Domain-Specific Reduction of Language Model Databases: Overcoming Chatbot Implementation Obstacles

Nicholas J. Kaimakis, Samuel Breck, & Benjamin D. Nye
Institute for Creative Technologies, Univ. of Southern California
Los Angeles, California
{kaimakis, breck}@usc.edu, nye@ict.usc.edu

Dan M. Davis
HPC-Education/USC
Long Beach, California
dmdavis@acm.org

ABSTRACT

This paper addresses optimization of language model databases for use in range of chatbots, *e.g.* virtual conversational mentors. The proliferation and practical use of chatbots depends on the ability of users to conversationally retrieve information or activate events. These conversations ideally help the user efficiently access a broad set of computational functions accurately and in as few turns as possible. However, there is a concomitant pressure to reduce size, increase speed, and enable offline capabilities. Such offline and online dialog capabilities are particularly vital for DoD applications, which are increasingly being developed to produce intelligent agents that help explain complex data and operate in low-resource environments (*e.g.*, no reliable internet). Unfortunately, not only is natural language processing a computationally expensive task, but language models are often heavyweight, with large storage space footprints and challenges for keeping data in-memory that affect search and retrieval times. A chatbot must be able to understand and respond to a variety of unique user inputs and do so within the limits of conversationally tolerable latencies. The difficulty comes in trying to balance an effective chat agent while optimizing storage for peak performance. This paper applies and analyzes two methods for reducing language models, using the Google News Word2vec model as an example. These methods were implemented with the goal to support natural a language dialog system without the storage overhead of the significantly larger comprehensive models. Two metrics were applied to reduce the language model: word frequency and relevance to the domain-specific information that the agent can discuss. We will document and analyze similar efforts, set forth potential approaches, discuss solutions within our environment, quantify impacts of the implementation of these approaches, outline future applications, and suggest topics for further research. The paper closes with ways in which others can adopt this approach to their own efforts.

ABOUT THE AUTHORS

Nicholas Kaimakis is active in research at the Institute for Creative Technologies of the University of Southern California. His current research thrusts are in the use of computer generated avatars or video clips, animated and directed by natural language optimized Artificial Intelligence (A/I) programs that present a life-like dialogue capability to interact with remote users via the internet. He has demonstrable success in designing efficient project structures, teaching programming to varied audiences, and managing multi-faceted teams. His current project is funded by the Navy and is designed to help improve knowledge of STEM fields across varied demographics with the development of an interactive interface that makes STEM information more accessible on-line. He is studying Computer Science and Business Administration (CSBA) at the Viterbi School of Engineering at the University of Southern California.

Dan M. Davis is a consultant for the University of Southern California, focusing on large-scale distributed DoD simulations and virtual conversational computer agents and avatars. Pre-retirement, he was the Director of USC's JESPP project at JFCOM for a decade. As the Assistant Director of the Center for Advanced Computing Research at Caltech, he managed Synthetic Forces Express, bringing HPC to DoD simulations. Prior experience includes serving as a Director at the Maui High Performance Computing Center and as a Software Engineer at the Jet Propulsion Laboratory and Martin Marietta. He has served as the Chairman of the Coalition of Academic Supercomputing Centers and has taught at the undergraduate and graduate levels. As early as 1971, Dan was writing

programs in FORTRAN on one of Seymour Cray's CDC 6500's. He saw duty in Vietnam as a USMC Cryptologist and retired as a Commander, USN. He received B.A. and J.D. degrees from the University of Colorado in Boulder.

Samuel Breck is a Research Assistant at the Institute for Creative Technologies of the University of Southern California. His major research interests are any platforms involved with natural language processing and the upcoming infrastructural challenges smart cities will face. He has been an intern at several tech companies including PlayStation and lives in the Bay Area. He anticipates receiving a BS degree in Computer Science from the Viterbi School of Engineering, University of Southern California in May 2019.

Benjamin Nye, Ph.D. is the Director of Learning Science at the University of Southern California, Institute of Creative Technologies (USC ICT). Ben's research tries to remove barriers to development and adoption of adaptive and interactive learning technology so that they can reach larger numbers of learners. He serves as the principal investigator for the MentorPal project. He is the membership chair for the International Artificial Intelligence in Education (IAIED) Society and holds memberships in Educational Data Mining Society (EDM), and Association for the Advancement of Artificial Intelligence (AAAI). He also co-chairs the FLAIRS Learning Technologies track (2015-2017). He earned a B.S. degree in Computer Science from Trinity College Hartford Connecticut and a PhD in Electrical and Systems Engineering from the University of Pennsylvania.

Domain-Specific Reduction of Language Model Databases: Overcoming Chatbot Implementation Obstacles

Nicholas J. Kaimakis, Samuel Breck, & Benjamin D. Nye
Institute for Creative Technologies, Univ. of Southern California
Los Angeles, California
{kaimakis, breck}@usc.edu, nye@ict.usc.edu

Dan M. Davis
HPC-Education/USC
Long Beach, California
dmdavis@acm.org

INTRODUCTION AND BACKGROUND

Enhancements in Artificial Intelligence (A/I) and Natural Language Processing (NLP) have now achieved the stage in which personal conversational interfaces are possible. Acceptance of these means that large-scale implementation issues are increasingly seen as imminent challenges (Traum et al., 2015). That, coupled with the popularity of hand-held devices such as tablets, MP3 players, and smart-phones has brought to the fore the issue of not over-burdening these limited resources with data storage and data transfer requirements that would tax their memory storage limits and bandwidth constrictions. The Institute for Creative Technologies (ICT) of the University of Southern California (USC) has been engaged by the Office of Naval Research to conduct research on MentorPal, a computer-generated mentor capable of sustaining a conversational series of responses to secondary school students who are considering careers, especially t careers in one of the Science, Technology, Engineering, and Math (STEM) disciplines. An early version is shown in Figure 1. The prototype for this interactive mentor targeted the Microsoft Surface® line of tablet computers. The limited power, RAM and secondary memory of these small portable devices mandates optimal sizing of all processes and data storage and transfer (Nye et al., 2017). This design challenge led to the guiding problem behind the model reduction techniques explained here.

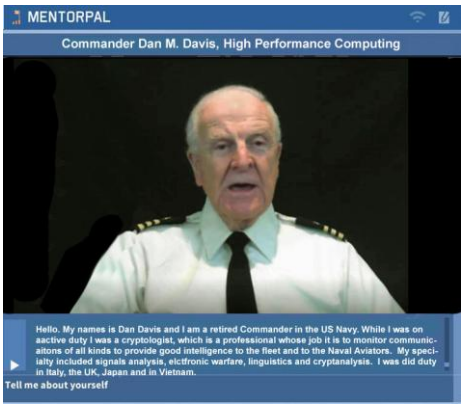

Figure 1 - MentorPal Display Screen Shot

The authors see the paper's insights as being extensible into a broad range of applications that require more and more processing power and data management capabilities. The authors assert that the insights gained and the techniques employed apply to model reductions in practically any specific domain, rather than only to chatbots. Part of this assertion is recognizing the need to prepare for operations under conditions of limited computational power and data capacity, while at the same time designing code to make it amenable to scaling if a less constricted environment is available. These constrained environment issues are known to be common in much of the work currently being conducted at ICT (ICT, 2018).

The basic objective of the project is to provide a proof-of-concept version of career mentoring to students who may otherwise have no, or severely limited, access to advice or mentoring as they face the daunting task of selecting a career. The ultimate goal is to increase the number of technically trained personnel available. They might be used in the uniformed services or as civilian researchers for the DoD. To accomplish this, prospective career candidates may make other choices if they do not have someone to whom to speak who is experienced in and feels positively toward the desired careers. Students who do not have access to a technical mentor, may not consider and therefore not select a STEM career. The reason for such lack of access may be due to a number of reasons, the two most prevalent of which the authors hold to be: geographic remoteness and low Socio-Economic Status (SES) (Crisp, 2009).

The solution advanced by ICT, and the focus of this paper, was to produce a computer-generated mentor to be available via any Windows operating system with the appropriate software installed. The student user would be

presented with a computer screen interface via which he could engage a mentor in a conversation-like exchange about issues of concern. Input from the student would be by text entry or audio speech recognition software. This mentor was designed to be both compelling and engaging. The major *raison d'être* for ICT was its coordination with the creative community in the LA Basin. ICT advanced the position that the mentors would be most compelling and engaging if they were real people (Traum et al., 2015), recorded live and demonstrated screen presence.

First, a list of pertinent questions (500-1500) would be generated. This requires an excellent knowledge of both the questions that are relevant to the mentor's profession and of the questions that the target users are likely to ask. In the MentorPAL project, the team considered input from several members who had technical and Navy experience in order to generate a thorough list of questions. A previous paper (Nye et al., 2017) discussed this process at greater length. This process is somewhat lengthy and requires a significant amount of staff and mentor commitment.

The mentors would then record a large number of (~500) video clips relating to their experience in the Navy and their background in their own STEM professions. These clips are then stored in a standardized database. When the user poses a question the program selects the most appropriate clip to play. This must be done very rapidly to sustain the desired "conversational" effect of the exchange. Based on the team's experience with briefing high school students, the clips were designed to be on the order of 90 seconds or less. This technology rests on previous research into the use of Virtual Humans acting as program interfaces in counseling (Morbini, et al., 2012) and in history capture (Artstein et al., 2014). The team found that these videotaped mentors are difficult to classify easily: sometimes being referred to as Virtual Humans, but they are actually intelligent agents that sequence and present recordings of real people. This has been termed "time-offset interaction" in some related work (Artstein et al., 2014). There is, of course, an issue as to the relative costs and efficacies of fully animated or "live" avatars. In actual implementations, the data management issues remain approximately the same. There still needs to be a substantial corpus of carefully collected questions and answers available to the system that can be answered within the latency limits that circumscribe the feel of conversational dialogue.

In the chart below, the flow begins at the top left and proceeds in a clockwise fashion. *Caveat*: It should be noted that the icons are notional only and should not deceive the reader as to their importance, weight or size, *e.g.* the storage icons approximately the same size, whereas, in fact they differ by three orders of magnitude in data size. The important steps are the data flows from the storage locations into and out of the computational functions.

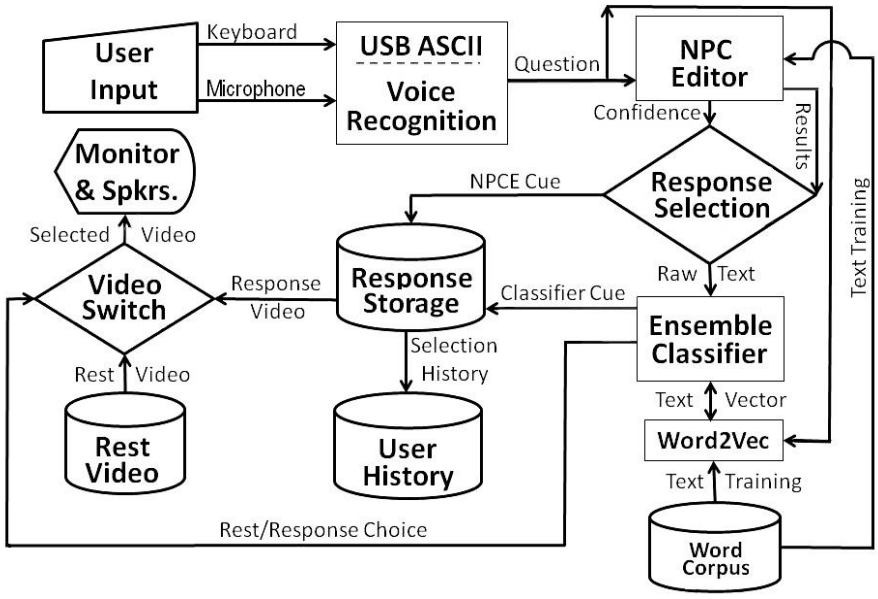

Figure 2 - MentorPal Notional Flow Chart

Model Reduction Needs

Recent advancements have been both incremental and transformative. These lead to commoditized services for previously niche computational fields: natural language processing, artificial intelligence, machine learning, neural networks, and human-computer interaction. This paper outlines the certain dilemmas that result from scaling of advances in the areas above. Specifically, we aim to address the tradeoffs between large data versus local data, which have practical implications on the performance and potential of natural language processing applications. The techniques used to improve the performance these apply not only to functions of the same field, but might be further translated to the operations of fields with similar limiting conditions.

Common limitations in the space of intense computation and data processing can be briefly summarized to two elementary components: time and size. The correlation of the two is such that applications which require the processing of greater amounts of data also require more time to complete. Expensive powerful hardware has the ability to direct computational power to resource demanding tasks, but everyday applications hosted by low-cost, but inferior hardware face a fundamental obstacle for providing efficacious services.

The team was tasked with the production of an interactive virtual mentorship platform within which users could discover career information and probe Navy STEM professionals about their experiences, advice, and knowledge within the domain of the mentor's background, career, and education. Beginning in fall 2016, the team developed a desktop based graphical user interface with the functionality of a live question and answer platform. Using previously recorded and transcribed video interviews, the system would receive textual user input, ideally a question for the mentor, and then play the best matching recorded video in response. All of the interaction and matching happened in context, and the data storage and processing occurred locally on the machine being used. In this case, that machine was usually a Windows Surface 3 with 2 GB of RAM, four cores, and 1.6 GHz processing power. The limited resources of the utilized machine reflect the constraints faced by countless practical applications of computationally expensive operations.

The classification system designed by our team implements a combined logistic regression, long short-term memory (Hochreiter & Schmidhuber, 1997), and skip gram model using Google's Word2Vec vector representation (Mikolov, Sutskever, Chen, Corrado, & Dean., 2013). With this model, it is necessary to utilize a corpus, essentially a dictionary, of words and their representative meanings in order to make comparisons for classification. The corpus used for the purposes was Google's free online news vector, composed of 3 million words and phrases from Google's dataset of close to 100 billion words (McCormick, 2016). The original vector has a size of about 3.64 GB, making it very time intensive to load when being utilized for classification, especially when the operating system stores some of the model within pagefiles, essentially secondary memory as an extension to RAM, to hold the model. Upon initial preprocessing, the corpus took approximately 5 minutes to load on the machines with which we were working.

Given the context of our mission in creating a live interactive question and answer system, time optimization was of utmost importance. Short lag times are essential for creating a seamless user experience that effectively represents the live communication of a user with his or her mentor. Engagement is easily undermined by system delays, particularly within the target audience of pre-college students who are likely accustomed to working with applications that offer a near immediate response time. The resultant extra time necessary to load the model was enough to deter students from the beginning and delineated a fundamental obstacle preventing us from finalizing an adequate product. The need to improve the time bottleneck was epitomized during our primary round of user testing, and we found it necessary to look towards scaling down the size of the data intensive Google news vector.

Previous Approaches

Our initial research uncovered that, although this was a particularly specific problem, others had faced it before, and tested various methods in amelioration. The most extensive research done by another party was by Jurgovsky *et al.* in their 2016 publication of *Evaluating Memory Efficiency and Robustness of Word Embeddings*. The team analyzed "a dimensionality-based, a parameter-based and a resolution-based method to obtain parameter-reduced embeddings" (Jurgovsky *et al.*, 2016). Through their reduction and analysis, the authors were able to determine the trade offs for each method as measured by model accuracy and memory efficiency.

The first strategy discussed, linear transformation, utilizes the quality of the data being stored within data vectors with a certain initial number of features. The unmodified Google News dataset contains word vectors of size 300 (McCormick, 2016). By employing linear transformation to map the complete set of feature vectors to vectors of fewer features, it is possible to curtail the amount of memory necessary to hold the descriptive vectors for their respective words. This results in an inherent loss of accuracy, as vectors linearly transformed to have fewer features are less detailed and specific in their representative description of words. This strategy is most effective in the case that the higher feature structure contains redundant information, *i.e*. the differences between words within the corpus do not have significantly *specific* differences in meaning. This method of reduction was found to be the least effective of the three methods tested by Jurgovsky *et al.* in improving memory efficiency as compared to end loss in accuracy. In our case, the performance loss was further disadvantageous due to the precise and limited domain of language within which our model functions. In effect, because the questions, answers, and language in our system are particularly specific to STEM and the Navy, reducing the specific accuracy of the model by reducing vectors by features resulted in especially considerable losses in classification accuracy.

The second approach evaluated by Jurgovsky *et al.* is pruning: essentially eliminating a subgroup of features based upon a threshold of the significance of that feature by setting that feature to zero. Pruning reduces memory overhead by discarding features within the matrix composed of all words that are determined to be insignificant as compared to the selected threshold. The paper concluded that "up to a certain pruning level, the inaccuracy induced by Pruning has no qualitative effect on word vector arithmetic and word similarity computations" (Jurgovsky *et al.*, 2016). This method proves effective by deliberately shifting word vectors closer towards their coordinate axes and disqualifying the memory required from storing the features with lesser influence. Although pruning was determined to be more effective than linear transformation up to a certain point, we recognized difficulties similar to that of the previous approach due to the specificity of our domain; the smaller impact features often were those that enabled us to extricate minimal differences in meaning between words.

The final technique investigated by Jurgovsky *et al.* is bit truncation. This procedure avoids attrition by discriminating features or eliminating them altogether, but instead proportionally pares them by describing them with fewer bits. After shifting the vector features to contain only positive numbers, the approach shortens the range of the features proportionally to a max value of 1, and produces an eventual model by multiplying the values by 2^B, B equating to the number of final bits to be described, and storing the concluding representations to a 32-bit integer type (Jurgovsky *et al.*, 2016). This method retains the relational accuracy of words due to its proportionate nature of reduction. The advantage of bit truncation is that meanings are not entirely lost, they are simply proportionally abbreviated into fewer bits until they become too insignificant to be accounted for. Jurgovsky *et al.* found this means of reduction to result in the slightest loss of accuracy of the three approaches analyzed; "Bit-Truncation does not cause any loss on any of the evaluated datasets up to 75 % reduction (24Bit)" (Jurgovsky *et al.*, 2016). This model was the best of the three addressed to apply to our situation, but despite this impressive outcome, the strategy was nonetheless not optimized for domains of our particularity.

The aforementioned resolutions for optimizing accuracy of vector language models for their memory performance are effective in many cases, however, we hypothesize that further compelling methods exist for language models

that do not require exhaustive language coverage. In the following section, we address our unique need for a domain based reduction of the utilized vector language model.

THE DATA REDUCTION EFFORT

New Approach

According consideration to the concentrated scope of the corpus, the intention of our reduction process was to minify the memory required for the use of the language model with the minimum resultant effect on the accuracy of the MentorPal classification system. This illustrates the specific need of applications to optimize the robustness of their language processing technology within the context of limited resources and definitive domains. Since other approaches that can be effective in general contexts suffer from the drawbacks discussed earlier, we suggest approaches that may be more appropriately applied to cases similar to ours. Two primary schemes were implemented and analyzed in our studies: reduction by word frequency and reduction by domain relevance.

Word frequency can be closely modeled using Zipf's Law, with frequency of words in a random text being inversely proportional to its index in the frequency distribution (Kingsley, 1932). Applying this principle within the context of the Google language vector, it is possible to discard words based on their frequencies relative to the Google News data with the frequency rankings of the 3 million words included.

The other refining process adopted was a filter by relevance. Because our language domain was unique, it was possible for us to compare the unfiltered language model with our corpus, comprised of all words transcribed from the recorded content. A simple yet suitable measure of word similarity can be calculated using the cosine similarity of the word vectors (Mihalcea *et al.,* 2006). Adopting this metric, we compared each word in the Google News corpus with words in our corpus, storing the maximum relevance match found for each word in the Google News corpus. We then chose thresholds for cosine similarity such that we discard all words with maximum relevance values below this threshold. By doing so, we keep only the words that are pertinent to at least one word in our corpus.

Results

The method implemented for measuring accuracy was a leave-one-out paraphrase test; one of the paraphrases from our corpus of questions would be removed from the training dataset and used only within the testing phase to determine precision. We readily recognize that this binary assessment of classification matching does not optimally describe the fit of responses to question input, as myriad answers returned are related or effective, yet not exact fits. Despite this admission, we accepted the binary strategy of testing reduced model effectiveness as a sufficient metric for comparison for the purposes of this research. The following data, (Table 1), depicts the models created using our reduction procedure with varied thresholds for cosine similarity and frequency, subsequently being tested with 382 leave-one-out paraphrases to provide final percentages of perfect match accuracy.

Table 1 – Model Performance by Cosine Similarity and Frequency Reduction

Cosine Similarity	Number of Words by Frequency	Model Size (MB)	Accuracy (%)
0.000	3,000,000	3750.0	36.65
0.400	250,000	1000.0	33.25
0.400	25,000	801.0	32.98
0.450	25,000	372.3	31.68
0.475	25,000	265.0	31.68
0.500	250,000	435.5	31.94
0.500	25,000	199.5	30.37
0.550	25,000	138.3	29.58
0.550	0	89.5	29.06
0.600	250,000	365.5	31.41
1.000	250,000	361.6	31.41

Although complete generation and analysis of model reduction has not yet been completed, improvements can be seen comparing minor modifications in cosine similarity thresholds. For example, the increase from cosine similarity threshold 0.45 to 0.475 results in a model size reduction from 372.3 MB and 265 MB respectively, while retaining the same accuracy, which the team considered adequate at this point. Even small enhancements such as this can have important impacts on the performance of domain oriented systems. The smallest model generated, only retaining words that met a cosine similarity of at least 0.55, resulted in an 89.5 MB (97.5% reduction) model with only an approximate 7 %decrease in perfect matching accuracy. The resulting progress in speed of our system made it noticeably more usable in practice.

A great deal of work has yet to be done on the analysis of varying the size of model generation, but the multi-day training time for each model makes preparation a laborious process. An additional practical evaluation would be to find the point at which accuracy becomes exponentially depressed as model size decreases: at this point the utility of filtering becomes trivial. Despite the need for further development, the results of our research were very promising and reflect the potential of reduction based on relevance to domain specific use cases.

FURTHER IMPACT AND FUTURE RESEARCH

Impacts of MentorPal

The authors assert that these observations and research portend future impacts in two dimensions: 1) the expansion of the use of the technology of MentorPal, *i.e*. the provision of conversational avatars to replace or assist human mentors or advisors; 2) the extension of the process of optimizing or minifying databases for other uses, particularly where computing assets are limited for one reason or another.

The concept of a conversational interface in a computer is familiar with most people in the industrial world. The last few decades of the 20th Century was replete with fictional representations of such interfaces, albeit without a video representation of the avatar. Two of the most familiar are the HAL computer in Arthur C. Clark's *2001, A Space Odyssey* (Kubrick & Clarke, 1968) and the device addressed simply as "Computer" in the *Star Trek* series (Rodenberry, 1966-69). Voice recognition and response generation have now caught up with the earlier science fiction vision of computers with whom one can talk. These are most familiar in terms of Alexa, Siri *et alii* (Hoy, 2018). The above cited research ostensibly goes significantly farther, achieving a more conversational exchange and providing a more "human" sensitivity, as well as generating constructive advice. The authors feel that, as the costs and the decreasing availability of qualified personnel constrain the provision of these two important assets to students, there will be an expansion of the use of computers to provide them. The above project focuses on academic and career counseling, but other projects have focused on delivering patient-centered counseling for PTSD patients (Morbini *et al.*, 2012) and on capturing the memories of holocaust survivors (Traum, *et al.*, 2015). The need for mentoring in many fields has also been established, particularly in the services (Dougherty & Dreher, 2007). Other types of counseling and therapies come readily to mind. Were these to be implemented and ported to increasingly small personal computing devices, the need for optimization of data usage seems assured, even when advances in hardware are considered. There will very likely be a concomitant increase in demand for more extensive, facile and efficient data capabilities. On-going future research is anticipated in the all of the areas addressed above.

Within this project, the future is here already. There is an ever-present need for more speed (users are uncomfortable with awkward latencies) (Nah, 2004) and an evolving dynamic of a responsiveness to the change in context depending on what has already been said. Another area of growth is the ability to recognize the body language symbology, voice tone, and facial expression of users and implement response modifications to appropriately address those perceptions. These are all inputs that are important to live mentors or counselors, though they will

require additional data resources which could be enabled by the minification work discussed above. There is also the area of realistic modification and projection of the mentor and his or her setting as a field of research that would seek out the possible impacts of the appearance of the mentor (age, SES, language, accent, gender, ethnicity, *etc*.) (Kendricks, Nedunuri & Arment, 2013) and the nature of the surrounding background (office, workplace, casual rest area, *etc*.). All of these indicate that there will be an incessant need for efficacy in the data manipulation.

Future Research

Minifying databases such as the Word2Vec Model done for the purposes of the MentorPal project relied substantially on techniques such as reduction of word frequency and filtering out irrelevant topics. However, the future of minifying these types of models should not result in the loss of generalization, especially given that the nature of a fully conversational interface would involve being able to capture and classify varying types of input. Some common edge cases could include one word, extraordinarily specific, vulgar, irrelevant, unrecognizable, or vague user queries.

A further approach by Facebook research teams was the recent open-sourcing of their own text classification, FastText, which makes use of linear classification instead of the two-layered neural network used by other traditional text classifiers such as Word2Vec (Joulin, Grave, Bojanowski, & Mikolov., 2016). Word2Vec training and execution is relatively slow compared to FastText, which achieves the same training in a couple of seconds compared to the Word2Vec neural network training, which takes several days. The ability to tailor the FastText training data to our particular language domain would unlock an entirely new level of customization that Word2Vec simply could not offer considering its lengthy training times. Aside from providing these obvious training improvements, FastText also takes great strides towards improving efficiency in execution. Notably, the FastText method executes similar approaches to those which we took in order to reduce our Word2Vec model. Similar to our first method of reducing the size of the Word2Vec model by filtering for higher frequency words associated with our particular language domain, FastText takes advantage of that fact that certain categories of words are more frequent than others. Because FastText uses a data structure similar to that of a binary search tree, it manages to use a hierarchical structure that creates a shallower depth for a category that is more popular (Joulin *et al.*, 2016).

The improvements and iterations upon the open sourced FastText have the potential to ultimately solve the issues that the MentorPal team dealt with in regard to the sizing and efficiency of contextualizing input without the loss of generalization.

In regard to the future research surrounding text classification models, multilingual classifications will become increasingly important as these applications are scaled worldwide. One of the current challenges that MentorPal faces is the language barrier that might deter future users from the opportunity to benefit from interaction with the application. Given that the target audiences are people without access to readily available mentors, the subset may involve communities within which English is not the primary language. Having a multilingual system would further ensure the inclusiveness and scalability of MentorPal and similar projects. (Avdelidis, 2002) Another opportunity for further research is analyzing the performance of reduced models with a multifaceted testing set of matches with varied levels of suitability.

CONCLUSIONS

The major issue is: “Can the data be minified to make the program effective in even the physically and computationally constrained confines of a small tablet computer or smart phone?” This must be done without unduly increasing the error rate experienced by the user. The first conclusion that will be offered is that the minification effort was surprising in its ability to reduce data sizes without dramatically increasing the error rate, as was adduced

in text and tables above. The reduction of the model database size by an order of magnitude was observed to produce only a 15% decrement in the number of correct answers to questions. An associated issue is more subjective: "Were the results of these attempts detrimental to the users' experience and did they decrease the users' inclination to continue the dialogue?". Even a modest degradation of performance may make a significant difference in the users' experience. While further A/I algorithm training and user experiences will be necessary to respond effectively to this issue, the preliminary results cited in Table 2 above seem to indicate the likelihood that this approach will be effective. (Lazar, *et al*., 2006)

A second conclusion is similarly subjective. Unanticipated and irregular interruptions in conversation flow were caused by large language model sizes. It was hypothesized that these interruptions were likely due to the system running out of useable Random-Access Memory (RAM) and falling back on paging functions requiring writing to and reading from secondary memory. (Caulfield, *et al*., 2010)

The third conclusion is that these techniques are extensible to many data model efforts. The authors have all programmed on various platforms and participated in various programs. They see the applicability of this work in many fields of simulation, virtual humans, language processing, data visualization, and information fusion. Based on these observations, the authors are confident that the above mitigation and evaluation techniques would be applicable and useful in a wide range of disciplines which are facing the same challenges (Ceornea, 2011).

ACKNOWLEDGEMENTS

The authors want to acknowledge the leadership and support from one of the co-Principal Investigators, William Swartout, PhD. His counsel for us and patience with us were invaluable. He contributed ideas and advice that were included in the analysis above. Further credit is due to Julia Campbell, Ed. D., Madhusudhan Krishnamachari, Kayla Carr, and Joseph Gunderson for their contributions to the project. Much of the work described above was conducted in response to Office of Naval Research contract: MentorPal: Growing STEM Pipelines with Personalized Dialogs with Virtual STEM Professionals, N00014-16-R-FO03 as well as NPCEditor and PAL3, under Army contract W911NF-14-D-0005. The opinions expressed herein are the authors' own and do not necessarily reflect those of the Department of the Navy, the Department of the Army or the U.S. Government.

REFERENCES

Artstein, R., Traum, D., Alexander, O., Leuski, A., Jones, A., Georgila…. & Smith, S. (2014, February). Time-offset interaction with a Holocaust survivor. In *Proceedings of the 19th international conference on Intelligent User Interfaces* 163-168. ACM.

Avdelidis, K., Dimoulas, C., Kalliris, G., Bliatsiou, C., Passias, T., Stoitsis, J., & Papanikolaou, G. (2002). Multilingual automated digital talking character. In *Proceeding of the International Broadcasting Convention*, Amsterdam.

Caulfield, A. M., Coburn, J., Mollov, T., De, A., Akel, A., He, J., … & Swanson, S. (2010, November). Understanding the impact of emerging non-volatile memories on high-performance, io-intensive computing. In *High Performance Computing, Networking, Storage and Analysis (SC), 2010 International Conference for* (pp. 1-11). IEEE.

Ceornea, V. (2011).Master's thesis, KTH Royal Institute of Technology, *Applying Next Generation Web Technologies in the Configuration of Customer Designed Products*. Retrieved 03 Feb 2018 from DiVA-Portal, on from: http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A450333&dswid=9721

Crisp, G., Nora, A., & Taggart, A. (2009). Student characteristics, pre-college, college, and environmental factors as predictors of majoring in and earning a STEM degree: An analysis of students attending a Hispanic serving institution. *American Education Research Journal*, 924-942.

Dougherty, T. W., & Dreher, G. F. (2007). Mentoring and Career Outcomes. *The Handbook of Mentoring at Work: Theory, Research and Practice*, 51-93. Thousand Oaks, California: Sage Publications,.

Forman, G. H., & Zahorjan, J. (1994). The Challenges of Mobile Computing. *Computer, 27(4)*, 38-47.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735-1780.

Hoy, M. B. (2018). Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Medical Reference Services Quarterly, 37(*1), 81-88.

ICT. (2018). *Research Project One-Sheets*, Institute for Creative Technologies, University of Southern California. Retrieved on 17 January 2018 from: http://ict.usc.edu/research/pdf-overviews/

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification in the *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*.

Jurgovsky, J., Granitzer, M., & Seifert, C. (2016). Evaluating memory efficiency and robustness of word embeddings. *In European Conference on Information Retrieval* (pp. 200-211).

Kendricks, K. D., Nedunuri, K. V., & Arment, A. R. (2013). Minority student perceptions of the impact of mentoring to enhance academic performance in STEM disciplines. *Journal of STEM Education: Innovations and Research, 14*(2), 38.

Kingsley, Z. G. (1932). *Selective studies and the principle of relative frequency in language*. Boston, Massachusetts: Harvard University Press.

Kubrick, S., & Clarke, A. C. (1968). *2001, A Space Odyssey*. United States of America: Metro-Goldwyn-Meyer. Culver City, California.

Lazar, J., Jones, A., & Shneiderman, B. (2006). Workplace User Frustration with Computers: An exploratory Investigation of the Causes and Severity. *Behaviour & Information Technology*, 25(03), 239-251.

Leuski, A., & Traum, D. (2011). NPCEditor: Creating virtual human dialogue using information retrieval techniques. *AI Magazin*e, *32*(2), 42-56.

Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 3*8*(6), 1842-1845.

McCormick, C. (2016). *Google's trained Word2Vec model in Python*. Retrieved on 05 January, 2018, from: mccormickml.com/2016/04/12/googles-pretrained-word2vec-model-in-python/.

Mihalcea, R., Corley, C., & Strapparava, C. (2006, July). Corpus-based and knowledge-based measures of text semantic similarity. *In AAAI. Vol. 6*, 75-780.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint* arXiv. 1301-3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. 3111-3119.

Morbini, F., Forbell, E., DeVault, D., Sagae, K., Traum, D. R., & Rizzo, A. A. (2012). A Mixed-initiative Conversational Dialogue System for Healthcare. In Proceedings of the *13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 137-139.

Nah, F. F. H. (2004). A Study on Tolerable Waiting Time: How Long are Web Users Willing to Wait?. *Behaviour & Information Technology*, *23*(3), 153-163.

Nye, B., Swartout, W., Campbell, J., Krishnamachari, M., Kaimakis, N. and Davis, D. (2017). MentorPal: Interactive Virtual Mentors Based on Real -Life STEM Professionals. In the Proceedings of the *Interservice/Industry Simulation, Training and Education Conference*.

Roddenberry, G. (Writer), (1966-60). *Star Trek*. : Norway Productions and DesiLu Productions (Producers). Los Angeles California: National Broadcasting Company.

Traum, D., Jones, A., Hays, K., Maio, H., Alexander, O., Artstein, R., ... & Jungblut, K. (2015). New Dimensions in Testimony: Digitally preserving a Holocaust survivor's interactive storytelling. In *International Conference on Interactive Digital Storytelling*, pp. 269-281. Cham, Switzerland: Springer.

Zhang, S., Yang, Y., Fan, W., & Winslett, M. (2014). Design and implementation of a real-time interactive analytics system for large spatio-temporal data. *Proceedings of the VLDB Endowment*, *7*(13), 1754-1759.