**Presenters:**

Theodore LaGrow

Jacob Bieker

**Abstract Title:**

Do You Know Where Your Research Is Being Used?

An exploration of the National Science Foundation’s databases using Natural Language Processing.

**Abstract Body:**

We have developed software to decipher the uses of previous technology and topics relating to research used in other scholars’ research. In such a complex and dynamic field as computer science, it is of interest to understand what resources are available, how much the resources are used, and for what the resources are used. We demonstrated the feasibility of automatically identifying resource names on a large-scale from scientific literature in arXiv’s database and showed that the generated data can be used for exploration of software and topics. While scholarly literature surveys can provide some insights, large-scale computer-based approaches to identify mentions of technology and methods from primary literature is needed to automate systematic cataloguing and facilitate the monitoring of usage in a more effective method. We developed a pdf parser to extract text from articles in the database that we then trained using Natural Language Processing to evaluate if the article relates to the technology and methods in question. . The articles were then passed using n-grams of length 15 using a dictionary of words as a center point that usually directly correspond to technology and methods. The n-grams were then passed using parts-of-speech based on tokenization of nouns, titles, verbs, adjectives, conjunctions, and interjections. We took all of the noun-phrases and added them to a dictionary and counted all of the iterations to see what would be most common. With this frequency dictionary, we were then able to evaluate a trend of technology and methods used in each specific niché of science. As we continue to expand this software, we will analyze the researchers’ sentiment about the technology and methods.

Introduction:

NLP Overview:

Natural language processing (NLP for short) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages.

Modern NLP algorithms are based on machine learning, especially statistical machine learning. The paradigm of machine learning is different from that of most prior attempts at language processing. Prior implementations of language-processing tasks typically involved the direct hand coding of large sets of rules. The machine-learning paradigm calls instead for using general learning algorithms — often, although not always, grounded in statistical inference — to automatically learn such rules through the analysis of large corpora of typical real-world examples. A corpus (plural, "corpora") is a set of documents (or sometimes, individual sentences) that have been hand-annotated with the correct values to be learned.

In 1950, Alan Turing published an article titled "Computing Machinery and Intelligence" which proposed what is now called the Turing test as a criterion of intelligence.

The Georgetown experiment in 1954 involved fully automatic translation of more than sixty Russian sentences into English. The authors claimed that within three or five years, machine translation would be a solved problem.[2] However, real progress was much slower, and after the ALPAC report in 1966, which found that ten-year-long research had failed to fulfill the expectations, funding for machine translation was dramatically reduced. Little further research in machine translation was conducted until the late 1980s, when the first statistical machine translation systems were developed.

Up to the 1980s, most NLP systems were based on complex sets of hand-written rules. Starting in the late 1980s, however, there was a revolution in NLP with the introduction of machine learning algorithms for language processing. This was due to both the steady increase in computational power (see Moore's Law) and the gradual lessening of the dominance of Chomskyan theories of linguistics (e.g. transformational grammar), whose theoretical underpinnings discouraged the sort of corpus linguistics that underlies the machine-learning approach to language processing.

Methodology:

Results:

Conclusion:

Sources: (MLA)

Hucka, Michael, and Matthew J. Graham. "Software Search Is Not a Science, Even among Scientists." ArXiv (2016). ArXiv. Web. 13 May 2016.