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

Nicholas J. Kaimakis,
Samuel Breck, & Benjamin D. Nye
Institute for Creative Technologies (ICT)
Univ. of Southern California (USC)

Dan M. Davis
HPC-Education
& USC

MODSIM WORLD 2018
ENGINEERING THE FUTURE AND ENABLING READINESS: VIRTUALLY READY FOR ANY REALITY
Norfolk, Virginia • April 24-26, 2018

Major Thesis

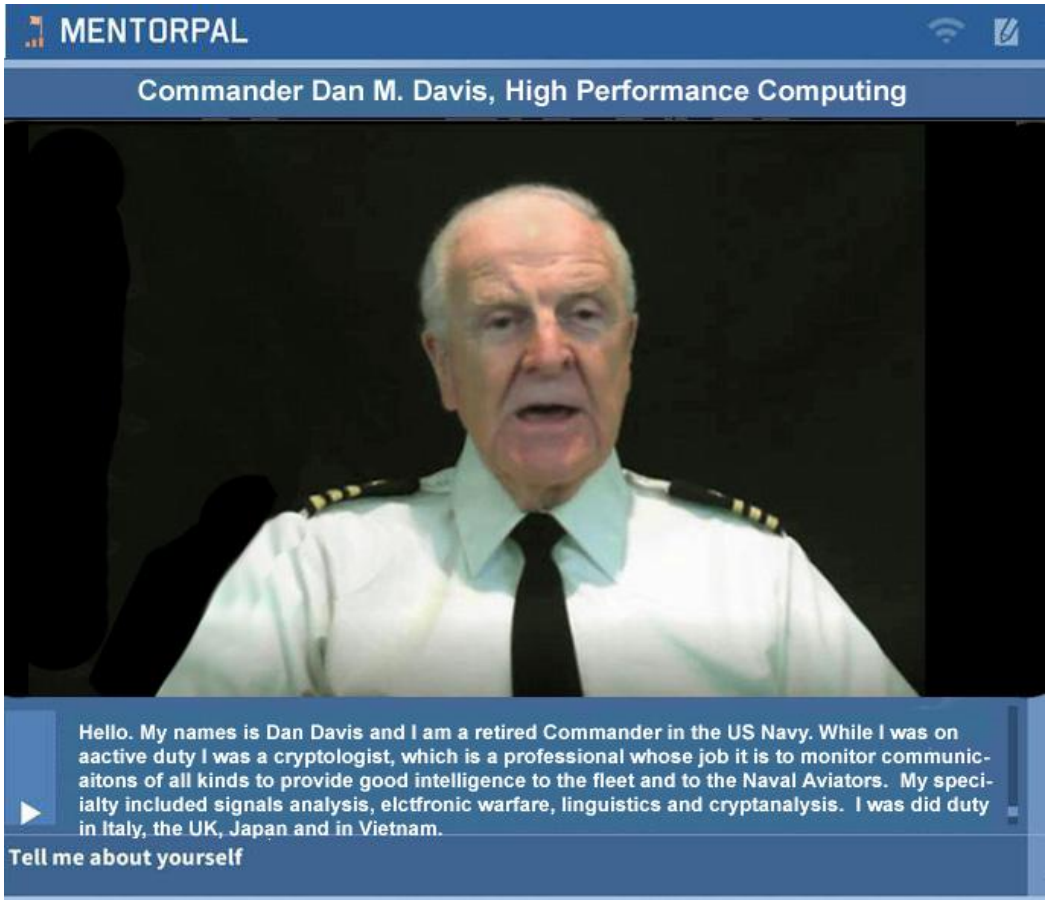
- Chatbot technology is proliferating
- Computer generated and sustained conversation is central to these programs
- Most of the capabilities depend on language data
- Efficiently creating, storing and using this data is critical
- Minimizing data merits research and optimization
- Improvement in this area would be a valuable contributor to this technology

Existing Problem

- Personnel costs and other communication frictions make computer generate conversations vital
- These capabilities depend on Artificial Intelligence (AI) and Natural Language Processing NLP
- Exacerbating the issue is the trend to small devices
- ICT/USC is working on programs in this area
- Any advances made here would likely be extensible into many similar area

Current Research

- One USC program is MentorPal
- Computer generated conversation is use to counsel students
- Responses must be rapid, germane and engaging to retain student interest



Steps in Creating MentorPal

- Hand generation of germane question list (5-1.5K)
- Appropriate personnel are recruited to respond
- About 20 hours of taping is accomplished
- Responses are then machine transcribed, hand edited and carefully analyzed for utility and attractiveness
- Many of these steps require machine analysis using language data bases
- Size, speed, and access are all important parameters

Evaluating Progress

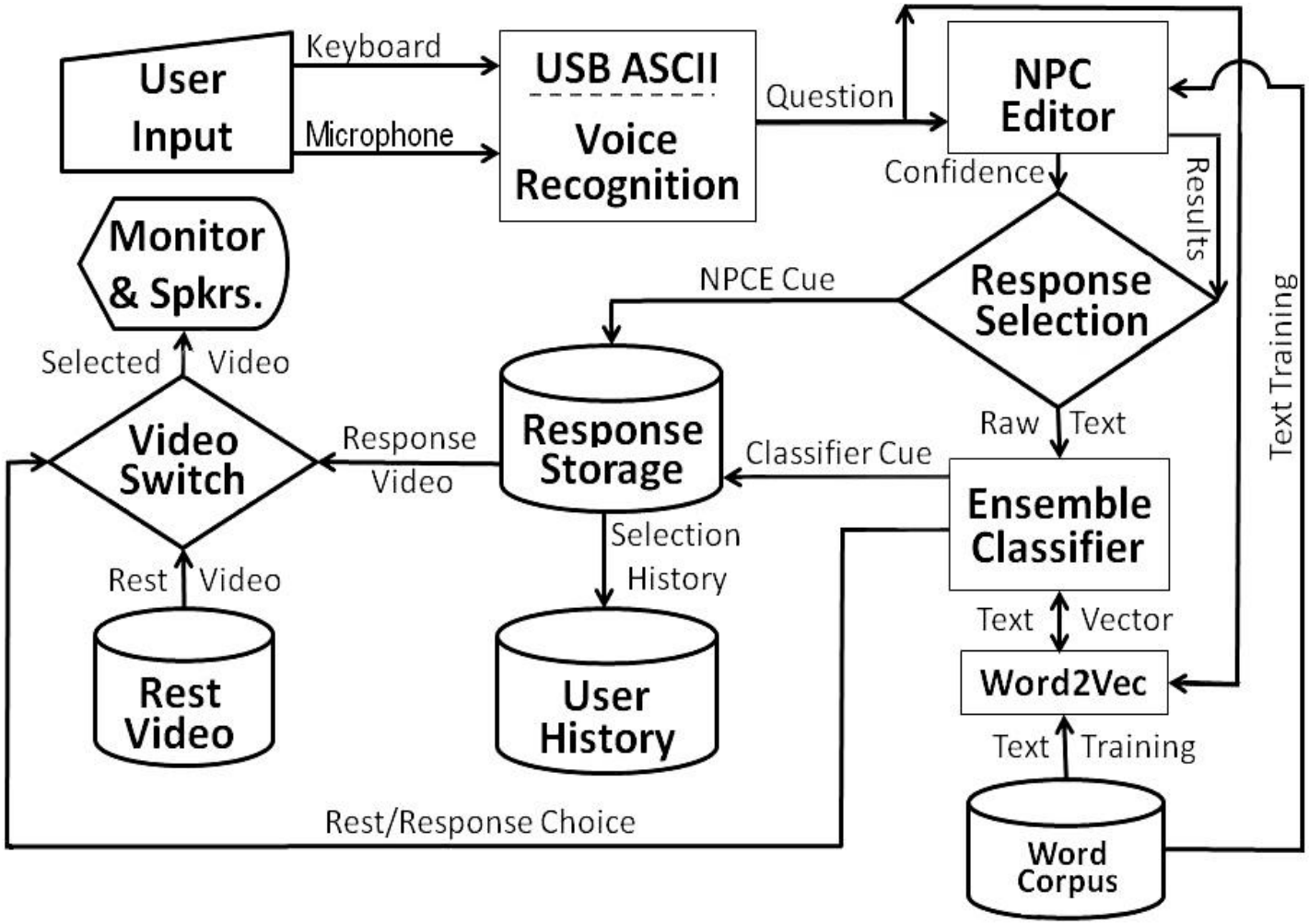
- Program is then tested on students
- Subjects are monitored to assess “issues”
- Issues of concern:
  - Responsiveness of mentor
  - Conversational quality
  - Student interest
  - New questions

- Students have trouble formulating good questions
- Major issue still is speed of data retrieval & responses

Data Flow Through System

- Student enters question or picks from list
- Either ASCII from keyboard or voice recognition from Microphone
- Submitted to NPC Editor and to Word2Vec
- Input analyzed for critical central points
- Word corpus is engaged to parse out meaning
- Response program compares input to answers
- MentorPal data is activated to play video response
- All has to be accomplished in < 500 msec

Notional Flow Chart



Limitations and Challenges

- Basic limitation categories
  - Size (especially critical for small devices)
  - Time (input, access, and retrieval)
- Limitations are synergistic, impacting each other
- Further constrained by physical size and cost issues
- Classifications systems used in MentorPal
  - combined logistic regression
  - long short-term memory
  - skip gram
  - Word2Vec

Word Corpus

- Word2Vec 3M: words out of Google 100B word data
- Vector data size: 3.4 GB
- Paging became a disruptive factor
- Loading data required five minutes on boot up
- Time delays were observed to quickly and invariably destroy student engagement
- Mandated investigation into optimizing system by reducing data size
- Would impact both time and size constraints

Approaches

- Previous work on this topic
  - Dimensionally based
  - Parameter based
  - Resolution based
- Analyzed trade offs Avoid redundant information
- Linear transformation: map vectors to fewer features
- Least effective
- Pruning: eliminating less important features; better
- Bit truncation: reducing descriptions till degradation
- Truncation was best method

New Approach

- Goal: minimum memory with minimum impact
- Two schemes implemented – reduction by:
  - Word frequency
  - Domain relevance
- Word frequency and Zipf's law
- Discard words based on frequencies in Google News
- Relevance – compare language model with corpus
- Cosign similarity measures angle between vectors
- Kept high scores, deleted rest

Results from Minifying Effort

- Measured by leave-one-out paraphrase test
- Best fit or most productive may be different
- More sophisticated testing was not supported by funding and would be constrained by time
- Data generated was nevertheless useful
- Future training is anticipated to generate better response rates without impacting data sizes
- Data on next slide represents various cosign values
- Represents 382 leave-one-outs

Performance by Cosign 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

Analysis

- 0.45 to 0.475 reduction results in a model size reduction from 372.3 MB and 265 MB respectively
- smallest model generated, retained words: a cosine similarity ≥0.55
- yielded a 89.5 MB model and ~7 percent decrease in perfect matching accuracy
- Need more research in varying models
- Hampered by training time for each new model
- Will seek point where accuracy drops exponentially

Demonstration of Program

Further Impacts

- Should directly impact success of MentorPal
- Extensible to other uses of virtual conversationalist
- Kubrick and Clarke had fully conversational computers projected for 2001's HAL and "Computer"
- Current Chatbots are proliferating
- ICT projects have found uses from Holocaust survivor archiving and PTSD treatment therapies
- All will need minified data sets.

Future Research

- Assessing utility of FaceBook's FastText
- Trains is seconds, rather than in days
- FastText uses approaches similar to the ones above
- Multi-lingual databases are looming and problematic
- A bi-lingual MentorPal will surely be critical
- Applying this approach to other programs plagued by size and time constrains may bear fruit
- This is a new field of optimization

Conclusions

- Virtual conversations are burgeoning and vital
- Minification of databases are necessary for success
- Minification has been shown to be possible
- Degradation has been tolerable or trivial
- Large data sizes are disruptive and cause paging
- Needs will only increase and demands for smaller device sizes will only become more urgent
- This work should be extensible to other areas

Acknowledgements & Caveats

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.