Shannon meets Shortz: A Probabilistic model of Crossword Puzzle Difficulty

Miles Efron
School of Information, University of Texas
Sanchez Building 564, 1 University Station D7000
Austin, TX 78712-0390
miles@ischool.utexas.edu

Abstract

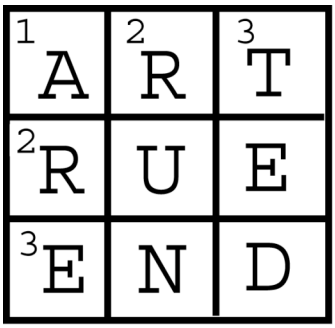
This paper is concerned with the difficulty of crossword puzzles. A model is proposed that quantifies the difficulty of a puzzle P with respect to its clues. Given a clue-answer pair (c,a), we model the difficulty of guessing a based on c using the conditional probability $P(a \mid c)$; easier mappings should enjoy a higher conditional probability. The model is tested by two experiments, each of which involves estimating the difficulty of puzzles taken from *The New York Times*. Additionally, we discuss how the notion of information implicit in our model relates to more easily quantifiable types of information that figure into crossword puzzles.

1 Introduction

This paper is concerned with the difficulty of crossword puzzles. Given a puzzle P we ask, how difficult is it to complete the puzzle? Of course crossword difficulty is subjective, depending on the skill and knowledge of an individual solver. But regular solvers of *The New York Times'* crossword are acquainted with a more objective notion of puzzle difficulty. It is well known (and well publicized) that *The Times'* puzzles get harder as the week goes on (Newman and Lasswell, 2006). Monday puzzles are the easiest, while Saturday puzzles are the hardest.

This paper asks, what distinguishes a Monday puzzle from a Tuesday puzzle, or for that matter from a Saturday puzzle? We seek some function $f(P)$ that captures this phenomenon, returning the day (or an equivalent measure of difficulty) that the puzzle is suited for. We focus on *The New York Times'* crossword puzzle for several reasons. From a practical standpoint, the *Times'* weekly puzzle schedule provides a well structured learning problem. Moreover, *The New York Times'* crosswords are highly regarded by puzzle enthusiasts. Nonetheless, the model derived in this paper would suit other American-style crosswords.

Modeling crossword puzzle difficulty is interesting for several reasons. First, recent literature has seen an increase of research on textual affect and style: aspects of meaning other than the denotational valence of statements. Recent studies have analyzed the semantic orientation of words (Turney and Littman, 2002), the political leanings of blogs (Efron, 2006), and the opinions of movie reviewers (Pang and Lee, 2005). The difficulty of guessing an answer based on a crossword puzzle clue presents another aspect of textual semantics. While a tremendous literature on textual difficulty (with respect to reading level, in particular) exists, crosswords present an interesting modeling challenge insofar as their difficulty is intentional and artfully deployed. Crossword puzzles must be *entertainingly* difficult, and this is the phenomenon we hope to describe here.



Across: 1 But is it ___?
2 Regret. 3 Finish.
Down: 1. Latin 101 suffix.
2 A good clip. 3 Senator Kennedy

Figure 1: A Simple Crossword Puzzle

Analyzing crossword puzzles also provides a useful laboratory for revisiting the worn, but still unsettled question: what is information in the context of information science? Definitions of information abound, and this paper makes no effort to offer another. But the proposed model of crossword difficulty is predicated on the notion that information is quantifiable and measurable. Given a particular answer *a*, the crossword author manages the difficulty of his puzzle by crafting a clue *c*. If *c* is highly informative with respect to *a*, the solution is easy. But if the author offers a less informative clue, finding the answer is difficult. Given a clue *c* with an answer *a*, we provide a model for answering the question: how informative is *c* with respect to *a*?

A caveat is in order here. This paper makes no claim for the cognitive validity of the model it presents. The human process of solving puzzles is obviously far more complex and creative than this paper can account for. However, the approach presented here is intended to demonstrate a novel way that a highly subjective information problem can be modeled quantitatively.

2 Background

Crossword puzzles are games comprised of a set of clues and an $n \times n$ grid that contains answers to its clues. Answers in the grid are interlaced; they appear horizontally (across) and vertically (down), as in Figure 1, a very simple, completed puzzle. Realistic puzzles have a few blank squares to aid in creative puzzle construction.

Several rules apply to American crossword puzzles:

- The pattern of blank squares is symmetrical
- Entries must be at least three letters long
- Every square on the grid must be reachable from every other square without interference of a blank.

This paper does not treat British style crosswords or other, similar games. Thus when we refer to a crossword puzzle, we are speaking of a set of clues and a grid of correct responses.

Crossword puzzles have already attracted a good deal of research in the artificial intelligence literature. However, the majority of these treatments focus on puzzle construction (Mazlack, 1976; Ginsberg et al., 1990) or automatic puzzle solving (Ernandes et al., 2005; Goldschmidt and Krishnamoorthy, 2004; Littman et al., 2002, 1999). This paper is concerned with describing the

difficulty a solver will face when attacking a given puzzle. In particular, we will develop a model that allows us to quantify and predict the difficulty of a particular crossword puzzle clue.

3 Crossword Puzzles and Information

Information enters into crossword puzzle solving in several distinct ways. In the context of this paper, we categorize them as follows:

1. Clue information: Each clue provides information about the correct corresponding answer.
2. Structural information: As a solver completes the puzzle, his solutions in one direction yield information about answers in the other direction.
3. Extrinsic information: Crosswords require a great deal of cultural and linguistic expertise. Cultural references, wordplay, and knowledge of crossword puzzle conventions all come into play during a solution.

For the purposes of this paper, we do not explicitly consider what we have called extrinsic information. We touch on information of type 2 in Section 8. But we focus mainly on information of type 1: the bulk of our analysis treats “clue information." We leave detailed analysis of structural information to future research. However, structural information is readily quantifiable, so we begin our discussion by considering it, in efforts to lay the framework for theorizing clue information.

3.1 Structural Information in Crosswords

Consider the simple puzzle fragment in Figure 2. Since our clue is blank all we know is that the response contains three letters. How difficult is this puzzle?

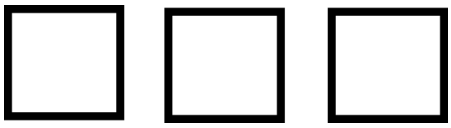


Figure 2: A Crossword Puzzle Fragment

One way to address the question of difficulty is to ask, what is the probability, given the information at hand, that we could correctly guess the solution?

The first edition of *The Official Scrabble Player’s Dictionary* (1978) contains 908 three-letter words[1]. If we assume (for purposes of exposition) that the answer comes from this dictionary and that each of these words is equally likely to be the correct answer (a naive, but not unreasonable assumption for a puzzle), we have a one in 908 chance of guessing the right answer.

[1] An electronic version of the dictionary is available online at http://www.dcs.shef.ac.uk/research/ilash/Moby/mwords.html. The electronic version was used in this paper.

Figure 3: A Crossword Puzzle Fragment

The uncertainty we face given this puzzle may be modeled by a quantity proportional to the probability of guessing the answer correctly. Throughout this paper, we use the term "uncertainty" to refer to the problem of guessing the value of a random variable. Let P be the event that we guess the puzzle's answer correctly:

$$Pr(P) = \frac{1}{908} = 0.0011 \tag{1}$$

But if we learn something about the correct answer our chances of solving the puzzle change. For instance, if we learn that the second letter of the solution is `R` as in Figure 3, our chances of success increase. The dictionary contains only 66 three-letter entries with an `R` in the middle position. Thus advancing from puzzle 2 to 3 reduced our uncertainty by a large margin:

$$Pr(P|P_2) = \frac{1}{66} = 0.015 \tag{2}$$

where p_2=*R* is the event that the second letter of the answer is an `R`. Learning that the middle letter is `R` reduced our uncertainty by a factor of $\frac{908}{66}$=13.76.

On the other hand, if before learning the middle letter, we learned that the first letter is `A` we find a different scenario. The dictionary contains 253 three-letter words starting with `A` so the conditional probability $Pr(P \mid p_1 = \texttt{A}) = \frac{1}{253} = 0.004$. This marks a large improvement over the uncertainty we face when we know none of the letters, but it is less informative than the knowledge of the second letter.

If we know that the first two letters are `A` and `R`, respectively, the dictionary contains only 16 possible solutions. As a solver completes a puzzle, his *across* answers reveal information about the correct solutions for the *down* clues and vice versa. This is akin the simplified scenario pursued here. Each supplied letter reduces the probability of guessing incorrectly. The important, though obvious, point for our discussion is that the reduction in uncertainty that accompanies gaining information about a ⟨*clue*→*answer*⟩ mapping makes guessing the answer easier. Given an unknown answer A and some knowledge about the solution K, we model the difficulty of

guessing $A=a$ as the conditional probability $Pr(A=a \mid K)$. The more K reduces our uncertainty about A, the easier it is to guess a correctly.

3.2 Clue Information

In the preceding section we tried to guess the correct response to a simplified puzzle using only structural information to reduce our uncertainty. However, crosswords offer additional aid to the solver in the form of clues. Thus we might find the simplified puzzle of Figure 4. How difficult is this puzzle?

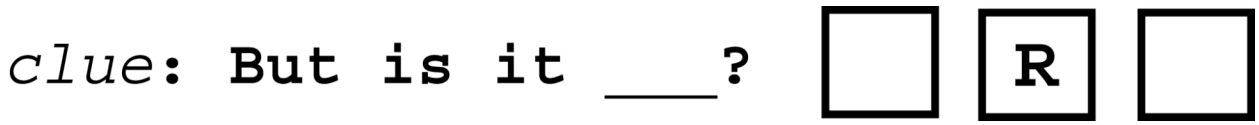


Figure 4: A Crossword Puzzle Fragment

The clue gives us information about the correct answer, but not in any form that is easy to measure. Intuitively, however, this type of information is quantifiable, insofar as certain clues are much easier to solve than others. The clue in Figure 4 is likely to be more informative to casual solvers than, say, `Klimt's kunst`, though this is of course debatable. In the remainder of this paper we pursue a model that attempts to formalize this intuition.

4 Modeling Puzzle Difficulty

As we described above, it is convenient to consider puzzle difficulty in terms of probabilities. Given a puzzle P comprised of a set of answers **A** and a set of clues **C**, what is the probability of guessing **A** given that **C** is known? This gives rise to our basic model of puzzle difficulty:

$$D(P) = Pr(\mathbf{A} \mid \mathbf{C}) \tag{3}$$

Intuitively, an easy puzzle should yield a higher probability of success than a hard puzzle.

To help us operationalize this model we begin by noting the conditional probability of answer a_i given clue c_i

$$Pr(a_i \mid c_i) = \frac{Pr(a_i,\ c_i)}{Pr(c_i)}. \tag{4}$$

If we assume that the clues are independent, we have the model

$$Pr(A \mid C) = \prod_{i \in P} \frac{Pr(a_i,\ c_i)}{Pr(c_i)}. \tag{5}$$

In other words, we model the difficulty of a puzzle P as the product of the conditional probabilities $Pr(a_i \mid c_i)$, the likelihood of guessing the i^{th} answer given the i^{th} clue.

It is worth addressing here a potential complaint about this model, namely that it takes no account of the likelihood of the answer, $Pr(a)$. Might it not be better to model the difficulty of a clue-answer pair as the mutual information between them, where our estimate would be conditioned on the marginal probability of the answer?

To motivate the lack of conditioning on $Pr(a)$ it helps to consider that a puzzle solver is given a clue and must produce the unknown answer. Thus there is a practical directionality to the difficulty of a clue-answer pair. Since in general $Pr(a|c) \neq Pr(c|a)$, the proposed model captures this directionality.

Furthermore, from a practical standpoint conditioning on the marginal probability of an answer is not consistent with the problems that face crossword solvers. Consider the following clue-answer pair:

`Neither ___ nor there.  →  HERE`

which is very easy. But the fact that the answer `HERE` is a common word has little to do with this ease. Consider:

`Villain to Shakespeare's Othello.  →  IAGO`

which (for anyone who remembers high school English) is also easy, although `IAGO` is a much rarer word than `HERE`.

Likewise, common words can prove difficult to guess, as this example shows:

`Can one defy exegesis?  →  CODE`

The word `CODE` is relatively common. But to solve this clue, one must realize that the clue itself is a code (not a very good one; the first letters of the clue spell the answer).

This example demonstrates a stark simplification in our theory: in this paper, difficulty relates only to the probability of generating a particular answer based on a clue. This model clearly imputes a rationality to puzzle solving that fails to include more nuanced workings of human language. For our model, what matters in assessing the difficulty of guessing an answer based on a clue is the conditional probability. How likely is it that a "statement" in the language at large that contains the clue also contains the answer?

4.1 Operationalizing the Model

To make our terms concrete we must first operationalize the notion of the language at large. To accomplish this we use a very large corpus of documents (in this case the World Wide Web) as a sample of the general language. With respect to the notion of a "statement" we choose the document as our unit of discourse. Other choices such as the paragraph or the sentence would of course yield different (probably better) results. But even a corpus as large as the Web will not

include enough data to derive adequate counts of most term co-occurrences at the level of smaller linguistic units.

To estimate $D(P)$, the difficulty of puzzle P, we begin with the maximum likelihood estimate

$$\hat{Pr}_{ml}(a \mid c) = \frac{|a,\ c|}{|c|} \tag{6}$$

where $|a, c|$ is the number of documents (in some corpus) in which answer a and clue c both appear. Likewise, $|c|$ is the number of documents in the corpus containing the clue. Equation 6 is the maximum likelihood (unbiased) estimate in a binomial model. From this, the definition of the maximum likelihood estimate $\hat{D}_{ml}(P)$ of the difficulty of puzzle P follows naturally: it is simply the product of the conditional probabilities of the i clue-answer pairs.

To obtain term count information, the experiments reported here relied on the Internet search engine Yahoo![2] , using their API to count the frequency of clues and co-occurrences of clue-answer pairs. This approach has the advantage that commercial search engines maintain corpora so large that we can obtain useful statistics, even on relatively rare phrases. The disadvantage of using a third-party database is the lack of transparency. When searching for multi-word phrases, there is some ambiguity about how the matching engine determines the number of hits[3].

From an algorithmic standpoint, we estimate the difficulty of a clue-answer pair with Equation 6, approximating the quantity $|x|$ by the number of results returned for a Yahoo! search for the words that comprise x.

5 Improving the Model

The maximum likelihood estimator described above is subject to several problems. Most obviously, we would be reluctant to assign 0 probability to a phrase, even if Yahoo! has no records that match it. After all, we have seen the phrase in the puzzle so it must have non-zero probability. With this in mind, we simply add 1 to both the joint and marginal frequencies, a simple form of Laplace smoothing.

Perhaps more problematic, though, is the maximum likelihood estimate's sensitivity to unimportant details of the clues and answers. Since the search engine ANDs terms together, longer clues will shrink the counts in our estimate, even if several of the search terms give little information about the nature of the clue-answer relationship. In other words, we suspect that the maximum likelihood estimator will display a (possibly) problematic variance.

---

[2] http://search.yahoo.com

[3] In this study multi-word clues were searched without any constraining quotation marks, allowing the search engine to return documents that matched parts of the clue. Many answers are also multiple words, though this fact is not obvious due to the necessary lack of spaces between words. Answers were handled in the following fashion. If they were not contained in the WordNet database (Fellbaum, 1998) we assumed that they were likely to be multiple words conflated together. We thus issued a spell-check query to Yahoo! . Final counts on these answers were obtained by averaging the results from a search with the raw answer and with the spell-checked version.

To mitigate this, we use Bayesian updating to improve our estimates. A Bayesian approach is logical here. Since each clue-answer pair appears in a puzzle (i.e. a collection of clues and answers), we should take into account the overall difficulty of the puzzle in efforts to update our estimates of the individual clue-answer difficulties.

In a Bayesian framework we begin by positing some prior belief about the parameter of interest's distribution. In this case our prior belief is, informally, that $Pr(a \mid c)$ is a small, but non-zero quantity. Later we will make this more explicit. For now, it suffices that when we encounter a maximum likelihood estimate of $Pr(a \mid c)$, we would like to condition it against our prior belief.

With this in mind we begin by noting that we can understand Equation 6 as an estimate of a binomial proportion. Let $i=|a, c|$. That is, i is the number of documents in which both the clue and the answer appear. Likewise let $n=|c|$, the number of documents containing the clue. Thus $Pr(a \mid c)$ is $\frac{i}{n}$. We thus have i successes in n trials, giving the maximum likelihood estimate for p, the probability of success in a binomial distribution.

We hope to derive the maximum a posteriori (MAP) estimate of the parameter p. Using Bayes' rule, we note that the posterior distribution of p given the data is

$$Pr(p \mid i,\ n) = \frac{\mathcal{L}(i \mid p,\ n)\ Pr(p)}{Pr(i \mid n)} \propto \mathcal{L}(i \mid p,\ n)\ Pr(p) \tag{7}$$

where $\mathcal{L}$ is the likelihood function. In other words, the posterior is proportional to the likelihood times the prior. We have all the information we need to write the binomial likelihood function in terms of the data

$$\mathcal{L}(i \mid p,\ n) = \binom{n}{i} p^i (1-p)^{n-i} \propto p^i (1-p)^{n-i}. \tag{8}$$

We ignore the binomial coefficient since it doesn't depend on p.

All that remains to find for our MAP estimate is the prior distribution. We shall model the prior distribution of the binomial parameter p using the beta distribution with hyperparameters α and β:

$$Pr(p) \propto p^{\alpha-1}(1-p)^{\beta-1} \tag{9}$$

We use the beta distribution because it is the conjugate prior of the binomial distribution that constitutes our maximum-likelihood model. That is, the beta distribution is of the same family as the binomial distribution. The hyperparameters α and β formalize our prior, so choosing them will be of utmost importance. To do them justice, we defer discussion of their selection until later.

Following Equation 7 we multiply Equations 8 and 9. Thus we have the posterior distribution of p

$$Pr(p \mid i,\ n) \propto p^{\alpha+i-1}(1-p)^{\beta+n-i-1} \tag{10}$$

which is the kernel of a new beta distribution. The maximum a posteriori estimate is thus $\arg\max_p \Pr(p \mid i,n)$. Differentiating Equation 10 and setting the derivative to zero, we find that the MAP estimate is given by

$$\arg\max_p Pr(p \mid i,\ n) = \frac{\alpha + i - 1}{\alpha + \beta + n - 2} \tag{11}$$

which gives us an estimate of the binomial parameter p, updated to include information about our prior belief and the data (Carlin and Louis, 1996).

6 Puzzle Data

To assess the model described here, a sample of 840 puzzles was obtained from *The New York Times*. This sample contained all daily puzzles (Monday through Saturday) published in 2004, 2005 and during the first nine months of 2006, with several puzzles removed due to file corruptions.

For each puzzle we used the Yahoo! search API to estimate frequency statistics for each clue. Based on these statistics, we obtained the maximum likelihood estimate of the difficulty of each clue. Following Equation 5 we multiply these estimates to derive an estimate of the puzzle's overall difficulty. However, this approach suffers one defect. Puzzles in the sample have different numbers of clues, varying from 52 to 80 with a mean of 73.74. The number of clues is only weakly related to the puzzle's difficulty; in fact having more clues implies shorter answers (and an easier puzzle) since all puzzles have a 15×15 grid. But because our maximum likelihood estimates are small, puzzles are "penalized" for having many clues. While in a strictly probabilistic sense, this is correct, it doesn't jibe with intuition or the reality of the puzzles. As such, we substitute the geometric mean of the clue estimates for their product when assessing puzzle difficulty:

$$\hat{D}_{ml} = \exp[\sum_i \frac{\log(\hat{d}^i_{ml})}{n}] \tag{12}$$

where $\hat{d}^i_{ml}$ is the ML estimate of the i^{th} clue's difficulty out of n clues in the puzzle.

The same process of finding the difficulty of a puzzle by the geometric mean of its clues was used for the Bayesian estimates described below.

6.1 Priors for the Puzzle Data

To model our prior belief about the distribution of p, the difficulty of a crossword puzzle, as described above we need to assign values to the beta distribution's shape parameters α and β. This decision involves two considerations. First we must decide on the shape of the distribution. Is it symmetrical? Is it skewed towards zero or maybe towards one? Second, we must decide

how strongly to account for our prior belief. A weak prior will cause the MAP estimate to converge on the ML estimate, while a strong prior weakens the influence of the data.

Choosing a shape for our prior distribution is fairly easy. We suspect (and will show later) that our observed binomial proportions skew to the left, falling near zero in the majority of cases. Thus we know that α should be less than β, probably by a large margin.

In the context of updating our estimate of a binomial proportion, a prior that is distributed *beta*(α,β) is equivalent to a data set with α-1 successes and β-1 failures. The mode of the data in Figure 7 appears to be close to 0.01 (the median is 0.011), and we would like our prior to reflect that. This desire leads us to choose α and β such that: $\frac{\alpha\text{-}1}{\beta\text{-}1}=0.01$.

We must also decide on the strength of our prior belief. Choosing α=2 and β=101 gives the same effective ratio as α=10001 and β=10000001. But putting these parameters into Equation 11 clearly gives different results. As the magnitude of the prior hyperparameters decreases, the data overwhelms our prior belief. Since the number of hits for most terms in our puzzles is large (on the order of a million), we need a very strong prior to move the MAP estimate even a small distance from the ML estimate.

With these considerations in mind, we calculated MAP estimates using four parameterizations of the prior, as shown in Table 1. Parameterizations *B*1 and *B*2 were both "subjective" in the sense that the values were chosen to conform with informally derived prior beliefs: both *B*1 and *B*2 correspond to an effective prior parameterization of p=0.01, but *B*2 is a much weaker prior than *B*1. On the other hand methods *B*3 and *B*4 used the method of empirical Bayes (Carlin and Lewis, 1996). For *B1*, α=39600 is the median count obtained for $|a, c|$ in our 840 puzzles. That is, it is the median occurrence of a document containing both the clue and the answer. Likewise β=1650000 is the median value for $|c|$, the frequency of the clue. Finally, *B*4 was also fitted using the median joint and marginal frequencies. However, in this case, the median was calculated at the puzzle level. In other words, a different prior parameterization was chosen for each puzzle based on its own median joint and marginal frequencies.

	Subjective		Empirical	
	B1	B2	B3	B4
α	100,000	10,000	39,600	local
β	10,000,000	1,000,000	1,650,000	local

Table 1: Hyper-parameters for Beta Prior Distribution

To help us understand the implications of the models proposed in this section, Figure 5 shows the estimated difficulty of Monday and Saturday puzzles in our sample. The *x*-axis of each panel in the figure is simply each puzzle's time of publication (1 is the first puzzle, i.e. from January 2004), while the *y*-axis shows the estimated difficulty under a given model. Ideally we would like

to see a clear separation for Monday and Saturday puzzles, with the probabilities derived from Monday puzzles significantly higher than Saturdays. Additionally, we would like the variance for a given day's estimates to be low; each week's Monday puzzle shouldn't be too much easier or harder than another Monday puzzle.

Two facts are apparent from Figure 5. First, the Monday puzzles appear to have much higher variance than the Saturday puzzles do. Second, the various hyperparameters of our Bayesian priors show a strong influence on the posterior difficulty estimates. Our strong, subjectively chosen prior *B*1 actually appears worse (both in terms of separation and in terms of variance) than the *ML* model. However, the other Bayesian estimates appear to mitigate the high variance of Monday's puzzles to a certain extent. Which among our four estimators is actually the best is the subject of the next section.

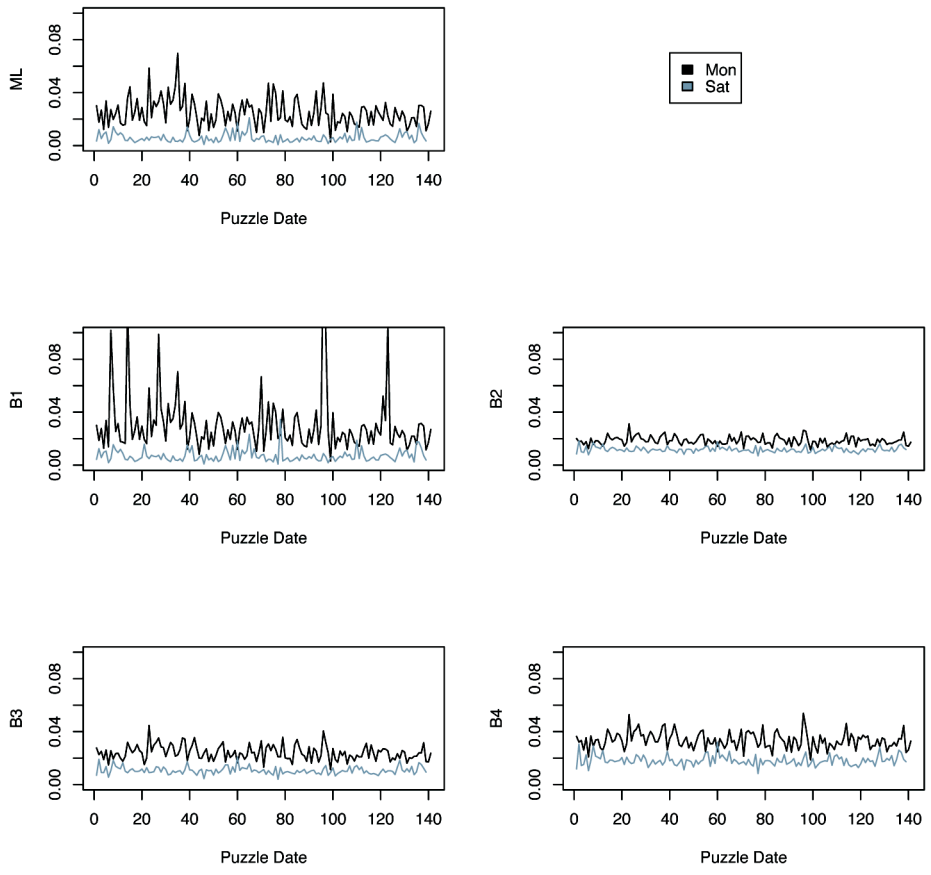


Figure 5: Estimated Difficulty for Monday and Saturday Puzzles

7 Predicting Puzzle Difficulty

To assess the model described in Section 4 we would like to be able to predict the difficulty of a puzzle P based on the estimator $\hat{D}(P)$. For the purposes of experimentation, we operationalize the

difficulty of a puzzle by the day of the week on which *The New York Times* published it. That is, we assume that Monday's puzzles are harder than Tuesdays', which are harder than Wednesdays', etc. The goal of the experiments reported in this section is to predict when a puzzle was published by analysis of its estimated difficulty.

We approach this problem as both a regression and as a classification problem. Our first experiment assigns a numeric value to each day of the week (*Mon*=1, *Tue*=2, ... , *Sat*=6). In this scenario, we wish to predict the day of the week for puzzle *P* based on $\hat{D}(P)$. In our second experiment, we approach prediction as a classification problem. In this case we bin the days of the week into three mutually exclusive categories: *easy, medium*, and *hard*. Again the goal is to predict the puzzle's category based on $\hat{D}(P)$.

7.1 Regression

Let *D* be the day on which a puzzle *P* was published, such that *Mon*=1, *Tue*=2, ... , *Sat*=6. We construct a linear model under the assumption that $D = \beta_0 + \beta_1 D(P) + \varepsilon$, where ε is distributed $N(0,\sigma)$. We seek the linear function $\hat{d}_i = f(\hat{D}_i)$ that minimizes the squared error $\sum_i (\hat{d}_i - d_i)^2$ over our sample. Using the standard least-squares estimators we built four models, one for each of difficulty estimators: the maximum likelihood estimate and the Bayesian estimates updated with the prior hyperparameters shown in Table 1.

	R^2	95% CI
ML	0.4232	(0.33, 0.506)
B1	0.2315	(0.193, 0.285)
B2	0.4836	(0.439, 0.5265)
B3	0.5782	(0.539, 0.616)
B4	0.5646	(0.524, 0.604)

Table 2: R^2 for Regression Models

Table 2 summarizes the results of our regression modeling. In terms of point estimates, model *B*1 was by far the worst, while model *B*3 was best. In addition to point estimates for R^2, Table 2 gives 95% confidence intervals for R^2 on each model. The confidence intervals were calculated via the bootstrap-*t* method described by Efron and Tibshirani (1993), using 1000 bootstrap samples. At the 95% confidence level, we judge that models *B*3 and *B*4 were significantly better than the maximum likelihood estimator. However, our handpicked priors fared poorly. Model *B*1 is significantly worse than *ML*, and model *B*2 appears slightly, but not significantly better than *ML*.

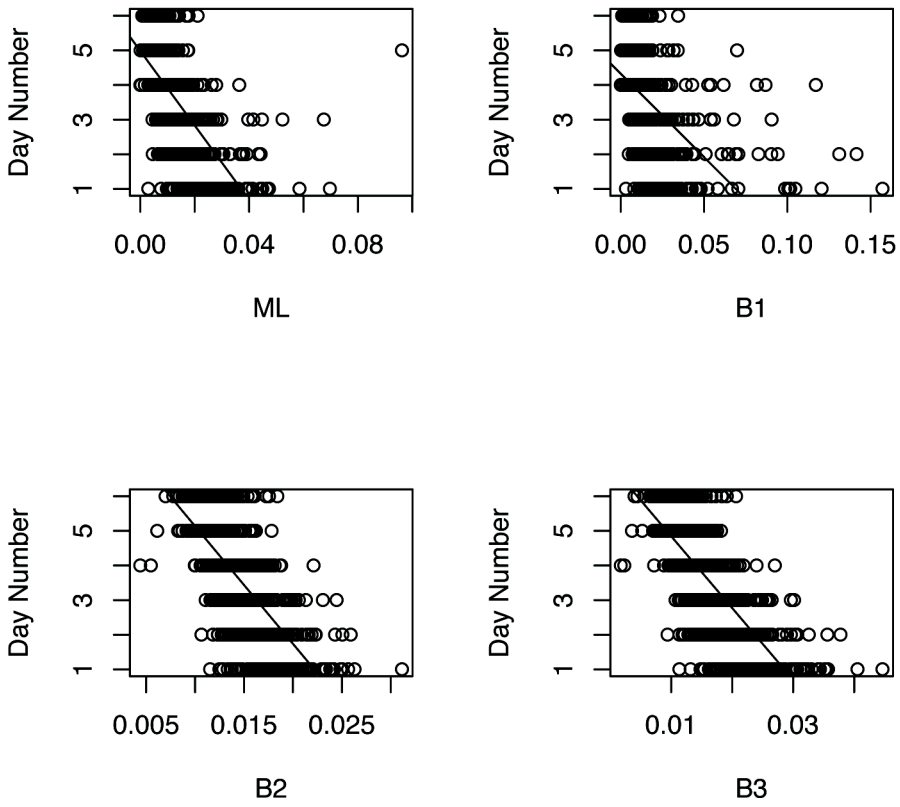


Figure 6: Linear Regressions on Puzzle Difficulty

Figure 6 plots the results of the linear regressions. Data for *B*4 is omitted from the figure due to the nearly identical performance of *B*3 and *B*4. The figure shows a given difficulty estimator on each *x*-axis and the day of the week (numerically coded) on the *y*-axis, along with the fitted regression line. For both *ML* and especially *B*1 the data exhibit so much variance that distinction between the days is difficult. Furthermore, both models *ML* and *B*1 suffer because when using them a high number of puzzles score near 0, a result that pulls the regression line downward. This leads to the high error evinced by each model's R^2. Models *B*2 and *B*3, on the other hand, show markedly less tendency to cluster puzzles near the origin. They also have lower variance. Though not pictured here, the same holds for *B*4.

These results suggest that Bayesian updating is appropriate in this setting, but that the prior parameters are of utmost importance. Especially interesting is that statistically significant improvements over the maximum likelihood estimator were seen for *B*3 and *B*4, the models with hyperparameters chosen by various empirical Bayes methods. The *B*2 model was statistically indistinguishable from *ML* while *B*1 incurred a significant loss of predictive power versus *ML*. Our empirical Bayes hyperparameters categorically outperformed (both in the regression and classification problems) those parameters chosen with more heuristic motivation.

Nontheless, several of the models performed quite well, considering the difficulty inherent in the problem. We will withold discussion of whether or not our results are "good" until after an alternative evaluation of our estimators' predictive power.

7.2 Classification

Instead of a regression, we may consider the problem of estimating puzzle difficulty as a classification challenge. For this experiment, we divided the week into three bins; Monday and Tuesday puzzles are considered *easy*, Wednesday and Thursday are *medium*, and Friday and Saturday are *hard*. Our goal here is to find a function $f(P)$ that assigns the puzzle P to the correct difficulty class.

Since this classifier operates in a single dimension, the problem amounts to finding two points on the number line that divide the interval [0,1] into three sections. We then assign puzzle P to whichever class corresponds to the region on [0,1] where $\hat{D}(P)$ lies.

Given the low dimensionality of the problem, the choice of classification method doesn't have a great impact on model performance. In the interest of simplicity all classification reported here entailed application of a simple naive Bayes classifier.

Table 3 shows the results of a 10-fold cross-validation run. The columns of the tables show the variant of $\hat{D}$ that was used for a particular model. Thus the first column (ML) shows results obtained by using the maximum likelihood estimator of D, while the second row (B1) uses our Bayesian estimate with the first set of hyperparameters on the prior. Rows of the table show the percentage of correctly classified puzzles for a given run, plus or minus the standard deviation of the correct percentage.

	(ML)	(B1)	(B2)	(B3)	(B3)
1	62.86±3.31	48.21±2.46	62.74±3.55	68.81±5.32 °	67.14±5.24
2	62.86±4.85	47.98±3.02 •	62.86±4.98	68.81±4.69 °	66.79±3.34
3	62.02±4.84	48.81±4.52 •	62.86±3.63	68.93±2.77 °	66.79±4.43
4	62.98±3.70	48.33±6.23 •	63.33±4.38	68.21±5.50 °	67.38±6.99
5	62.38±5.12	48.69±4.71 •	63.21±5.25	68.57±5.15 °	66.90±4.72 °
6	62.86±3.84	48.93±3.95 •	62.86±6.04	68.69±4.96 °	66.79±6.06 °
7	62.86±2.23	48.93±2.06 •	62.98±3.20	69.05±3.41 °	67.26±4.21
8	62.26±3.60	48.81±3.85 •	63.10±5.64	69.05±5.72 °	67.14±6.15 °
9	63.45±3.23	48.93±3.57 •	62.74±6.05	69.05±4.38 °	66.67±4.63
10	62.62±2.76	48.81±3.12 •	62.74±5.30	68.93±3.82 °	66.79±4.36

°, • statistically significant improvement or degradation, respectively

Table 3: Results from 10-fold Cross-Validation

Measures of statistical significance in Table 3 use the maximum likelihood estimator of D as a baseline; all improvements or degradations are with respect to maximum likelihood. The test used was the corrected resampled t-test (Witten and Frank, 2005, p. 157). Significant improvements or degradations imply that $p<0.05$.

Table 3 suggests that the hyperparameters of our first Bayesian model degrade predictive accuracy; using $B1$, all ten runs of the cross-validation led to decreased performance at 95% confidence. Our second set of hand-picked hyperparameters, $B2$, yield results that are statistically indistinguishable from the maximum likelihood estimates. However, the two empirical Bayes models, $B3$ and $B4$ show improvement over the non-Bayesian estimator. Model $B4$ outperformed the maximum likelihood estimator on three of the ten folds. Model $B3$ led to statistically significant improvements on all ten folds.

From Table 3 we may infer that of our tested models, using Bayesian updating with globally obtained empirical hyperparameterization improves classification accuracy over our raw ML estimator. The results of the classification roughly mirror those of our regression, suggesting not only the suitability of empirical Bayes for hyperparameterization, but also that the two methods of prediction, classification and regression, are describing the same underlying phenomenon.

Because the globally obtained empirical Bayes model $B3$ outperformed our other models decisively, the remainder of this discussion details only the performance of this particular model.

Predicted				
e	m	h		
190	78	14	easy	Actual
61	144	74	medium	
3	44	232	hard	

Table 4: Confusion Matrix from 10-Fold Cross-Validation

Table 4 shows the confusion matrix obtained from one iteration during cross-validation. As we would expect, few *easy* puzzles are misclassified as *hard* and vice versa. The majority of misclassifications involve the class *medium*. Over our 10-fold cross-validation, the F-measure (harmonic mean of precision and recall) for the *easy*, *medium*, and *hard*, classes was 0.709, 0.528, 0.775, respectively. In other words, puzzles of middling difficulty were the hardest to classify. This result reflects the ordinal nature of our three-class problem.

But perhaps more interestingly, our results raise the question of how high a level of accuracy is actually obtainable in the puzzle classification problem. We might ask, what level of accuracy with respect to day of publication could we expect from humans? We suspect that even human judges might disagree on whether certain puzzles are easy, medium or hard.

Without input from human judges it is difficult to address the question of how good our classification results actually are. However, we may approach the question of our estimator $\hat{D}$'s value obliquely by comparing its predictive power to the power of other features of crossword puzzles. To enable such analysis, we defined the set of 17 features shown in Table 5.

Info. Gain	Feature
0.7532	num. clues
0.6674	pct. ?'s
0.5935	B2
0.5715	B3
0.5553	ML
0.5497	B4
0.5448	pct. black
0.5224	pct. fill-in-the-blank
0.4699	B1
0.2866	pct. refers to another clue
0.187	pct. is word
0.1712	pct. 3-letter answers
0.0945	var(ML)
0.0908	var(B2)
0.0868	var(B3)
0.0419	var(B1)
0.0288	var(B2)

Table 5. Information Gain for17 Puzzle Features

Table 5 lists the features in decreasing order by their information gain on the 3-way classification variable (Mitchell, 1997, p.57)[4] . Prior to calculating each variable's information gain, the variables were discretized using the procedure outlined by Fayyad and Irani (1993).

Among our features, the number of clues that a puzzle contains gives the most information about a puzzle's difficulty (hard puzzles tend to have few clues with long answers). The next most informative feature is the percentage of clues that end with a question mark, a signal that a particular clue is difficult. However, after these two features, the ensemble of $\hat{D}$ estimators are the most informative among our sample. Models *B*2 and *B*3 provide nearly identical amounts of information on the class variable, with *ML* and the remaining Bayesian estimators trailing.

The last four rows of Table 5 show the information gain for the variance of each $\hat{D}$. While the Bayesian estimators tended to have very low (and uniform) variance across puzzles, it was

[4] The information gain *Gain*(*X*,*Y*) is the expected reduction in entropy of *X* if we know the outcome of *Y*.

surprising how little information the variance of *ML* carried, given its putatively informative profile on the puzzles we examined in Figure 5.

Table 5 suggests that the estimators of puzzle difficulty that we propose are strong indicators of a puzzle's actual level of difficulty.

8 Discussion

What is missing from our analysis is any reconciliation of two types of information we defined in Section 3, structural information and clue information. When solving a puzzle, we know at least two things: the number of letters in the correct response, and the clue intended to lead us to that response. To understand how information works in crossword puzzles, we must relate these information sources.

In fact, marrying clue information and structural information is not difficult. The relation follows naturally from the model we have proposed. Let us return to the example shown in Figure 4. Here we have a three-letter answer for the clue `"But is it ___?"`, with the correct response `ART`. Earlier we noted that our dictionary contains 908 three-letter words. However, armed with clue information, our chances of guessing the right answer must be better than $\frac{1}{908}$. The question is how much the clue improves our chances of success.

To answer this question, we consider that our structural knowledge — the answer is three letters — constrains the answer space to one of 908 words. However, this structural knowledge gives no information about the distribution across this constrained space; each of the candidate words has a $\frac{1}{908}$ chance of being the right answer.

Clue information, on the other hand, gives us a non-uniform distribution for candidate answers. We may understand the significance of this change in information-theoretic terms. Since the uniform distribution has the highest entropy of all distributions on a given interval (Cover and Thomas, 1991, p. 27), the clue-induced posterior must reduce our uncertainty about the correct answer.

As an example, consider a simplified scenario. Let us say that the answer to a particular question is very long; our dictionary only contains 10 possible solutions. With no extra knowledge, our chances of guessing correctly are $\frac{1}{10}$. We thus consider the random variable A whose entropy is

$$H(A) = -\frac{1}{10}\sum_{1}^{10}\log\frac{1}{10} = 3.32 \tag{13}$$

But if we learn that A follows a non-uniform distribution over its 10 possible outcomes, its entropy will change. This is how clues inform the solution process. As we have seen, given a particular clue, some answers are, linguistically speaking, more likely than others. Thus clues give information on the distribution over the possible outcomes of A.

$Pr(a_1 \mid c)$	$Pr(a_2 \mid c)$	$Pr(a_3 \mid c)$	$Pr(a_4 \mid c)$	$Pr(a_5 \mid c)$	$Pr(a_6 \mid c)$	$Pr(a_7 \mid c)$	$Pr(a_8 \mid c)$	$Pr(a_9 \mid c)$	$Pr(a_{10} \mid c)$
0.7	0.01	0.01	0.22	0.01	0.01	0.01	0.01	0.01	0.01

Table 6: Probabilities for a Clue-Answer Pair

Continuing our example, let c be the clue that accompanies A. Further, let the conditional probabilities $Pr(A=a|c)$ be given in Table 6. We see that two possible answers, a_1 and a_4 are more likely than the others. Using the conditional probabilities of Table 6, we calculate the specific conditional entropy of A given $C=c$

$$H(A|c) = -\sum_{1}^{10} Pr(a_i|c) \log Pr(a_i|c) = 1.37 \tag{14}$$

By considering clue information we have gone from 3.32 bits to 1.37 bits, a 40% reduction in uncertainty. From our clue, we learn about the distribution of A, gaining knowledge that reduces the uncertainty we face when guessing the answer in the puzzle.

Let us consider a more realistic example. *The New York Times* crossword of Monday, January 2, 2006 contains the clue-answer pair ⟨*WILD→FERAL*⟩. The Scrabble player's dictionary contains 8258 five-letter words. If we lacked the clue `Wild` and were armed only with our Scrabble dictionary, the entropy of this problem is, rounded to two digits, $H(FERAL) = -\frac{1}{8258}\sum \log\frac{1}{8258} = 13.$

But because we have the clue, our uncertainty is lower than 13 bits. As we have seen, structural information and clue information improve our chances of guessing a correct answer in different ways. Structural information about a clue parameterizes a uniform distribution. If we know the answer is five letters, we have a uniform distribution over 8258 outcomes. If we learn that the third letter is `R` the cardinality of the outcome space changes (there are 1160 possibilities), but the distribution is still uniform.

On the other hand, clue information alters the probability distribution over the outcomes of the random variable A. It informs us that certain answers a_i are more likely than other a_j answers. Insofar as the clue-conditional distribution of A departs from uniformity, the clue reduces our uncertainty about the correct answer.

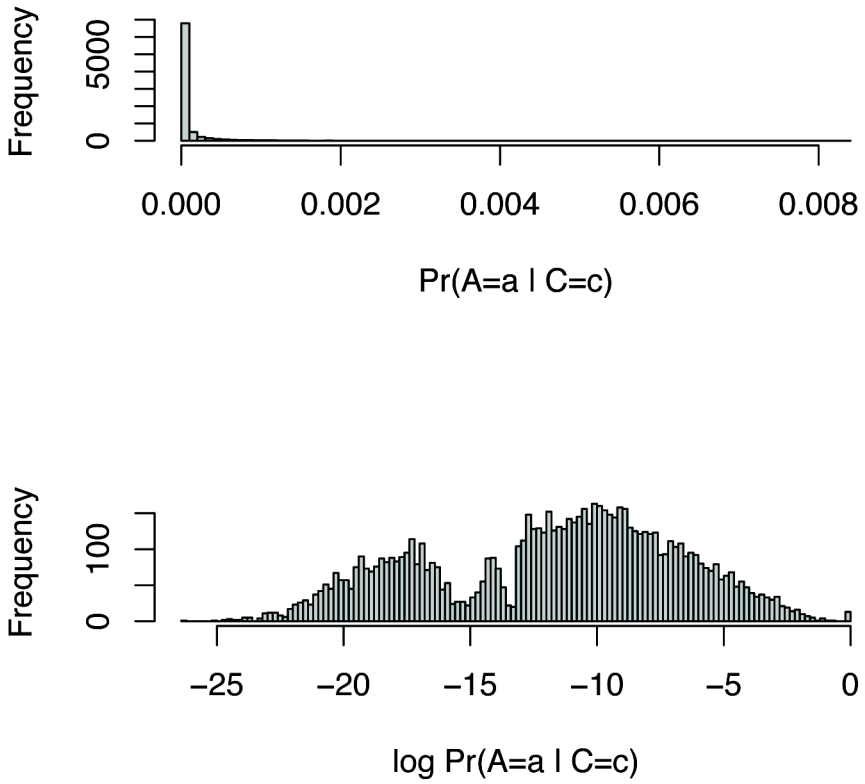


Figure 7: Conditional Distribution of 5-Letter Answers Given Clue `Wild`

Figure 7 shows the empirical distribution of $Pr(A|C=c)$. That is, for each of the 8258 possible five-letter answers a, we computed the conditional probability $Pr(a|c)$ as in our difficulty model (using the simple maximum likelihood estimate). From the figure it is clear that all candidate five-letter answers are not equally probable, given the clue `Wild`. The vast majority have near zero probability, with a few words having high probability. Thus the uniform distribution that we assume without clue information (i.e. based on the puzzle's structure) constitutes a poor model for choosing an answer to this question.

Precisely how informative a particular clue is can be understood as the Kullback-Leibler divergence between the uniform distribution $Pr(A)$ and the clue-conditioned $Pr(A|c)$. Let P and Q be two discrete distributions. The KL divergence is

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}. \tag{15}$$

If we take the logarithm to the base 2, the KL divergence gives the number of bits we gain by using P instead of Q. In this case we are interested in measuring the information gained by using clue-conditional information over uniform, structural information. Using the convention that $0\log\frac{0}{q}=0$, we compute the KL divergence $D_{KL}(\Pr(A \mid C = c) \| \Pr(A)) = 2.73$. Thus by going from our uniform distribution (induced by knowledge that the answer is five letters) to the distribution obtained from our clue, we have reduced our uncertainty by 2.73 bits.

This fact is readily observable if we note that the entropy $H(u)$ of a uniform distribution with n outcomes is simply $\log(n)$. Thus $H(A)$ in this case, rounding to two decimal places, is $\log 8258=13$. Meanwhile, the entropy of our $H(\Pr(A \mid C = c)) = -\sum_i \Pr(a_i \mid c)\log\Pr(a_i \mid c) = 10.27$. Subtracting $H(Pr(A|C=c))$ from $H(A)$ we have 13-10.27=2.73.

The relationship between structural information and clue information in crosswords can thus be understood as a difference in probability models for the answer space of a clue-answer pair. Knowing that the answer is five letters long gives us a uniform distribution over 8258 outcomes (using our Scrabble dictionary). If we solve part of the puzzle and learn that the first letter of the solution is `F` and the last letter is `L`, our chances of guessing correctly rise over the initial state, but all remaining candidate solutions are still equally probable. Learning that the clue for this solution is `Wild` allows us to revise the estimated probability of each of the n candidate answers according to the model proposed in Section 4. We may thus say that a particular clue is informative with respect to an answer to the extent to which the clue-induced distribution diverges from the initial uniform distribution obtained by the structural information we've gleaned.

9 Conclusion

Given a puzzle P that contains the clue set **C** and the answers **A**, we have argued that the difficulty of completing P can be understood probabilistically. The difficulty of a particular $\langle clue \rightarrow answer \rangle$ mapping, we argue, depends on the extent to which the answer and the clue co-occur in the language at large. Furthermore, the difficulty of P is modeled here as the product of its independent clues' difficulties. This intuitive model performed well in the experiments reported in Section 7.

However, a number of unresolved issues raised here demand further research. In particular, the structure of a crossword puzzle guarantees that its constituent answers are not independent. A solver need only complete half of the answers (i.e. all those answers in one direction) to solve the puzzle. This assumption surfaces here in our need to predict difficulty at the puzzle level via the geometric mean instead of the simple product of a puzzle's clue difficulties.

It is unclear whether our assumption of clue independence impedes the predictive power of the model. Perhaps if we wish to estimate the difficulty of a puzzle, assuming independence is a

tolerable simplifying assumption. But if we ask more probing questions, this model may prove inadequate.

In particular, in future work I hope to pose the question, how much information is in a particular puzzle? Information theory provides a framework for deriving a sensible answer to such a question, but such a derivation will demand a model that accounts explicitly for puzzle redundancy (in a fashion more complete than our discussion of structural information presented here).

Nonetheless, the work presented here shows promise in the context of the larger endeavor to develop techniques for modeling non-denotational aspects of textual meaning. This paper argues that in crosswords, difficulty operates as a function of the conditional distribution between two types of information–clues and answers. It will be an interesting challenge to generalize this argument to documents whose structure obeys different rules than crossword puzzles.

References

The Official Scrabble Players Dictionary, first Edition. (1978). Merriam-Webster, Springfield, MA.

Carlin, Bradley. and Louis, Thomas. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, New York.

Cover, Thomas M. and Thomas, Joy. A. (1991). *Elements of Information Theory*. Wiley-Interscience, New York.

Efron, Bradley. and Tibshirani, Robert. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Efron, Miles. (2006). Using cocitation information to estimate political orientation in web documents. *Knowl. Inf. Syst.*, 9(4):492–511.

Ernandes, Marco, Angelini, Giovani, and Gori, Marco (2005). Webcrow: A web-based system for crossword solving. In *Proceedings of the 20th National Conference on Artificial Intelligence*. July 9-13, 2005, Pittsburgh, PA. Cambridge: MIT Press. pages 1412–1417.

Fayyad, Usama M. and Irani, Keki B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Aritifical Intelligence*. Chambery, France. August 28-September 3,1993. Morgan Kaufmann, pages 1022–1029.

Fellbaum, Christiane, Ed. (1998). WordNet: An Electronic Lexical Database. Cambridge: MIT Press.

Ginsberg, M. L., Frank, M., Halpin, M. P., and Torrance, M. C. (1990). Search lessons learned from crossword puzzles. In *Proceedings of the 8th National Conference on Artificial Intelligence*. Boston, MA, July 29-August 3, 1990. Cambridge: MIT Press, pages 210–215.

Goldschmidt, David E. and Krishnamoorthy, Mukkai S. (2004). Solving crossword puzzles via the google api. In *Proceedings of the IADIS International Conference WWW/Internet*. Madrid, Spain. IADIS, pages 382–389.

Littman, Michael L., Keim, Greg A., and Shazeer, Noam M. (1999). Solving crosswords with proverb. In *Proceedings of the 16th National Conferrence on Artificial Intelligence and 11th Conference on Innovative Applications of Artificial Intelligence,* July 18-22, 1999, Orlando, Florida. Cambridge: MIT Press, pages 914–915.

Littman, Michael L., Keim, Greg. A., and Shazeer, Noam M. (2002). A probabilistic approach to solving crossword puzzles. *Artif. Intell.*, 134(1-2):23–55.

Mazlack, Lawrence. J. (1976). Computer construction of crossword puzzles using precedence relationships. *Artif. Intell.*, 7(1):1–19.

Mitchell, Thomas. (1997). *Machine Learning*. McGraw Hill.

Newman, Stanley and Lasswell, Mark. (2006). *Cruciverbalism: A Crossword Fanatic's Guide to Life in the Grid*. Collins, New York.

Pang, Bo and Lee, Lillian. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,* 25-30 June 2005, Ann Arbor, Michigan. The Association for Computational Linguistics, pages 115–124.

Turney, Peter D. and Littman, Michael L. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus. In *National Research Council, Institute for Information Technology, Technical Report ERB-1094*. National Research Council Canada. Retrieved online January 14, 2008 http://arxiv.org/pdf/cs.LG/0212012 .

Witten, Ian and Frank, Eibe. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann.