Sampling Representative Phrase Sets for Text Entry Experiments: A Procedure and Public Resource

Tim Paek, Bo-June (Paul) Hsu
Microsoft Research
One Microsoft Way, Redmond WA 98052 USA
{timpaek | paulhsu}@microsoft.com

ABSTRACT
Text entry experiments evaluating the effectiveness of various input techniques often employ a procedure whereby users are prompted with natural language phrases which they are instructed to enter as stimuli. For experimental validity, it is desirable to control the stimuli and present text that is representative of a target task, domain or language. MacKenzie and Soukoreff (2001) manually selected a set of 500 phrases for text entry experiments. To demonstrate representativeness, they correlated the distribution of single letters in their phrase set to a relatively small (by current standards) corpus of English prior to 1966, which may not reflect the style of text input today. In this paper, we ground the notion of representativeness in terms of information theory and propose a procedure for sampling representative phrases from any large corpus so that researchers can curate their own stimuli. We then describe the characteristics of phrase sets we generated using the procedure for email and social media (Facebook and Twitter). The phrase sets and code for the procedure are publicly available for download.

Author Keywords
Text entry, experiment, sampling, relative entropy.

ACM Classification Keywords
H.5.2. Information interfaces and presentation: Evaluation / methodology.

General Terms
Experimentation.

INTRODUCTION
Text entry experiments evaluating the effectiveness of various input techniques often employ a procedure whereby users are prompted with natural language phrases which they are instructed to enter as stimuli. Although it may seem more "natural" to have users enter whatever phrases come to mind, the variability engendered by spontaneous text production poses a threat to the internal validity of experimental conclusions [10, 7]. It is difficult, for example, to claim that one input technique performed better than another if participants using one or the other technique just happened to produce longer, more complicated phrases. To preserve internal validity, researchers typically control text production using pre-selected phrases as stimuli.

For stimuli, many researchers utilize the MacKenzie and Soukoreff (2001) phrase set [9], which contains 500 phrases that are intended to be moderate in length, easy to remember and representative of general English. In order to demonstrate representativeness, MacKenzie and Soukoreff correlated the distribution of single letters in their phrase set to the letter frequencies reported in tables by Mayzner and Tresselt [12]. Unfortunately, the letter frequencies derive from a dated corpus containing publications (newspapers, magazines, and books) prior to 1966, which may not reflect the style of text input today. Furthermore, the vocabulary size is about 20,000 words, which is relatively small by today's standards. Indeed, with the phenomenal growth of the Internet and advances in computational natural language processing, organizations such as the Linguistic Data Consortium (LDC) [8] offer researchers publicly accessible corpora of considerably larger scale for a plethora of tasks, domains and languages. For example, through LDC, Google has published a web corpus of 5-grams from a corpus of approximately 1 trillion tokens [4]. The Microsoft Web N-Gram Services [15, 13] even allows researchers to programmatically access word popularity on the Internet.

In this paper, we propose a procedure for sampling representative phrases from any large corpus so that text input researchers can curate their own stimuli for tasks, domains and languages they wish to target using publicly accessible resources. The procedure is based on grounding the notion of representativeness in terms of information theory. This paper consists of three contributions. First, we explain the mathematical concept of relative entropy and discuss how it relates to representativeness. Second, we propose an information-theoretic procedure for sampling representative phrases from a large corpus. Third, we describe the characteristics of phrase sets we generated using the procedure for email and social media (Facebook and Twitter).

ENTROPY AND REPRESENTATIVENESS
Information theory was developed in the 1940s by Claude Shannon [14] as a method to quantify "information" and determine the mathematical limits of signal processing

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

tasks such as data compression and reliable transmission through a noisy communication channel. The basic measure of information in the theory is *entropy*, which is computed for a random variable X over a discrete set of symbols χ (e.g., characters) with probability distribution $p(x)$ as:

$$H(X) = -\sum_{x \in \chi} p(x) \log_2 p(x)$$

The logarithm can be any base, but with log base 2, the resulting value is measured in bits. While entropy is often explained as the lower bound on the number of bits it would take to encode an outcome of X, or as the expected uncertainty of X, it can also be viewed in terms of the *Twenty Questions* game. If χ is the set of English characters, and we are given 20 yes/no questions to ascertain characters in sequence (e.g., "Is the next character a vowel?"), then entropy measures how many questions on average are needed to identify each character with total certainty [11]. In other words, $H(X)$ measures the size of the search space consisting of possible values of X and their probabilities.

We can leverage entropy to quantify and compare the "representativeness" of samples of phrases with respect to a source corpus. Suppose $p(x)$ and $q(x)$ are probability distributions of random variable X over the same set of English characters χ. In particular, let $p(x)$ be the probability distribution for a sample of a corpus, and $q(x)$ be the probability distribution for the source corpus itself. The difference between these probability distributions can be measured by *relative entropy*, or the *Kullback-Leibler divergence*:

$$D(p||q) = \sum_{x \in \chi} p(x) \log_2 \frac{p(x)}{q(x)}$$

The relative entropy is always non-negative and $D(p||q) = 0$ only when the distributions are identical. Hence, the more representative the sample distribution $p(x)$ is of the source distribution $q(x)$, the closer the relative entropy will be to zero. Note that we must be careful about interpreting relative entropy; although it may appear to be a distance metric, technically it is not because $D(p \parallel q) \neq D(q \parallel p)$ and it fails to satisfy the triangle inequality [2].

For representativeness, MacKenzie and Soukoreff (2001) [9] compute the correlation coefficient between two normalized letter frequencies. However, this approach essentially treats the frequencies as two random variables and measures the linear dependence between them. For comparing probability distributions, relative entropy constitutes a more mathematically principled method.

Perplexity

Related to relative entropy is the concept of cross-entropy. Suppose we have a true probability distribution $p(x)$ that generates some data and we wish to produce from that data a model distribution m that approximates p. To have as accurate of a model as possible, we need to minimize the relative entropy $D(m \parallel p)$, but of course we do not know p. Fortunately, cross-entropy allows us to still compare m with competing models.

By treating language as a stationary and ergodic (see [5] for more details) stochastic process consisting of a sequence of symbols $(X_i) \sim p(x)$ drawn from p, the *cross-entropy* between the empirically observed sequence (X_i) and a model m is given by:

$$H(X_i, m) = -\lim_{n \to \infty} \frac{1}{n} \log_2 m(x_1\, x_2 \ldots x_n)$$

If n is sufficiently large, we can drop the limit and approximate cross-entropy as:

$$H(X_i, m) \approx -\frac{1}{n} \log_2 m(x_1\, x_2 \ldots x_n)$$

Because $H(p) \leq H(p, m)$, the cross-entropy provides an upper bound on the true entropy of p. Hence, the more accurate m is, the closer the cross-entropy will be to the true entropy. This provides a way to compare which model of p is better without having to know p: just pick the model with the lowest cross-entropy.

To make cross-entropy easier to interpret, researchers in statistical natural language processing often utilize perplexity, which is just:

$$perplexity(X_i) = 2^{H(X_i, m)} = m(x_1\, x_2 \ldots x_n)^{-\frac{1}{n}}$$

Perplexity is the default measure of *n-gram* language models [11], which predict a set of symbols (e.g., words, characters, grammatical constituents, etc.) based on the history of previous symbols. Just as entropy provides a measure of the size of a search space, the perplexity of a language model expresses the average branching factor of the prediction search space – in particular, the average number of equi-probable symbols that can follow any given symbol. By building language models for a particular task, domain or language, the lowest perplexity we can obtain for the language model tells us how difficult the prediction task is for that task, domain or language. For example, a coin flip has a perplexity of 2 whereas a die roll has a perplexity of 6. Predicting what word follows any previous two words (i.e., trigram language model) for the Wall Street Journal has a perplexity of 170 [5]. We report the perplexities of language models we learned for our email and social media corpora in the last section.

SAMPLING PROCEDURE

Given the information-theoretic concepts we discussed in the previous section, we can now propose a procedure for sampling representative phrases from a large corpus. For text entry experiments, because the most significant point of differentiation is the mechanism to enter characters, words and phrases [9], we have geared our procedure towards text entry of characters. However, it should not be difficult for researchers to adjust the procedure to whatever units of input they are examining. In addition, because

Fitt's Law is commonly invoked in many character entry experiments, we have focused the procedure on obtaining representative character *bigrams* (2 character sequences) to capture distances typically travelled between keystrokes. Finally, like MacKenzie and Soukoreff [9], we sought to procure phrases of "moderate length" – about 4 words, though this can be easily adjusted as well. While we did not attempt to select phrases that were "easy to remember", which was manually done in [9], as a future direction, we could explore using entropy for selection as well, especially in light of prior research demonstrating a correlation between entropy rate and reading comprehension [6].

The procedure operates as follows:

1. Obtain a large corpus of interest. Many corpora can nowadays be obtained online for a wide variety of tasks, domains and languages (e.g., [8]).
2. Clean the text to remove any undesired portions of text (e.g., MIME encoding for emails, punctuations, etc.).
3. Collect all word 4-grams (4 word sequences) from the text, keeping around duplicates as separate entries. To adjust the phrase length, change the n-gram order n to the desired number of words.
4. Randomly generate an index number between 1 and the total number of 4-gram entries.
5. Find the 4-gram at the generated index and store this as a phrase for sample phrase set S_i.
6. Repeat steps 4-5 until M sample phrase sets of the desired size (e.g., 500 in [9]) are obtained.
7. Using the source corpus C, for each sample phrase set S_i, compute the relative entropy $D(S_i \parallel C)$ (see equation in previous section) of the character bigram distributions.
8. Select the sample phrase set S^* with the lowest relative entropy:

$$S^* = \underset{S_i}{\operatorname{argmin}}\, D(S_i \parallel C)$$

Note that steps 3-5 can also be performed efficiently using *inverse transform sampling* [3]. Furthermore, the number of sample phrase sets S_i that researchers may wish to collect depends on how much time they can devote to collecting candidates for computing relative entropy. For the three phrase sets we discuss next, we selected the best from among 100 samples.

In using character bigrams, it is worth remembering that n-grams only provide an empirical estimate of the true probabilities. While using higher order n may allow better modeling of the underlying distributions, it requires more data for estimation. Furthermore, larger samples may be required to achieve reasonable representativeness.

THREE PHRASE SETS

Using the procedure described in the previous section, we generated phrase sets for email and social media. The source corpus for the email phrase set is the Enron Email Data Set, a subset of which can be publicly accessed at [1], containing approximately 517,431 Enron employee emails from 1999 to 2002. For social media, we tapped into public feeds from Facebook and Twitter. For Facebook, we collected the subset of status update messages on August 15, 2010 marked as English, for a total of 295M words across 18.7M messages. For Twitter, we collected all tweets on September 1, 2010 marked as English, yielding a total of 164M words across 12.6M tweets.

	Enron	Facebook	Twitter
# unique unigrams	102	8,609	7,864
unigram perplexity	30.70	28.05	33.36
# unique bigrams	8,281	125,657	95,659
bigram perplexity	21.26	19.97	24.50
# unique word 4-grams	84M	239M	126M

Table 1. Characteristics of the Enron Email, Facebook and Twitter corpora used to generate phrase sets

In order to demonstrate the importance of using the procedure to find representative phrases, in Figure 1, we plot the mean and standard deviation of the relative entropy over various phrase set sizes on the Enron corpus, computed over 100 random samples. As shown, the average relative entropy decreases as we increase the phrase set size, since there are less quantization effects due to sampling. Whereas the standard deviation for large phrase set sizes are negligible, it is quite significant for smaller phrase set sizes. In other words, it is much more important to make sure that a sample phrase set is "representative" of the source corpus when there is a small number of phrases (e.g., 500) than when there is a large number (e.g., 10K). We suspect that most text entry experiments only utilize a small number of phrases.

Table 1 reports the characteristics of the three corpora. Looking at the number of unigrams, or unique characters, the reason why Facebook and Twitter has over 7000 is due to the inclusion of non-English text. Despite the greater

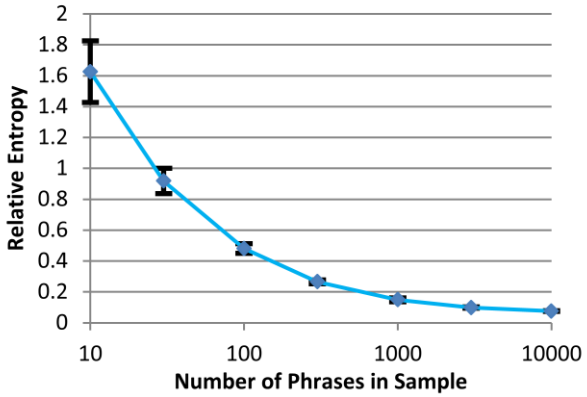

Figure 1. Relative entropy decreasing as the number of phrases in the sample increases on the Enron corpus.

Enron	Facebook	Twitter
just let me know	I'm lonely and I'm	waitin on your call
world's most liquid trading	there LOL. I learned	NOW FOLLOWIN >> @
We have the power	what im gonna do	you could win a
only 20% of stocks	that I spend with	Bout to get my
administration's national energy policy	text them to me	checked...n forwarded also..!! :)
market dipped sharply right	make things work (:	am not even kidding
the late House Speaker	of pics I have	happy hour with a
because it was your	creeppiest dream.. EVER o.o...	me. Take me, free
and penchant for double-burgers,	don't even do that	little bit of fun
happy to exploit any	me and my fam	wan go home! i

Table 2. Sample phrase sets from the Enron Email, Facebook and Twitter corpora.

number of unigrams, if we look at the perplexity, predicting the next character in social media is at just about the same level of difficulty (about 30) as in email. Similarly, email and social media have about the same bigram perplexity; on average, predicting the next character given its previous character has a branching factor of about 20. For generating the phrase sets, we had a lot more uniq 4-grams to sample from in the Facebook corpus (239 million) than in the other corpora. Table 2 presents sample phrases from the three corpora. Note that in generating the phrase sets, we may not have fully scrubbed the text of all offensive language. For Facebook and Twitter especially, this is a difficult problem due to the variety of creative ways in which users express profanity. We recommend that researchers manually inspect the phrase sets and remove any undesirable text. Phrase sets of varying sizes for all three corpora and sample code for the procedure are publicly available for download at http://research.microsoft.com/phrasesets/.

CONCLUSION

In this paper, we presented a procedure for sampling representative phrases from a large corpus so that text input researchers can curate their own stimuli targeting specific tasks, domains and languages. The procedure relies on measuring representativeness as the relative entropy between the probability distributions of the sample phrase set and the source corpus. Researchers can modify the procedure to obtain phrase sets of the desired unit of input, word length, etc. We applied the procedure to create three phrase sets for email and social media, and discussed their perplexities. The phrase sets and code for the procedure are available as a public resource.

REFERENCES

1. 2001 Topic Annotated Enron Email Data Set. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2007T22
2. Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. Wiley-Interscience, 1991.
3. Devroye, L. *Non-Uniform Random Variate Generation*. Springer-Verlag, 1986
4. Google Web 1T 5-Gram Corpus. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13
5. Jurafsky, D. and Martin, J. *Speech and Language Processing, Second Edition*. Prentice Hall, 2009.
6. Keller, F. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. *Proc. of EMNLP* (2004), 317-324.
7. Liebert, R. M. and Liebert, L. L. *Science and Behavior: An Introduction to Methods of Psychological Research*. Englewood Cliffs, NJ, Prentice Hall, 1995.
8. Linguistic Data Consortium. http://www.ldc.upenn.edu
9. MacKenzie, I. S. and Soukoreff, R. W. Phrase sets for evaluating text entry techniques. *Extended Abstracts of CHI 2003*, 754-755
10. MacKenzie, I. S. and Tanaka-Ishii, K. *Text Entry Systems: Mobility, Accessibility, Universality*. Morgan Kaufmann, 2007.
11. Manning, C. and Schütze, H. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.
12. Mayzner, M. S. and Tresselt, M. E. Table of single letter and digram frequency counts for various word-length and letter-position combinations, *Psychonomic Monograph Supplements I* (1965), 13-32.
13. Microsoft Web N-Gram Services. http://research.microsoft.com/web-ngram
14. Shannon, C. E. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.
15. Wang, K., Thrasher, C., Viegas, E., Li, E. and Hsu, P. An overview of Microsoft Web N-gram corpus and applications. *Proc. of NAACL-HLT* (2010), 45-48.