Real Life NLP Project for Military Assessments

Andre Almeida Reggiani, Neel Davar, Rivers Jenkins, Phillippe Phanivong, Jia Xi Tay

Pitch

ComGen is a program that translates plain English achievements of airmen into the formatted language of Enlisted Performance Report (EPR) Comments. We provide an easy to use interface for supervisors to create correctly formatted bullet point comments with strong impact verbs and automatically converted acronyms & abbreviations to speed up the report writing process. ComGen saves time and money while enabling supervisors to support their subordinates through stronger comments in the performance reports.

Introduction

The Enlisted Performance Report (EPR) is a form of evaluation required of all enlisted airmen in the United States Air Force (USAF). The EPR is filled out by an airman's direct superior usually once a year, although it may be completed more often if the airman changes assignments or duties. An EPR is designed to provide a comprehensive picture of what the airman has accomplished over the evaluation period. An airman's past EPRs are also referred to when it comes time for promotion and honors.
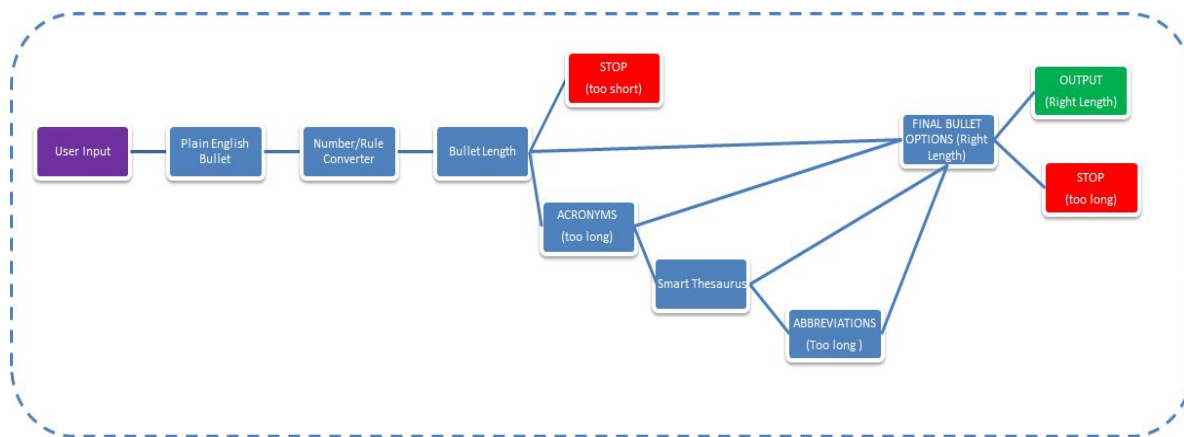
In order to paint the full picture of an airman on the EPR, the supervisor is required to leave 8-10 comments in different sections of the form. These comments must be in bullet format and satisfy strict formatting regulations. The comments must be within three spaces of the length of the line in the text box without running onto the subsequent line. They should contain as much useful information as possible while avoiding fluff and filler material. There are many accepted abbreviations and acronyms which can be used to fit the length requirements.

These stringent requirements make filling out EPRs tedious and time consuming. One airman's EPR can take 5-10 hours to complete after all revisions are done. Constructing the bullets so that they fit the length requirements currently involves large amounts of trial and error, substituting words and acronyms to try to find the perfect bullet which conveys the correct information while simultaneously satisfying all formatting and length requirements. This precipitates demand for a tool that could speed up this arduous process and alleviate the burden of airmen, so they can turn their attention to more significant matters.

Solution

Our solution to this problem is ComGen, a program that supervisors could use to convert plain English descriptions of achievements into the EPR formatted bullet points. Supervisors provide inputs in the standard 3-part "Action/Impact/Results" format and the program determines if an EPR bullet can be created from it. If it is possible the program curates a list of possible bullets that can be used.

ComGen uses a multi-step process to generate comments outlined in the architecture below.

User Input → Plain English Bullet → Number/Rule Converter → Bullet Length

STOP (too short)

ACRONYMS (too long)

Smart Thesaurus

ABBREVIATIONS (Too long)

FINAL BULLET OPTIONS (Right Length)

OUTPUT (Right Length)

STOP (too long)

After receiving the plain english bullet from the user, ComGen applies strict formatting rules to the bullet to ensure that it follows all regulations and conventions. It then tests the length of the bullet and terminates the process with an error message if the bullet is too short. Bullets that are too short are indicative to the user that insufficient content is provided for a meaningful comment bullet. Otherwise, ComGen proceeds to generate all possible acronym contractions, synonym substitutions, and abbreviations to give the largest number of correctly formatted combination of bullets possible to the user. If ComGen is unable to make any bullet option fit within the length requirements, the program terminates the process with a message letting the user know the bullet input was too long.

ComGen relies on acronym and abbreviation tables stored in CSV format, so that they may be easily modified by the user if necessary. This provides a straightforward method should there be updates to commonly accepted acronyms and abbreviations used in the USAF.

In addition, ComGen also uses natural language tools from Python's NLTK package package to find synonyms and transform words to fit in context. ComGen first recognizes the part-of-speech of each word from the user's inputs, before using that to specifically look up for possible word replacements through the Wordnet package. ComGen has the capability to rank word replacements based on the frequency with which certain synonyms are used within a select corpus of USAF manpower doctrines. The possibility of using internet-based thesaurus such as Thesaurus.com through the py_thesaurus package was also explored. But based on user needs, an offline system was required, thus precluding the use of such packages. Once synonyms have been generated, ComGen restores the original form of the word - such as tenses before outputting it as one of the possibilities.

To determine bullet length, ComGen relies on a lookup table for character length when rendered in Times New Roman 12pt font within a pdf document. This was necessary to avoid rendering differences between the pdf and Python. ComGen's number converter is able to distinguish fine differences in context, so that it may accurately convert numbers in english to numerals. For example, "moved cargo--fifty-two tons" translates correctly to "moved cargo--52 tons", or "one hundred and two cats and dogs" translates to "102 cats and dogs" correctly distinguishing "and" within a number versus outside of it.

Future Work

Although ComGen is currently operational and already able to assist in bullet construction, there are some improvements that we think could make it an invaluable asset to any airman.

The smart thesaurus currently uses natural language processing models trained on default common bodies of work. In the future, we hope to greatly improve its accuracy by tailoring it specifically to USAF terminology and writing conventions. This would be done by first incorporating a larger body of text into the corpus - specifically using personnel appraisals, award descriptions and performance reports. In addition, a word-2-vec model could be trained from these reports to generate high similarity words, rather than using the stock Wordnet database which could help to generate better word replacements. Secondly, the incorporation of feedback loops as described below would aid in improving word recommendations.

To make the user experience easier and more fluid, we hope to add a 'copy' button to the right of each bullet output to allow for the user to easily copy and paste the bullet into the EPR.

In addition to copying the output bullet to the clipboard of the user, this button would also add the user's input and selected output to a database containing past bullet inputs and outputs. This database could be used in the future to enhance word recommendations and also enable further analyses and studies. These could include providing summary statistics of the types of duties commonly performed in the USAF, the total tonnage of cargo transported, sorties flown, or similar variables divided up by squadron, group, or wing.

Mentor

SSGT Stephen Gibson is the Non-Commissioned Officer in Charge (NCOIC) of Squadron Safety in the 22nd Airlift Squadron in Travis Air Force Base. He serves as a C-5 Flight Engineer. SSGT Gibsons was helpful in charting out clear user requirements and providing us with a good start on understanding the EPR. SSGT Gibsons did not have a technical background in NLP or Python and could not provide us with assistance in that area.

Group Work Allocation

All of the team members contributed to the structure of ComGen. Each participated in brainstorming and early meetings with the mentor to determine needs for the program. Specific contributions are as follows:

- Rivers Jenkins - Led the project creating the code for the several of the major functions and the GUI, and coordinating efforts of other team members.

- Neel Davar - Led the creation of the final presentation and several of the diagrams used to explain the process.

- Phillippe Phanivong - Coded several of the functions, cleaned some of the data, and contributed to the presentation and final paper.

- Jia Xi Tay - Coded several of the functions, translated many of the inputs back into plain English, contributed extensively to the presentations and final paper.

- Andre Almeida Reggiani - Contributed to the initial idea phase of the project. Worked on several pieces of code but was not available close to the end of the project and did not participate in the final presentation or paper.

Link to GitHub Repo

https://github.com/SlicedBadge/Data-X_Project