Representatively Memorable: Sampling the Right Phrase Set to Get the Text Entry Experiment Right

Luis A. Leiva and Germán Sanchis-Trilles
PRHLT Research Center, Universitat Politècnica de València
{luileito,gsanchis}@dsic.upv.es

ABSTRACT

In text entry experiments, memorability is a desired property of the phrases used as stimuli. Unfortunately, to date there is no automated method to achieve this effect. As a result, researchers have to use either manually curated English-only phrase sets or sampling procedures that do not guarantee phrases being memorable. In response to this need, we present a novel sampling method based on two core ideas: a multiple regression model over language-independent features, and the statistical analysis of the corpus from which phrases will be drawn. Our results show that researchers can finally use a method to successfully curate their own stimuli targeting potentially any language or domain. The source code as well as our phrase sets are publicly available.

Author Keywords

Text Entry; Sampling; Memorability; Representativeness

ACM Classification Keywords

H.5.2 Information interfaces and presentation: User interfaces—*Input devices and strategies, Evaluation/Methodology*

INTRODUCTION

In text entry experiments, participants are prompted with phrases (short sentences) that must be entered as quickly and accurately as possible. Although it may seem more natural to have users enter free text and increase thus the external validity[1] of the experiment, it is critical to make the text entry method the only independent variable in the experiment, and increase thus its internal validity.[2] Indeed, if users were asked to type "as fast as possible" they would introduce rather biased text. Hence, researchers typically use pre-selected phrases, measuring the dependent variables (e.g., input speed or error rates) in a text-copy task. This eliminates noise and facilitates the comparison of text input techniques.

In general, copy-tasks should prefer memorable stimuli [7, 8, 13]. Unfortunately, to date there is no automated method to achieve this effect. Researchers resort to using manually curated English-only phrase sets, which are typically small according to modern standards, or rely on sampling procedures that cannot guarantee memorability. In contrast, today text is entered into mobile devices in many different languages, where text entry methods might perform very differently (c.f., English vs. Arabic). This fact evidences the necessity of an adequate sampling method, aimed at exploiting the huge amount of text corpora available in many languages.

We present a method for sampling memorable *and* representative phrase sets, based on a multiple regression model over language-independent features, so that it can generalize to other languages, and the statistical analysis of the (large) corpus from which phrases will be drawn. An interesting property of our method is that, being data-driven, phrases may contain unusual vocabulary as long as it is representative of the task or domain. Our method is validated in two user studies, showing that researchers can now gather their own stimuli for a given language or domain. It is available at `http://personales.upv.es/luileito/memrep/`.

RELATED RESEARCH

For the past decade, text entry researchers have predominately used the MacKenzie and Soukoreff phrase set [8], which contains 500 phrases that were manually selected according to three criteria: moderate in length, easy to remember, and representative of general English. More recently, Vertanen and Kristensson [13] released a phrase set based on genuine mobile emails. Unfortunately, both phrase sets are *1)* manually curated, *2)* only available in English, and *3)* relatively small according to today's standards. In contrast, repositories like the Linguistic Data Consortium, Data Wrangling LLC, or the ELRA catalog provide a plethora of large multilingual corpora that could be curated to automatically build phrase sets tailored to specific tasks or languages.

Paek and Hsu [10] devised a procedure for creating representative phrase sets by randomly sampling sets of n-grams and choosing the set with less entropy with regard to the original dataset. Although mathematically sound, this procedure does not guarantee that sampled phrases are memorable. Moreover, the phrase set generated in this way (the NGRAM dataset) contains incomplete sentences and near two thirds of the words are out of vocabulary,[3] sometimes with extremely unusual punctuation symbols. This might pose a threat to the internal validity of text entry experiments.

In a text-copy task, phrases can be briefly shown at the start or left visible throughout. Kristensson and Vertanen [6] observed that in the latter case entry rates are consistently higher

[1] The extent to which the observed effect is generalizable.

[2] The extent to which the observed effect is due to the test conditions.

[3] Most words are not regular English, according the NEWS dataset.

CHI 2014, April 26–May 1, 2014, Toronto, Ontario, Canada.

ACM 978-1-4503-2473-1/14/04..$15.00.
http://dx.doi.org/10.1145/2556288.2557024

at the cost of higher error rates and longer task times. They reported that the NGRAM phrase set is memorable, but we suspect it is because all sentences are exactly 4 tokens long.

On Entropy and Memorability

Genzel and Charniak [3] postulated in their "entropy rate" principle that speakers tend to produce sentences with similar entropy, so that they can be easily understood. Therefore, too informative sentences (high entropy) should be harder to process and, in consequence, less memorable. However, entropy has been shown to be a weak predictor of processing effort, being the latter better correlated with word length and word frequency [5]. Danescu-Niculescu-Mizil *et al.* [2] analyzed the memorability of movie quotes, concluding that "stand-alone" sentences built on common syntactic scaffolding are likely to be memorable. In general, shorter and frequent words take less time to read and therefore are easily understandable [4]. These observations are key to our work.

METHOD

We are interested in a sampling method to select, from a fairly large text corpus $\mathcal{C}$, those sentences that are good candidates for conducting text entry experiments. Such a method should select phrases that are:

1. **representative** of the task, domain, or language, for ensuring the external validity of the experiment;
2. **memorable**, for ensuring internal validity;
3. **complete**, since fragmented phrases are often confusing.

We assume that the corpus $\mathcal{C}$ contains text that users are likely to write in a real-world context. So, a statistical analysis should uncover commonalities of $\mathcal{C}$, such as phrase length or word frequencies, and the sampled subset should present similar characteristics. In contrast, memorable phrases should be easy to process and write out, which implies that *1)* phrases must be moderate in length, and *2)* words should not be infrequent overall. Then, representativeness and memorability can be pictured as antagonist forces, since e.g. phrases that are very short and contain short and common words will be very memorable, but will rarely be representative either of general language or the task. For instance, a dataset containing only phrases like "this is it" or "and so on" would not be very appropriate for conducting a text entry experiment. We therefore devise a single-pass procedure where each phrase is assigned an expected memorability value that is compensated with a representativeness score.

Modeling Memorability

We first looked at readability tests such as the Coleman-Liau index [1] or ARI [12]. However, they were designed to gauge the understandability of whole documents and are overly simplistic; i.e., they only contemplate sentence and word lengths, and systematically retrieve extremely short phrases. Fortunately, the accuracy (or quality) of a transcribed sentence, measured as the character error rate (CER), is widely accepted to be a good proxy of memorability [6, 7, 13]. Therefore, we focus on modeling CER to predict memorability.

Together with the previously discussed observations, we will attempt to express CER as a function of language-independent features, among which we chose the following:

- **`Nw`**: Number of words in the phrase. Longer phrases are generally harder to process.
- **`OOV`**: Number of infrequent words, relative to **`Nw`**. The higher this ratio, the harder the phrase is to process.
- **`Mchr`**: Average number of characters per word. Shorter words are easier to process.
- **`SDchr`**: Standard deviation of the number of characters per word. Higher variability leads to higher processing effort.
- **`Ppl2`** and **`Ppl3`**: Perplexity (or cross-entropy) of the phrase using word bigrams and trigrams, respectively. Lower perplexities may indicate that the phrase is easier to process.
- **`LProb`**: Natural logarithm of the probability of the phrase. Phrases with high probability are likely to be more usual expressions and so they should be easier to memorize.

Ideally, to compute **`OOV`**, **`Ppl2`**, **`Ppl3`** and **`LProb`** as accurately as possible we would need full knowledge of the "universe" of the target language or domain. Being this impossible, we resort to using a sufficiently large corpus $\mathcal{U}$ and consider that such corpus is descriptive of the desired language, task, or domain. In our experiments, $\mathcal{U}$ is the NEWS corpus[4] (Table 1). Then, *1)* infrequent words are those that do not appear in $\mathcal{U}$; *2)* perplexities are measured with respect to a language model built on $\mathcal{U}$; and *3)* word frequency counts within **`LProb`** are estimated on $\mathcal{U}$. Note that $\mathcal{C} \nsubseteq \mathcal{U}$, otherwise it would lead to over-trained estimations for the phrases in $\mathcal{C}$ (e.g., no word would be considered infrequent).

Language	Sentences	Running words	Vocabulay size	Singletons
English	68.5M	1.6G	3.4M	1.8M
Spanish	13.4M	0.4G	1.2M	0.6M

Table 1: Statistics of the NEWS corpus.

Model Estimation

The CER model was fitted according to generalized linear regression [9], since it allows for any distribution of the model features. We suspected that not all of the features described above, besides being intuitively useful, would be strong CER predictors. For instance, linear models assume that factors are independent of each other. So, in order to decide the best features for the model, we used the Bayesian Information Criterion (BIC) approach [11]. BIC was chosen over other criteria in the literature because it tends to build a simple model and converges as the number of observations increases.

We trained the model with the data released by Vertanen and Kristensson [13], which was drawn from the ENRONMOBILE dataset and provides 22,390 CER-labeled sentence memorization tasks completed by 386 MTurk workers. All workers were from the United States and India, either native speakers or having a competent English level. Workers had to memorize a sentence and then type it after pressing a continue button. Having 10 observations per phrase, intra-phrase observations that exceeded 1.5 times the interquantile CER range were considered outliers (e.g., a worker being distracted or

[4] `http://statmt.org/wmt13/translation-task.html`

writing a completely different sentence), and the remaining observations were averaged on a phrase basis. Eventually 2,390 "data points" were considered.

BIC revealed that the significant features to predict CER were only **Nw**, **OOV**, **SDchr**, and **LProb**. The model yielded a good fit (adj. $R^2 = 0.63$), with the following combination of features:

$$\textbf{CER} \approx -11.65 + 0.83 \cdot \textbf{Nw} + 0.48 \cdot \textbf{SDchr} + 6.94 \cdot \textbf{OOV} - 1.00 \cdot \textbf{LProb} \quad (1)$$

As observed, **Nw**, **SDchr**, and **OOV** have positive weights, implying that CER increases when sentences get longer, words are more infrequent, and words have a highly variable amount of characters. Conversely, **LProb** having a negative weight implies that the more likely the sentence, the less prone users are to make mistakes. In addition, **OOV** receives a much higher weight than the rest of the features, indicating that infrequent words are specially correlated with high CER values.

Ensuring Representativeness

Presumably, selecting sentences with the lowest CER estimates would yield the most memorable ones, although such sentences would end up being those with few and short words. To compensate this effect, we also need to ensure that sentences are representative either of general language, the sentence corpus, or the desired task for the text entry experiment. To achieve this, we estimate the empirical probability of the features present in Eq. (1), and reward those phrases that have a higher probability according to such prior distribution. Assuming Gaussian distributions, we can estimate the mean μ_i and standard deviation σ_i of each feature h_i on the target corpus $\mathcal{C}$. Then, the representativeness of a sentence is given by:

$$\textbf{Repr} = \prod_i P_{\mathcal{C}}(h_i; \mu_i, \sigma_i) \quad (2)$$

where $P_{\mathcal{C}}(h_i; \mu_i, \sigma_i)$ is the probability distribution of each feature h_i according to $\mathcal{C}$. Moreover, adjusting the meta-parameters μ_i and σ_i allows text entry researchers to fine-tune the kind of phrases that will be eventually used in the text entry experiment.

Finally, since we want to retrieve phrases with high memorability (low **CER**) and high representativeness (high **Repr**), we define the final score assigned to a phrase by the following expression:

$$\textbf{sc}(\text{phrase}) = \frac{\textbf{CER}}{\textbf{Repr}} \quad (3)$$

so the lower the score, the better. This way, our sampling method provides a closed-form solution to sample the (hopefully) right phrases to get the text entry experiment right.

EVALUATION

We tapped into the MACKENZIE dataset (500 sentences) and the NGRAM dataset (500 4-gram sentences). Both datasets have been reported to be memorable by native English speakers [6], so we replicated the analysis with non-natives. We also sampled 500 sentences at random from the public EUROPARL dataset (1.8M sentences from the proceedings of the European parliament), and 500 more following our method. Because EUROPARL contains overly long sentences (28 words per sentence on average) we restricted both samplings to lowercased phrases of 3–10 words, otherwise the random condition would be placed at a disadvantage; and punctuation symbols were removed (as in MACKENZIE). Moreover, $P_{\mathcal{C}}(\textbf{Nw})$ was set to that of MACKENZIE, resulting in phrases of 5–6 words, since estimating μ_i for **Nw** according to EUROPARL would be heavily skewed and therefore would tend to retrieve the longest sentences. Even though such sentences would be representative of this specific topic (parliamentary proceedings), they may not be reflective of everyday language and thus would not be very appropriate for conducting text entry experiments, where memorability is important.

We recruited 20 native Spanish speakers aged 28–38 using the available University's mailing lists. All participants had a qualified intermediate or advanced English degree according to CEFR.[5] Each participant was shown a phrase for 5 seconds or until the first keystroke, whatever happened first. Next, the phrase disappeared and users had to write it (as much as they could remember) with a physical QWERTY keyboard. Participants entered 20 phrases in a randomized order from each dataset, resulting in 1,600 annotated phrases in total. The results are show in Table 2. Together with CER (in %), we report the words per minute (WPM) to give an overview of the participants performance in terms of input speed. We also report the time since the phrase was loaded until the first keypress (T_k, in seconds), which provides an estimate of the time spent memorizing each phrase.

Dataset	WPM	CER	T_k
MACKENZIE	67.90 (12.23)	2.04 (2.58)	2.62 (0.74)
NGRAM	65.70 (11.34)	3.07 (2.45)	2.64 (0.74)
EUROPARL random	64.80 (13.63)	6.63 (4.68)	3.47 (0.80)
EUROPARL ours	65.49 (11.47)	1.40 (1.48)	3.27 (0.90)

Table 2: Evaluation results for English. SDs in parentheses.

An ANOVA test revealed that differences between datasets were statistically significant for CER [$F_{3,72} = 10.75, p < .0001, \eta_p^2 = 0.31$] and memorization time [$F_{3,72} = 5.89, p = .0012, \eta_p^2 = 0.20$]. Post-hoc pairwise t-tests (Bonferroni-Holm corrected) revealed that the phrase set derived by our method compares favorably to state-of-the-art phrase sets (no differences were found), and that participants performed significantly worse in the random sampling condition, both in terms of CER and memorization time. Overall, we observed more variability in our data in comparison to previous literature [6], which motivates the need to provide participants with sentences in their native language.

In light of the previous study, we generated 3 phrase sets of 500 sentences each by tapping into the Spanish version of EUROPARL. We used random sampling, the n-gram sampling procedure (with $n = 6$), and our method. In all cases, we used lowercased phrases of 3–10 words with punctuation symbols removed. We then repeated the same experiment with the same participants. The results are show in Table 3. Again, differences between datasets were found to be statistically

[5] http://www.coe.int/lang-CEFR

significant for CER [$F_{2,54} = 5.90, p = .0048, \eta_p^2 = 0.18$] and memorization time [$F_{2,54} = 20.94, p < .0001, \eta_p^2 = 0.44$]. Post-hoc pairwise t-tests (Bonferroni-Holm corrected) revealed that the random and n-gram conditions performed equally similar, and that our method performed significantly better than the other sampling procedures. This study shows that our method generalizes well to Spanish, which is a language quite different from English.

Dataset	WPM	CER	T_k
EUROPARL random	70.21 (13.70)	2.33 (1.96)	3.32 (0.60)
EUROPARL ngram	72.04 (15.71)	3.10 (2.53)	3.20 (0.66)
EUROPARL ours	74.51 (14.19)	0.85 (1.55)	2.24 (0.37)

Table 3: Evaluation results for Spanish. SDs in parentheses.

In both studies, a common complaint from participants was that "some phrases are incomplete [...] they are distracting and difficult to memorize." This observation backed up our intuitions regarding the desired properties of a phrase set for text entry experiments (see Method section).

DISCUSSION AND CONCLUSION

Many text entry methods require constant visual attention, such as eye typing or dialing a contact while driving. For experiments trying to emulate these or similar situations, memorability is critical since it can be difficult for participants to consult often the reference text. Memorability is also desirable to unburden the participants and let them exclusively focus on the text entry method. Until now, sampling methods aimed to select "representative" phrases, but memorability was largely ignored. This work therefore significantly contributes to HCI by making it possible to curate large text corpora in potentially any language and any task or domain.

Our findings are of special relevance to text entry researchers interested in conducting experiments tailored to the linguistic capabilities of their participants. Memorability was found to be correlated with sentence length, word variability, word frequency, and ratio of infrequent words. In sum, shorter phrases with frequent vocabulary are easier to remember. These findings were consistent in all of our experiments, and ultimately are aligned to previous key findings in the literature. Table 4 provides an overview of the type of phrases that can be deemed as being either of good or bad quality, as scored by our method. This illustrates a means to filter (un)desired phrases from the phrase set used as stimuli, in order to ensure the validity of the text entry experiment.

We have shown that our method can generalize to different domains and languages different from English. This provides text entry researchers with a scalable and unprecedented capability. However, it must be noted that our features may not be applicable to every possible language. For instance, Chinese words are not even formed by individual letters. Therefore, there is still an opportunity for future work. Thus one possibility would be experimenting with other memorability predictors, which could lead to better models that would in turn improve our sampling method. We hope that this work will be suitable for use in a variety of text entry evaluations.

Phrase set	Quality	Sample phrases
MACKENZIE	+	mary had a little lamb
	+	february has an extra day
	-	if you come home late the doors are locked
	-	rent is paid at the beginning of the month
NGRAM	+	will receive a unit
	+	is likely to propose
	-	& cyndi clark [mailto:bcclarks.att.net
	-	and skills–attend critical information-packed=20
EUROPARL	+	i forgot to mention it
	+	all human life is valuable
	-	two questions still remain however
	-	secondly the economic and financial crisis

Table 4: Good (+) and bad (-) phrase examples to conduct text entry experiments, according to Eq. (3). Good examples were scored the lowest.

ACKNOWLEDGMENTS

This work is supported by the 7th Framework Program of the European Commision (FP7/2007-13) under grant agreements 287576 (CASMACAT) and 600707 (tranScriptorium).

REFERENCES

1. Coleman, M., and Liau, T. L. A computer readability formula designed for machine scoring. *Journal of Applied Psychology 60*, 1 (1975).
2. Danescu-Niculescu-Mizil, C., Cheng, J., Kleinberg, J., and Lee, L. You had me at hello: How phrasing affects memorability. In *Proc. ACL* (2012).
3. Genzel, D., and Charniak, E. Entropy rate constancy in text. In *Proc. ACL* (2002).
4. Just, M. A., and Carpenter, P. A. A theory of reading: From eye fixations to comprehension. *Psychological Review 87*, 1 (1980).
5. Keller, F. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *Proc. EMNLP* (2004).
6. Kristensson, P. O., and Vertanen, K. Performance comparisons of phrase sets and presentation styles for text entry evaluations. In *Proc. IUI* (2012).
7. MacKenzie, I. S., and Soukoreff, R. W. Text entry for mobile computing: Models and methods, theory and practice. *Human-Computer Interaction 17*, 2 (2002).
8. MacKenzie, I. S., and Soukoreff, R. W. Phrase sets for evaluating text entry techniques. In *Proc. CHI EA* (2003).
9. Nelder, J. A., and Weddrburn, R. W. M. Generalized linear models. *J. R. Statist. Soc. 135*, 3 (1972).
10. Paek, T., and Hsu, B.-J. P. Sampling representative phrase sets for text entry experiments: a procedure and public resource. In *Proc. CHI* (2011).
11. Schwarz, G. E. Estimating the dimension of a model. *Annals of Statistics 6*, 2 (1978).
12. Senter, R. J., and Smith, E. A. Automated readability index. Tech. Rep. AMRL-TR-6620, Wright-Patterson Air Force Base, 1967.
13. Vertanen, K., and Kristensson, P. O. A versatile dataset for text entry evaluations based on genuine mobile emails. In *Proc. MobileHCI* (2011).