Reproducing A Competitive Scrabble Agent

Casey Pore
Colorado State University
cpore@rams.colostate.edu

Problem Description

In this paper, I describe the process of building a Scrabble agent from scratch and evaluating its potential for competitive play. The unique properties of Scrabble make it difficult to apply existing generalized AI strategies common to other games, such as an adversarial search. In fact, even the best Scrabble agent (Brian Sheppard's Maven) does not utilize any local search techniques until the very end of the game, although it has been shown that the end-game search can be crucial for clinching a win against the best human players (Sheppard 2002b). As we will see, creating a competitive Scrabble player is more about adding features that incrementally improve the agent's chances for higher scoring, such as a combination of move selection heuristics and end-game search, than it is any one over-arching concept of Artificial Intelligence. This sentiment was expressed by Brian Shepard when he said his program, Maven, "is a good example of the 'fundamental engineering' approach to intelligent behavior" (Sheppard 2002b).

Scrabble is neither a game of perfect information nor a zero-sum game. Outcomes are highly dependent on the usefulness of the tiles a player randomly draws from the bag to make high scoring moves. Additionally, it cannot be known what tiles the opponent holds. This makes it difficult (or at least futile) for an agent to plan ahead or anticipate future states of the game from which to make informed decisions about how to play its tiles. This limits the options for quickly generating moves an agent may select to play. To accomplish the goal of generating possible moves quickly, a compact data structure along with an algorithm for identifying valid moves lie at the heart of my implementation, as well as agents that have come before mine. Once this fundamental technique was in place, I worked toward adding heuristics to improve the chances of maximizing the average move score in hopes of creating a more competitive agent.

After completing the move generation algorithm, the first heuristic implemented was to simply choose the move which produced the the highest score. I found that my implementation, though slightly slower than Appel and Jacobson's player from which I based my implementation, outperformed their player in regards to average move score by a significant margin (Appel and Jacobson 1988). Next, I implemented several heuristics (drawing inspiration from others' previous work) under the hypothesis that each heuristic would add a statistically significant increase to the average move score of my agent and increase its win ratio. An agent was created for each of these heuristics and pitted against an agent using only the basic maximum score heuristic to evaluate their effectiveness. Finally, after evaluating which of these individual heuristics led to a better outcome, I implemented several agents that use different combinations of the heuristics hypothesizing that they would lead to even better outcomes. These multi-heuristic agents were pitted against the maximum score agent to evaluate which combinations worked the best. I found that almost all of the multi-heuristic agents played better than any single-heuristic agent for creating strong Scrabble agent; however, not all of the single-heuristic agents bested the maximum score agent or produced conclusive results about player strength.

Previous Work

Techniques for creating strong AI Scrabble players have been investigated as far back as 1982. The earliest attempt I encountered was a paper by Shapiro and Smith (S&S) called "A Scrabble Crossword Game Playing Program" (Shapiro and Smith 1982). This attempt was limited by its move generation algorithm. Moves were generated by trying permutations of tiles and then using backtracking to check the validity of the attempted move. Move positions were selected by placing words across existing words only. Although there was no intelligence about forming words with adjacent tiles to make multiple words, the algorithm produced moves averaging a respectable 13.125 points, by selecting either the highest scoring move or the first move found above a set point threshold (Shapiro and Smith 1982). Despite its limitations, this paper laid the foundation for how future programs would achieve a fast search by representing the lexicon as a trie data structure combined with a backtracking algorithm to find valid words.

Though work to create Scrabble agents started over three decades ago, major advancements were few and far between. Scrabble as an AI topic is indeed a narrow niche and not popular among academics due to its clear lack of opportunities for improvement and general application. The next milestone in Scrabble agent improvement came in 1988 in a

paper by Appel and Jacobson (A&J) that improved the speed over the fastest Scrabble agents at the time by two orders of magnitude (Appel and Jacobson 1988). A&J expanded upon S&S's program in two ways. First, the trie is reduced to a Directed Acyclic Word Graph (DAWG). While this does not directly improve the speed in which moves are generated, it has the advantage of significantly reducing the amount of memory needed to represent the lexicon. This made it small enough to keep the entire lexicon in memory, improving speed. Speed improvement was an important enhancement considering the limitations of memory at the time, and is still important today in mobile applications. Second, they implemented a better backtracking algorithm. S&S's program simply used backtracking to check if a permutation of letters was valid word. A&J's program improves this by considering information about the board to prune invalid moves from the search. This includes pre-computing cross-checks so that no tiles would be selected from the rack unless they were part of a valid cross-word, and using an "anchor" to select where to start placing "left parts" of a word to play[1]. An anchor is simply an open square with a tile to its right that a new word could be connected to.

A&J's program generates moves by using backtracking to place tiles from the rack that form left parts first. Because the cross-checks have been computed and there is a tile to build the left part from, there is never any tile placed from the rack that is not part of *some* word. These two properties reduce the number of moves that need to be placed and evaluated significantly. Once the left part of each move is formed, the algorithm builds the right part of the word in the same way, except it accounts for tiles already on the board. The takeaway here is that A&J's algorithm tries to maintain move validity as it goes. It ensures placed tiles are always part of a word, even if it can't complete it, whereas S&S's algorithm tries an entire permutation, and then checks that the move is valid. Like S&S's program, A&J's implementation simply selected the highest scoring move it found.

The apex of Scrabble agent development seems to have come in the form of Brian Sheppard's Maven. Development of Maven actually started in 1986 before A&J's program was created and was already a strong agent capable of tournament-level play (Schaeffer 2001). However, it immediately switched to using A&J's superior move generation algorithm when Sheppard discovered it. Maven is the accumulation of Sheppard's long history of building a Scrabble agent, and his experience is evident in its multifaceted approach to play. Beyond A&J's fast move generation algorithm, Maven adds several features, many of which are applicable only to corner-cases, but important for catching every last bit of opportunity for selecting best moves. Sheppard has classified these features into 3 categories: rack evaluation, board evaluation, and search.

Rack evaluation is a mechanism by which the tiles left on the rack (known as a leave) for a given move are evaluated based on their utility to produce higher scoring subsequent moves. There are two main techniques used by Maven for rack evaluation. First, Maven maintains a static list of tile combinations, each resolving to a "learned parameter" which adjusts the utility of the move based on the leave. The learned parameters are not a feature per se, but is the process by which Sheppard gathered statistics about what values are optimal to use with different combinations for evaluation. This parameter learning is not part of Maven (it does not employ any machine learning or adjustment of parameters at run-time), but were gathered as the result of "a day's worth of self-play games," tested for their optimality, and applied to the various leave combinations (Sheppard 2002b). Beyond the learned parameters, Maven uses several heuristics (referred to as "extensions" by Sheppard) for rack evaluation. These heuristics include potential usefulness of tiles left in the bag, vowel/consonant balance, and holding onto a "U" tile if the "Q" tile hasn't been played yet.

Board evaluation is the process of using features of the game board to determine the utility of a move. Several board evaluation techniques were considered. For example, it was hypothesized that choosing a good opening move would prevent the opponent from playing on bonus squares. Or that you should always play a word on a bonus square, so that your opponent couldn't take it, even if it meant sacrificing a higher scoring move. In the end, it was found that almost all of the board evaluation techniques examined were irrelevant, except for one - triple word squares. Like the learned parameters for rack evaluation, Maven has a table of parameters to evaluate how many points could be compromised as a function of the placement of an opening move in terms of its location on the board. This is because bad opening moves could lead to missed opportunities to use the triple word square later in the game.

Lastly, Maven performs a search during the pre-endgame and endgame. Before the endgame search, Maven performs a pre-endgame search which begins when there are nine tiles left in the bag plus the seven left on the opponents rack. Sheppard explains very little about how the pre-endgame works other than it "is a one-ply search with a static evaluation function that is computed by an oracle," and admits that it may not be useful, except to prevent an opponent from fishing for bingos (using all seven tiles in one move). The endgame search commences when all the tiles have been drawn from the bag, so the remaining tiles on the opponent's rack can be deduced from what's on the board and the game becomes one of perfect information. The endgame search utilizes a B* algorithm that evaluates N-turn sequences of moves and returns intervals that correspond to the risk associated with the sequence, rather than a utility of the board state itself (Sheppard 2002b).

Approach Taken

One prerequisite that I had for this project is that I plan on later porting it to the Android platform, so I preferred to implement it in Java. I had hoped to find a Java implementation that I could use as a starting point to add heuristics and endgame search to. I found many implementations in various languages, but only a few in Java. I didn't feel that any of those where implemented in a way that wouldn't require

[1]Please note that for brevity, I only explain horizontal move generation throughout this paper. Vertical move generation is simply a transpose of these concepts.

significant refactoring to meet my needs for this project, so I started from scratch. The only code in this project that wasn't written by me, was the DAWG used to perform word validation. This was leveraged from a Github project called "android-dawg" (icantrap (Github User) 2013).

I referenced A&J's paper as a performance baseline in terms of speed and player strength. Creating a player that quickly and correctly generated moves comparable to A&J was the first (and biggest) milestone for me to reach for this project. Beyond implementing this agent, I drew inspiration from Sheppard and other sources to add features to my agent that could improve its level of competitiveness. These include combinations of Maven's rack evaluation, board evaluation, and extension heuristics, as well as a few heuristics described in a project called "sharpscrabble" (wsanville@gmail.com (Google Code User) 2010).

Move Generation

While A&J's program was my basis for comparing my program to, I did not implement my agent the same way as them. We share many common techniques for move generation, but my implementation ended up being more of a "middle ground" between the techniques used by S&S and A&J. This shows in regards to the speed in which my agent selects its move to play. On average, my agent selects a move in 4.5 seconds, which is about 3 times slower than A&J, but about 9x faster than S&S. Adjusted for modern hardware, S&S and I are probably on equal ground in regards to speed, but I achieve much more in terms of player strength. In this section, I describe the methods used for move generation in my agent and compare and contrast it with S&S and A&J's implementations.

My program uses the second edition of the Official Tournament and Club Word List (TWL06) as it's lexicon (North American Scrabble Players Association 2009). The TWL06 contains 178,691 words and is 1.8MB is size. The data structure used to represent the lexicon in memory is key to generating moves quickly for all Scrabble agents, including mine. As noted earlier, the data structure used is called a DAWG, which maintains the lookup properties of a trie, but within a more compact memory footprint. Thus, the DAWG was chosen for its compactness, not for any efficiency gains. In my case the lexicon was reduced from its initial 1.8MB to only 687.5KB. Both the trie and the DAWG have equivalent lookup times of $O(L)$, where L is the length of the word to be searched. The longest words in TWL06 are 15 letters long, so at most $L = 15$. The DAWG is finite state recognizer that must be built from the lexicon before it can be used. To build and use the DAWG, I used an implementation that I found on Github, called android-dawg (icantrap (Github User) 2013). This is the only code in the project that I did not implement myself.

Because I chose to use android-dawg for storing the lexicon, this limited how my algorithm was able to leverage the board state to identify valid moves. In A&J's program, the anchor, in combination with the pre-computed cross-checks, allowed their algorithm to prune many unnecessary tile combinations from their search. Android-dawg only allows for two functions, to look up a given word, and to find all subwords given a string of characters (in this case, our rack tiles and potentially tiles already placed on the board). Because of this, finding valid moves is more akin to S&S's implementation that tries permutations of tiles from the rack due to the dawg's ignorance of the board state. I do, however, implement all of the cross-checking and other board state awareness features present in A&J's implementation, but at the expense of added time. My algorithm for generating all valid moves during a player's turn can be broken down into three phases: identify open positions, try rack permutations, and move validation.

To identify open positions, my program uses the concept of an anchor similar to A&J, but my definition is slightly different and I refer it as a "slot", so as not to overload the term anchor. A&J's anchor is an open space with an adjacent tile on its right which it can connect with to form a word, whereas my slot is the position of a tile on the board that has open spaces either to the west or to the east (or north or south for considering vertical placement). The first step in my algorithm iterates over the entire board to collect a list of these slots to determine where it could make a possible move.

Once all the possible slots have been identified, the algorithm takes the tiles on the rack and generates a list of all possible permutations of the tiles, without repetition. For a seven tile rack, this list contains 5040 permutations, and takes 35ms on average to create. The high-level algorithm (in Pythonic pseudo-code) for placing placing permutations is as follows:

```
possible_moves = list()
possible_slots = get_possible_slots(
    game_board)
permutations = get_permutations(rack)

for permutation in permutations:
  for slot in possible_slots:
    #places to the right of the slot
    if slot.is_open_east():
      place_permutation_east(
          permutation)
      if placement_ok(permutation):
        move = get_utility(permutation)
        possible_moves.append(move)
    #places to the left of the slot
    if slot.is_open_west():
      place_permutation_west(
          permutation)
      if placement_ok(permutation):
        move = get_utility(permutation)
        possible_moves.append(move)
```

The `place_permutation` functions place the tiles on board relative to each open slot which are then tested for whether it is a valid move. How these permutations are placed is different depending on which side of the slot we are placing them. The `place_permutation_east` function is simple. It places all the tiles in the permutation to the right of the slot, skipping over tiles already on the board. Then it checks that the word formed from the tiles placed from

the rack, and tiles already placed on the board, is a valid move. The `place_permutation_west` function is easy to understand, but was a bit more difficult to implement. It works the same way as the `place_permutation_east` function with an additional step. Instead of just placing the permutation once, it must try placing the permutation once for each open tile to the left of the slot, up to the length of the permutation, or if it collides with an existing tile on the board it will terminate at that length.

After a permutation is placed, it is checked to verify whether or not it is a valid move. This includes checks to ensure that the word is in bounds of the board, that all tiles forming cross-words are valid, the formed word is legally connected, etc. If a placement is found to be a valid move, is evaluated by one or more heuristics to calculate its utility and added to a list of possible moves. The function for validating a horizontally placed permutation (a `move`) is shown below (simplified for publication).

```
#Check that the tiles do not extend
#past board boundaries
if not in_bounds(move):
  return false

#Handles case where a horizontal word
#is placed at the bottom of a vertical
#word (i.e. there are no existing tiles
#on the board in the horizontal word's
#path)
if not is_connected(move):
  return false

#Ensures that the main word formed is
#in the DAWG
if not main_word_ok(move):
  return false

#Does cross-checking to ensure that any
#vertical or adjacent words formed are
#in the DAWG
for tile in move:
  place(tile)
  if has_north_neighbor(tile):
    if not vertical_word_ok(tile):
      return false
  if has_south_neighbor(tile):
    if not vertical_word_ok(tile):
      return false

#Gets the final score for the move,
#including points for adjacent words
get_score(move)

return true
```

Move Selection

The final move to be played is selected from a list of possible moves that are sorted according to the utility assigned to it by various heuristics. The move with the highest utility is selected. The formula for calculating the utility is based on the actual points scored by the move plus or minus a value calculated by a heuristic. My agent doesn't implement the more advanced features of Maven, such a endgame search, but does employ heuristics that are sufficient for besting A&J's program. The heuristics described in this section were inspired by or directly pulled from Maven or sharpscrabble.

Sheppard is protective about the values of the learned parameters of his heuristics, and does not provide many details, so some of parameters used in the heuristics I implemented based off his description may not be optimal. Additionally, Sheppard does not give details about how his heuristics use the learned parameters to calculate the utility of a move and uses obscured language to say that for a given heuristic, "a linear function implements this concept," and offers little more (Sheppard 2002b). However in his PhD dissertation, he does offer some "Basic" learned parameters I was able to utilize, but still does not provide details about the function to utilize them properly. As such, for many of these functions it was a process of trial and error to determine what worked and what didn't. Despite lack of insight to Sheppard's learned parameters, I was able to achieve some positive results (Sheppard 2002a).

The `MaxScore` heuristic is a basic heuristic that is simply the number of points a move would receive if it were to be placed on the board. All Scrabble agents I came across in my research use it. It is self-explanatory, so won't go into much detail about this it, except to say that in my implementation, all other heuristics use it to add or subtract their utility to/from.

`SaveCommon` is the first of five rack evaluation heuristics. Rack evaluation heuristics take into consideration that it is better to hold onto or play certain tiles because the relative usefulness of these tiles will lead to better scoring words down the road. `SaveCommon` was sourced from sharpscrabble, and of the five, this one is the most aggressive. This heuristic works by checking the leave for common tiles, which are the letters A, E, I, N, R and S. For each of these letters found in the leave, 5 points are added to the move's score (wsanville@gmail.com (Google Code User) 2010).

The second rack evaluation heuristic, `TileTurnover`, was inspired by Maven. It tries to estimate how good a rack leave is based on the potential for drawing high value tiles from the bag to replace the played tiles. When a move is played, the player must draw N tiles to replace the played tiles. The utility of the N new tiles may be better than the tiles played, in which case we would prefer to play the move in order to get the potentially more useful tiles on our rack. The potential usefulness of the tiles drawn from the bag is calculated by averaging the points of the tiles that have been unseen on the board. This will include tiles that may be on the opponent's rack, but there is no way to determine which tiles are on the opponent's rack and which tiles are left in the bag, so it is assumed they are all in the bag. The leave is calculated as such with P being the sum of the tiles left on the rack and A being the average value of the tiles in the bag: $P + (N * A)$. We also calculate the value of the tiles on the rack before the leave and the lesser of the two is the

resulting utility (Sheppard 2002b).

`VowelConsonant`, our third rack evaluation heuristic, was also sourced from Maven. The goal here is to keep a good balance of vowels and consonants on the the rack. Maven has a list of values made from its learned parameters for every possible vowel/consonant combination. Sheppard does not provide this information, but does provide a description of a "Basic" version of the learned parameters in his PhD dissertation that was used in early versions of Maven. Below is a table of these "Basic" parameters. The number of vowels and consonants in the leave are counted and these values are added or subtracted from the base score of the move (Sheppard 2002a).

Table 1: Basic Vowel-Consonant Balance Table

		Consonants						
		0	1	2	3	4	5	6
Vowels	0	0	0.5	1.5	0	-3.5	-6	-9
	1	-0.5	1.5	1	0.5	-2.5	-5.5	
	2	-2	-0.5	0.5	0	-2		
	3	-3	-2	-0.5	1.5			
	4	-5	-4.5	-0.5	1.5			
	5	-7.5	-7					
	6	-12.5						

`UseQ`, the fourth rack evaluation heuristic, is a simple heuristic that I devised after noticing early on in my testing that many times, the game finished with the Q tile being left unplaced because there were no moves that could be made from it. This heuristic simply adds a large value to a move's score, if it uses a Q. This artificially inflates a move's utility to prioritize moves that make use of it, which forces the agent to use the Q at its first opportunity, even if it means sacrificing higher-scoring moves during its turn.

The final rack heuristic, `UWithQUnseen`, is borrowed from Maven. It is similar to `UseQ`, except it prioritizes the QU combination rather than just Q. The justification for this, according to Sheppard, is that "U-less Q costs 12 points, whereas a QU together is neutral" (Sheppard 2002b). Again, Sheppard gives no hints about how the utility of this calculated. My implementation checks the move to see if it's going to play a U, and if the Q has not been played yet. If this condition occurs, it penalizes the move by subtracting the 12 points that could potentially be lost from the score.

The last heuristic, `UseBonusSquares`, is a board evaluation heuristic. Though Sheppard contends that board evaluation heuristics are useless (save for triple word squares), I wanted to perform a limited investigation of this concept myself to confirm his assertion. The idea for `UseBonusSquares` was borrowed from sharpscrabble. The concept is that moves that cover a bonus square will have a higher utility because they will prevent the opponent from utilizing those squares in future turns. The fact that the average tile point value is 1.9 is used to approximate points that the opponent will not be able to score as the result of our agent covering a bonus square. This value is returned as the utility for double or triple letter squares. For double and triple word squares, we use the fact that the average scrabble word length is 3.5 letters to calculate the potential point loss for the opponent as $3.5 * 1.9 = 6.65$ and $2 * 3.5 * 1.9 = 13.3$, respectively (wsanville@gmail.com (Google Code User) 2010).

Evaluation and Analysis

Evaluating what makes a good Scrabble player is a bit tricky. At first, it would seem obvious that the number points scored in a game would be the best metric for evaluating skill, however this is not the case. The quality of your opponent makes a big difference in how how many points can be scored by a strong player. If a strong player is competing against a weaker one, typically the weaker opponent will use less tiles per move. In turn this leads to more moves per game, and thus the strong player has more opportunity to score points. This doesn't really tell us how *well* the agent played, it only reflects how poorly the opponent played. In contrast, when two strong players compete, tile usage is maximized, which leads to less and higher scoring moves in a game, but generally the end score is lower and closer to each other. Sheppard notes that average move score is not the final word as far as metrics go for evaluating player strength:

> We do not really need one single measure of strength; the important thing is to have measures that are appropriate for what we are trying to accomplish. In comparing consecutive versions of Maven, points-per-turn is used since it is easier to reach conclusions instantaneously. In comparing Maven versus humans, rating proves to be the most useful metric. (Sheppard 2002b)

Given this, my analysis investigates two metrics: average move score (Sheppard's points-per-turn), and win percentage. I reserve the right to cherry-pick the metric which better supports my hypothesis. Just kidding - kind of. As we will see, the results are mixed and sometimes difficult to interpret, but also seemingly positive depending on which metric is examined. It's also worth noting that `MaxScore` alone is already a very proficient Scrabble player, and can easily best most human opponents. Heuristics evaluated here are meant to incrementally improve upon `MaxScore`.

Experiment Design

To evaluate the strength of my Scrabble agent, experiments were run in two phases. In the first phase, an agent was created for each of the seven heuristics to use as its move selection strategy. These seven agents were then played against the `MaxScore` agent for 100 games each. The hypothesis was that there would be a statistically significant difference of at least 0.5 points between each agent's average move score and `MaxScore`'s average move score, though I don't make claims as to whether the difference will be better or worse. Additionally, I recorded the win percentage that each agent attained over `MaxScore`, as this could be significant in determining if the agent was a better player as well.

Once the results of the first phase had been evaluated, the heuristics that seemed to add value to an agent's strength were identified. To execute the second phase, agents were created using combinations of the best heuristics to evaluate which combinations led to the creation of a strongest player.

Similarly to the first phase, this new batch of agents were played against MaxScore for 100 games, under the same hypothesis. Again, win percentage was recorded and examined as well.

Finally, an informal comparison of MaxScore, my highest winning agent, A&J, and Maven is conducted to gauge where my program stands in comparison to existing ones.

Experiment Analysis

Table 2 below shows a summary of results for phase 1 of the experiments. It contains some confounding and a few seemingly contradictory results. I discuss each one individually. Spoiler: phase 1 was fairly inconclusive, but phase 2 produced some tangible results.

Table 2: Single Heuristic Results Summary

Agent	Agent Mean	Max Score Mean	p-value	Agent Win Pct
UseQ	24.42	23.2	0.0052493	71%
UWithQUnseen	23.31	24.84	0.1762468	56.25%
SaveCommon	17.68	23.69	0	30%
VowelConsonant	24.39	24.39	0.1936386	56%
TileTurnover	22.4	22.47	0.2086259	60%
UseBonusSquares	23.53	23.71	0.2530724	55%

UseQ gave the strongest result with a p-value of .005 and a win percentage of 71%, however there is one caveat here. For UseQ (and also UWithQUnseen), only the games in which the agent held a Q tile were recorded. While I have little doubt that using a Q early leads to better outcomes (because otherwise it may go unused), the average move score may be artificially inflated by the fact that simply having a Q to play, whether it is used now or later, may lead to better outcomes. Q is one of the highest valued tiles in Scrabble, and my results only reflect games where the agent had the opportunity to play it.

From this point forward, no heuristic achieved statistical significance for the average move score, however that does not mean they don't add value to player strength. In fact, all of them except one bested MaxMove.

Results for UWithQUnseen were a bit confusing. It actually performed *worse* than MaxScore on the average move score metric, yet still won more games over it. My assumption is that it is because UWithQUnseen had a wider variance in its average move score, with about 25 for MaxScore and about 29 for UWithQUnseen.

SaveCommon was an overwhelming failure. It performed significantly worse in both average move score and win percentage. As noted earlier, this is the most aggressive heuristic, and the aggressiveness did not pay off. This speaks to the the value of using learned parameters to tune heuristics.

VowelConsonant was interesting because both it and MaxScore had the same average move score. It won 56% of its games, but this is not significant enough to determine it a success. It is not clear whether this heuristic adds any value to the agent, despite it's learned parameters coming directly from Sheppard. The learned parameters are the "Basic" version from early versions of Maven, but still I expected them to perform better than they did.

TileTurnover is an oddity because, like UWithQUnseen, its average move score value was less than MaxMove. However, this was the strongest single heuristic with a 60% win percentage.

Finally, UseBonusSquares performed similarly to the VowelConsonant agent. That is to say, results were inconclusive. Average move score was very similar for both agents, and it had 56% win percentage.

After evaluating these results, the best heuristics were selected to use for the agents in phase 2. Since none of the average move scores for any of the heuristics produced statistically significant results (save for MaxQ, with caveats), I used the win percentage as my main criteria for selecting heuristics to use. If the win percentage was 56% or higher, it was used, except for UWithQUnseen because using it and MaxQ at the same time is not logically possible. So the final heuristics selected to create multi-heuristic agents with were VowelConsonant, TileTurnover, and UseQ. Using these three heuristics, I created 4 agents as follows:

- Multi1 - VowelConsonant, TileTurnover, and UseQ
- Multi2 - VowelConsonant and TileTurnover
- Multi3 - VowelConsonant, and UseQ
- Multi4 - TileTurnover, and UseQ

The computed utility for each of these agents is calculated by simply adding the sum of the individual heuristics to the base score. Phase 2 results are summarized in Table 3 below. Unlike the single heuristic agents, results for these agents were a bit clearer and mostly positive, though not completely conclusive.

Table 3: Multi-Heuristic Results Summary

Agent	Agent Mean	MaxScore Mean	p-value	Agent Win Percentage
Multi1	23.64	23.98	0.3567191	65%
Multi2	24.03	23.05	0.0026197	66%
Multi3	25.16	24.33	0.0182257	60%
Multi4	22.56	22.82	0.3104625	55%

Though Mulit1 did not achieve a significant difference in average move score, it was clear that it was a better player. It bested MaxScore 65% of time, though it actually fared a bit worse than MaxScore in average move score. These confusing results cause me to wonder if there is a mistake in the mechanism I used to record move scores (perhaps a floating point issue?), or if this is expected. I would have expected to see positive results given the win percentage.

Multi2 did very well, besting MaxMove by almost a full point. The p-value is significant at about 0.002. The win percentage echoes this result at 66%. This combination of heuristics seems to be the best, although Multi3 performed significantly better in the average move score metric.

Multi3 also did well and showed a significant gain in average move score, with a p-value of about 0.02. Though

not a strong correlation, this agent achieved the highest average move score of 25.16. However, the win percentage was lower than `Multi2` at 60%. It can be concluded that this was an effective agent, but the weak correlation for average move score is a good example of how using it as a single metric may not always be appropriate for the model.

`Multi4` was the only multi-heuristic agent that cannot be concluded a success. The average move score did not show a statistically significant gain, and the win percentage was an inconclusive 55%. In addition, the average move score was lower than nearly every other agent, even most of the single-heuristic ones.

The results of the multi-heuristic agents demonstrate what a subtle process it can be to create a competitive Scrabble player. When examining the data for the single-heuristic agents, one might conclude that it would not be worth utilizing them in an agent. When used together however, some real performance advantages may be realized.

I end this section with a rough comparison of my `MaxScore` agent and my best agent, `Multi2`, with the two agents I referenced when creating my program. Below, Table 4 lists some metrics that are useful for the comparison.

Table 4: Comparison to Existing Agents

Agent	**Moves/ Game**	**Avg Move Time**	**Avg Move Score**	**Avg Game Score**
A&J vs itself	22.4	1.3	16.8	376.3
MaxMove vs itself	15.59	4.5	24.015	379.1
Multi2 vs MaxMove	16.55	4.7	24.03	401.1
Maven vs itself	10.5	No Data	35.0	367.5

The goal of my project was to create an agent that was at least as good as A&J and hopefully approach the level of play of Maven. By looking at the number of moves per game and the average move score metric, it seems that my program lands in the middle of the two. This leads me to believe I would best A&J, but it's hard to say for sure. Perhaps my player is only strong when played against itself. In terms of speed, my program is noticeably slower that A&J for the reasons mentioned earlier. Though the average move score for `MaxMove` and `Multi2` is similar we can't draw any conclusions from this because of the problems with two weaker players competing against each other that was noted earlier in the paper. So take Table 3 with a grain of salt; outcomes are largely dependent on who a player is competing against.

Conclusions, Future Work

Early on, I had high hopes of creating an agent comparable to Maven. Even as I was implementing the heuristics, I was anticipating they would add a significant performance advantage. Though I feel that my efforts were successful in creating a stronger player than `MaxScore` alone or A&J's program, it was clear that there is much more room for improving the parameters of my heuristics. The heuristics used in Maven enjoy the advantage of using finely-tuned learned parameters to calculate a move's utility. Appropriate tuning of these parameters is the result of Sheppard's long history and experience working on Maven.

In addition to tuning the heuristics, The is a lot of opportunity for speed improvement as well. If I had more time, I would scrap the DAWG that I used in the project, so that I may implement the backtracking algorithm used by A&J and Maven. Even if I didn't do that, there is still opportunity to implement a better cross-checking algorithm. By pre-computing the cross-checks, I would be able to prune unnecessary permutations from my search for those permutations that had letters in the positions that are know to form invalid words in the crossword.

I came out of this study with a greater appreciation of the painstaking process that Sheppard went through to develop his program. My final thought on developing this program is that creating a good Scrabble agent is not so hard, but creating a *great* Scrabble agent is. Kudos to Brian Sheppard.

References

Appel, A. W., and Jacobson, G. J. 1988. The world's fastest scrabble program. *Commun. ACM* 31(5):572–578.

icantrap (Github User). 2013. android-dawg. https://github.com/icantrap/android-dawg. Accessed: 2015-05-01.

North American Scrabble Players Association. 2009. Official tournament and club word list. http://www.scrabbleplayers.org/w/Official_Tournament_and_Club_Word_List. Accessed: 2015-05-01.

Schaeffer, J. 2001. A gamut of games. *AI Magazine* 22(3):29.

Shapiro, S. C., and Smith, H. R. 1982. A scrabble crossword game playing program. Technical Report 119, Indiana University.

Sheppard, B. 2002a. *Towards Perfect Play of Scrabble.* Ph.D. Dissertation, Universiteit Maastricht.

Sheppard, B. 2002b. World-championship-caliber scrabble. *Artificial Intelligence* 134(1):241–275.

wsanville@gmail.com (Google Code User). 2010. scharpscrabble. https://code.google.com/p/sharpscrabble/. Accessed: 2015-05-01.