Andrew Sutardji
September 11, 2016
CSCI 4243W

Project Summary

SmartCaster automatically generates football summaries. It falls into the realm of information technologies. This approach retrieves the finalized data of football games provided by the NFL API, constructing headlines and brief content. This project requires knowledge of natural language processing, as well as machine learning, to better derive tone and rhetoric from other sports news authors. A dictionary and thesaurus must be provided for the sake of readability. The composition, in the eyes of the reader, must be nearly indistinguishable from the work of a sports columnist. A grammar must also be developed to ensure generated sentences follow the rules of English linguistics. The program takes articles, headlines, and updates as input, reading the frequency of words used in certain contexts, and make judgments on its own word choice based on these detectable nuances.

This Small Business Innovation Research Phase I project gives insight into how humans understand language. The process of coding a program to construct short multi-sentence updates, mimicking the styles of typically published sorts updates and articles, coupled with the programmer's own interpretation of the English language, can be likened to teaching. The difference arises when the "teacher" must set aside subjective teaching practices and conform to a logical process that forces the programmer to evaluate their own understanding of English. It becomes difficult when diversified word choice for the sake of eliminating redundancy arises, but the program should produce content that is informative while refraining from repetitiveness for the sake of emphasis. Words cannot be overused, but when implementing the thesaurus, the program must also be aware the synonymy does not equate interchangeability. If "bad" or poorly written content is mistakenly fed through the program as input, the "habits" and biases, groomed through processing "good" articles, should rectify any undesired deviations. Interpreting data is also an extremely complicated task, especially in the field of sports where a multitude of variables must be accounted for when trying to evaluate any given player and predicting their future performance. *SmartCaster* will bring a greater cognizance in the understanding of language, as well as providing insight into the motivations behind how authors choose to compose their work. Aside from being a practical reporting tool, the program will essentially be a reflection of the human thought process.

Rather than having analysts construct headlines and updates after football games, *SmartCaster* produces outputs of similar quality. Post-game summaries will provide accurate data, free of the biased commentary of sports columnists, while retaining a sentiment and an option of pandering to a specific audience within a particular team's culture. Ultimately, the goal is to guide an understanding of language from the bottom-up, where words are known by rigid definitions and sentences are composed according to the established grammatical framework. The program will becomes a technical tool in tracing how sentences relate to one another as subjects are placed in greater context. As a method of exploring and cataloging the complexity of language and rhetoric, this program has the potential to map the evolution of an individual's understanding in a "lead by example" format, with human-composed articles formatting its parameters.

Andrew Sutardji
September 25, 2016
CSCI 4243W

Elevator Pitch

The Customer. Describe the expected customer for the innovation. What customer needs or market pain points are you addressing?

SmartCaster serves the average-to-avid sports fan along with fantasy football players. Relaying information from live games to users is difficult without actually viewing the game. Audio-only commentators (such as radio announcers) bog down statistics and play details with semi-relevant references and input and hard statistical summaries are not often used, or even viewed, in the same manner. *SmartCaster* constructs brief summaries free of extraneous content, lined with some tone set to pander to the user's preferences.

The Value. What are the benefits to the customer of your proposed innovation? What is the key differentiator of your company or technology?

SmartCaster allows users to receive news on the players and teams they care about. In the same sentiment, the content presented is built around their preferences. Sports columnists inject their biases into their rhetoric, and commentators over-inform viewers with constant narration and sidetracking. Through natural language generation, tone can accompany small content. The wordiness of commentary can be eliminated to emphasize the importance of stats and core information, and readability can complement plain statistical fact. In this comfortable medium, the update provides a single clear thought that is comprehensible, concise, and informative.

The Innovation. Succinctly describe your innovation. This section can contain proprietary information that could not be discussed in the Project Summary. What aspects are original, unusual, novel, disruptive, or transformative compared to the current state of the art?

Automated storytelling is the future of reporting, especially when it comes to small, informative content. The time and effort put into constructing mere headlines and summaries is less imperative when expressing fact trumps creativity. Diversity in sentence structure and simple word usage is easily trained when the length of the content is constrained. The moment a football game concludes, straight data can be delivered in the form of a brief multi-sentence update.

Rather than piling on utilities into an all-in-one tool, or confusing users with extensive features, *SmartCaster* stands alone as a reporting tool. The delivered content consists of a single thought easily understood by the user. It is non-intrusive in its accessibility, and hopes to give as little distraction as possible. Rather than overstimulating the user with massive amounts of content, *SmartCaster* provides a fast and easy way of staying informed without distracting from the theatrics of the game and its media.

Andrew Sutardji
October 31, 2016
CSCI 4243W

Commercial Opportunity

SmartCaster possesses two characteristics rare to the sports industry: simplicity and clarity. Sports entertainment draws in fans by stirring fanaticism, drawing support from theatrics and injecting emotion into events to personalize teams. Sports reporting is primarily a self-promotional tool, surrounding primary content with contextual references and prodding the viewer into submerging themselves into the cultural mindset. Fantasy sports act in tandem with traditional sports entertainment, dependent on it for the data but more interactive and thought-oriented. As part of the entertainment industry, sporting events are meant to stimulate users and participants, provoking attachments and connection to distant faces and icons for a nominal fee.

SmartCaster, like other sports reporting outlets, looks to feed the user with the information they feel necessary to keep up with their teams. Participating in fandom is not a necessity, but a phenomenon that embraces inclusiveness for the sake of consumerism. Sports fans do not support their teams out of convenience, but rather out of passions inspired by socially imbued intrigue. Going to games, researching team facts, and buying official merchandize is not seen as a burden as much as it is a hobby. Being active in this section of culture turns the means of participation into another game, at the very least another form of entertainment.

SmartCaster occupies a niche often overlooked by the sports industry. This area is already saturated with outlets delivering information with the intent of captivating its audience. Rather than competing with these other mediums *SmartCaster* looks to work alongside them. As opposed to luring the user in and arousing them to consume more content, *SmartCaster* delivers brief summaries that inform the user without causing distraction. Its composition is centered around a single coherent thought. What is given is clear and concise, leaving little room for questioning or interpretation. It is almost passive in its purpose, non-conflicting with the rest of the user's informational intake, yet somehow necessary in gaining useful information. Preferences can be set to include only the teams and players the user wishes to be informed about, serving all from the lukewarm football fan to the avid fantasy team player.

SmartCaster becomes useful in the realm of fantasy football. As fantasy team owners look to find data on their players outside of whatever application they use for actual gameplay, *SmartCaster* provides them with a quick report easily internalized. Fantasy sports, more specifically fantasy football, are a growing topic in the United States. In 2014, there were 41.5 million participants in fantasy sports in the United States and Canada. In 2015, there were 56.8 million. Of those 56.8 million, 40 million players were involved in fantasy football. In 2016, there is projected to be 75 million fantasy football team owners. On average, each of those participants is expected to spend $107 within their league, primarily in the form of bets. DraftKings and FanDuel, two fantasy sports websites, took in $3 billion in entry fees and lost $400,000 in payouts to winning participants. Aside from being a highly lucrative gambling industry, the entertainment value of playing on a week-to-week basis over an already interactive platform creates a personal emotional investment that nearly trumps the financial side.

Alongside football games, ESPN and NFL feature fantasy football programs. Reporting includes statistics relevant to fantasy football calculations, and ties between player performance and fantasy team owner advice become pertinent. NFL RedZone, which features all Sunday football games simultaneously, includes continually updated fantasy player statistics as games progress.

Football media and fantasy football have become intertwined in American culture to the point where the relationship is, to some degree, mutually beneficial: football fans are drawn to fantasy football through peer pressure and stay for the entertainment value. Fantasy team owners dive into a sea of information and research to better their chances of wining week-to-week. Both ESPN and the NFL offer means of playing fantasy football online, as well as their own means of reporting, but these summaries lie among a sea of features that clutter their websites.

SmartCaster is another extraneous tool in the world of football, and it risks being overlooked. It is designed to be as small as possible, only pushing out stats the user cares for. Hesitancy about using a tool that performs such a minute function is expected, and once implemented it acts fairly passively. It can be framed as pure novelty in its capacity. But, in the same thread, accessibility and convenience becomes a key feature. Large bodies of text can be overwhelming to a reader doing thorough research. Many people merely skim massive analytical articles, if not skip them altogether. *SmartCaster* provides non-intimidating small content that bears easy on the reader's attention span.

Andrew Sutardji
October 31, 2016
CSCI 4243W

Broader Impact

Automated storytelling is the future of small content. Entity extraction and sentiment readers processing bodies of texts lends understanding to how literary composition should be constructed. In growing the product and development in the field of natural language processing, *SmartCaster* looks to deepen thought around the logic of language structure.

All languages have grammatical rules and syntax that guide the recipient's thought process as they parse through sentences. With the English language, there seem to be many exceptions, or creative liberties, that have the potential to convey variant feelings from the formal structure. From changing intent comes changing structure, and from changing structure comes changing tone. Although it is difficult to pinpoint the subjective emotions that these variances illicit, it is clear that the language itself is the guide for thought. In embodying the intent, *SmartCaster* can help pioneer the way into how human minds actually process language, from conceiving ideas to producing concrete structures communicable and naviagatable by other humans.

The implications of automated storytelling, however, threatens to render small content authors obsolete. As long as grammatical rules are obeyed, and sentence diversity prevails through multiple iterations, the efforts put forth by this venture can bring forth a new wave of reporting where all brief summaries can be reported instantly through automated means as soon as input is available. This industrial dimension can be applied to mass reporting and information dissemination in a clear, concise form. The informative medium that requires little attention is high in demand in the age of micro-blogging and single-utility applications.

The potential for misuse and mass misinformation is high. The real issue arises from training sentiments against specific content to generate biases. Along the same thread as Microsoft's Twitter AI, "Tay", and even IBM's "Watson", machine learning programs utilizing natural language generation can mirror the rhetoric and agenda of the propagator and its audience. Mindless reaffirmation could embed these biases into the AI and pander to communities under the guise of entertainment. With the issue of "fake news" reporting on the rise, and the revelation of how the mass audience is susceptible to these lies, automated storytelling with the purpose of pushing forth special agendas is dangerous. As with all culture and its media, communities need to be held accountable for the transgressions of its own creation. The overseer of AI can do as they please to reflect and validate their own beliefs, but it falls upon the community to be knowledgeable enough to discern passive propagation from human-composed content. Education and staying informative becomes the responsibility of the individual if they truly refuse to be duped by these easily published lies.

Andrew Sutardji
December 9, 2016
CSCI 4243W

Technical Discussion

Describe the key technical challenges and risks in bringing the innovation to market. Which of these will be your focus in the proposed Phase I project?

The primary technical challenge in bringing *SmartCaster* to market is the constant development required to keep it competitive against other automated storytelling machines. Many undergoing the beta stages of development are refined over the course of use and testing. The field is growing out of both intrigue and convenience. Much of the development will include styling, such as rhetoric and diversified syntax depending on the audience, and adaptation to meet the standard composition of various professional fields. Words, buzzwords, and jargon need to be recognized and understood by automated storytelling and publishing programs to make them more appealing to a wider range of audiences.

A large part of this challenge lies in sentence diversity, especially when recognizing the continued diversity of grammatical structure in writing. Apart from developing the natural language generation aspect to include a greater audience, English in general has many exceptions, especially when standalone clauses are permissible. Writing composition can always operate to a finer tune, and the linguistically aesthetic facet can always be further refined.

Describe the innovation in sufficient technical depth for a knowledgeable reviewer to understand why it is innovative and how it can provide benefits in the target applications. Supplement this description with any necessary background information.

SmartCaster is an automated storytelling program that, in its current form, generates brief summaries and analyses of American football games. It has the potential to include updates, given a proper API that can grab live data from NFL Game Center. It may also be expanded beyond football into the realm of general news or other content construction.

First, data is retrieved from a source, such as NFL Game Center, through some medium, such as the NFL API. The statistical data currently comes in the form of JSON files, allowing for easy access to specific information through header values.

This data is then analyzed and set against data from previous games to provide some contextual basis for the composition. Trends are identified, and drastic changes in the numbers on a week-to-week basis are picked out as subjects for the constructed content. In an effort to provide brief, concise summaries, only “key” information is highlighted to be included.

The actual textual composition will be implemented through parse trees. Parse trees may be the best data structure to ensure grammatical syntax is followed. While adhering to this syntax, the program may also determine where data may be inserted within the text, looking for trends in sample texts and understanding through example writing methodology.

The greatest issue with the current method of live reporting is the author’s liberty to draw their own conclusions, often formed from their own understanding and biases. Sometimes, this can’t be helped, as information needs to be pushed into public view. It must be written in a captivating manner to draw the attention of more readers. Slight misinformation and fanaticism begin rolling as the implications of the author's rhetoric easily slowballs in front of an audience thoroughly unfamiliar with the subject. Programs like *SmartCaster* have the ability to put forth headlines free of those biases, especially when the only input is the hard data.

As for more lengthy texts, professionals writing guides also can't help but use their own subcultural jargon when conveying their knowledge. Sometimes, words can't be substituted in the mind of someone who uses it often enough in their line of work, especially when it is understood by their peers and reinforced by repetition. *SmartCaster*, and other automated storytelling programs, may translate these subcultural nuances and bring academic words to a wider audience with more commonly understood vocabulary.

Describe the key objectives to be accomplished during the Phase I research, including the questions that must be answered to determine the technical and commercial feasibility of the proposed concept.

The key objectives to be accomplished during Phase I research are primarily linguistic. Defining and encoding English syntax, with development to include all possible sentence structures in the proper context, will be a never-ending venture. Word choice can always be refined, and word diversity does not merely apply in the context of a single text, but its repetition across multiple texts. That being said, the meaning of words wear thin with common and "loose" usage with respect to its definition, sometimes redefining it in its trend. *SmartCaster* must recognize these trends through constant textual input and evolution of its own dictionary, far beyond template writing. The perceived gracefulness throughout composition lends to the interest behind it.

With ongoing refinement of *SmartCaster* in the context of American football, the program must be sophisticated enough to write on a variety of subjects with only sample texts. Expansion of automated storytelling platforms is pertinent to their development and survival, especially into the news and entertainment industry. Its value then becomes dependent on its adaptability.

The marketability of automated storytelling does not hail solely from how interested the reader is in the content it produces, but how much of it they can bear reading before mentally withdrawing their attention. When it is known that a passionless machine produces text meant to illicit personal responses, it loses some degree of its personal touch. However, if the content is composed in such a manner that conveys information while maintaining aesthetic rhetoric, the reader's attention won't stray until their curiosity in the subject has been satisfied. User-friendly language and the actual informational content are the focus of the reading, and non-excluding rhetoric brings the user further in to program usage.

As said before, the widespread commercialization of such a product comes from its adaptability to fit the needs of various audiences. This ties in closely with its technical feasibility. News reports, sports columnists, technology analysts, among many others, have differing methods of conveying their thoughts. Their means of understanding vary in their fields, as well as in personal views, directing themselves into a discipline set in its own guidelines. These guidelines define the acceptable means of delivery, from language to professional jargon to personal sentiments. Reiterating, the defining aspect of how well an automated storytelling device can fend among others is not merely by its performance, but by how many audiences it can sufficiently pander to.

Once the program can achieve composition, independent of subculture-specific templates, it can easily move into other realms. The initial product is technically feasible, and can only be quantified against competing platforms and the quality of current reporting products.

Describe the critical technical milestones that must be met to get the product or service to market.

The technical milestones are all algorithmic, with respect to linguistics. Logistically, they consist mostly of discovering the most optimal ways to handle inputted information and the data required to compose and publish said information. When it comes to personalization of the product, the parameters must either be stored on the device or created through a corpus of example texts. The former would require an already defined knowledge and the ability to check against the used words, potentially requiring a large mass of storage space. The latter would be time-consuming, and although storage-friendly would cost ad-hoc allocated processing time and resources.

Implementing a parse tree is the first critical milestone. Templating is required for sentence structure, or at least an understanding of subject and word relationship within sentences. The program must be able to recognize and use a wide variety of sentence structures when appropriate. The parse tree must also come into consideration in paragraph structure, as well as article structure, to dictate flow and reduce repetition within larger bodies of text.

A thorough review of the text is another critical milestone to recognize. With regards to the previous, diversity also must be implemented in word usage. Synonymy and antonymic words are elements the program must become accustomed to, learning proper usages for each in whatever context they may appear.

Content determination will be an over-arching process throughout the development of *SmartCaster*. Not only must the program choose which points to highlight within the text, it must determine appropriate word, sentence, paragraph, and article structure with the content it possesses.

Research and Development Plan

1. Data Retrieval

The NFL API, written in Python, gathers data weekly from NFL Game Center and makes player, team, and game statistics available as JSON data. From these objects, specific data can be retrieved via attributes, contributing to a temporary corpus of only necessary information. If specified information is not available within certain contexts, the program should recognize the problem and proceed if it is seen as an anomaly.

Estimated time of completion: no more than 1 week.

2. Data Analysis

Once relevant data is aggregated, it must be evaluated. This comes in the form of averages, charting, trend recognition, and highlighting drastic points of change. The Python Data Analysis Library (pandas) has the means of parsing through this data and doing the tedious, menial labor of looking through all this simple information.

Estimated time of completion: ~4 weeks. May be done in tandem with content determination. Will require continual updating and correction of algorithm.

3. Dictionary & Thesaurus

In striving for the most proper way to compose text, *SmartCaster* must abide by English syntax. This means following the set of grammatical rules and tenses already rigidly defined in the world of authorship. The Natural Language Toolkit (NLTK) has its own method of conveniently tokenizing texts and identifying the parts of speech of words in their context. WordNet functions as a thesaurus, lending a hand in diversifying word usage in texts.

Estimated time of completion: ~4 weeks. Will be an ongoing process as the dictionary may be expanded to appeal to more markets.

4. Content Determination

Providing the user with concise and non-imposing content means determining what information is the most pertinent to the user. The need to weigh the importance of averages and highlights comes into question here, and will be determined by the frequency of some cases, including positive and negative trends and relevance of an event within the context of a game. This aspect will be built with pandas, but likely will have to find many alternative routes.

Estimated time of completion: ~6 weeks. May be done in tandem with data analysis. Will require continual updating and correction of algorithm.

5. Textual Composition

Parse trees will be the most effective structure to ensure the composition's adherence to English syntax. Markov chains would also come into play when dictating sentence diversity throughout the text. The thesaurus will factor in when imbuing the content with a desired sentiment, and the dictionary will weigh in against any minor changes to ensure grammatical correctness.

Estimated time of completion: ~8 weeks. Will be done alongside content determination. Will require continual correction of algorithm to ensure sentence diversity, as well as the inclusion of further audiences and wider ranges of rhetoric.

6. Textual Review

Entity extraction programs, available through NLTK and AlchemyLanguage, provide "scores" as to how relevant any given subject within the text is against its context. AlchemyLanguage also provides sentiment detection, useful in reviewing the nearly finished product before the reader receives it. These quantified "scores" can be evaluated as parameters, requiring a threshold for specified qualities before publication.

Estimated time of completion: ~2 weeks. Will be done alongside textual composition.

Major Changes Made to Document

1. Put more effort in keeping the text on subject to meet the defined document structure.
2. Less attention on over-expounding on more trivial aspects of the program.
3. More consistent word usage, and less vague pronouns when references different aspects of the project.