

# What Crossword Puzzles Teach us about Information

Miles Efron

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

January 19, 2007

## Abstract

This paper studies crossword puzzles as a vehicle for analyzing information in a rigorous yet meaningful fashion. The paper asks, how does information operate in the context 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  $P(a | c)$ ; easier mappings should enjoy a higher conditional probability. The model is tested by a corpus 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? Estimating crossword puzzle difficulty, we argue, involves quantifying how much information the puzzle offers the solver. The paper formalizes a model that describes how informative a puzzle's clues are with respect to its solution.

Due to their constrained structure, crossword puzzles provide 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 contributes to the discussion of what constitutes information insofar as it examines the relationship between probabilistic and semantic notions of information.

## 2 Crossword Puzzles and Information

Crossword puzzles are games comprised of a set of clues and an  $n \times n$  grid that contains answers to the 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.

<sup>1</sup> A	<sup>2</sup> R	<sup>3</sup> T
R	U	E
E	N	D

**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



Figure 2: A Crossword Puzzle Fragment

Crossword puzzles have 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.

Information enters into crossword puzzle solving in several distinct ways. In the context of this paper, we focus on two types of information. First we consider “structural information,” information communicated by the logical constraints of a puzzle. Second, we take up “clue information,” the verbal clues that guide the solver. So-called “clue information” is recognizable as information in the vernacular sense. On the other hand, “structural information” is readily quantifiable using results from information theory. The goal of this paper is to relate these two types of information formally.

## 2.1 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? 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* scr (1978) contains 908 three-letter words<sup>1</sup>. If we assume that the answer comes from this dictionary and that each of these words is equally likely to be the correct answer, we have a one in 908 chance of guessing the right answer.

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, our chances of success increase. The dictionary contains only 66 three-letter entries with an R in the middle position. Thus learning the second letter reduced our uncertainty by a large margin:  $Pr(P | p_2 = R) = \frac{1}{66}$  where  $p_2 = R$  is the event that the second letter of the answer is an R. Learning that the middle letter is R reduces our uncertainty by a factor of  $\frac{908}{66} = 13.76$ .

As a solver completes a puzzle his *across* answers reveal information about the correct solutions for the *down* clues and vice versa. The important, though obvious, point for our discussion is that the reduction in uncertainty that accompanies gaining information about a  $\langle \text{clue} \rightarrow \text{answer} \rangle$  mapping makes guessing the answer easier. Given an unknown answer  $A$  and some knowledge about its solution  $K$ , we model the difficulty

<sup>1</sup>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

clue : But is it ...?

Figure 3: A Crossword Puzzle Fragment

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

## 2.2 Clue Information

Besides structural information, crosswords offer additional aid to the solver in the form of clues. Thus we might find the simplified puzzle of Figure 3. How difficult is this puzzle?

The clue gives us information about the correct answer, but not in any form that is easy to measure. Yet it is still the case that some clues are easier to solve than others. The clue in Figure 3 is likely to be more informative to casual solvers than, say, *Klimt's kunst*, though this is of course debatable. The following section formalizes this intuition.

## 3 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  $\mathbf{A}$  and a set of clues  $\mathbf{C}$ , what is the probability of guessing  $\mathbf{A}$  given that  $\mathbf{C}$  is known? This gives rise to our basic model of puzzle difficulty:

$$D(P) = Pr(\mathbf{A} | \mathbf{C}). \tag{1}$$

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  $c_i$

$$Pr(a_i | c_i) = \frac{Pr(a_i, c_i)}{Pr(c_i)}. \tag{2}$$

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

$$Pr(A | C) = \prod_{i \in P} \frac{Pr(a_i, c_i)}{Pr(c_i)}. \tag{3}$$

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

### 3.1 Operationalizing the Model

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

$$\hat{Pr}_{ml}(a | c) = \frac{|a, c|}{|c|} \quad (4)$$

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. 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!<sup>2</sup>, using their API to count the frequency of clues and co-occurrences of clue-answer pairs<sup>3</sup>. From an algorithmic standpoint, we estimate the difficulty of a clue-answer pair with Equation 4, approximating the quantity  $|x|$  by the number of results returned for a Yahoo! search for the words that comprise  $x$ .

## 4 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've 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, since the search engine ANDs terms together, longer clues will lead to low probabilities regardless of difficulty. To mitigate this, we use Bayesian updating to improve our estimates. In this case our prior belief is, informally, that  $Pr(a | c)$  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  $Pr(a | c)$ , we would like to condition it against our prior belief.

With this in mind we begin by noting that we can understand Equation 4 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  $Pr(a | c)$  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

$$Pr(p | i, n) = \frac{\mathcal{L}(i | p, n) Pr(p)}{Pr(i | n)} \propto \mathcal{L}(i | p, n) Pr(p) \quad (5)$$

---

<sup>2</sup><http://search.yahoo.com>

<sup>3</sup>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 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

where  $\mathcal{L}$  is the likelihood function. Thus we have all the information we need to write the binomial likelihood function in terms of the data

$$\mathcal{L}(i | p, n) = \binom{n}{i} p^i (1-p)^{n-i} \propto p^i (1-p)^{n-i}. \quad (6)$$

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  $\alpha$  and  $\beta$ :

$$Pr(p) \propto p^{\alpha-1} (1-p)^{\beta-1}. \quad (7)$$

The hyperparameters  $\alpha$  and  $\beta$  formalize our prior, so choosing them will be of utmost importance. To do them justice, we defer discussion of their selection until later.

To find the posterior distribution of  $p$  we multiply Equations 6 and 7:

$$Pr(p | i, n) \propto p^{\alpha+i-1} (1-p)^{\beta+n-i-1} \quad (8)$$

which is the kernel of a new beta distribution. The maximum a posteriori estimate is thus  $\arg \max_p Pr(p | i, n)$ . Differentiating Equation 8 and setting the derivative to zero, we find that the MAP estimate is given by

$$\arg \max_p Pr(p | i, n) = \frac{\alpha + i - 1}{\alpha + \beta + n - 2} \quad (9)$$

which gives us an estimate of the binomial parameter  $p$ , updated to include information about our prior belief and the data.

## 5 Testing the Model

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 3 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. Because our maximum likelihood estimates are small, puzzles are "penalized" for having many clues. As such, we substitute the geometric mean of the clue estimates for their product when assessing puzzle difficulty:

	Subjective		Empirical	
	B1	B2	B3	B4
$\alpha$	100000	10000	39600	local
$\beta$	10000000	1000000	1650000	local

Table 1: Hyper-parameters for Beta Prior Distribution

$$\hat{D}_{ml} = \exp\left[\sum_i \frac{\log(\hat{d}_{ml}^i)}{n}\right] \quad (10)$$

where  $\hat{d}_{ml}^i$  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.

## 5.1 Priors for the Puzzle Data

To model our prior belief about the distribution of  $p$ , the difficulty of a crossword puzzle, we need to assign values to the beta distribution’s parameters  $\alpha$  and  $\beta$ . Choosing a shape for our prior distribution is fairly easy. We suspect that probabilities will be skewed left, falling near zero in the majority of cases. Thus  $\alpha$  should be less than  $\beta$ , probably by a large margin.

We must also decide on the strength of 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  $B1$  and  $B2$  were both “subjective” in the sense that the values were chosen to conform with informally derived prior beliefs. On the other hand methods  $B3$  and  $B4$  used the method of empirical Bayes. For  $E1$ ,  $\alpha = 39600$  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  $\beta = 1650000$  is the median value for  $|c|$ , the frequency of the clue. Finally,  $B4$  was fitted using the medians calculated at the puzzle level; a different parameterization was chosen for each puzzle based on its own median joint and marginal frequencies.

## 6 Predicting Puzzle Difficulty

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. We assume that Monday’s puzzles are harder than Tuesdays, which are harder than Wednesday’s, etc. The goal of the experiment reported in this section is to predict when a puzzle was published by analysis of its estimated difficulty.

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. Given the low dimensionality of the problem (the classifier is 1-dimensional), the choice of classification method doesn’t have a great impact on model performance. In the interest of simplicity, all classification reported here entailed a simple naive Bayes classifier.

Ten-fold cross-validation suggests that our model of puzzle difficulty works well, but that we must exercise care in choosing Bayesian priors. The maximum likelihood (ML) model yielded an average of 62.715% accuracy across the three classes. The Bayesian model B1 was significantly worse than ML at 95% confidence with respect to classification accuracy. On the other hand B3 was significantly more accurate than ML with 95% confidence, reaching 68.94% accuracy on average. Models B2 and B4 performed similarly to ML, with B4 significantly outperforming ML on three of the ten folds.

These results suggest that among our tested models, using globally obtained empirical estimates for hyperparameterization of our Bayesian prior improves classification accuracy over our raw ML estimator. 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		Actual
190	78	14	easy	
61	144	74	medium	
3	44	232	hard	

Table 2: Confusion Matrix from 10-Fold Cross-Validation

Table 2 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. 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, a result that reflects the ordinal nature of our three-class problem.

## 7 Discussion

What is missing from our analysis is any reconciliation of the two types of information we defined in Section 2, 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 3. 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 does the clue improve our chances of success?

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 space; each of the candidate words has a  $\frac{1}{908}$  chance of being right.

Clue information, on the other hand, gives us a non-uniform distribution over the 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.

For example, the New York Times crossword of Monday, January 2, 2006 contains the clue-answer pair  $\langle \text{Wild} \rightarrow \text{FERAL} \rangle$ . 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,  $H(\text{FERAL}) = -\frac{1}{8258} \sum \log \frac{1}{8258} = 13$ .

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.

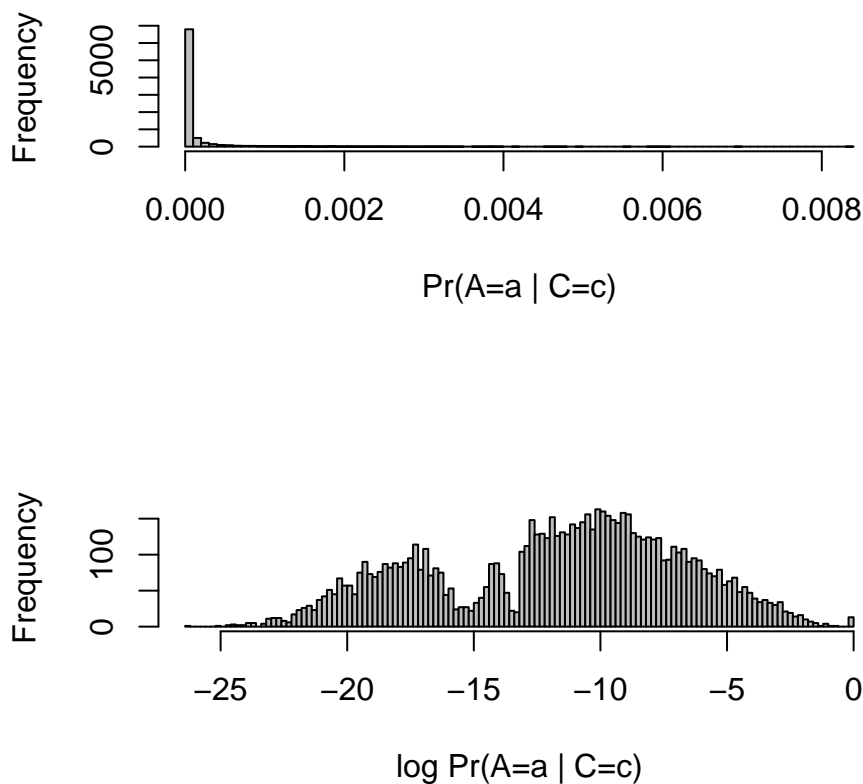


Figure 4: Conditional Distribution of 5-letter Answers given Clue Wild

Figure 4 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

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}. \quad (11)$$

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  $D_{KL}(Pr(A|C = c)||Pr(A)) = 2.73$ . 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  $H(Pr(A|C = c)) = -\sum_i Pr(a_i|c) \log Pr(a_i|c) = 10.27$ . 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 3. 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.

## 8 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  $\langle \text{clue} \rightarrow \text{answer} \rangle$  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 experiment reported in Section 6.

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 complete puzzle. 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 a more probing question, 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 joint 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

- (1978). *The Official Scrabble Players Dictionary, first Edition*. Merriam-Webster, Springfield, MA.
- Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley-Interscience, New York.
- Ernandes, M., Angelini, G., and Gori, M. (2005). Webcrow: A web-based system for crossword solving. In *AAAI*, pages 1412–1417.
- Ginsberg, M. L., Frank, M., Halpin, M. P., and Torrance, M. C. (1990). Search lessons learned from crossword puzzles. In *AAAI*, pages 210–215.
- Goldschmidt, D. E. and Krishnamoorthy, M. S. (2004). Solving crossword puzzles via the google api. In *ICWI*, pages 382–389.
- Littman, M. L., Keim, G. A., and Shazeer, N. M. (1999). Solving crosswords with proverb. In *AAAI/IAAI*, pages 914–915.
- Littman, M. L., Keim, G. A., and Shazeer, N. M. (2002). A probabilistic approach to solving crossword puzzles. *Artif. Intell.*, 134(1-2):23–55.
- Mazlack, L. J. (1976). Computer construction of crossword puzzles using precedence relationships. *Artif. Intell.*, 7(1):1–19.