Decoding Crosswords: Semantic Wordplay using NLP Techniques

PROJECT SYNOPSIS

OF MINOR PROJECT

BACHELOR OF TECHNOLOGY

COMPUTER SCIENCE AND ENGINEERING

SUBMITTED BY

BHARATDEEP SINGH	CHANDANBIR SINGH	DIVNEET KAUR
URN: 2104080	URN: 2104083	URN: 2104093
CRN:2115026	CRN: 2115029	CRN: 2115039

BATCH: 2021-25

GURU NANAK DEV ENGINEERING COLLEGE,

LUDHIANA

Table of contents

INTRODUCTION

Modern computational models have made great progress at handling a variety of natural language tasks that require interpreting rich syntactic and semantic structures. However, in NLP as in other areas of AI, machines still lag humans on tasks that require flexible problem solving, rapid learning of unseen tasks, and generalization to new domains. Hence, creative language games are a rich area for developing more flexible NLP models. Our project 'Decoding Crosswords: Semantic Wordplay using NLP Techniques' aims to enhance lexical understanding by solving English crossword puzzles. At the end it will be able to identify the answers based on clues given and will provide an environment that is easy for people to interact with, and eventually, our model will solve crossword puzzles.

Crossword puzzles are perhaps the world's most popular language game, with millions of solvers in the whole world. Crosswords test knowledge of word meanings, trivia, commonsense, and wordplay, while also requiring one to simultaneously reason about multiple intersecting answers. Consequently, crossword puzzles provide a testbed to study open problems in AI and NLP, ranging from question answering to search and constraint satisfaction. Hence, crossword puzzle solving is a useful benchmark for people developing natural language processing (NLP) systems that can apply linguistic and meta-linguistic knowledge in more imaginative, adaptable, and human-like ways. Creating such a system for solving crossword puzzles, can open up new possibilities for creative solutions in a variety of domains that need complicated interpretation and reasoning. Operating at the core of specialized fields, e.g. NLP, puzzle solving, our project addresses critical challenges within. The level of skill needed goes beyond traditional applications and explores the nuances of language processing abilities beyond traditional question answering.

The development of an automated crossword solver hinges on essential software tools and hardware resources. Key software requirements encompass NLP libraries like NLTK and Hugging Face Transformers, Python for algorithm implementation, and machine learning frameworks for model integration. Access to diverse crossword puzzle databases is crucial for algorithm training and testing. On the hardware side, computational resources with ample processing power and memory are necessary for efficient algorithm execution, particularly during training and testing phases. For the implementation purpose we will be using various neural and non-neural networks. Depending on the accuracies obtained we will be choosing the most appropriate model for our work. Below is a quick description of a few of them:

WordNet is a systematic lexical database that groups words into synsets to identify connections between words in the English language.

T5, a powerful transformer-based language model developed for a text-to-text framework, is proficient in a wide range of natural language processing (NLP) tasks since it has been pre-trained on a vast amount of textual data, which aids in the representation-learning process.

 While Rule-Based Systems can be helpful in certain situations, they might not be as adaptable and intelligent as more complex machine learning models, like T5. Rule-Based Systems use logic and preset rules to process information.

RATIONALE:

Automating crossword puzzle solving with natural language processing (NLP) is a noteworthy technological development with real-world applications. The NLP-driven system demonstrates how advanced language models can be used in real-world situations to handle challenging linguistic tasks. This technology will advance the creation of more intelligent and user-friendly artificial intelligence systems while also showcasing the capabilities of natural language comprehension. The crossword solver will offer a smooth and effective user experience for an otherwise laborious and cognitive endeavour. It will demonstrate how artificial intelligence (AI) may improve user experiences and make recreational activities more pleasurable and accessible. Furthermore, the incorporation of this technology into language-related apps, learning environments, or cognitive support tools will emphasize its flexibility and adaptability, providing users with novel methods to interact with and leverage NLP capabilities.

The creation of sophisticated language models and their incorporation into crossword solving programs will advance artificial intelligence (AI) in a larger technological framework. It will challenge the limits of what is possible in natural language understanding and will pave the way for further developments in language-centric applications in a range of fields.

OBJECTIVES:

Following objectives will be achieved for the accomplishment of the project:

1. To extract semantic information from the given hints for filling blank cells of given crossword puzzle.

2. To develop a mechanism to predict word based on extracted semantic information.

3. To develop a system for providing complete solution to crossword puzzle.

LITERATURE REVIEW

Our approach is inspired by the success of the Berkeley Crossword Solver [3], which demonstrated remarkable achievements in automatically solving crossword puzzles using innovative techniques. This paper presents a cutting-edge method of automatically resolving crossword puzzles by fusing local search strategies, loopy belief propagation, and neural question answering models. Showcasing its effectiveness, the system beats human rivals at the American Crossword Puzzle Tournament, achieving notable improvements in accuracy. By using cutting-edge technologies, the system demonstrates its proficiency in navigating the intricacies of crossword clues and establishes a new standard for precision and productivity in automated crossword solutions. The authors of the paper titled "Key to crossword solving: NLP" [2] provide an extensive collection of crossword puzzles from the New York Times covering 25 years. As baselines, the paper presents a variety of retrieval- and sequence-to-

sequence-based models and presents a non-parametric constraint satisfaction method for solving complete problems. Through the work, a useful benchmark for assessing NLP models on the crossword-solving problem is established. The paper [4] addresses the difficulty of interpreting natural language in crossword solving and highlights the importance of applying NLP knowledge-based notions. The automated solver uses natural language processing techniques to decode clues and output a possible set of solutions to users. The system enables users to select the best answer from the generated set by adhering to crossword solving criteria. This method maintains the inherent qualities of crossword solving while also demonstrating the potential of NLP to improve crossword solving skills. The paper [1] explores the potential of advancing NLP systems through the analysis of cryptic crosswords, a prevalent variety in the UK. Cryptic crosswords pose a promising challenge for NLP systems due to their semantically complex and highly compositional nature. The study introduces a novel dataset of cryptic clues, serving as a challenging benchmark for NLP systems aspiring to process compositional language in a more creative, human-like manner. Thus, in near future, we plan to broaden our investigation to the difficult field of cryptic crosswords. Given that they often contain highly compositional, semantically difficult text that is intended to stump human solvers, cryptic crosswords present a unique opportunity to enhance NLP systems.

FEASIBILITY STUDY:

Technical Viability: The technological feasibility is great because it will make use of pre-existing NLP models and will incorporate neural question answering techniques for hint interpretation. Having access to well-established NLP frameworks guarantees steady progress.

Monetary Viability: A financial viability of the project is guaranteed through the use of open-source NLP libraries and technologies. The project is financially feasible due to the accessibility of computational resources and the lack of restrictions imposed by proprietary software.

Practicality for Operations: Operational feasibility is strong because the project fits within the present capabilities of NLP research and development teams. The automated crossword solver will be implemented and maintained effectively due to the NLP expertise that already exists.

User Importance: Crossword puzzles are incredibly popular, and automating the solution process will help users to overcome the difficulties presented by difficult clues. The development of the project is crucial because there is a clear need for an accurate and efficient crossword solver.

An automatic crossword solver is a helpful tool that will help the crossword solvers overcome obstacles presented by difficult clues. The importance of this tool is in improving user satisfaction and encouraging crossword puzzle participation.

METHODOLOGY/ PLANNING OF WORK:

The project adopts an applied research approach, focusing on the development and implementation of an automated crossword solver using advanced NLP technique. NLTK and

Hugging Face Transformers are popular libraries that will be used for activities like tokenization and model integration in order to collect and analyse data for NLP. We choose Python as our main programming language in order to guarantee flexibility and NLP tool compatibility. We will use libraries like BeautifulSoup4 (bs4), requests in our web scraping efforts to precisely and effectively extract data from online pages. We will be making HTTP requests to the target domains using Python's requests module. With the help of bs4, we would be able to parse HTML and XML data and systematically browse the webpage's structure. The combination of bs4 and requests will prove to be a powerful web scraping solution, offering a quick and efficient way to collect data from many internet sources. We will be scraping our data from 'https://www.theguardian.com/crosswords'.

The gathered data will be altered and structured in accordance with the guidelines provided in the paper [1]. Following appropriate data scraping and cleansing, the submitted data will be divided into three categories: training, validation, and testing. Later on, we'll analyse the accuracy of several NLP techniques (models) and draw conclusions about which model to utilize based on that comparison. When we have achieved a good level of accuracy, we will create an environment that is easy for people to interact with, and eventually, our model will solve crossword puzzles.

FACILITIES REQUIRED FOR PROPOSED WORK

The proposed development of the automated crossword solver primarily requires access to essential software tools and hardware resources. Software requirements include NLP libraries such as NLTK and Hugging Face Transformers for language understanding, Python as the primary programming language for algorithm implementation, and machine learning frameworks for model integration. Additionally, access to crossword puzzle databases is crucial to compile diverse datasets for algorithm training and testing. On the hardware front, computational resources with sufficient processing power and memory are necessary for efficient algorithm execution, with the capacity to handle training and testing phases effectively.

EXPECTED OUTCOMES:

Following are expected outcomes of the project:

1. It will solve the given crossword puzzle based on the given hints.

2. It will be able to solve various American Crosswords and British Crosswords.

3. It will provide a more enjoyable and engaged crossword puzzle solving experience to users.

4. 4. It will advance the fields of creative problem-solving and language comprehension more broadly.

5. It will be able to enhance lexical understanding.

REFERENCES:

[1] J. Rozner, C. Potts, and K. Mahowald, "Decrypting cryptic crosswords: Semantically complex wordplay puzzles as a target for NLP," *arXiv [cs.CL]*, 2021.

[2] "Key to crossword solving: NLP," 2017.

[3] E. Wallace et al., "Automated crossword solving," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022.

[4] Down and Across: Introducing Crossword-Solving as a New NLP Benchmark Saurabh Kulshreshtha. Olga Kovaleva.

[5] G. Angelini, M. Ernandes, and M. Gori, "Solving Italian crosswords using the web," in Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 393–405.

[6] Fine-tuning T5 and RoBERTa Models for Enhanced Text Summarization and Sentiment Analysis. The Great Lakes Botanist, 2023.

[7] "Crosswords," The guardian.