Tedi Dika (txd190008) and Ritvik Regulapati(rrr180003)

ACL Paper Summary

In this essay we summarize the paper “Automated Crossword Solving” by authors Eric Wallace, Nicholas Tomlin, Albert Xu, Kevin Yang, Eshaan Pathak, Matthew Ginsberg, and Dan Klein. The authors of the paper are affiliated with the University of California, Berkeley NLP Group and the publisher is the Association for Computational Linguistics. The authors have prior works covering game playing-agents and analyzing various NLP models. In 2019, Eric Wallace wrote the paper “AllenNLP Interpret” which won the best demo paper award. The paper discussed AllenNLP Interpret, a toolkit he created that provides interpretation primitives for any AllenNLP model and task, a suite of built-in interpretation methods, and a library of front-end visualization components. The Automated Crossword Solving paper addresses the problem of solving crossword puzzles using machine learning. The Berkeley Crossword Solver (BCS) is described as the world's best crossword solver, and the first ever computer program to beat all human competitors at the world’s top crossword solving tournament. BCS makes significant innovations in Question Answering and Post-Processing capabilities. The paper begins by discussing how, compared to traditional QA benchmarks, crossword puzzles provide a more novel challenge as the clues are less literal, span different reasoning types, and cover diverse linguistic phenomena like wordplay. BCS has three major steps: first-pass QA, constraint resolution, and local search. One of the unique contributions of the paper is that the authors created a custom dataset to train their QA model. The dataset contains over six million question-answer pairs and serves as a unique and challenging testbed as it varies in authorship, spans over 70 years of pop culture, and contains examples that are difficult for even expert humans. In the initial step, the QA model generates likely answer candidates to given questions. The authors built their QA model based on a bi-encoder architecture and used two neural

network encoders trained with DPR. The authors decided to use a bi-encoder architecture because it is able to score answers efficiently and learn using fewer examples. Previous crossword solvers, such as “Dr.Fill " use TFIDF-like scoring in its QA model. When tested, BCS’s bi-encoder model significantly outperformed Dr.Fill’s TFIDF model and improved top 1000 recall from 84.4% to 94.6%. Next, in the constraint resolution step, BCS uses a Loopy Belief Propagation that takes in answer candidates produced by the QA model and creates a puzzle solution that satisfies the letter constraints. Belief Propagation was chosen because it prioritizes solutions with the highest expected overlap with the ground-truth solution, rather than the solution with the highest likelihood under the QA model. As a result, less anomalies and ‘distractor answers’ generated by the QA model are chosen. In the final step, local search is used to make small letter corrections on the puzzle solution. Corrections are done by scoring alternate answer proposals that are a small edit distance away. The paper discusses how during this step previous crossword solvers often made the mistake of scoring every proposal. This creates risk of abnormalities like nonsensical character flips that lead to higher model scores. To avoid this, BCS only scores proposals that are within a 2-letter edit distance and have nontrivial likelihoods according to Belief Propagation or a dictionary. To recap, the paper presents new unique methods for crossword solving based on neural question answering, structured decoding, and local search. This program outperforms expert human competitors and can solve crossword puzzles from various domains.

In terms of evaluating the effectiveness of the BCS in solving crossword puzzles, there were points in which important metrics were taken note of to observe the effectiveness of the solver as a whole. The first of the evaluations was done while looking at the first pass of the BCS. In this step, the question-answer pairs are pulled from crosswords that came from various

publishers. To evaluate the effectiveness of this step, the dataset of question-answer pairs was then split into a validation and test set where the validation set only consisted of crossword puzzles published by the New York Times from 2020 and 2021. This constraint was set due to factors of both temporal distributions as well as the fact that the New York Times puzzles are the most popular as well as the most validated crossword publisher. Another important statistic that was measured was the Top-k recall of the QA model. As it turns out, this value was highly correlated with downstream solving performance. Using this statistic in comparison with previous crossword solvers such as Dr. Fill, it was found that the QA model is far more effective with a Top-K recall of 94.6% while Dr. Frill only had a Top-K recall of 84.4%. Furthermore, to compare the end-to-end results of the QA model in comparison with other solvers, other accuracy metrics were used. These accuracy metrics that were compared were perfect puzzle, word, and letter. Perfect puzzle as a metric refers to answering every clue in the puzzle correctly while word and letter as metrics refer to getting their respective answer correct given a clue. When using these accuracy metrics, it was discovered that the QA model outperformed Dr. Fill in every way at differing levels. Last but not least, the QA model was evaluated by being placed into the American Crossword Puzzle Tournament in order to see how it would perform. Here it was recorded that the model outperformed all human participants and achieved a score of 13,065 which corresponds with getting 6 out of 7 puzzles perfect and only getting 1 letter wrong on one puzzle.

Overall, this paper has only been cited once. The direct result of this paper is proof that solving crossword puzzles using natural language procession can be done to a near-perfect level and in fact, can perform better than humans can when given time. Although the implications of

this are not that far-reaching, the progression of the field as a whole and its testing of limits make this paper an interesting case study of a very specific problem.