Machine Learning for Memorable Word Encodings

Nathaniel J. McAleese
University of Cambridge
Cambridge, UK
nm583@cam.ac.uk

ABSTRACT

Many interactions with computers require the recall of arbitrary information. Previous approaches have often neglected to take a human-centred approach to this problem, resulting in phone numbers, passport numbers, IP addresses and passwords that are difficult for humans to recall. This paper proposes a backwards compatible encoding scheme that allows arbitrary information in such domains to be encoded as memorable sequences of words. A controlled experiment on human participants verifies that these sequences do indeed improve short-term recall.

CCS Concepts

•Human-centered computing → Empirical studies in HCI; Empirical studies in interaction design;

Author Keywords

memory, memorability, information, encoding, passwords

INTRODUCTION

Society requires users to recall and transmit information. Often this includes unique identifiers, such as national insurance numbers (< 34 bits of entropy), bank account details (< 46 bits), or phone numbers (< 34 bits).

It should be immediately apparent that not all encodings of information are equally memorable to humans because our cognitive systems are heavily adapted to certain tasks that were relevant to survival in the ancestral environment [6].

The goal of this work is to demonstrate one way in which machine learning (ML) can exploit this fact. An encoding scheme is developed that maps a large space of random word strings (encoding n bits of information or more) into a smaller required space (of size m bits where $m < n$) and exploits the redundancy in such an encoding to seek the word strings that are most memorable.

CHI'16, May 07–12, 2016, San Jose, CA, USA

ISBN 123-4567-24-567/08/06. . . $15.00

DOI: http://dx.doi.org/10.475/123_4

RELATED WORK

There are several related areas of work. Classical psychology has focused on the nature of memory for some time, and detailed analysis has been done of the "chunking" phenomenon that was first proposed in "The Magic Number Seven, plus or minus two" [17]. The key observation is that the capacity of human short-term memory increases almost linearly with the amount of information stored in each *chunk* of information presented. The question of what constitutes a single chunk of information to a given user is difficult to answer in the general case, but specific examples include digits, characters and words. Note that the true information content of these chunks varies widely, and thus learning to chunk information (e.g. by converting binary digits to hex) is one of the most effective means of improving short-term recall of arbitrary information [17].

There are further complexities that effect memorability. Layout and kinesthetic clues both influence users' recall ability [14, 22], but more critical influences for short word and character sequences are semantic and acoustic similarity [4], with acoustic similarity producing a particularly strong effect.

Interaction designers have capitalised on these effects before. In 1998, the PGPfone alphabet [9] was developed to satisfy the need to communicate complex binary information over a noisy human channel (telephone conversations) by using a phonetic alphabet designed to minimise acoustic similarity between words, however, its construction used a relatively primitive phonetic model and it was not evaluated with any user study.

Work in HCI and memorability has also investigated passwords in particular detail. Interesting approaches have used approximate techniques that rely on the recall of vague impressions of images or sequence statistics left by previous exposure [25], but these techniques are not currently appropriate for encoding arbitrary information. "Persuasive" techniques that interactively prompt the user to improve the security of their passwords have also been investigated, as has the effect of such schemes on memorability. Results suggest that, given a memorable phrase, users can be prompted to increase its security with no effect on recall [7].

Conflicting results show both that passphrases[1] are easier to recall and usually entered accurately [10], and are no easier to recall and less accurately entered than random passwords

[1] long passwords consisting of real words, such as "correct horse battery staple"

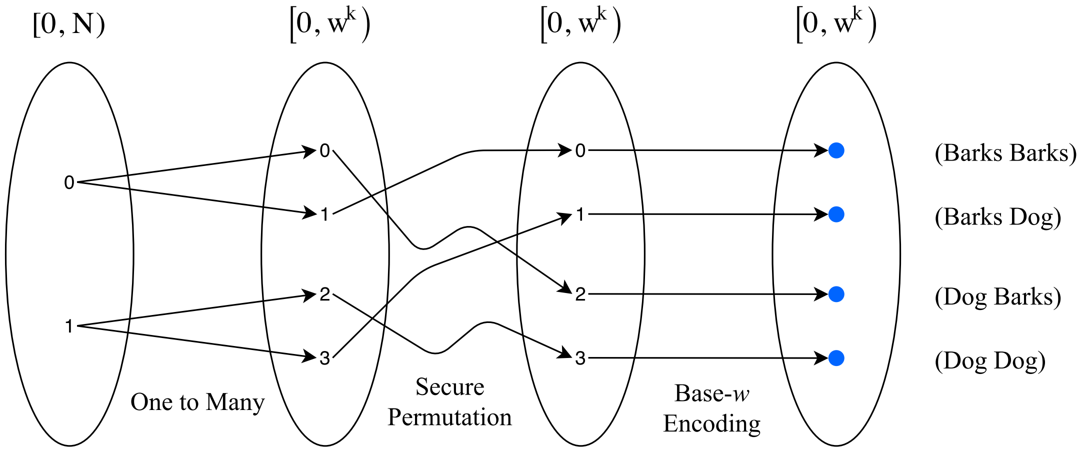

Figure 1. This toy example illustrates the approach for $N = 2$ with a vocabulary consisiting of {Barks, Dog}. In a practical application the vocabulary is much larger — see Table 1 for realistic encodings of a number.

[11]. However, these reports do not account for the advantages of using a fixed vocabulary for encoding information. Entry errors in [11] were classified as being typographic when the entered text was below a certain Levenshtein ratio from the true text - with appropriate selection of a fixed vocabulary, all these errors can be corrected automatically on entry without reducing the information content of the string.

The machine learning techniques that make motivate modelling memorability are advances in "sequence to sequence" learning [21] and results indicating that these approaches are particularly applicable to both acoustic and semantic language modelling, for example, text-to-speech and machine translation systems [23, 1]

ENCODING METHOD

Consider a finite space of N identifiers, such as the numbers 0 to $10^{10} - 1$. Given a vocabulary of w words, it is straightforward to map each number to a sequence of words - we may simply consider the number in base w. Thus for three-word encodings, we may enumerate w^3 different encodings, and in general for k word encodings w^k. If w^k is larger than an integer multiple of N, then we may have more than one encoding for each number in the space 0 to N. For example, with a vocabulary size of 2, a word encoding size of 2 and $N = 2$, we could map each number $0 \leq n < 2$ to 2 distinct word encodings. This toy example is exactly illustrated in Figure 1.

Most naive methods of constructing this mapping provide very little diversity in the set of codes produced. Fortunately, cryptography provides a convenient primitive that we may exploit; by using a block cipher on an arbitrary finite domain [2, 5] we may construct a secure permutation on the integers $0 \leq n < w^k$. This then gives us a constant time reversible mapping between the domain to be encoded (for example telephone numbers) and individual word encodings; the multiple available encodings for a particular number may also be quickly enumerated. Table 1 shows a portion of the encodings of zero under the vocabulary that was used for the experiment.

Having obtained a diverse set of word encodings for every number in the desired domain, we may then score them using any of a number of appropriate techniques. In this work, we focus on using a language model trained on a corpus of jokes to score the provided encodings (see Section 4). The hypothesis is that word encodings likely under such a model are more memorable.

Encodings of 0
returned shell implementation
engineer menu well
prerequisite puerto corp
horrible attacked gross
suspected philip differ
...

Table 1. By using a secure permutation, we can easily ensure that the available encodings of each number in the domain are diverse. Some encodings of zero are shown using the same vocabulary as the experiment.

This method also much more efficient in terms of the number of words required for each code (k) when compared to previous approaches. This further improves memorability [17]. Consider the PGPhone alphabet and related designs, in which phonetically similar words are removed from the vocabulary of available words. When we remove x words from the corpus due to phonetic similarity to each other, we hugely reduce the number of available encodings from w^k to $(w - x)^k$. By contrast, dynamically eliminating phonetically similar encodings from the diverse space of available options for each number to be encoded has much less of an impact on the number of words needed - intuitively because the first approach eliminates all codes containing a particular word, instead of just those codes that contain the word and are insufficiently phonetically diverse. A precise statement of this effect is included in Appendix A.

LANGUAGE MODEL

A character level language model [12] was used to score the encodings. This takes a sequence of characters and estimates the log probability of its occurrence in the training corpus. In particular, a recurrent neural network (RNN) was used

Decimal	Most Likely	Least Likely
10-96-94-33-97	*forget worse step*	*equilibrium versus dude*
10-97-15-77-40	*wrong five should*	*diameter live bibliographic*
91-25-37-36-30	*block instant magical*	*concerts litigation tongue*
97-70-55-91-95	*shake issues given*	*baltimore reliability commonwealth*

Table 2. The number encoded and the highest and lowest scoring encodings under the system. By using a language model to rank the available encodings for each number, we natually prefer shorter encodings that use common words. The approach also promotes the use of more gramatically correct encodings.

based on the gated recurrent unit [3]. 128 dimensions were used for the character embeddings, and 512 for the hidden representation. AMSGrad [19] was used as an optimizer. A complete specification of model hyperparameters is publically available alongside the rest of the implementation of the word encoding system [15].

The model was then trained on *n*-grams extracted from a large corpus of jokes [13]. Note that a classical *n*-gram language model would not have been appropriate in this case, unless an extremely large corpus is available. This is because many of the encodings presented for scoring will not have been previously available in the training data. A naive *n*-gram model would not be able to exploit morphology to provide a ranking in these instances. This presents no issues to a character-level model, for further discussion see [12].

PHONETIC MODEL

A character-level RNN was also used to embed the vocabulary into a phonetic space. This was done with an identical architecture to the language model that was trained on the CMU phonetic dictionary [24] to predict the pronunciation of a word from its spelling. The hidden representation of the model may then be used as a word embedding and the phonetic variation in a given encoding may be defined as the minimum distance between the embedding of any two words in the encoding.

EXPERIMENTAL DESIGN

Experiments that evaluate the behaviour of long-term memory are complex and difficult to run. Ideally, they take the form of long-term ethnographic studies that observe the behaviour of users over several months [8, 26]. Due to practical limitations, this sort of study was not viable, and so in order to provide initial support for the proposed idea, experiments were conducted to investigate the effect of the proposed encoding on short-term recall. Short term manipulations are useful in and of themselves, and short-term memorability is at least somewhat correlated with long-term memorability, even if the relationship is complex [18].

A small web application was used to test the memorability of the different encoding schemes. Upon visiting the site (after registration of informed consent) each user is automatically assigned a unique identifier. They then proceed through three testing phases. In each, they are asked to study a particular encoding of their identifier. They are then presented with a distractor task, before being asked to input the code from memory. In order to account for the possibility of ordering effects, the order of the three encodings is randomised for each participant.

Query	Phonetically Similar		
microsoft	microwave	microphone	ministers
health	healthy	heath	wealth
mexico	mexican	monaco	medicine
obituaries	abilities	obesity	officially
kick	pick	click	nick

Table 3. Qualatative evalutation of randomly selected words suggests that the phonetic model captures similarity well.

The distractor task was selected in the style of [20] and occupied the participants for a median time of just under 20 seconds. It required the participants to click on recognisable images that moved between random static positions on the page.

Participants were not supervised during the experiment, and thus there was the potential for cheating. This might reduce the external validity of the test, and thus some steps were taken to mitigate the possibility of corrupted results. To help reduce cheating a prompt was added to the initial instructions asking participants not to physically record the codes in any way. Copy and paste were also programmatically disabled. To aid in identifying cheating, a question was added to the end of the test that explicitly stated *"There is no penalty for answering yes. Did you write down or otherwise record any of the codes?"*.

A pilot study was carried out with student participants from the computer science department, and the initial results influenced the design of the larger study. In particular, the initial version of the experiment required users to go back and retrieve codes that they entered incorrectly. This was intended to increase external validity by mirroring a real interaction in which the user cannot proceed until the code is entered correctly, but this component was removed after it was found to dramatically increase the ordering effect of the questions — after a mistake, participants would spend longer on each code and error rates dropped to near zero. The pilot study was also used to tune the length of the distractor task, targeting 30 seconds of engagement.

RESULTS & DISCUSSION

88 participants were recruited over the internet to take the test; 25 confessed to cheating in the final question. Thus the remaining 63 were considered in the analysis.

Figure 2 shows the results of the human trial. The exact McNemar test [16] was used to determine whether the probability of successful recall varied between the encoding types. The

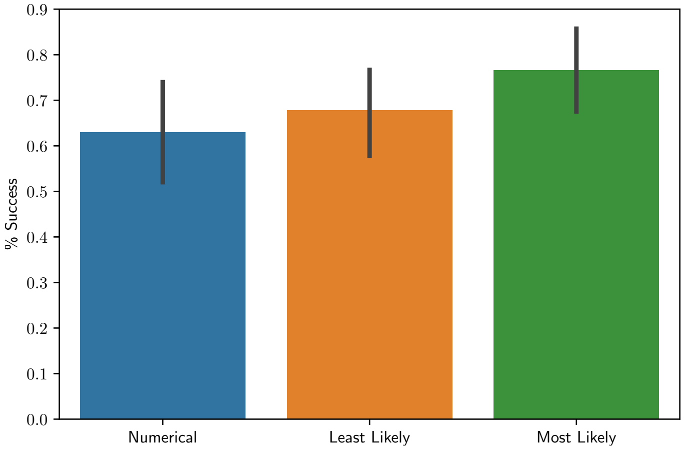

Figure 2. The percentage of successful recall for each of the code types. The error bars show 95% confidence interval from a normal approximation to the binomial.

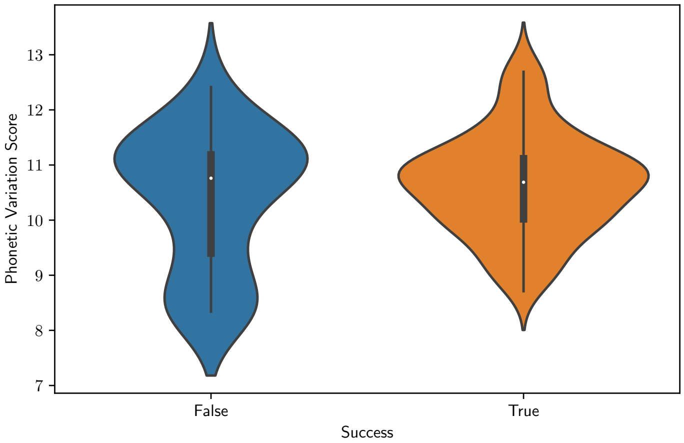

Figure 3. The data were inconclusive about the influence of phonetic variation on code memorability. Whilst the literature would lead us to expect that greater phonetic variation (x axis) would lead to a larger proportion of successful recalls, this trend does not appear in the data.

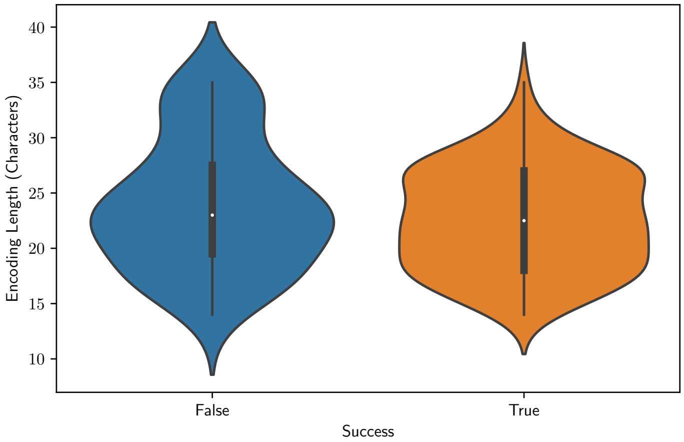

Figure 4. The difference in recall is not accounted for by variation in the number of characters in the encoding.

model selected "best" encoding shows a statistically significant improvement in successful recall over both the naive numerical encodings ($p = 0.016$) and the least likely word encodings ($p = 0.049$).

In general, we would anticipate that longer encodings (e.g. those with more letters, due to the selection of longer words) would have a lower score under the language model. Figure 4 shoes that the improved recall was not simply due to the language model selecting shorter encodings. This is a promising result as if the total length of the words in the encoding was the primary influence on memorability, then the use of a sophisticated model would have been unnecessary.

This demonstrates the validity of the core idea, but we can also examine whether phonetic variation influences the probability of successful recall amongst the collected data. The use of such models initially seemed promising, because phonetic variation is associated with ease of recall within the existing literature, but as shown in Figure 3 this is not borne out by the data. This result merits further investigation — inspecting the model output for randomly selected words, as is done in Table 3 quite strongly suggests that this is not due to a failure of the phonetic model. With further experimentation, it seems likely that such measures of phonetic similarity could be incorporated into the system to improve overall recall.

In many ways, the approach described so far represents the minimum viable demonstration of the proposed idea. With additional resources, it would be extremely interesting to explore other models of human memory. Approaches might include:

- Models explicitly trained to predict memorability.
- Personalised models, that adapt to a specific user.

Other approaches might circumvent the need for modelling memory at all by showing the user some or all of the available encodings and allowing them to choose. This is not always appropriate, however, for several reasons. If the set of available encodings is large, because $N << w^k$, then the amount of choice may be impractical for the user to consider. Requiring interaction also eliminates some popular methods of assigning identifiers, such as by post or email[2].

In general, as improvements move the system away from proxying memorability from existing sources of information it will become increasingly interesting from an interaction perspective. Indeed, schemes that directly predict memorability are in some sense doing offline inference that can predict later user behaviour; a useful paradigm for modelling systems that seek to chose, ahead of time, what will make the user's life easier.

FUTURE WORK & CONCLUSION

This work demonstrates that ML can be used to improve the ability of users to recall arbitrary information. Encoding schemes such as the one proposed here could ease numerous existing interactions that people have with computers by directly replacing hard-to-recall strings of digits with carefully

[2] The author doesn't endorse these methods, but nonetheless compatibility with them is desirable.

selected codes, and the backwards compatibility of the proposed approach would allow for straightforward deployment through a website or browser extension.

The system, however, is far from perfect. The use of a language model to proxy memorability is plainly sub-optimal, and further investigation should almost certainly be able to improve the scoring of encodings to further aid memorability.

One exciting avenue for work would be an active learning approach - websites must provide fallback mechanisms for users who forget their personal information, and once a system like this was in place this could be used as feedback to improve the model of memorability (for publically available identifiers). Personalisation and theming also offer the potential to improve the system in its current form.

Modern machine learning techniques allow us to better model the way in which humans understand the world. Exploiting this fact will constitute a great deal of exciting future work in interaction design.

ACKNOWLEDGMENTS

We would like to thank the numerous experimental participants, without whom this work would not have been possible, and Angus Hammond who noted the size of the efficiency gain when compared to word elimination, specified in Appendix A.

REFERENCES

1. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
2. John Black and Phillip Rogaway. 2002. Ciphers with arbitrary finite domains. In *CryptographersâĂŹ Track at the RSA Conference*. Springer, 114–130.
3. Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
4. R Conrad, A D Baddeley, and A J Hull. 1966. Rate of presentation and the acoustic similarity effect in short-term memory. *Psychonomic Science* 5, 6 (1966), 233–234. DOI:`http://dx.doi.org/10.3758/BF03328368`
5. Sashank Dara and Scott Fluhrer. 2014. FNR: Arbitrary length small domain block cipher proposal. In *International Conference on Security, Privacy, and Applied Cryptography Engineering*. Springer, 146–154.
6. Charles Darwin. 1859. *On the Origin of Species by Means of Natural Selection*. Murray. or the Preservation of Favored Races in the Struggle for Life.
7. Alain Forget and Robert Biddle. 2008. Memorability of Persuasive Passwords. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. 3759. DOI: `http://dx.doi.org/10.1145/1358628.1358926`
8. PG Inglesant and M Angela Sasse. 2010. Studying password use in the wild: practical problems and possible solutions.
9. Patrick Juola. 1996. Whole-word phonetic distances and the pgpfone alphabet. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, Vol. 1. IEEE, 98–101.
10. Mark Keith, Benjamin Shao, and Paul Steinbart. 2009. A Behavioral Analysis of Passphrase Design and Effectiveness. *Journal of the Association for Information Systems* 10, 2 (2009), 63–89. DOI: `http://dx.doi.org/Article`
11. Mark Keith, Benjamin Shao, and Paul John Steinbart. 2007. The usability of passphrases for authentication: An empirical field study. *International Journal of Human Computer Studies* 65, 1 (2007), 17–28. DOI: `http://dx.doi.org/10.1016/j.ijhcs.2006.08.005`
12. Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-Aware Neural Language Models.. In *AAAI*. 2741–2749.
13. Rohit Kulkarni. 2017. Urban Dictionary Words And Definitions. `https://www.kaggle.com/therohk/urban-dictionary-words-dataset`. (2017). (Accessed on 03/23/2018).
14. Svenja Leifert. 2011. The influence of grids on spatial and content memory. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11*. 941. DOI: `http://dx.doi.org/10.1145/1979742.1979522`
15. Nat McAleese. 2018. wordify/language_model.py at master Âů N-McA/wordify. `https://github.com/N-McA/wordify/blob/master/wordify/language_model.py`. (3 2018). (Accessed on 03/22/2018).
16. Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12, 2 (1947), 153–157.
17. George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.
18. Bennet B Murdock. 1974. *Human memory: Theory and data.* Lawrence Erlbaum.
19. Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. 2018. On the convergence of adam and beyond. In *International Conference on Learning Representations*.
20. T Shallice, P Fletcher, CD Frith, P Grasby, RSJ Frackowiak, and RJ Dolan. 1994. Brain regions associated with acquisition and retrieval of verbal episodic memory. *Nature* 368, 6472 (1994), 633.
21. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

22. Desney S Tan, Randy Pausch, Jeanine K Stefanucci, and Dennis R Proffitt. 2002. Kinesthetic cues aid spatial memory. *extended abstracts on Human factors in computer systems CHI 02* (2002), 806–807. DOI: http://dx.doi.org/10.1145/506443.506607

23. Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

24. Robert L Weide. 1998. The CMU pronouncing dictionary. *URL: http://www. speech. cs. cmu. edu/cgibin/cmudict* (1998).

25. Daphna Weinshall and Scott Kirkpatrick. 2004. Passwords you'll never forget, but can't recall. *In Proceedings of CHI 04 extended abstracts* (2004), 1399–1402. DOI: http://dx.doi.org/10.1145/985921.986074

26. Jeff Yan, Alan Blackwell, Ross Anderson, and Alasdair Grant. 2004. Password memorability and security: Empirical results. *IEEE Security & privacy* 2, 5 (2004), 25–31.

Appendices

COMPARISON TO ELIMINATING WORDS

Supposing there are x phonetically distinct words in the vocabulary, each of which is phonetically similar to n words. This uniform distribution is the worst possible case for a comparison between the proposed method and word elimination. For encodings of length k, word elimination gives us a space of size x^k, whereas selecting distinct encoding such that no encoding contains two or more phonetically similar words provides a space of size:

$$xn \times (x-1)n \times ... \times (x-k)n = n^k \frac{x!}{(x-k)!} \tag{1}$$

For practical values of $x = 10, n = 1000, k = 3$ this results in a space roughly a thousand times bigger, allowing the encoding of 10 digit numbers as opposed to 7. As noted in the acknowledgements, thanks is provided to Angus Hammond, who wrote down the above formula as a triviality.