Issues in Information Systems
Volume 25, Issue 1, pp. 24--32, 2024

DOI: https://doi.org/10.48009/1_iis_2024_103

Text summarization using the relationship structure of a document

George Ray, *Shepherd University, gray@shepherd.edu*

Abstract

A new approach for textual analytics is introduced that creates a relationship structure for a document by using the relationship words in the system of basic English. Relationship words interlock concepts in sentences. The author of a document proposes certain relationships between concepts and uses relationship words in sentences to present a pattern of those relationships. In this paper, the concept of a document's relationship structure is used for automating summaries for documents. In addition to developing an automated approach to analyze, understand and present knowledge in documents, this article also considers how to validate the correctness of an artificial intelligence algorithm.

Keywords: artificial intelligence, text summarizer, Rouge scoring, validating AI, thinking machines

Introduction

Artificial Intelligence (AI) is the replication of human analytical ability. One of the major domains of AI is Natural Language Processing (NLP), which enables intelligent systems to both understand and generate human language (Cuantum, 2023). NLP has applied rule-based methods, statistical methods, machine learning and deep learning to perform its tasks of part-of-speech (POS) tagging, named-entity-recognition (NER), sentiment analysis, text summarization, machine translation and question answering. Text summarization analyzes a document, selects the key information in the document and generates a clear summary. This article is a concept paper for a new method of text summarization. The seminal work using modern computers for text summarization was by H.P. Luhn, a researcher at IBM. In a concept paper in the IBM Journal, Luhn (1958, p 159) describes and justifies an algorithm "to facilitate quick and accurate identification of the topic of published papers." He then explained the results obtained after applying a computer program based on that algorithm to a document.

The following sections describe a novel algorithm that is implemented in R and then applied to documents. First, a background is provided to justify the approach taken through a review of the long train of scholars exploring the nature of language and its use in defining the boundaries of human knowledge. The review begins in the 13th century with the invention of a computing device to make logical deductions. The concept of validating the outputs of an artificial intelligence process is then explored. Ramon Llull was a 13th century theologian and philosopher who developed a machine to make deductions that he and his followers, over time, believed would produce knowledge (Stagnaro, 2016). His work showed that aspects of human thought can be replicated on a machine. Llull's contribution was in relating different concepts by use of a "thinking machine." One of his followers was Gottfied Leibniz.

Leibniz worked through his life to employ Llull's methods "to formally define the extent of possible knowledge" (Schmidt-Biggemann, 2018, p. 56). He followed up Llull's work with his alphabetum

cogitationum humanarum, an alphabet for human thought, and a syntax for combinations; he sought a basic set of linguistic building blocks that would be the foundation for knowledge.

Our ability to understand a situation, perform research, or solve problems is bounded by the words we have at our disposal. The statistician Karl Pearson in his work the *Grammar of Science* argued that the "obscurities of modern science" are the result of an anemic orthology; orthology meaning the accurate use of language (Pearson, 1900, p. 515). C.K. Ogden, a prominent linguist in the 20th century, took up this orthology challenge and developed a Panopticon algorithm to generate orthologically correct sentences (Ogden, 1934, pp 305-308).

Ogden's Panopticon is a derivative of Llull's Wheels of Combination (Crovitz, 1970), but while Llull used his device to make syllogisms, Ogden looked to make unambiguous sentences. Ogden developed a short form of English called Basic English and organized it into sets of words that represented comparison, direction, action, qualities, and joining. He gave examples of how to use this approach for religion, science, politics, literature and commerce (Ogden, 1934). Using his approach, one could combine concepts into meaningful sentences. Crovitz (1970) called Ogden's Basic a heuropticon in that it generates concrete examples of an abstract concept to aid discovery, a heuristic approach.

The idea of recombining concepts as the basis for innovation and cultural advance was proposed by anthropologist Homer Barnett (Barnett, 1953). With recombination, the qualities of two objects that are similar but compare differently on some desirable performance trait are recombined. The weaker performer adapts the desired quality from the better performer. Wallace (1963, pp122-123) gave an example of how this type of combination was used in American submarine design during the cold war: submarines with a ship shape move slowly underwater while sharks with a shark shape move quickly; therefore, submarines with a shark shape should move more quickly underwater. Freeman and Newell (1971) considered this approach from an artificial intelligence perspective, and extended recombination to include not only new features to be added and existing features to be removed but also what existing features should be increased or decreased.

Crovitz (1970) developed his own Panopticon algorithm that he called the Relational algorithm. He created a set of words from Ogden's sets that he termed relationship words. The idea is that a statement can contain a relationship between one concept and another if they are connected by one of the relationship words. His algorithm, through recombination of relationship words, was a means of automating the process of invention, discovery or generation of basic ideas.

Cultivating attention is another approach for generating interesting ideas that also applies the process of combination (Davis, 1971). Davis holds there are 12 categories of assumption behind ideas. Challenging one or more of these assumptions can generate an interesting idea that garners attention. As an example, functionality is one of the assumptions listed by Davis. If the assumption is dysfunction, that something is dysfunctional, look instead to the possibility that it is functional. An example of this might be Volvo. The generally held view in the 1950s and 1960s was that automobiles were dysfunctional at safety. Nader's (1965) popular work *Unsafe at Any Speed* reflected this view. Volvo did not accept this assumption of dysfunctionality and pursued a set of product development strategies around the combination of safety and an automobile.

Kim and Marbourne (2005) likewise apply the process of recombination to generate new ideas. Their focus is generating ideas about product features to improve market share in commerce. They give Cirque du Soleil as an example, where the concept of a circus is combined with the concept of theater. Hargadon (2003) also follows a similar application of recombination but applies it to improving production processes. Henry

Ford changed the automobile industry when he combined the concept of a meat packing assembly line with the production of cars.

Creative Systems

A document is a creative work that interconnects concepts in order to develop and present the ideas posed by the author. A line of thought for the past 700 years has explored different systems that connect concepts to create new ideas or reasoning that leads to discovery. We can look for such structures in a document in an automated manner as a method of natural language processing. Applying relationship word analysis is a rich field for study with many avenues to pursue. This paper will look at applying the relationship word set to document summarization. The premise is that concepts can be connected with types of words to generate new ideas or arguments. The concepts are interconnected with other concepts through these words. An author preparing a paper connects and compares concepts and their qualities using relationship words that interconnect the varying concepts used to make the point of the document. These relationship words can be used by a text summarizer to generate a meaningful summary of a document.

Extractive text summarizers work by selecting what they consider to be the most representative sentences in the document as the summary (Madhuri and Kumar, 2019). The approach to text summarization taken here is to rank the sentences based on their relationship interlocking - their employment of the basic relationship words to associate concepts to develop the author's argument. The goal is for such a summarization to provide key concepts of the document in a few sentences. Another type of summarization is abstractive summarization that composes a summary rather than selecting representative sentences (Kouris, Alexandridis and Stafylopatis, 2021).

A principal advantage of abstractive summarization is the cohesion of the paraphrased summary rather than the disjointed nature of extractive summaries. Abstractive summarization also avoids problems with redundancy. Neural techniques can be used for both extractive and abstractive approaches (Kryściński,, Keskar, McCann, Xiong & Socher, 2019). Abstractive techniques may also use large language models (LLM) to understand and paraphrase the text (Basyal and Sanghvi, 2023).

Generally, extractive summarizers perform better than abstractive summarizers (Veningston, Venkateswara and Ronalda, 2023; Kouris, et al, 2021; Allahyari, Pouriyeh, Assefi, Safaei, Trippe, Gutierrez & Kochut, 2017). Because of this, many existing abstractive systems often start with a preprocessing phase that is an extractive summarization (Allahyari, et al, 2017). Extractive summarizers need fewer resources and provide greater accuracy. Their accuracy makes them more suitable for legal documents or technical papers where precision and exactness are more important. Abstractive summarizers, including those that use LLM, are more suitable for presenting readable summaries to the general public.

AI Validation

As the vision for a software project is set, it's then time to begin verification and validation (V&V) work. Thayer and Sommerville (2002) recommend that V&V be conducted in all phases of software development. In a similar manner, Musa (1999) recommends defining the necessary reliability for a system in the earliest phases of the life cycle. The nature of V&V in AI must be discussed. Validating AI systems may require different approaches than those taken with traditional software systems. Software systems started in areas with well-defined formal notations such as the mathematics of engineering calculations or the mathematics of accounting systems. Software systems are now applied to newer areas where the problems are not as well defined, the domains are highly dynamic, and there may be no clear correct or incorrect results (Partridge, 1998; Smith, Black, Davenport, Olszewska, Rossler and Wright, 2022).

Many AI systems will change behavior when new information is ingested, including during tests. Furthermore, in many cases, AI systems are trained on data rather than based on a formal specification that defines the required operation of the system. In such systems, there may be no clear specification. In addition, as in the case of document summary systems, there may be no "correct" answer to determine if the system succeeded or failed in a trial.

AI systems employ heuristics, which are strategies that may work, but without guarantees (Partridge, 1998; Feehan, Owen, McKinnon and DeAngelis, 2021). Traditional verification and validation (V&V) assume the outcome is known in advance. This is difficult with dynamic systems such as artificial intelligence that rapidly change (Partridge, 1998; Smith, et al, 2022). Moreover, for artificial intelligence, the modular design representations of the system often include black boxes, and it is not intuitively obvious what processing is done in the black box and how it is related to the domain problem being addressed (Partridge, 1998; Feehan, Owen, McKinnon and DeAngelis, 2021).

The verification stage is to prove the algorithm is the correct procedure for automating the original specification. This is difficult in AI based on neural networks because mapping the varying weightings in different layers to a business problem is not intuitive (Partridge, 1998). Validation is difficult in AI – who is to say that one summary is correct, as even different people will come up with different summaries. With AI systems there may not be a correct output for a given input so that in these cases there isn't the basis for traditional software testing.

For text summarization algorithms, an automated tool for evaluating them is available. Rouge is a tool that calculates the recall, precision and fmeasure of a generated summary compared to a reference summary. Recall is the proportion of words in the reference summary that are also in the candidate summary. Precision is the proportion of words in the candidate summary that are also in the reference summary, or a measure of how much of the candidate summary is relevant to the reference summary. Fmeasure is a composite of the recall and precision factors (Ganesan, 2017).

Lin (2004) notes that manually evaluating automatically generated summaries is not feasible because such an approach is expensive and time consuming. Lin conducted a series of tests on Rouge to determine how well it correlated with manually generated reference summaries. The Rouge measures have a high correlation with human judgments (Lin, 2004). A Rouge package is available in R that provides a comparison of a reference summary to a summary from a proposed text summarizer. There are no studies on the nature of the Rouge recall distribution so it cannot be assumed to be normal.

Methodology

The methodology of this research is to generate a summary based on the relationship word structure of a document and then compare its performance with a widely distributed and accepted text summarizer. The summaries from the standard text summarizer already published in R and a summarizer based on the relationship word structure of the document will both be compared to reference summaries using Rouge. The set of Rouge performance metrics for the two will then be analyzed to determine if there is a significant difference between them in terms of their fmeasure.

If they have similar Rouge scores, in that there is no statistical difference in the Rouge scores even though the summaries are different, then the proposed relationship word summarizer is considered to be reasonable. lexRankr::lexRank is a standard text summarizer that is available in the R language. The vector of sentences

for a document is converted into a string, which lexRankr uses to prepare a summary. The length of the summary is adjustable, and in this study, a three-sentence summary was used.

The relationship word summary function was developed by the author to extract summaries based on the number of relationship words in a sentence. This function parses each sentence in the vector, counts the number of relationships words and stores a sentence identifier and count in a list. The complexity level of the summary can be selected.

The author added routines to read documents from the Gutenberg online library as well as pdf or text documents on local or remote file systems. There are standard programming libraries available to read pdf documents, which are the common format from online document libraries such as EBSCO or JSTOR. During the extraction, transformation and loading phase of analysis, all these formats are transformed into text files that the data analytic functions require. The documents were further prepared by removing publisher annotations so that only the original text was read.

An article about smoking in Appalachia was selected as the document to be summarized. The prepared article text was submitted to seven online text summarization services. A manual reference summary was also prepared by the author. Each of these summaries became a reference summary used by Rouge to compare with the summary generated by lexRankr as well as with the summary generated by the relationship word summarizer. The summary from lexRankr was treated as a candidate summary, which was compared to each reference summary one at a time to calculate the Rouge scores. The higher the score, the higher Rouge ranks the candidate solution. Likewise, the summary from the relationship word analysis function was treated as a candidate summary and compared to each reference summary to calculate its Rouge scores.

The comparison between the set of Rouge scores generated for lexRankr and the Rouge scores generated for the relationship word summarizer was used to determine if there was a significant difference between them. If lexRankr scored significantly higher than the relationship word summarizer then we would reject the null hypothesis that the relationship word summarizer has comparable performance to an existing, widely used summarizer. The two candidate summaries were separately compared to the reference summaries, both the online services and the manually generated, using Rouge. The resultant recall, precision and fmeasure were calculated.

Although the t-test is often used to determine the differences between such groups of scores (Jackson, 2005), the t-test assumes that the distributions are bell shaped. The author could not find a study to confirm this assumption, so a Kolmogorov-Smirnov (KS) test was used to determine if the two sets of scores were different. Lin (2004) recommends using the fmeasure score to determine the similarity between summaries. These scores were used in the KS test to calculate if the scores are significantly different. For the KS test, the null hypothesis is that the two sets of Rouge fmeasure scores come from the same population while the alternative hypothesis is that they are different enough to form two separate distributions.

Results

The results of generating the two candidate summaries, eight reference summaries and calculating the Rouge scores comparing each candidate summary against all eight reference summaries are shown in Table 1. Higher scores denote better performance. At first glance, the widely distributed lexRankr and the proposed relationship word summarizer have similar performance across all three metrics. In some cases, lexRankr outperforms the proposed summarizer (relWord columns in Table1) and in other cases the proposed summarizer outperforms lexRankr.

Table 1: Rouge score results

Reference Source	Recall		Precision		fmeasure	
	relWord	lexRankr	relWord	lexRankr	relWord	lexRankr
Paraphraser	0.451613	0.419355	0.323077	0.330508	0.376682	0.369668
Ahrefs	0.5	0.571429	0.215385	0.271186	0.301075	0.367816
Resoomer	0.530612	0.438776	0.4	0.364407	0.45614	0.398148
TLDR This	0.403846	0.5	0.161538	0.220339	0.230769	0.305882
Semrush	0.6	0.511111	0.207692	0.194915	0.308571	0.282209
Text Compactor	0.3	0.39	0.230769	0.330508	0.26087	0.357798
Summary Generator	0.459459	0.486486	0.261538	0.305085	0.333333	0.375
Manual effort	0.535714	0.553571	0.230769	0.262712	0.322581	0.356322
Average	0.472656	0.483841	0.253846	0.284958	0.323753	0.351605

As noted above, the fmeasure score will be used for further analysis. A visual comparison of the fmeasures for lexRankr and the relational word summarizers is shown in Figure 1. The plot shows that the two summarizers appear to have similar performance, with slight differences.

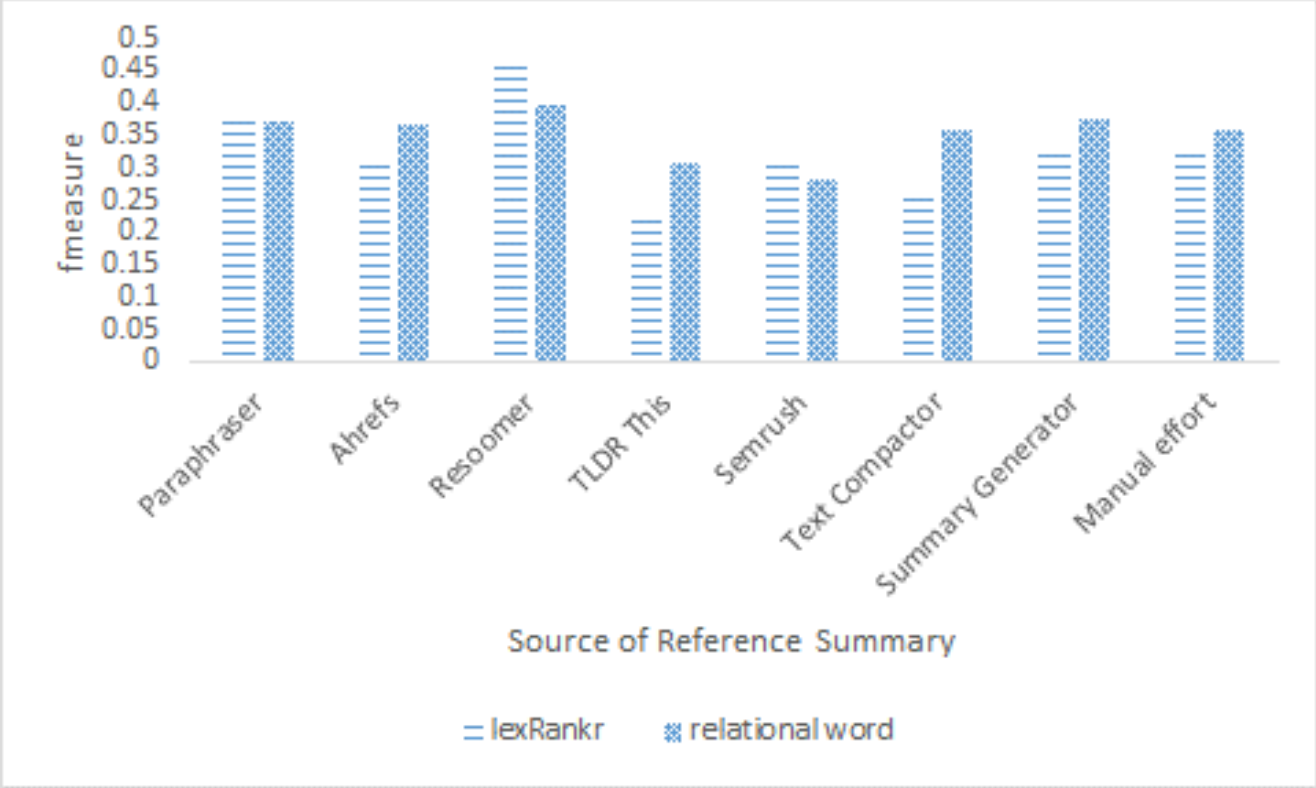

Figure 1: Comparison Plot of fmeasure Scores.

A test was performed to discover if there is a significant difference between the two sets of scores. The KS is a two-sample test used to determine the similarity of two samples. Our interest is to determine if the Rouge scores for the lexRankr summarizer are like the relationship word summarizer. The implementation of the KS test takes the two fmeasure samples as parameters. The KS test on the fmeasure scores resulted in a D statistic of .5 with a p-value = 0.2827. Since the p-value is > .05, we cannot reject the null, which is that the two sets of fmeasures are similar enough to have come from the same population. The D statistic represents the maximum distance between the cumulative distributions for the lexRankr Rouge and the relationship word fmeasures. The results show that the Kolmogorov-Smirnov test found no statistically significant performance difference between lexRankr and the relationship word summarizer.

Discussion

Relationship words associate the concepts in a document together to form a line of reasoning. The proposition of this research is that one application area for such a relationship word analysis is text summarization. A straightforward application program in the R programming language implemented this idea and was applied to produce a test summary. An approach was also developed for evaluating the performance of the new text summarizer. It utilized the Rouge scores for the relationship summarizer against a manually prepared summary as well as seven summaries generated by summarizers available online. To provide a reasonable baseline to evaluate the performance of the relationship summarizer, the Rouge scores for the widely distributed lexRankr summarizer were also calculated. The result is that the two summarizers, lexRankr and the relationship word summarizer, performed similarly. There is no statistically significant difference between the Rouge scores of the two. This provides support that the relationship word summarizer is at least as effective as an accepted summarizer.

The major contribution to the field of artificial intelligence by this study is that it takes an alternative approach to natural language processing based on the relationship structure of a document. Historically, such an approach has been applied in the areas of linguistics and scientific inquiry, but it has not been implemented on modern computer systems with their powerful computational ability. Although the approach in this study was applied to extractive summarization, analyses other than text summarization can be automated. As noted previously, relationship words can be used to identify the concepts the author wants to interrelate.

Another approach for using relationship words in textual analytics is to develop a statistic for a document's relationship complexity by doing frequency counts to get the number of sentences by number of relationship words. The resulting distribution of x_1 sentences with one relationship word, x_2 sentences with two relationship words up through x_n with n could be used to compare different documents to determine differences between genres. A multivariate analysis on such distributions could be done over a series of time periods to see if and how the relationship complexity of documents has changed over time. Furthermore, documents may cluster based on the distribution of relationship words. These are opportunities for future research.

There are further applications for deriving an author's thoughts and ideas using different structural viewpoints to understand the author's document. Ogden (1934) and Crovitz (1970) identified action words, relationship words, and qualities that are used to generate documents. A document's structural model can be its organization with respect to these types of word sets. For example, an action structure would base sentence selection for a summarizer using the set of action words. This paper traced a line of reasoning on using automated methods that analyze language to generate insights. Based on this line of reasoning, the author created a text summarizer that leverages the relationship structure for a document to produce a

summary. The results show such a summarizer compares favorably to an existing, accepted summarizer that is available in programming language libraries.

A limitation of the research is the difficulty in validating artificial intelligence products. The test approach taken considers the number of words in summary X that are also in summary Y (recall) and the number of words in Y that are also in X (precision). This approach showed that the resultant summarizer was as effective as a summarizer already published.

References

Allahyari, M., S. Pouriyeh, M. Assefi, S. Safaei, E. Trippe, J. Gutierrez & K. Kochut (July 28, 2017). Text Summarization Techniques: A Brief Survey. Cornell University. Retrieved from https://arxiv.org/abs/1707.02268

Barnett, H. (1953). Innovation: The Basis of Cultural Change. New York, NY: McGraw-Hill.

Basyal, L. & M. Sanghvi (October 17, 2023). Text Summarization Using Large Language Models: A Comparative Study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models. Cornell University. Retrieved from https://arxiv.org/pdf/2310.10449

Crovitz, H. (1970). Galton's Walk. Harper & Row.

Cuantum Technologies (July 6, 2023). Introduction to Natural Language Processing with Transformers. Independently published.

Davis, M. (1971). That's Interesting! Towards a Phenomenology of Sociology and a Sociology of Phenomenology. Philosophy of Social Science, 1(1971), 309-344.

Feehan, M., L. Owen, I. McKinnon & M. DeAngelis (October 2021). Artificial Intelligence, Heuristic Biases, and the Optimization of Health Outcomes: Cautionary Optimism. Journal of Clinical Medicine. 2021, 10(22), 5284; DOI: https://doi.org/10.3390/jcm10225284

Freeman, P. & A. Newell (1971) A model for functional reasoning in design. *International Joint Conference on Artificial Intelligence Proceedings (1971)*, 621-640

Ganesan, K. (January 25, 2017). An intro to ROUGE, and how to use it to evaluate summaries. Retrieved from https://medium.com/free-code-camp/what-is-rouge-and-how-it-works-for-evaluation-of-summaries-e059fb8ac840

Hargadon, A. (2003). *How Breakthroughs Happen.* Boston, MA: Harvard Business School Press.

Jackson, S. (2005). Statistics. Thompson Wadsworth.

Kim, W. & R. Mauborgne (2005). Blue Ocean Strategy: How to Create Uncontested Market Space and Make Competition Irrelevant. Boston, MA: Harvard Business Review Press

Kouris, P., G. Alexandridis & A. Stafylopatis (July 28, 2021). Abstractive Text Summarization: Enhancing Sequence-to-Sequence Models Using Word Sense Disambiguation and Semantic Content Generalization. retrieved from https://direct.mit.edu/coli/article/47/4/813/106774/Abstractive-Text-Summarization-Enhancing-Sequence

Kryściński, W., N. Keskar, B. McCann, C. Xiong & R. Socher (August 23, 2019). Neural Text Summarization: A Critical Evaluation. Cornell University. Retrieved from https://arxiv.org/abs/1908.08960

Lin, C. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Luhn, H.P. (April 1958). The Automatic Creation of Literature Abstracts. The IBM Journal.

Madhuri, J. & R. Kumar (2019). Extractive Text Summarization Using Sentence Ranking. 2019 International Conference on Data Science and Communication (IconDSC), Bangalore, India, 2019, pp. 1-3, doi: 10.1109/IconDSC.2019.8817040.

Musa, J. (1999). Software Reliability Engineering. McGraw-Hill.

Nader, R. (1972). Unsafe at Any Speed. Grossman

Ogden, C.K. (1934). The System of Basic English. Harcourt Brace.

Partridge, D. (1998). Artificial Intelligence and Software Engineering. Routledge

Pearson, K. (1900). The Grammar of Science. Dent original; Cosimo reprint. Reprint publication date 2007.

Schmidt-Biggemann, W. (2018). Llull, Leibniz, Kircher, and the history of Lullism in the Early Modern Era. In A. Vega, P. Weibel & S. Zielinski (Eds.), Dia-Logos. (pp. 38-61). ZKM Karlsruhe.

Smith, A., R. Black, J. Davenport, J. Olszewska, J. Rossler & J. Wright (2022). Artificial Intelligence and Software Testing. BCS

Stagnaro, A. (June 30, 2016). Bl. Raymond Llull and the World's First Computer. National Catholic Register. Retrieved from https://www.ncregister.com/blog/bl-raymond-llull-and-the-worlds-first-computer

Thayer, R. & I. Sommerville (2002). Software Engineering, Volume II. IEEE.

Veningston, K., R. Venkateswara & M. Ronalda (2023). Personalized Multi-document Text Summarization using Deep Learning Techniques. In Procedia Computer Science Volume 218, 2023, Pages 1220-1228. Retrieved from https://www.sciencedirect.com/science/article/pii/S187705092300100X

Wallace, A. (1963). *Culture and Personality.* New York, NY: Random House.