soccer ball⇒S
...	...
23	basketball court⇒BK
24	**football field⇒F**
25	soccer field⇒S
...	...

(a) sports

Order	Item
1	research⇒P
...	...
16	**school⇒F**
17	requirement⇒C
18	**grade⇒C**
19	science⇒C
...	...
37	**school⇒F**
38	**grade⇒C**
...	...

(b) webkb

Order	Item
1	yacht
2	**rowboat**
3	paddle boat
4	casino
5	steam liner
6	warship
7	aircraft carrier
...	...
11	clipper ship
12	**rowboat**

(d) boats

Order	Item
1	automobile
2	truck
3	**train**
4	boat
5	**train**
6	airplane
7	bicycle
8	motorcycle
9	minivan
10	bus

(c) vehicles

Table 1. Example human-generated lists, with repeats in bold red. (a) Feature volunteering for sports articles: BS=baseball, F=football, S=soccer, and BK=basketball. (b) Feature volunteering for academic web pages: P=project, F=faculty, and C=course. (c,d) Verbal fluency tasks, from healthy and brain-damaged individuals, respectively.

item is replaced but with its probability discounted. This allows us to model order and repetition simultaneously. In Section 2, we formally define SWIRL and provide the maximum likelihood estimate to learn model parameters from human-generated lists. Though not in closed form, our likelihood function is convex and easily optimized. We present a machine learning application in Section 3: feature volunteering for text classification, where we incorporate SWIRL parameters into down-stream text classifiers. We compare two frameworks: (*i*) Generalized Expectation (GE) for logistic regression (Druck et al., 2008) and (*ii*) informative Dirichlet priors (IDP) for naïve Bayes (Settles, 2011). We then present a psychology application in Section 4: verbal fluency, where SWIRL itself is used to classify healthy vs. brain damaged populations, and its parameters provide insight into the different mental processes of the two groups.

2. Sampling With Reduced Replacement (SWIRL)

Let $\mathcal{V}$ denote the vocabulary of items for a task, and $\mathbf{z} = (z_1, \ldots, z_n)$ be an *ordered* list of n items where $z_t \in \mathcal{V}$ and the z's are not necessarily distinct. The set of N lists produced by different people is written $\mathbf{z}^{(1)} = (z_1^{(1)}, \ldots, z_{n^{(1)}}^{(1)})$, $\ldots$, $\mathbf{z}^{(N)} = (z_1^{(N)}, \ldots, z_{n^{(N)}}^{(N)})$.

We now formally define SWIRL. Assume that humans possess an unnormalized distribution over the items for this task. Let $s_i \geq 0$ be the "initial size" of item i for $i \in \mathcal{V}$, not necessarily normalized. One would select item i with probability proportional to s_i. Critically, the size of the selected item (say, z_1) will be *discounted* by a factor $\alpha \in [0, 1]$ for the next draw: $s_{z_1} \leftarrow \alpha s_{z_1}$. This reduces the chance that item z_1 will be selected again in the future. To make it a full generative model, we assume a Poisson(λ) list length distribution. The process of generating a single list $\mathbf{z}$ is specified in Algorithm 1.

Algorithm 1 The SWIRL Model

Parameters: $\lambda, \mathbf{s} = \{s_i \mid i \in \mathcal{V}\}, \alpha$.
$n \sim \text{Poisson}(\lambda)$
for $t = 1, \ldots, n$ **do**
  $z_t \sim \text{Multinomial}\left(\frac{s_i}{\sum_{j \in \mathcal{V}} s_j} \mid i \in \mathcal{V}\right)$
  $s_{z_t} \leftarrow \alpha s_{z_t}$
end for

2.1. Relation to Other Sampling Paradigms

Setting $\alpha = 1$ recovers sampling with replacement, while $\alpha = 0$ differs subtly from sampling without replacement. Consider an "urn-ball" model where balls have $|\mathcal{V}|$ distinct colors. Let there be m_i balls of color i, each with size s_i. The probability of drawing a ball is proportional to its size. The chance that a draw has color i is $P(\text{color } i) = s_i m_i / (\sum_{j \in \mathcal{V}} s_j m_j)$. We contrast several sampling paradigms:

Sampling without replacement and all colors have the same size $s_i = s_j$. If a draw produces a ball with color i then $m_i \leftarrow m_i - 1$ for the next draw. Let $\mathbf{m} = (m_1, \ldots, m_{|\mathcal{V}|})^\top$ be the counts of balls in the urn and $\mathbf{k} = (k_1 \ldots k_{|\mathcal{V}|})^\top$ be the counts of balls drawn, then $\mathbf{k}$ follows the multivariate hypergeometric distribution $\text{mhypg}(\mathbf{k}; \mathbf{m}, 1^\top \mathbf{k})$.

Sampling without replacement when the sizes may be different. The distribution of $\mathbf{k}$ follows the multivariate Wallenius' noncentral hypergeometric distribution $\text{mwnchypg}(\mathbf{k}; \mathbf{m}, \mathbf{s}, 1^\top \mathbf{k})$, which is a generalization to the multivariate hypergeometric distribution (Wallenius, 1963; Chesson, 1976; Fog, 2008). "Noncentral" means that the s_i's may be different. Note that after drawing a ball of color i, we *subtract* s_i from the "total size" of color i.

SWIRL. Each color has only one ball: $m_i = 1$, but the sizes s_i may differ. Balls are replaced but trimmed: m_i stays at one but $s_i \leftarrow \alpha s_i$. This results in a geometric (rather than a Pólya urn process-style arithmetic)

change in that color's size. We are interested in the probability of the ordered sequence of draws (i.e., $\mathbf{z}$) rather than just a count vector $\mathbf{k}$. The salient differences are illustrated by the following example.

Example (Non-exchangeability)**.** Sampling without replacement *is well-known to be exchangeable. This is not true for SWIRL. Let there be two colors* $\mathcal{V} = \{A, B\}$*. Consider two experiments with matching total size for each color:*

(Experiment 1) There are $m_1 = a$ *balls of color* A *and* $m_2 = b$ *balls of color* B*. Let* $s_1 = s_2 = 1$*, and perform sampling without replacement. Let* z_i *be the color of the* i*th draw. It is easy to show that* $P(z_i = A) = P(z_j = A) = \frac{a}{a+b}; \forall i, j$*.*

(Experiment 2) There is $m_1 = 1$ *ball of color* A *with size* $s_1 = a$*, and* $m_2 = 1$ *ball of color* B *with size* $s_2 = b$*. We perform SWIRL with discounting factor* α*. Then* $P(z_1 = A) = \frac{a}{a+b}$*, but* $P(z_2 = A) = (\frac{\alpha a}{\alpha a+b}\frac{a}{a+b}) + (\frac{a}{a+\alpha b}\frac{b}{a+b})$*. In general,* $P(z_1 = A) \neq P(z_2 = A)$*. For instance, when* $\alpha = 0$ *we have* $P(z_2 = A) = \frac{b}{a+b}$*.*

2.2. Maximum Likelihood Estimate

Given observed lists $\mathbf{z}^{(1)} \dots \mathbf{z}^{(N)}$, the log likelihood is

$$\ell = \sum_{j=1}^{N} n^{(j)} \log \lambda - \lambda + \sum_{t=1}^{n^{(j)}} \log P\left(z_t^{(j)} \mid z_{1:t-1}^{(j)}, \mathbf{s}, \alpha\right),$$

where $n^{(j)}$ is the length of the jth list, and λ is the Poisson intensity parameter. The key quantity here is

$$P\left(z_t^{(j)} \mid z_{1:t-1}^{(j)}, \mathbf{s}, \alpha\right) = \frac{\alpha^{c(z_t^{(j)}, j, t)} s_{z_t^{(j)}}}{\sum_{i \in \mathcal{V}} \alpha^{c(i,j,t)} s_i} \qquad (1)$$

where $c(i, j, t)$ is the count of item i in the jth prefix list of length $t-1$: $z_1^{(j)}, \dots, z_{t-1}^{(j)}$. This repeated discounting and renormalization couples α and $\mathbf{s}$, making the MLE difficult to solve in closed form.

The λ parameter is independent of $\mathbf{s}$ and α, so the MLE $\widehat{\lambda}$ is the usual average list length. To simplify notation, we omit the Poisson and focus on the log likelihood of $\mathbf{s}$ and α, namely $\ell(\mathbf{s}, \alpha) =$

$$\sum_{j=1}^{N} \sum_{t=1}^{n^{(j)}} c(z_t^{(j)}, j, t) \log \alpha + \log s_{z_t^{(j)}} - \log \sum_{i \in \mathcal{V}} \alpha^{c(i,j,t)} s_i.$$

We transform the variables into the log domain which is easier to work with: $\boldsymbol{\beta} \equiv \log \mathbf{s}$, and $\gamma \equiv \log \alpha$. The log likelihood can now be written as $\ell(\boldsymbol{\beta}, \gamma) =$

$$\sum_{j=1}^{N} \sum_{t=1}^{n^{(j)}} c(z_t^{(j)}, j, t)\gamma + \beta_{z_t^{(j)}} - \log \sum_{i \in \mathcal{V}} \exp(c(i, j, t)\gamma + \beta_i).$$

Due to the log-sum-exp form, $\ell(\boldsymbol{\beta}, \gamma)$ is concave in $\boldsymbol{\beta}$ and γ (Boyd & Vandenberge, 2004). Note the initial sizes $\mathbf{s}$ are scale invariant. We remove this invariance by setting $s_{MFI} = 1$ where MFI is the most frequent item in $\mathbf{z}^{(1)} \dots \mathbf{z}^{(N)}$. Equivalently, $\beta_{MFI} = 0$.

The complete convex optimization problem for finding the MLE of SWIRL is

$$\min_{\boldsymbol{\beta}, \gamma} \quad -\ell(\boldsymbol{\beta}, \gamma) \qquad (2)$$

$$\text{s.t.} \quad \beta_{MFI} = 0 \qquad (3)$$

$$\gamma \leq 0. \qquad (4)$$

The MLE is readily computed by quasi-Newton methods such as LBFGS, where the required gradient for (2) is computed by

$$\frac{\partial \ell}{\partial \beta_i} = -c(i) + \sum_{j=1}^{N} \sum_{t=1}^{n^{(j)}} \frac{\exp(c(i, j, t)\gamma + \beta_i)}{\sum_{i' \in \mathcal{V}} \exp(c(i', j, t)\gamma + \beta_{i'})}$$

$$\frac{\partial \ell}{\partial \gamma} = \sum_{j=1}^{N} \sum_{t=1}^{n^{(j)}} -c(z_t^{(j)}, j, t) + \frac{\sum_{i \in \mathcal{V}} \exp(c(i, j, t)\gamma + \beta_i) c(i, j, t)}{\sum_{i' \in \mathcal{V}} \exp(c(i', j, t)\gamma + \beta_{i'})},$$

where $c(i)$ is the total count of occurrences for item i in the lists $\mathbf{z}^{(1)} \dots \mathbf{z}^{(N)}$, for $i \in \mathcal{V}$.

2.3. Maximum A Posteriori Estimate

Additionally, it is easy to compute the MAP estimate when we have prior knowledge on $\boldsymbol{\beta}$ and γ. Suppose we believe the initial sizes should be approximately proportional to $\mathbf{s}_0$. For example, in the vehicle verbal fluency task, $\mathbf{s}_0$ may be the counts of various vehicles in a large corpus. We can define $\beta_{0_i} \equiv \log(s_{0_i}/s_{0_{MFI}}); \forall i \in \mathcal{V}$, and adopt a Gaussian prior on $\boldsymbol{\beta}$ centered around $\boldsymbol{\beta}_0$ (equivalently, $\mathbf{s}$ follows a log-Normal distribution). Since this prior removes the scale invariance on $\boldsymbol{\beta}$, we no longer need the constraint $\beta_{MFI} = 0$. Similarly, we may have prior knowledge of γ_0. The MAP estimate is the solution to

$$\min_{\boldsymbol{\beta}, \gamma} \quad -\ell(\boldsymbol{\beta}, \gamma) + \tau_1 \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + \tau_2(\gamma - \gamma_0)^2 \qquad (5)$$

$$\text{s.t.} \quad \gamma \leq 0\,,$$

where τ_1, τ_2 are appropriate regularization terms to prevent overfitting.

3. Application I: Feature Volunteering for Text Classification

We now turn to a machine learning application of SWIRL: training text classifiers from human-volunteered *feature labels* (rather than documents). A feature label is a simple rule stating that the presence of a word or phrase indicates a particular class label.

For example, "touchdown⇒football" implies that documents containing the word "touchdown" probably belong to the class "football." Some prior work exploits a *bag* of volunteered features from users, where each has an equal importance. Although such feature labels help classification (Liu et al., 2004), the order of volunteered words is not taken into account. The order turns out to be very useful, however, as we will show later. Other prior work solicits labeled features through a form of feature *queries*: the computer, via unsupervised corpus analysis (e.g., topic modeling), proposes a list of high-probability candidate features for a human to label (Druck et al., 2008).

Departing from previous works, we point out that the human teacher, upon hearing the categorization goal (e.g., the classes to be distinguished), can *volunteer* an ordered list of feature labels without first consulting a corpus; see Table 1(a,b). This "feature volunteering" procedure is particularly attractive when a classifier is promptly needed for a novel task, since humans can be recruited quickly via crowdsourcing or other means, even before a corpus is fully compiled. Another possibility, which we recommend for practitioners[1], is to treat feature volunteering as a crucial first step in a chain of progressively richer interactive supervision, followed by queries on both features and documents. Queries can be selected by the computer in order to build better classifiers over time. Such a combination has been studied recently (Settles, 2011).

In this section, we show that feature volunteering can be successfully combined with two existing frameworks for training classifiers with labeled features: (*i*) Generalized Expectation (GE) for logistic regression (Druck et al., 2008) and (*ii*) informative Dirichlet priors (IDP) for multinomial naïve Bayes (Settles, 2011). We also show that "order matters" by highlighting the value of SWIRL as a model for feature volunteering. That is, by endowing each volunteered feature with its size s_i as estimated in Section 2, we can build better classifiers than by treating the volunteered features equally under both of these machine learning frameworks.

3.1. Generalized Expectation (GE)

Let $y \in \mathcal{Y}$ be a class label, and $\mathbf{x} \in \mathbb{R}^{|\mathcal{F}^+|}$ be a vector describing a text document using feature set $\mathcal{F}^+$, which is a super set of volunteered feature set $\mathcal{F}$. Consider the conditional probability distributions realizable by multinomial logistic regression

$$\Delta_\Theta = \{p_{\boldsymbol{\theta}}(y \mid \mathbf{x}) \mid \boldsymbol{\theta} \in \mathbb{R}^{|\mathcal{F}^+| \times |\mathcal{Y}|}\}, \tag{6}$$

where

$$p_{\boldsymbol{\theta}}(y \mid \mathbf{x}) = \frac{\exp(\theta_y^\top \mathbf{x})}{\sum_{y' \in \mathcal{Y}} \exp(\theta_{y'}^\top \mathbf{x})}. \tag{7}$$

Generalized Expectation (GE) seeks a distribution $p^* \in \Delta_\Theta$ that matches a set of given reference distributions $\widehat{p_f}(y)$: distributions over the label y if the feature $f \in \mathcal{F}$ is present. We will construct $\widehat{p_f}(y)$ from feature volunteering and compare it against other constructions in the next section. Before that, we specify the matching sought by GE (Druck et al., 2008).

We restrict ourselves to sufficient statistics based on counts, such that $x_f \in \mathbf{x}$ is the number of times feature f occurs in the document. Let $\mathcal{U}$ be an unlabeled corpus. The empirical mean conditional distribution on documents where $x_f > 0$ is given by

$$M_f\left[p_{\boldsymbol{\theta}}(y \mid \mathbf{x})\right] \equiv \frac{\sum_{\mathbf{x} \in \mathcal{U}} \mathbb{1}\{x_f > 0\} p_{\boldsymbol{\theta}}(y \mid \mathbf{x})}{\sum_{\mathbf{x}' \in \mathcal{U}} \mathbb{1}\{x'_f > 0\}}. \tag{8}$$

GE training minimizes a regularized objective based on the KL-divergence of these distributions:

$$p^* = \operatorname*{argmin}_{p_{\boldsymbol{\theta}} \in \Delta_\Theta} \sum_{f \in \mathcal{F}} \mathrm{KL}\Big(\widehat{p_f}(y) \,\Big\|\, M_f[p_{\boldsymbol{\theta}}(y \mid \mathbf{x})]\Big) + \frac{\|\boldsymbol{\theta}\|^2}{2}. \tag{9}$$

In other words, GE seeks to make the reference and empirical distributions as similar as possible.

3.2. Constructing GE Reference Distributions $\widehat{p_f}(y)$ with SWIRL

Recall that in feature volunteering, the N human users produce multiple ordered lists $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$. Each item z in these lists is a $f \Rightarrow y$ (feature⇒label) pair. It is possible that the same f appears multiple times in different z's, mapping to different y's (e.g., "goalie⇒soccer" and "goalie⇒hockey"). In this case we say feature f co-occurs with multiple y's.

For each list $\mathbf{z}$, we split it into $|\mathcal{Y}|$ sublists by item labels. This produces one ordered sublist per class. We collect all N sublists for a single class y, and treat them as N lists generated by SWIRL from the yth urn. We find the MLE of this yth urn using (2), and normalize it to sum to one. We are particularly interested in the size estimates $\mathbf{s}_y = \{s_{f \Rightarrow y} \mid f \in \mathcal{F}\}$. This is repeated for all $|\mathcal{Y}|$ urns, so we have $\mathbf{s}_1, \ldots, \mathbf{s}_{|\mathcal{Y}|}$.

We construct reference distributions using

$$\widehat{p_f}(y) = \frac{s_{f \Rightarrow y}}{\sum_{y' \in \mathcal{Y}} s_{f \Rightarrow y'}};\ \forall f \in \mathcal{F}\,, \tag{10}$$

where $\mathcal{F}$ is the union of features appearing in the lists $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$ and $s_{f \Rightarrow y} = 0$ if feature f is absent from

[1] Feature volunteering and feature query labeling are complementary ways of obtaining feature labels. The former provides a way to capture importance of feature labels (e.g., by their order), while the latter can consult extra resources (e.g., unlabeled data) to harvest more feature labels, as pointed out by Liu et al. (2004).

the yth list. For example, imagine "goalie⇒soccer" appears near the top of a list, so $s_{\text{goalie}\Rightarrow\text{soccer}}$ is large (say 0.4), "goalie⇒hockey" appears much later in another list, so $s_{\text{goalie}\Rightarrow\text{hockey}}$ is small (say 0.1), and "goalie" is never associated with "baseball". Then by (10), $\hat{p}_{\text{goalie}}(\text{soccer}) = 0.8$, $\hat{p}_{\text{goalie}}(\text{hockey}) = 0.2$, and $\hat{p}_{\text{goalie}}(\text{baseball}) = 0$.

In our experiments, we compare three ways of creating reference distributions. **GE/SWIRL** is given by Equation (10). **GE/Equal** is defined as

$$\widehat{p_f}(y) = \frac{\mathbb{1}\{s_{f\Rightarrow y} > 0\}}{\sum_{y' \in \mathcal{Y}} \mathbb{1}\{s_{f\Rightarrow y'} > 0\}}; \;\; \forall i \in \mathcal{F}, \tag{11}$$

which is similar to (10), except that all features co-occurring with y have equal size. This serves as a baseline to investigate whether order matters. **GE/Schapire** is the reference distribution used in previous work (Druck et al., 2008):

$$\widehat{p_f}(y) = \begin{cases} q/m & \text{if feature } f \text{ co-occurs with } y \\ (1-q)/(|\mathcal{Y}| - m) & \text{otherwise,} \end{cases} \tag{12}$$

where m is the number of distinct labels co-occurring with feature f in $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$, and q is a smoothing parameter. We use $q = 0.9$ as in prior work.

3.3. Informative Dirichlet Priors (IDP)

Another way to use feature volunteering is by training multinomial Naïve Bayes models, where feature⇒label rules are adapted as informative priors on feature parameters. The class distribution $p(y)$ is parametrized by π_y, and the likelihood of a document $\mathbf{x}$ given a class label y is modeled as $p(\mathbf{x} \mid y) = \prod_f (\phi_{fy})^{x_f}$, where x_f is the frequency count of feature f and ϕ_{fy} is multinomial parameter for feature f under class y. Dirichlet priors are placed on each class-conditional multinomial parameter, where the hyperparameter is denoted by d_{fy} for phrase f under class y.

We estimate π_y by class proportion of labeled documents, and ϕ_{fy} by posterior expectation as follows:

$$\phi_{fy} \;\propto\; d_{fy} + \sum_{\mathbf{x}} p(y \mid \mathbf{x}) x_f, \tag{13}$$

where ϕ_{fy} is normalized over phrases to sum to one for each y. When learning from only labeled instances, $p(y \mid \mathbf{x}) \in \{0, 1\}$ indicates the true labeling of instance $\mathbf{x}$. When learning from both labeled and unlabeled instances, we run EM as follows. First, initialize ϕ_{fy}'s by (13) using only the d_{fy} hyperparameters. Second, repeatedly apply (13) until convergence, where the second summation term is over both labeled and unlabeled instances and $p(y \mid \mathbf{x})$ for unlabeled instance $\mathbf{x}$ is computed using Bayes rule.

Corpus	Class Labels	N	$\lvert\mathcal{F}\rvert$	$\lvert\mathcal{F}^+\rvert$
sports	baseball, basketball, football, hockey, soccer	52	594	2948
movies	negative, positive	27	382	2514
webkb	course, faculty, project, student	56	961	2521

Table 2. Domains in the feature volunteering application.

3.4. Constructing IDP Priors d_{fy} with SWIRL

We compare two approaches for incorporating prior knowledge into naïve Bayes by feature volunteering. **IDP/SWIRL** sets the hyperparameters as follows: $d_{fy} = 1 + k n_y s_{f\Rightarrow y}$, where f is a feature, k a parameter, and n_y is the number of unique features in y's list. Again, we compute $s_{f\Rightarrow y}$ via SWIRL as in Section 3.2.

Note that replacing $s_{f\Rightarrow y}$ with $\mathbb{1}\{s_{f\Rightarrow y} > 0\}/n_y$ recovers prior work (Settles, 2011). In this method, only the association between a feature f and a class y is taken into account, rather than relative importance of these associations. This baseline, **IDP/Settles**, sets $d_{fy} = 1 + k\mathbb{1}\{s_{f\Rightarrow y} > 0\}$ and serves to investigate whether order matters in human-generated lists.

3.5. Experiments

We conduct feature volunteering text classification experiments in three different domains: **sports** (sports articles), **movies** (movie reviews), and **webkb** (university web pages). The classes, number of human participants (N), and the number of distinct list features they produced ($|\mathcal{F}|$) are listed in Table 2.

Participants. A total of 135 undergraduate students from the University of Wisconsin-Madison participated for partial course credit. No one participated in more than one domain. All human studies in this paper were approved by the institutional review board.

Procedure. Participants were informed of only the class labels, and were asked to provide as many words or short phrases as they thought would be necessary to accurately classify documents into the classes. They volunteered features using the web-based computer interface illustrated in Figure 1 (shown here for the sports domain). The interface consists of a text box to enter features, followed by a series of buttons corresponding to labels in the domain. When the participant clicks on a label (e.g., "hockey" in the figure), the phrase is added to the bottom of the list below the corresponding button, and erased from the input box. The feature⇒label pair is recorded for each action in sequence. The label order was randomized for each subject to avoid presentation bias. Participants had 15 minutes to complete the task.

Figure 1. Screenshot of the feature volunteering interface.

Data cleaning. We normalized the volunteered phrases by case-folding, punctuation removal, and space normalization. We manually corrected obvious misspellings. We also manually mapped different forms of a feature to its dictionary canonical form: for example, we mapped "lay-up" and "lay up" to "layup." The average (and maximum) list length is 39 (91) for sports, 20 (46) for movies, and 40 (85) for webkb. Most participants volunteered features at a fairly uniform speed for the first five minutes or so; some then exhausted ideas. This suggests the importance of combining feature volunteering with feature and document querying, as mentioned earlier. This combination is left for future work.

Unlabeled corpus $\mathcal{U}$. Computing (8) requires an unlabeled corpus. We produce $\mathcal{U}$ for the **sports** domain by collecting 1123 Wikipedia documents via a shallow web crawl starting from the top-level wiki-category for the five sport labels (e.g., "Category:Baseball"). We produce a matching $\mathcal{U}$ for the **movies** and **webkb** domains from the standard movie sentiment corpus (Pang et al., 2002) (2000 instances) and the WebKB corpus (Craven et al., 1998) (4199 instances), respectively. Note that $\mathcal{U}$'s are treated as unlabeled for training our models, however, we use each $\mathcal{U}$'s in a transductive fashion to serve as our test sets as well.

Training with GE. We define the feature set for learning $\mathcal{F}^+$ (i.e., the dimensionality in (6)) to be the union of $\mathcal{F}$ (volunteered phrases) plus all unigrams occurring at least 50 times in the corpus $\mathcal{U}$ for that domain. That is, we include all volunteered phrases, even if they are not a unigram or appear fewer than 50 times in the corpus. $|\mathcal{F}^+|$ for each domain is listed in Table 2. We construct the reference distributions according to **GE/SWIRL**, **GE/Equal**, and **GE/Schapire** as in section 3.2, and find the optimal logistic regression models $p^*(y \mid \mathbf{x})$ by solving (9) with LBFGS for each domain and reference distribution.

Corpus	SWIRL	Equal	Schapire	FV
sports	**0.865**	0.847	0.795	**0.875**
movies	**0.733**	**0.733**	**0.725**	0.681
webkb	**0.463**	0.444	0.429	0.426

(a) GE accuracies for logistic regression

Corpus	SWIRL	Settles	FV
sports	**0.911**	0.901	0.875
movies	**0.687**	0.656	**0.681**
webkb	**0.659**	**0.651**	0.426

(b) IDP accuracies for multinomial naïve Bayes

Table 3. Text categorization results.

Training with IDP. We use the same $\mathcal{F}^+$ in GE training, construct IDP hyperparameters according to **IDP/SWIRL** and **IDP/Settles**, and learn MNB classifiers as in section 3.3 using uniform π_y. Following prior work, we apply one-step EM with $k = 50$.

Feature Voting Baseline (FV). We also include a simple baseline for both frameworks. To classify a document $\mathbf{x}$, FV scans through unique volunteered features for which $x_f > 0$. Each class y for which $f \Rightarrow y$ exists in $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$ receives one vote. At the end, FV predicts the label as the one with the most votes. Ties are broken randomly. Accuracy of FV is measured by averaging over 20 trials due to this randomness.

Results. Text classifiers built from volunteered features and SWIRL consistently outperform the baselines. Classification accuracies of the different models under GE and IDP are shown in Table 3(a,b). For each domain, we show the best accuracy in bold face, as well as any accuracies whose difference from the best is not statistically significant[2]. Under the GE framework, **GE/SWIRL** is the best on movies and webkb, and is indistinguishable from the best on sports. Under the IDP framework, **IDP/SWIRL** consistently outperforms all baselines.

The fact that both **GE/SWIRL** and **IDP/SWIRL** are better than (or on par with) the baselines under both frameworks strongly indicates that *order matters*. That is, when working with human-generated lists, the item order carries information that can be useful to machine learning algorithms. Such information can be extracted by SWIRL parameter estimates and successfully incorporated into a secondary classification task. Although dominated by SWIRL-based approaches, **FV** is reasonably strong and may be a quick stand-in due to its simplicity.

A caveat: the human participants were only informed of the class labels and did not know $\mathcal{U}$. Mildly amusing mismatch ensued. For example, the webkb corpus was collected in 1997 (before the "social media" era), but the volunteered feature labels

[2]Using paired two-tailed t-tests, $p < 0.05$.

included "facebook⇒student," "dropbox⇒course," "reddit⇒faculty," and so on. Our convenient but outdated choice of $\mathcal{U}$ quite possibly explains the low accuracy of all methods in the webkb domain.

4. Application II: Verbal Fluency for Brain Damaged Patients

One human list-generation task that has received detailed examination in cognitive psychology is "verbal fluency." Human participants are asked to say as many examples of a category as possible in one minute with no repetitions (Glasdjo et al., 1999).[3] For instance, participants may be asked to generate examples of a semantic category (e.g., "animals" or "furniture"), a phonemic or orthographic category (e.g., "words beginning with the letter F"), or an ad-hoc category (e.g., "things you would rescue from a burning house").

Verbal fluency has been widely adopted in neurology to aid in the diagnosis of cognitive dysfunction (Monsch et al., 1992; Troyer et al., 1998; Rosser & Hodges, 1994). Category and letter fluency in particular are sensitive to a broad range of cognitive disorders resulting from brain damage (Rogers et al., 2006). For instance, patients with prefrontal injuries are prone to inappropriately repeating the same item several times (*perseverative* errors) (Baldo & Shimamura, 1998), whereas patients with pathology in the anterior temporal cortex are more likely to generate incorrect responses (*semantic* errors) and produce many fewer items overall (Hodges et al., 1999; Rogers et al., 2006). Despite these observations and widespread adoption of the task, standard methods for analyzing the data are comparatively primitive: correct responses and different error types are counted, while sequential information is typically discarded.

We propose to use SWIRL as a computational model of the verbal fluency task, since we are unaware of any other such models. We show that, though overly simplified in some respects, SWIRL nevertheless estimates key parameters that correspond to cognitive mechanisms. We further show that these estimates differ in healthy populations versus patients with temporal-lobe epilepsy, a neurological disorder thought to disrupt semantic knowledge. Finally, we report promising classification results, indicating that our model could be useful in aiding diagnosis of cognitive dysfunction in the future.

Participants. We investigated fluency data generated from two populations: a group of 27 patients with temporal-lobe epilepsy (a disorder thought to disrupt semantic abilities), and a group of 24 healthy controls matched to the patients in age, education, sex, nonverbal IQ and working-memory span. Patients were recruited through an epilepsy clinic at the University of Wisconsin-Madison. Controls were recruited through fliers posted throughout Madison, Wisconsin.

Procedure. We conducted four category-fluency tasks: animals, vehicles, dogs, and boats. In each task, participants were shown a category name on a computer screen and were instructed to verbally list as many examples of the category as possible in 60 seconds without repetition. The audio recordings were later transcribed by lab technicians to render word lists. We normalize the lists by expanding abbreviations ("lab" → "labrador"), removing inflections ("birds" → "bird"), and discarding junk utterances and interjections. The average (and maximum) list length is 20 (37) for animals, 14 (33) for vehicles, 11 (23) for dogs, and 11 (20) for boats.

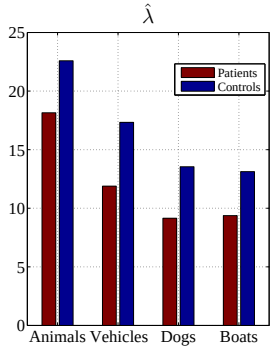
Results. We estimated SWIRL parameters $\lambda, \mathbf{s}, \alpha$ using (2) for the patient and control groups on each task separately, and observed that:

1. *Patients produced shorter lists.* Figure 2(a) shows that the estimated Poisson intensity $\hat{\lambda}$ (i.e., the average list length) is smaller for patients on all four tasks. This is consistent with the psychological hypothesis that patients suffering from temporal-lobe epilepsy produce items at a slower rate, hence fewer items in the time-limit.

2. *Patients and controls have similar lexicon distributions, but both deviate from word usage frequency.* Figure 2(b) shows the top 10 lexicon items and their probabilities (normalized $\mathbf{s}$) for patients ($\mathbf{s}_P$) and controls ($\mathbf{s}_C$) in the animals task, sorted by $(\mathbf{s}_{Pi} + \mathbf{s}_{Ci})/2$. $\mathbf{s}_P$ and $\mathbf{s}_C$ are qualitatively similar. We also show corpus probabilities $\mathbf{s}_W$ from the Google Web 1T 5-gram data set (Brants et al., 2007) for comparison (normalized on items appearing in human-generated lists). Not only does $\mathbf{s}_W$ have smaller values, but its order is very different: horse (0.06) is second largest, while other top corpus words like fish (0.05) and mouse (0.03) are not even in the human top 10. This challenges the psychological view that verbal fluency largely follows real-world lexicon usage frequency. Both observations are supported quantitatively by comparing the Jensen-Shannon divergence (JSD) [4] between the whole distributions $\mathbf{s}_P$, $\mathbf{s}_C$, $\mathbf{s}_W$; see Figure 2(c). Clearly, $\mathbf{s}_P$ and $\mathbf{s}_C$ are relatively close, and both are far from $\mathbf{s}_W$.

[3] Despite the instruction, people still do repeat.

[4] Each group's MLE of $\mathbf{s}$ (and hence p) has zero probability on items only in the other groups. We use JSD since it is symmetric and allows disjoint supports.



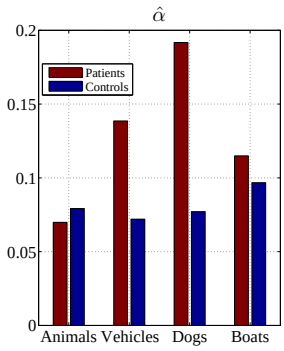
(a) list length $\hat{\lambda}$

Item	$\mathbf{s}_P$	$\mathbf{s}_C$	$\mathbf{s}_W$
cat	.15	.13	.05
dog	.17	.11	.08
lion	.04	.04	.01
tiger	.04	.03	.01
bird	.04	.02	.03
elephant	.03	.03	.01
zebra	.01	.04	.00
bear	.03	.03	.03
snake	.02	.02	.01
horse	.02	.03	.06
⋮	⋮	⋮	⋮

(b) top 10 "animal" words

Task	Pair	JSD
animals	$(\mathbf{s}_P, \mathbf{s}_C)$	.09
	$(\mathbf{s}_P, \mathbf{s}_W)$	.17
	$(\mathbf{s}_C, \mathbf{s}_W)$	.19
vehicles	$(\mathbf{s}_P, \mathbf{s}_C)$	.18
	$(\mathbf{s}_P, \mathbf{s}_W)$	.23
	$(\mathbf{s}_C, \mathbf{s}_W)$	.25
dogs	$(\mathbf{s}_P, \mathbf{s}_C)$	.15
	$(\mathbf{s}_P, \mathbf{s}_W)$	.54
	$(\mathbf{s}_C, \mathbf{s}_W)$	.55
boats	$(\mathbf{s}_P, \mathbf{s}_C)$	.15
	$(\mathbf{s}_P, \mathbf{s}_W)$	.54
	$(\mathbf{s}_C, \mathbf{s}_W)$	.52

(c) distribution comparisons



(d) discount factor $\hat{\alpha}$

Figure 2. Verbal fluency experimental results. SWIRL distributions for patients, controls, and general word frequency on the World Wide Web (using Google 1T 5-gram data) are denoted by $\mathbf{s}_P$, $\mathbf{s}_C$, and $\mathbf{s}_W$, respectively.

3. *Patients discount less and repeat more.* Figure 2(d) shows the estimated discount factor $\hat{\alpha}$. In three out of four tasks, the patient group has larger $\hat{\alpha}$. Recall that a larger $\hat{\alpha}$ leaves an item's size relatively unchanged, thus increasing the item's chance to be sampled again — in other words, more repeats. This is consistent with the psychological hypothesis that patients have a reduced ability to inhibit items already produced.

Healthy vs. Patient Classification. We conducted additional experiments where we used SWIRL parameters to build down-stream healthy vs. patient classifiers. Specifically, we performed leave-one-out (LOO) classification experiments for each of the four verbal fluency tasks. Given a task, each training set (minus one person's list for testing) was used for learning two separate SWIRL models with MAP estimation (5): one for patients and the other for healthy participants. We set $\tau_1 = \infty$ and $\boldsymbol{\beta}_0 = \mathbf{1}$ due to the finding that both populations have similar lexicon distributions, and $\tau_2 = 0$. We classified the held-out test list by likelihood ratio threshold at 1 for the patient vs. healthy models. The LOO accuracies of SWIRL on animals, vehicles, dogs and boats were 0.647, 0.706, 0.784, and 0.627, respectively. In contrast, the majority vote baseline has accuracy $27/(24+27) = 0.529$ for all tasks. The improvement for the dogs task over the baseline approach is statistically significant[5].

5. Conclusion and Future Work

Human list generation is an interesting process of data creation that deserves attention from the machine learning community. Our initial foray into modeling human-generated lists by sampling with reduced replacement (SWIRL) has resulted in two interesting applications for both machine learning (effectively combining SWIRL statistics with modern feature-labeling frameworks) and cognitive psychology (modeling memory in healthy vs. brain-damaged populations, and predicting cognitive dysfunction).

Learning from human-generated lists opens up several lines of future work. For example: (*i*) Designing a "supervision pipeline" that combines feature volunteering with feature label querying, document label querying, and other forms of interactive learning to build better text classifiers more quickly. (*ii*) Identifying more applications which can benefit from models of human list generation. For example, creating lists of photo tags on Flickr.com, or hashtags on Twitter.com, can be viewed as a form of human list generation conditioned on a specific photo or tweet. (*iii*) Developing a hierarchical version of SWIRL, so that each human has their own personalized parameters λ, $\mathbf{s}$, and α, while a group has summary parameters, too. This is particularly attractive for applications like verbal fluency, where we want to understand both individual and group behaviors. (*iv*) Developing structured models of human list generation that can capture and learn from people's tendency to generate "runs" of semantically-related items in their lists (e.g., pets then predators then fish).

Acknowledgments

The authors were supported in part by National Science Foundation grants IIS-0953219, IIS-0916038, IIS-1216758, IIS-0968487, DARPA, and Google. The patient data was collected under National Institute of Health grant R03 NS061164 with TR as the PI. We thank Bryan Gibson for help collecting the feature volunteering data.

[5]Using paired two-tailed t-tests, $p < 0.05$.

References

Baldo, J. V. and Shimamura, A. P. Letter and category fluency in patients with frontal lobe lesions. *Neuropsychology*, 12(2):259–267, 1998.

Boyd, S. and Vandenberge, L. *Convex Optimization*. Cambridge University Press, Cambridge UK, 2004.

Brants, T., Popat, A.C., Xu, P., Och, F.J., and Dean, J. Large language models in machine translation. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

Chesson, J. A non-central multivariate hypergeometric distribution arising from biased sampling with application to selective predation. *Journal of Applied Probability*, 13(4):795–797, 1976.

Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., and Slattery, S. Learning to extract symbolic knowledge from the world wide web. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, pp. 509–516. AAAI Press, 1998.

Druck, G., Mann, G., and McCallum, A. Learning from labeled features using generalized expectation criteria. In *Proceedings of the International ACM SIGIR Conference on Research and Development In Information Retrieval*, pp. 595–602, 2008.

Fog, A. Calculation methods for Wallenius' noncentral hypergeometric distribution. *Communications in Statistics - Simulation and Computation*, 37(2): 258–273, 2008.

Glasdjo, J. A., Schuman, C. C., Evans, J. D., Peavy, G. M., Miller, S. W., and Heaton, R. K. Normas for letter and category fluency: Demographic corrections for age, education, and ethnicity. *Assessment*, 6(2):147–178, 1999.

Hodges, J. R., Garrard, P., Perry, R., Patterson, K., Bak, T., and Gregory, C. The differentiation of semantic dementia and frontal lobe dementia from early alzheimer's disease: a comparative neuropsychological study. *Neuropsychology*, 13:31–40, 1999.

Liu, B., Li, X., Lee, W.S., and Yu, P.S. Text classification by labeling words. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, pp. 425–430. AAAI Press, 2004.

Monsch, A. U., Bondi, M. W., Butters, N., Salmon, D. P., Katzman, R., and Thal, L. J. Comparisons of verbal fluency tasks in the detection of dementia of the alzheimer type. *Archives of Neurology*, 49(12): 1253–1258, 1992.

Pang, B., Lee, L., and Vaithyanathan, S. Thumbs up: Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79–86. ACL, 2002.

Rogers, T.T., Ivanoiu, A., Patterson, K., and Hodges, J.R. Semantic memory in alheimer's disease and the fronto-temporal dementias: A longitudinal study of 236 patients. *Neuropsychology*, 20(3):319–335, 2006.

Rosser, A. and Hodges, J.R. Initial letter and semantic category fluency in alzheimer's disease, huntington's disease, and progressive supranuclear palsy. *journal of Neurology, Neurosurgery, and Psychiatry*, 57: 1389–1394, 1994.

Settles, B. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1467–1478. ACL, 2011.

Troyer, A.K., Moscovitch, M., Winocur, G., Alexander, M., and Stuss, D. Clustering and switching on verbal fluency: The effects of focal fronal- and temporal-lobe lesions. *Neuropsychologia*, 36(6), 1998.

Wallenius, K.T. *Biased Sampling: The Non-Central Hypergeometric Probability Distribution*. PhD thesis, Department of Statistics, Stanford University, 1963.