AI Magazine Volume 29 Number 1 (2008) (© AAAI)

A Web-Based Agent Challenges Human Experts on Crosswords

Marco Ernandes, Giovanni Angelini, Marco Gori

■ *Crosswords are very popular and represent a useful domain of investigation for modern artificial intelligence. In contrast to solving other celebrated games (such as chess), cracking crosswords requires a paradigm shift towards the ability to handle tasks for which humans require extensive semantic knowledge. This article introduces WebCrow, an automatic crossword solver in which the needed knowledge is mined from the web: clues are solved primarily by accessing the web through search engines and applying natural language processing techniques. In competitions at the European Conference on Artificial Intelligence (ECAI) in 2006 and other conferences this web-based approach enabled WebCrow to outperform its human challengers. Just as chess was once called "the Drosophila of artificial intelligence," we believe that crossword systems can be useful Drosophila of web-based agents.*

Games and puzzles reproduce the complexity of the real world with "the smallest initial structures" (Minsky 1968), thus making their study important for both theoretical and practical issues. Since the birth of artificial intelligence (AI), games and puzzles have received much attention. The game that has captured most of the attention of computer scientists is chess. The founding fathers of AI such as McCarthy, Simon (Simon and Schaeffer 1992), Samuel, Shannon (Shannon 1950), Turing, and Von Neumann have all been involved in automatic chess playing. After decades of unsuccessful attempts (Mittman and Newborn 1980, Munakata 1996), the IBM machine Deep Blue achieved the astonishing result of defeating world champion Gary Kasparov in May 1997 (Campbell, Hoane, and Hsu 2002).

Games play the role of a laboratory where machines can safely be tested by a direct competition with humans. Along with the development of successful game-playing programs, the investigation should also include a methodological discussion on how this performance is achieved. Deep Blue heavily relied on computational power joined with a search algorithm based on static evaluation functions to assess the different configurations of the game. Instead of relying on a preprogrammed approach, Tesauro conceived TD-

gammon (Tesauro 1994), a strong automatic backgammon player that exploits the idea of reinforcement learning. These studies on TD-gammmon had a significant impact in other fields and application domains.

An important dichotomy exists between closed-world and open-world games. In the first case, the agents have to select their legal actions from a finite set, whereas in the latter the possible legal actions are not clearly limited, which seems to be more common in most real-world problems. The shift from closed to open-world games corresponds with a shift from a computational to a cognitive framework of complexity, namely from nondeterministic polynomial time (NP) (Cook 1983) to AI-completeness (Shapiro 1992). In most closed-world games, intelligent agents based on search algorithms reach or even overtake human masters (for example, chess, checkers, and Othello), even though there are remarkable exceptions like the Asian board game Go (Bouzy and Cazenave 2001, Gelly and Silver 2007). Among others, linguistic games are ideal to appreciate the border of closed-world games because of their natural grouping into word and language games (Littman 2000). Scrabble and Boggle are popular word games in which letters and words are merely used as symbols of a cryptanalytic problem in which no semantics is involved. Computers already display superhuman performance for this class of problems (see, for example, Scrabble [Sheppard 2002]). On the contrary, the solution of language games, such as Trivial Pursuit or ordinary crosswords, deeply depends on the meaning of the words involved. In principle, the answers to crossword clues are in a finite dictionary, but the number of different concepts that can be expressed by a clue is not clearly limited, and therefore searching the solution involves semantics. Some clues can be answered rather directly, like <*Cutter or clipper*[4]: ship>, while others are formulated in ambiguous ways, making it more challenging to find the solution, as in <*Bottle filler*[4]: ship>. Answers to clues also include compound words, phrases, neologisms, abbreviations, acronyms, name initials, or some kind of semantic or linguistic metamorphosis (for example, <*Tic-tac-toe loser*[3]: xxo>). Moreover, a crossword solver must exploit the constraints of the puzzle to find a smart strategy on how to fill the grid. This involves the identification of the clues to be answered first for a gradual uncovering of the others. The experiential, linguistic, and encyclopedic knowledge used by humans to crack crosswords clearly indicates that devising intelligent agents for such a task requires a paradigm shift towards the ability to handle semantics, in contrast to other celebrated games (for example, chess).

To the best of our knowledge, the Proverb project (Littman, Keim, and Shazeer 2002; Keim et al. 1999) is the only significant attempt to address the problem of crossword solving. For each clue a list of candidate solutions was generated, and subsequently, the "best" fill of the grid was produced on the basis of a model similar to belief network inference. Proverb made use of a very large crossword database as well as several expert modules, each mainly based on domain-specific databases (for example, movies, writers, and geography). In addition, there were generic-word list generators and clue-specific modules, for example, addressing fill-in-the-blank problems, like <*A Rose for* ___ *Maria*[3]: ana>.

Proverb work included a foray into web mining, but the approaches used were almost entirely undocumented. Web access is at the heart of WebCrow.[1] It relies strongly neither on large crossword databases nor on many domain-specific expert modules but primarily on an intelligent agent that operates online on the multilingual web repository of information. Interestingly, the dynamic evolution of the web affects the program behavior through time, thus reacting automatically to the evolution of human culture. Moreover, because of its web-centered architecture, WebCrow is better suited for dealing with different languages. WebCrow's core feature is in fact the web search module (WSM; see figure 2) that encompasses a special form of web-based question answering. The module relies on a nucleus of information-retrieval techniques including the search engine service and on natural language processing (Aravind 1991, Manning and Schütze 1999). Instead of relying mainly on expert modules, the aim of the WSM is to deal with the encyclopedic knowledge required during the clue-answering process, thus emphasizing the role of question answering, which is a key point in many language games such as Trivial Pursuit or Jeopardy. The idea of using of a web-based knowledge agent is also adopted in Lam et al. (2003), which tackles the quiz-show game "Who wants to be a Millionaire?" In that game the agent has to select one answer out of a small number of possible choices, a feature that moderates the complexity of the task. Despite displaying poor results on "easy for humans" questions such as "How many legs does a fish have?" the machine is still able to perform at average human level with as many as 70 percent correct answers. While the web is of fundamental importance for the answering process, it is of little help for covering and organizing "obvious everyday life notions" such as "men have two ankles" or "fish have no legs." In fact, despite its size, the web typically fails to cover these basic concepts, simply because the web community is made of humans that do not usually exchange trivial concepts.

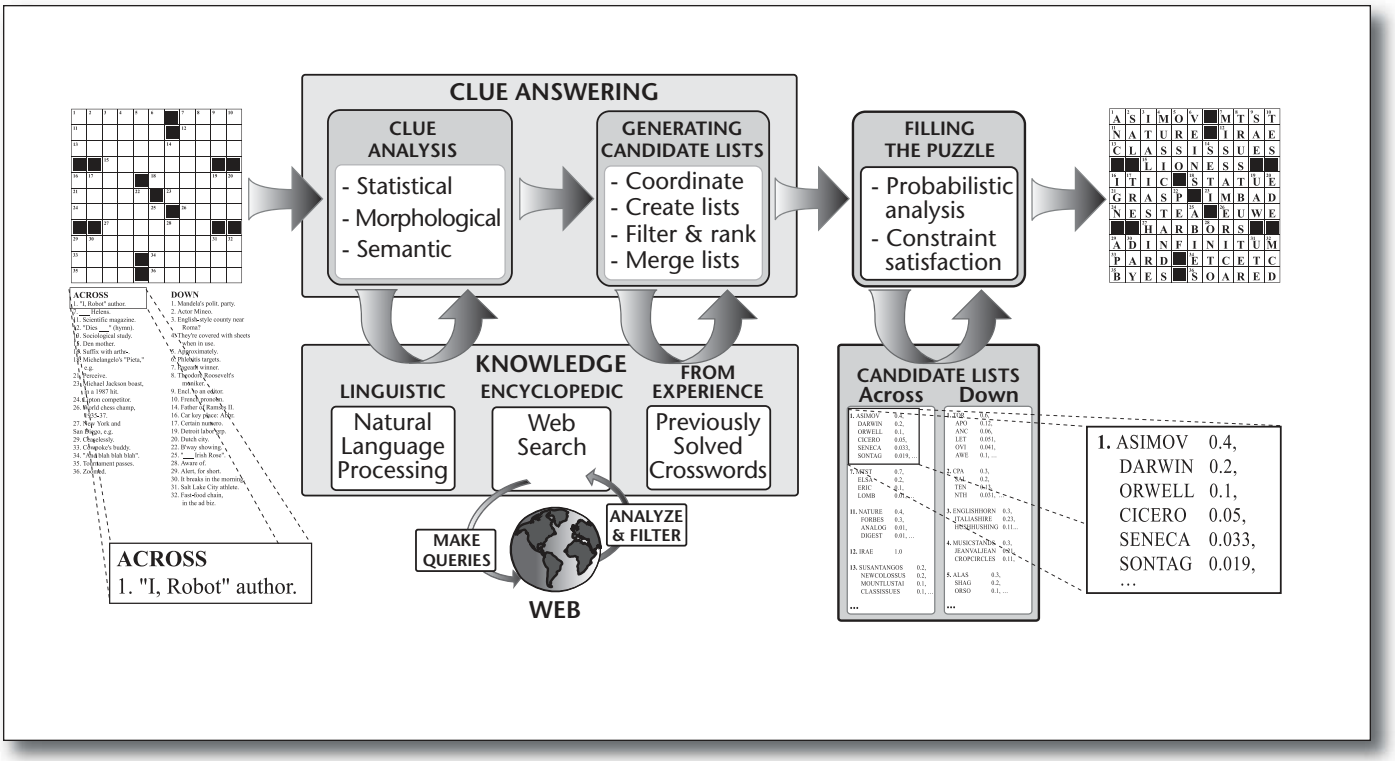

Figure 1. Overview of WebCrow.

Crossword solving is carried out by two sequential tasks: (1) answer the clues by creating a list of candidates; (2) fill out the puzzle with the best combination of answers. During the clue answering, WebCrow analyzes the clues using natural-language programming NLP techniques, which are used to coordinate the knowledge resources and to rank all the retrieved candidates for each clue. The program uses the web (querying Google) in order to discover fresh encyclopedic knowledge. In addition, there are a number of other clue-answering experts that appeal to linguistic and experiential knowledge. Once these expert modules have processed all the clues, the lists of weighted answers are merged (yielding one single list per clue) and the grid-filling module is triggered. During the filling process a final score is associated to each candidate by propagating the probabilities, provided by the clue-answering step, over the constraint network defined by the crossword grid. Finally, A* is used to generate the filled configuration with the highest sum of word posterior probabilities.

Inside WebCrow

WebCrow is based on the sequential combination of "clue answering" and "grid filling" (figure 1), a solution, inspired by Proverb, that is radically different from a human's approach. During clue answering, WebCrow analyzes the linguistic features of a crossword clue so as to yield a morphological and semantic classification of each answer. For example, the answer to the clue <*Phlebitis targets* [5]: veins>, could be morphologically classified as: common noun plural: 70 percent, common noun singular 25 percent, proper noun 5 percent. Likewise, its semantic classification could be: medicine 65 percent, biology 25 percent, economics 10 percent. This information is used in the answer lists to boost those candidates that are associated with the most probable morphological and semantic categories. This approach could be more aggressive by pruning candidates that correspond to categories with low probability, as done in standard "question answering" (Voorhees 2003) and human decision making. The drawback is that this would make the clue-answering step brittle, possibly causing the deletion of the correct answer from the candidate list, which would harm the grid-filling process. Much of the charm and challenge in crosswords comes from answering the clues and, thus, finding those paths that lead to the linked solutions. Unlike standard question answering, clues have no standard interrogative form; they possess an intrinsic and intentionally conceived ambiguity, as their topic can be either factoid <*Crimean conference site* [5]: yalta> or nonfactoid <*Whisper, as a secret* [7]: breathe>.

In order to generate the list of candidates, WebCrow exploits its expert modules, which appeal to experiential, linguistic, and encyclopedic knowledge. Experiential knowledge contains crossword-specific information that humans achieve by experience. It should be stressed that there is a consistent percentage of solutions that appear with high frequency (for example, *eyre* in American crosswords), along with a number of clue-solutions pairs that are somehow similar to others already seen. On the other hand, there are clues that fol-

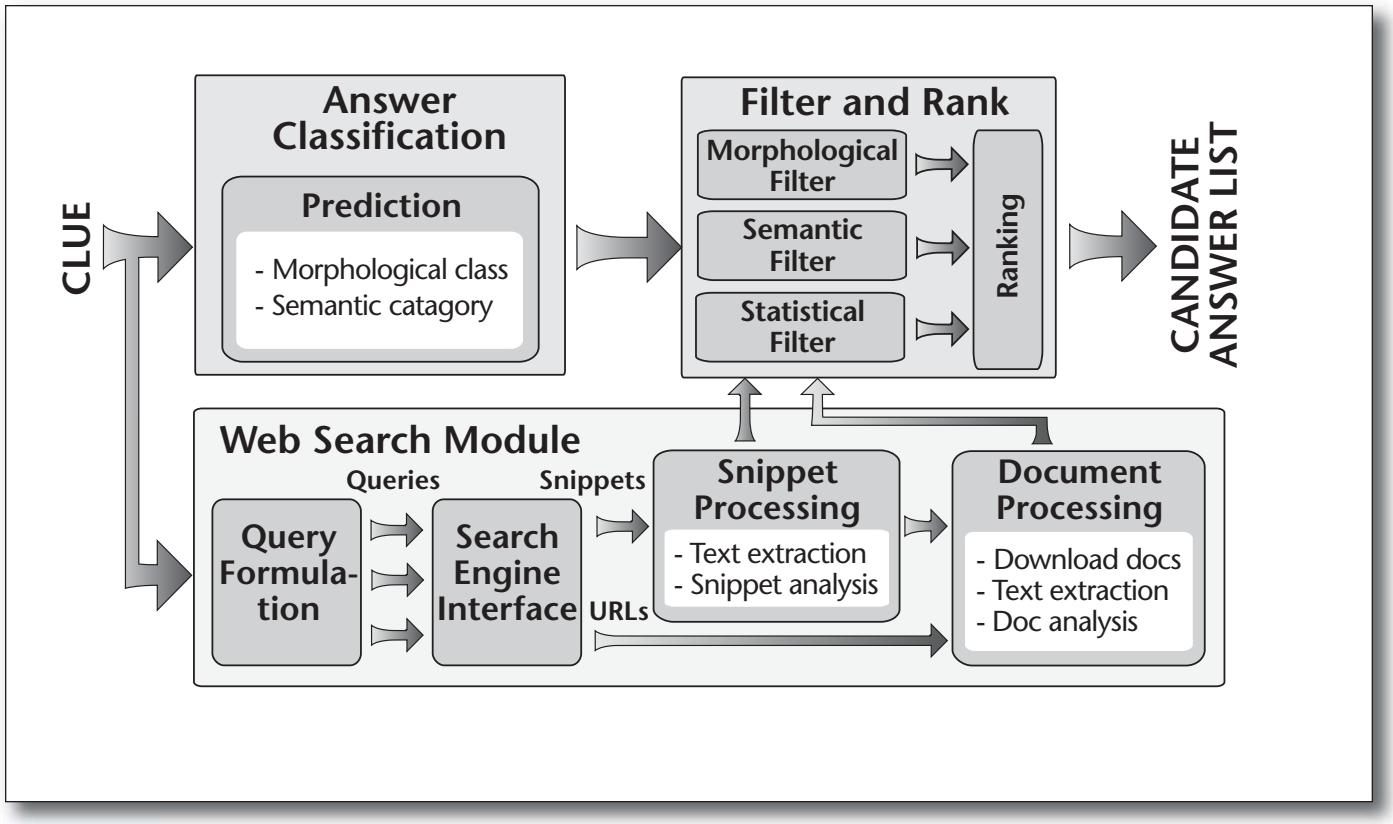

Figure 2. WebCrow; inside the Web Search Module (WSM).

For each clue WebCrow generates a small number of reformulations. At first, the WSM downloads a significant number (ranging from 20 to 30) of snippets. If the answers that are extracted are unsatisfactory then full-document download is triggered. All possible answers are stored and weighted using statistical, morphological, and semantic analysis.

low crossword-specific rules (for example, fill-in-the-blanks). To deal with these cases, WebCrow is equipped with a crossword database (CWDB) of previously solved clue-answer pairs and with a number of pattern-based answering experts that return the answer even in the case of an approximate matching of the clues in the database. The WebCrow database is currently made of about 5000 American and 900 Italian solved puzzles. Linguistic knowledge is handled using several computational linguistics technologies, such as a machine-trained part-of-speech tagger (Abney 1997), a lemmatizer,[2] and a lexical database with semantic relations between concepts expressed by WordNet (Miller 1995), along with the Italian extension MultiWordNet and WordNet Domains (Magnini and Cavaglia 2000). In addition, WebCrow uses language-specific dictionaries that are accessed by inverted indexes. Finally, large lexicons are used in order to increase the probability of retrieving the correct answer. Encyclopedic knowledge is handled exclusively using the web, thus relying on its autonomous growth. Clues are converted into queries to a search engine, and the returned results are automatically filtered so as to extract candidates, which are ranked by taking into account statistical, morphologic, and semantic features (see figure 2).

The grid-filling process can be seen as a probabilistic constraint-satisfaction problem (Littman, Keim, and Shazeer 2002; Kumar 1992). WebCrow takes into account the probabilities associated with each clue answer and the puzzle constraints using a belief network (Pearl 1988). The filling task is accomplished using a standard optimization algorithm (Hart, Nilsson, and Raphael 1968) that finds the configuration with the highest probability. Special attention is given to the case in which the answer is missing from the relative candidate list. Taking inspiration from humans, a loop-back approach is adopted to enrich the lists by a second answer-filling process. This is currently handled by generating *n*-gram probabilities for all the unseen sequences of characters that implicitly appear in the grid.

Clue Answering Using the Web

Figure 2 gives a sketch of how WebCrow makes use of the web to answer crossword clues. The input is

given by the scheduler, which assigns a set of clues to the web search module. For each clue a small number of reformulations are processed in order to make the search more effective. Similarly to the SMU Falcon system (Harabagiu et al. 2000), WebCrow's reformulations follow two different query expansion approaches: one that enriches the morphological forms of the keywords (for example, varying the number and gender of a noun or the tense of a verb or even introducing nominalizations of verbs) and one that subsequently enriches the query with synonyms and hypernyms from WordNet. After the query reformulation step, the WSM downloads a significant number (ranging from 20 to 30) of snippets. The statistical distribution of the candidates inside the snippets is used for scoring: if there are answers extensively distributed throughout the different snippets then their confidence is given a high score. Snippet answering contributes substantially to dealing with straightforward clues, especially when they deal with named entities (for example, proper names of persons, title of movies, and so on). For example, in clues such as <*Hotelier Helmsley* [5]: leona> the answer can rapidly and precisely be found by simply checking the snippets instead of downloading the full document. If the snippet processing is unsatisfactory, WebCrow triggers a number of full-document downloads. Depending on the established time limits and the available bandwidth, the volume of downloads can vary substantially.[3] Moreover, very small (for example, < 1 KB) and very big (for example, > 200 KB) documents are discarded to increase the efficiency of the answering process. For standard question answering, online full-document web search is typically reputed to be too slow (Kwok, Etzioni, and Weld 2001). However, it is a successful approach for crossword clues, since they can be tackled in parallel, thus reducing the latency of HTTP responses to a minimum. Moreover, the WSM can be launched simultaneously with all the other answering modules, since the CPU usage of the WSM is significantly smaller. Full document answering does not guarantee the same precision that is obtained by snippets, but the download of full documents greatly increases the coverage. In a clue like <*As an old record* [8]: scratchy> the focus of the question is not sufficiently explicit; thus the snippets are likely to miss the answer. On the other hand, when searching inside one of the retrieved documents, a passage like "[...] those old scratchy classics [...]" could do the trick. All the possible answers from either snippets or full documents are stored and weighted using statistical (Chakrabarti 2002), morphological, and semantic analysis. The statistical ranking exploits features such as the TF-IDF[4] and the distance between the keyword occurrences and the answer itself by using an ad hoc variation of mean-square distance (Ernandes, Angelini, and Gori 2005). The morphological ranking uses the output of the PoS (part of speech) tagger on the web passages to match the predicted morphological class of the answer as determined by the morphological classifier. The semantic ranking uses the domains associated with each synset (using WordNet Domains) appearing in the text passages in order to boost the words that are closer to the domain profile that is expected for the clue answer. At the end of the clue-answering process the relative output lists for each clue are merged on the basis of their confidence and of a set of module-specific parameters, which are learned from examples.

Interestingly, when suppressing the web search module, our experiments show that there is a significant performance drop (see figures 3 and 4 and Ernandes, Angelini, and Gori [2005] for more details). This is much more evident for long (that is, > 6 letters) words than for short ones, that can be more easily found by accessing the crossword database (CWDB). Providing a good coverage of long answers has a direct impact on the letter-filling step, since a higher number of correct letters guarantees a strong and more effective constraint propagation over the grid. The effect of removing the WSM is also evident by observing the coverage of correct answers in the candidate lists (see table 1), but it is even more significant when analyzing the grid-filling results. On the ECAI-06 competition dataset (see table 2) the average number of correct words dropped by 49 percent for Italian puzzles, by 15 percent for American ones, and by 33 percent for bilingual ones. The performance drop is more significant for Italian than American puzzles for at least three reasons: (i) Italian puzzles contain more encyclopedic clues; (ii) WebCrow accesses a smaller crossword database for Italian than for American crosswords; (iii) Italian puzzles contain a number of unconstrained cells (see discussion below), increasing the complexity of the grid-filling.

WebCrow Competitions

WebCrow has been tested on American and Italian crosswords and additionally on bilingual crosswords with the joint presence of American and Italian clues (see table 2). Leaving aside some minor differences, these crosswords share the same structure, which allows us to restrict the main architectural differences to the linguistic and experiential knowledge modules. In bilingual crosswords, a machine-trained language recognizer is also used to correctly address the clues and launch the appropriate modules.

The human-machine challenge requires that the machine and the human participants attack the

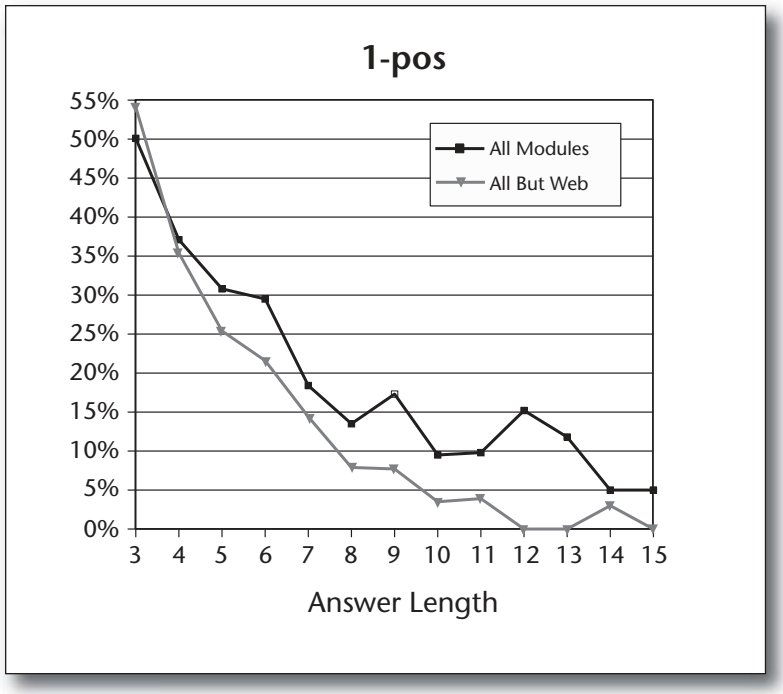

Figure 3. Clue Answering with and without Web Search.

The answering performance is measured by observing the success rate at 1-position (the probability of finding the correct answer in the first position) with different answer lengths.

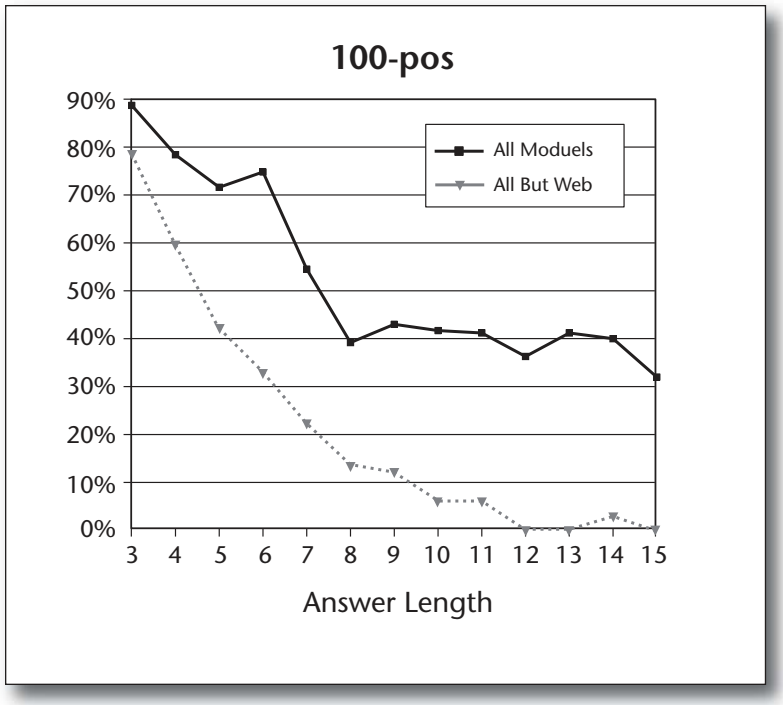

Figure 4. Clue Answering with and without Web Search.

The answering performance is measured by observing the success rate at 100-position (the probability of finding the correct answer in the first 100 positions).

same set of puzzles under the same rules. A number of different crossword competitions involving WebCrow have been organized[5] starting from a challenge against 181 Italian crossword solvers in which WebCrow placed 55th (February 2006).

The ECAI-06 Crossword Competition

On 30 August 2006, during the European Conference on Artificial Intelligence (ECAI-06), 25 volunteer conference attendees challenged WebCrow to an official human-machine contest. In addition, 53 crossword lovers voluntarily joined the contest through the Internet (the event was publicly promoted by *La Repubblica*, an online newspaper and crossword publisher). Online challengers were allowed to use any information source, including the Internet and encyclopedias. The participants were all requested to solve five different crosswords (two Italian, ITA-1 and ITA-2, two American, USA-1 and USA-2, and one bilingual puzzle with clues in American and Italian) given 15 minutes for each puzzle. WebCrow was launched on the five puzzles in a real-time challenge against the human competitors. As shown in table 2, WebCrow displayed steady performance over all five puzzles, with a minimum of 75 points on ITA-2 and a maximum of 96 points on USA-2. WebCrow won three of the four challenges, reporting an overall score of 440, whereas the best human participant only reached 319. In the Italian ranking, the program ended up at the 21st position out of 74 participants.[6]

Data Set

An initial pool of 55 crossword puzzles was gathered, and WebCrow was tested on all of them. For each category the best and the worst results were removed, leaving a final pool of 45 crosswords.

Sources

For the Italian crosswords, we made use of two of the main Italian crossword publishers: *La Settimana Enigmistica* (ITA-1) and *La Repubblica Online* (ITA-2). The ITA-1 crosswords were regenerated in order to give them more novel clue-answer pairs. The American crosswords were extracted from the databases of the *Washington Post* and the *New York Times* (Monday–Wednesday puzzles), both well known for their high quality. All puzzles were published between January and May 2006. The bilingual crosswords were prepared by Douglas Eck (Université de Montréal) using Crossword Compiler, commercial software for crossword editing, fed with all the previously mentioned Italian and American sources. All the bilingual puzzles were required to contain at least 40 percent of both Italian and American clue-answer pairs. The ECAI pool was randomly selected by five conference attendees.

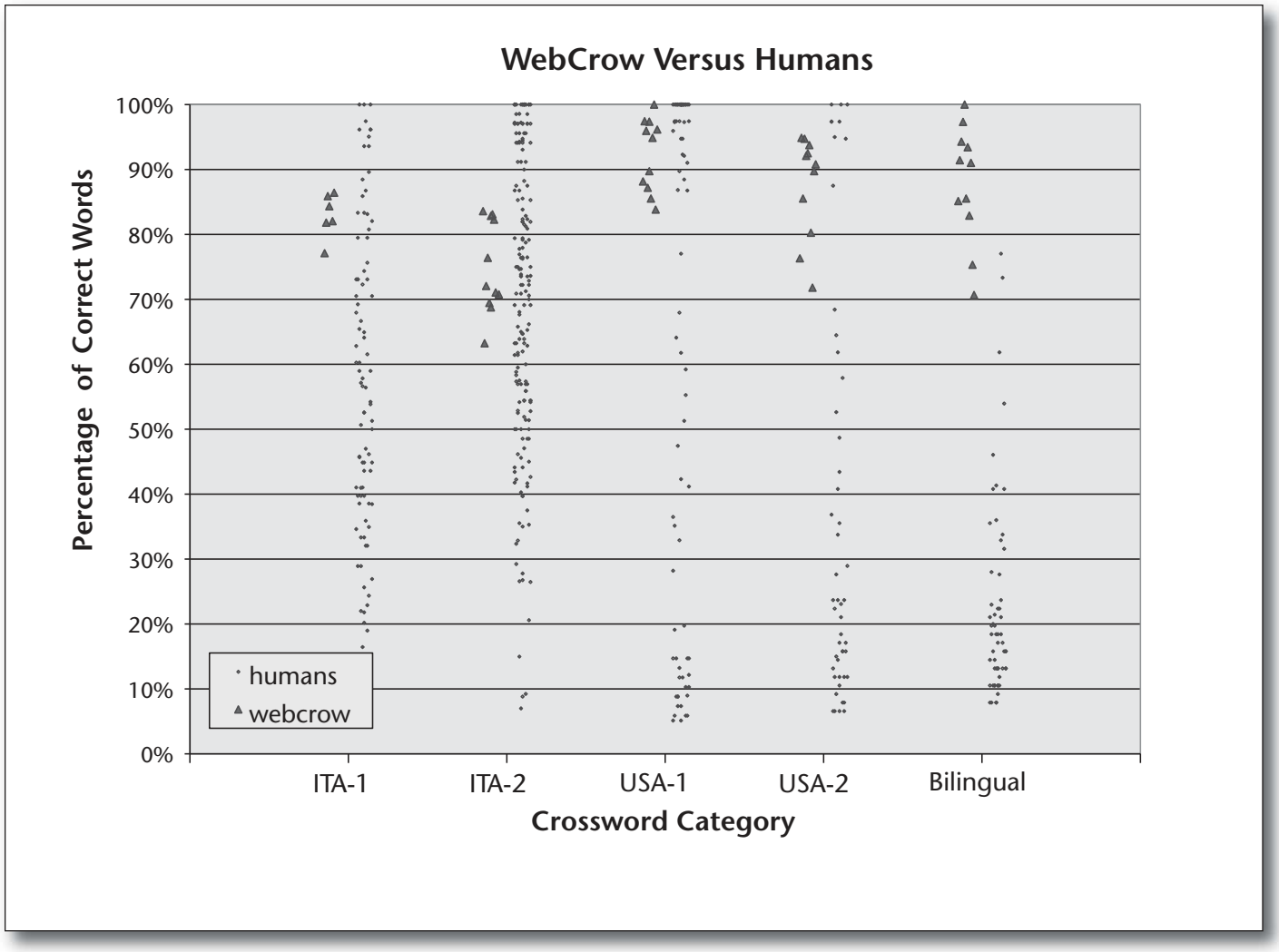

Figure 5. WebCrow Versus Humans

Experimental results carried out after the ECAI-06 competition, which include data collected mostly from crossword masters from Cruciverb.com. Each dot of the plot represents a single crossword solution. The triangles show WebCrow's results on the 45-puzzle pool selected for ECAI-06, while the small circles indicate the results obtained on the same puzzles by 137 crossword experts.

Results

Each cell in table 2 shows the ranking of WebCrow along with the number of participants (including WebCrow) in the corresponding category. For each crossword, the score was computed on the basis of the percentage of words correctly inserted. A "quality" bonus of 10 points was assigned to completely correct puzzles, along with a "time" bonus for quick solutions of 1 point per minute early. Competitors were divided into separate onsite (in a conference hall) and online (from all over the world) groups.

It is likely that the different performance achieved in the American and Italian competitions is mainly dependent on the limited number of native American-English speakers involved in the

Answer Coverage	Using the Web	Without the Web
First 5 positions	57 percent	51 percent
First 50 positions	72 percent	62 percent
First 1000 positions	87 percent	75 percent
Full answer list	96 percent	92 percent

Table 1. Impact of the Web on Clue Answering in the ECAI-06 Competition.

The number of correct answers given by WebCrow decreases substantially if we suppress the web search module. This effect becomes more evident as we observe the coverage of correct answers within the first 50 and 1000 candidates.

Data Set Category Label	Crossword Publisher Source	Initial Pool	Selected Pool	ECAI Competition Set
ITA-1	*La Settimana Enigmistica* (modif.)	7	5	1
ITA-2	*La Repubblica Online*	12	10	1
USA-1	*Washington Post*	12	10	1
USA-2	*New York Times* (Monday–Wednesday)			
Bilingual	All above and new—mixed	12	10	1
Total		55	45	5

Results WebCrow Rank	ITA	USA	Bilingual	Overall
Onsite Competition	5th (21)	1st (22)	1st (19)	1st (26)
Onsite Competition	17th (54)	1st (39)	1st (39)	1st (54)
Onsite Plus Online	21st (74)	1st (58)	1st (59)	1st (79)

Table 2. ECAI Crossword Competition.

	Seen Answers	Seen Clues	Seen Clue-Answer Pairs
Washington Post '06	94.44 percent	52.70 percent	45.25 percent
New York Times '06 Monday–Saturday	90.32 percent	44.95 percent	34.42 percent
New York Times '06 Sunday	88.52 percent	5.59 percent	35.13 percent
USA-1 (ECAI-06)	83.82 percent	48.53 percent	20.59 percent
USA-2 (ECAI-06)	90.79 percent	52.63 percent	46.05 percent
AI Magazine	59.24 percent	7.14 percent	3.36 percent

Table 3. Puzzle Novelty.

The novelty of a puzzle is measured by the number of answers, clues, and clue-answer pairs that can be observed in previously seen crosswords (CWDB).

competition. In order to get a richer data set we decided to open an online competition from September to October of 2006 by inviting the main Internet crossword-lover community (Cruciverb.com) to join in; 137 crossword experts answered the challenge. The results of this extended competition are summarized in figure 5 and provide additional insight on the actual performance of the program. With minor discrepancies between the different categories, WebCrow outperformed more than 70 percent of human experts on average, whereas the program still out-performed all human competitors on the bilingual competitions.

Additional Crossword Competitions

After the ECAI-06 conference, two additional human-machine crossword competitions have been organized with the goal of gathering useful data for a broader evaluation of the system's performance. The first contest was held in Italy (Machine Creativity Festival, Florence, December 2006), where WebCrow outperformed 59 of its 90 human challengers on a set of difficult Italian crosswords from *La Settimana Enigmistica*. The second one took place online during the IJCAI-07 conference (Hyderabad, India, January 2007). This contest was based on four unpublished crossword puzzles with clues extracted from the *New York*

	Dimensions	Average Answer Length	Number of Clues	Average Word Connection	Black Cells	Constrained Cells
Washington Post '06 (average)	15 x 15*	5.06	75.45	13.56 percent	15.87 percent	100 percent
New York Times '06 Monday–Saturday (average)	15 x 15*	5.20	73.51	14.36 percent	15.51 percent	100 percent
New York Times '06 Sunday (average)	21 x 21*	5.27	141.27	7.53 percent	16.63 percent	100 percent
USA-1 (ECAI-06)	15 x 15	5.79	68	17.18 percent	12.44 percent	100 percent
USA-2 (ECAI-06)	15 x 15	5.03	76	13.24 percent	15.11 percent	100 percent
AI Magazine	30 x 30	3.98	238	2.49 percent	33.89 percent	59.33 percent

Table 4. Puzzle Topology.

Several indices of a crossword puzzle's topology are reported. The "Dimensions" column displays the grid dimensions of the puzzles (* indicates that there are a small number of *Washington Post* and *New York Times* puzzles that differ from the standard grid size). "Average Answer Length" is the average length of the answers, and "Number of Clues" is the number of different clues in a puzzle. "Average Word Connection" quantifies the average amount of down answers crossed by an across slot and vice versa. Keeping the average answer length fixed, the word connection typically decreases in inverse proportion to the increase in the area available. The "Black Cells" column indicates the percentage of nonletter cells in the grid. "Constrained Cells" is the percentage of cells that belong to both an across and a down slot.

Times (Monday–Wednesday and Thursday–Saturday) and from the *Washington Post*. WebCrow placed second out of 25 participants (including 20 native speakers). These additional experiments confirmed the crossword skills displayed by WebCrow at the ECAI-06 competition: the system steadily outperforms around three quarters of human challengers.

Case Studies

In order better to understand the strengths and weaknesses of WebCrow, in this section we discuss its performance on three different crosswords: two from the ECAI-06 competition (USA-1 and USA-2) and one published in *AI Magazine* in the winter 2006 issue.

What makes a puzzle difficult to solve? Crosswords are difficult when clues are hard to answer and the crossing answers are less helpful in disclosing missing solutions.

Tables 3 and 4 illustrate these two aspects. In table 3 we get a hint of the novelty introduced in the crossword clues and answers. Specifically, the degree of novelty indicates the percentage of clues (clue-answer pairs) that were not previously seen in other crosswords. We compared crosswords published in 2006 with our crossword database, which includes crosswords published by the *New York Times* and the *Washington Post* before 2006. As pointed out by grandmaster chess players (Ross 2006), experts rely on specific knowledge and intensive training. This also holds for crosswords that become more difficult as more uncommon clues and answers are introduced.

Table 4 takes into account the topology of the puzzle. In general, the introduction of black cells reduces the connectivity between words and cells. This is inevitable, as the majority of the answers are short (3 to 6 letters long). In our analysis, full connectivity is given by a puzzle containing no black cells. In this case all across answers cross all down answers (that is, 100 percent word connection). In addition, table 4 shows the percentage of cells crossed by both an across answer and a down answer. This information is important, since cells without this constraint are more difficult to answer.[7] Generally, *New York Times* and *Washington Post* crosswords do not contain unconstrained cells. On the other hand, this problem does apply to Italian crosswords, which usually contain no more than 10 percent unconstrained cells. This turns out to be another explanation of the better performance of WebCrow on American than on Italian crosswords.

This solution shows a frequent problem. All the misplaced letters are clustered in a specific area of the puzzle (marked in gray). This happens when candidates of that zone have few highly ranked correct solutions; thus one or more incorrect

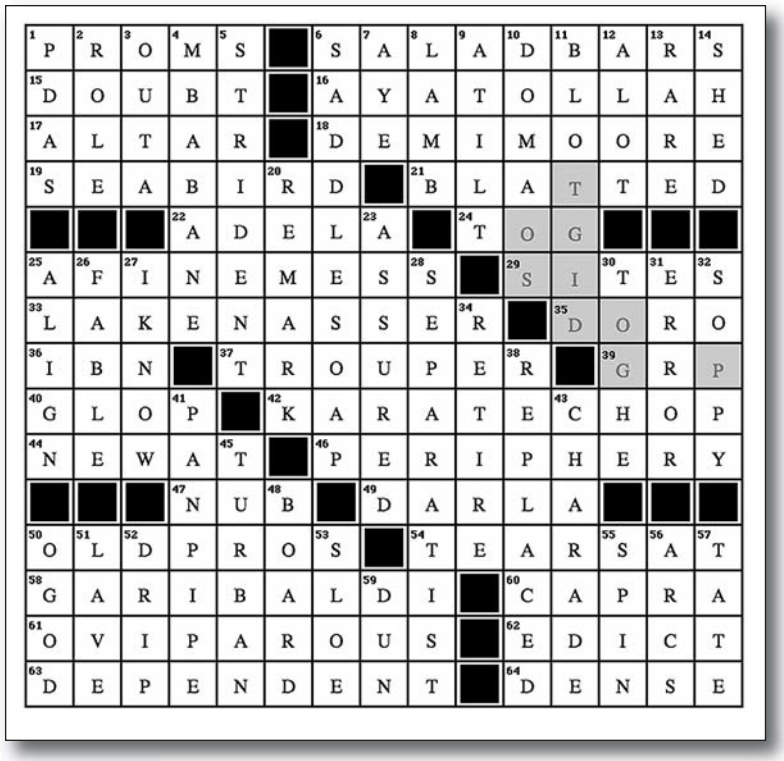

Figure 6. WebCrow Solution of USA-1 at ECAI-06.

Dimensions: 15 x 15. Time: 14 minutes, 13 seconds. Correct Words: 59 / 68 (86.76 percent).

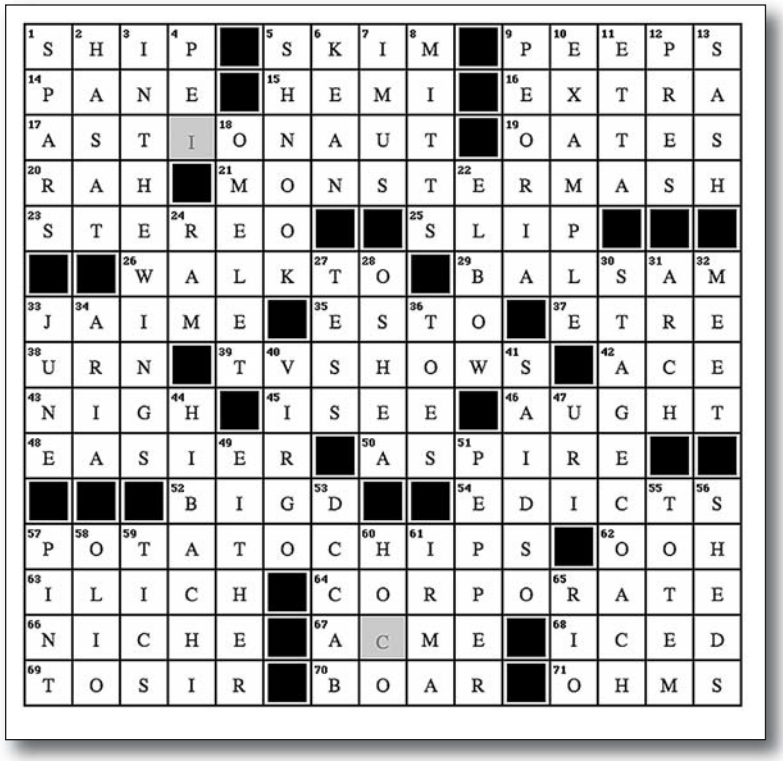

Figure 7. WebCrow Solution of USA-2 at ECAI-06.

Dimensions: 15 x 15. Time: 13 minutes, 25 seconds. Correct Words: 72 / 76 (94.74 percent).

answers were selected during the grid-filling phase.

WebCrow misplaced two letters, each corresponding to two crossing answers. In both cases the two clue-answer pairs were extremely challenging for the system, as they required a high level of semantic and logical competence. For example, for the misplaced letter *C* there are two clue-answer pairs: 67-A <*"What is to be done?!"*: ahme> and 60-D <*Popular cream-filled cake:* hoho>. In addition, the implicit module tends to infer answers when inserted candidate answers have a low probability. This can sometimes lead to errors, as in the case of the word *ACME*.

American puzzles at the ECAI-06 competition. In the ECAI-06 competition WebCrow solved a Washington Post puzzle (USA-1) and a New York Times puzzle (USA-2).

Based on tables 3 and 4, the difficulty of USA-2 can be evaluated as average, while USA-1 can be considered harder to solve. In the latter, the presence of long and unusual words yielded to a higher word connectivity. As a validation, many participants at the ECAI competition confirmed the difficulty of the *Washington Post* crossword. WebCrow had 86.8 percent correct words on USA-1 and 94.7 percent correct words on USA-2. The solutions provided by WebCrow are shown in figures 6 and 7. In the case of USA-2, just two letters were misplaced, while in USA-1 WebCrow was unable to correctly fill a portion of the puzzle.

The AI Magazine crossword puzzle. The crossword that appeared in the 2006 winter issue of *AI Magazine* is harder than standard crosswords for several reasons. First, as we can see in table 3, its novelty is very high when compared to an ordinary crossword. Since it is on the specific topic of artificial intelligence, almost none of the clues are in the crossword database.[8] For this reason systems that mainly depend on precompiled knowledge are likely to perform poorly on this puzzle, whereas WebCrow managed to put the correct solution in 95.8 percent of its candidate lists (23.1 percent in the first position, compared to 35–40 percent in ordinary crosswords).

Table 5 shows WebCrow's ability to deal with a wide range of knowledge. The three errors highlight the difficulty of disambiguating the correct answer. In particular, WIN and WON differ only by one letter and its corresponding puzzle cell is unconstrained.

In addition to the high clue-answer novelty, another main difficulty comes from the extremely low connectivity of the puzzle (see table 4). In the *AI Magazine* puzzle more than one third of the cells are black, leaving a high number of unconstrained letters (over 40 percent) and a very low level of word connectivity (2.5 percent). This situation greatly affected the grid-filling phase, where prop-

Clue	Answer	WebCrow Answer
26-A. Robot soccer event.	ROBOCUP	ROBOCUP
60-A. An early AT&T tablet computer.	EO	EO
86-A. Mitchell.	TOM	SAM
106-A. Engelmore.	BOB	BOB
188-A. A bloody good expert system.	MYCIN	MYCIN
206-A. The science of artificial ___.	INSEMINATION	INTELLIGENCE
30-D. Had "controlled hallucination" vision.	CLOWES	CLOWES
54-D. 13th Century theologian.	LLULL	LLULL
160-D. What Stanley did in 2005.	WON	WIN
179-D. A Nobel prize winner.	SIMON	SIMON
184-D. A joint conference.	IJCAI	IJCAI

Table 5. Some Clue-Answer Pairs from the AI Magazine Puzzle.

	Clue-Answering Success Rate			Final Results after Grid Filling	
	First Position	Within 10th Position	Within 100th Position	Correct Words	Correct Letters
Full WebCrow	23.1 percent	48.7 percent	71.8 percent	32.4 percent	45.9 percent
WebCrow without WSM	12.6 percent	24.4 percent	37.8 percent	6.7 percent	15.6 percent

Table 6. WebCrow Performance on the AI Magazine Puzzle

WebCrow's performance, with and without the web search module, measured in terms of clue answering and grid filling.

agating the probabilities through the constraint network usually enhances list positions of the correct solutions. As a result (see figure 8), WebCrow processed the crossword puzzle in around 52 minutes, producing 32.4 percent correct words (77 / 238), 45.9 percent correct letters (273 / 595), 50.8 percent wrong letters (302 / 595), and 3.4 percent blanks. Although the performance is inferior to that obtained on ordinary crosswords, it is still quite interesting, owing mainly to the web search module. For example, WebCrow correctly answered the clue <*A bloody good expert system:* mycin>, whereas it failed by proposing "intelligence" as the answer to <*The science of artificial* ____*:* insemination>.

As in our other experiments, in order to validate the role of the web search module we launched WebCrow on the *AI Magazine* crossword after having removed the web search module. The results are reported in table 6. Interestingly, the number of correct words dropped to 6.7 percent (16 / 238).

Conclusions

Apart from the intriguing challenge of crossword solving in itself, the experiments carried out with WebCrow illustrate benefits derived from a fruitful interaction with the web and the resultant capability to attack problems that for humans require deep semantic knowledge. WebCrow is not critically dependent on expert modules and can handle different languages. This is due to the fact that it relies strongly on its ability to mine encyclopedia knowledge from the web.

WebCrow is an example of the vision of AI systems drawing on the web for their knowledge. This is an area of active investigation in a number of AI areas. As Tom Mitchell wrote, "for the first time in history every computer has access to a virtually limitless and growing text corpus" (Stone and Hirsh 2005), providing great opportunities for mining by by AI systems (for example, Etzioni [2007]), and helping to address the

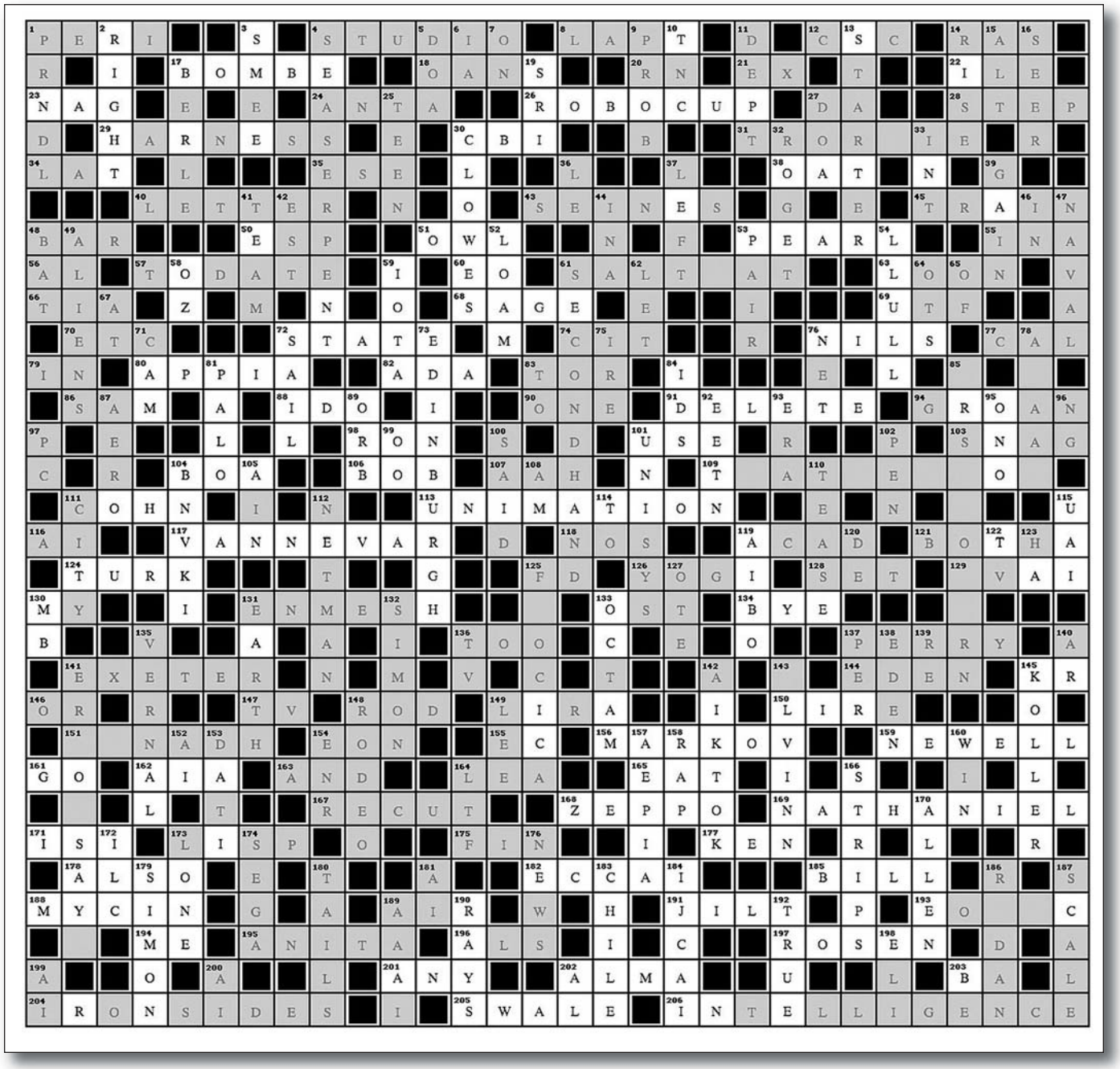

Figure 8. WebCrow Solution of the AI Magazine Winter 2006 Crossword.

Dimensions: 30 x 30. Time: 51 minutes, 53 seconds. Correct Words: 77 / 238 (32.35 percent)

knowledge acquisition bottleneck (Hendler 2005).

The architecture adopted in WebCrow might suggest solutions to relevant problems from different disciplines, ranging from information extraction to cognitive science, or from economics to biology. Studies on genes using computational linguistics (Searls 2002) underline the interdisciplinary interest of such methods. Moreover, intelligent web-based agents could have a strong impact on supporting scientific research in different disciplines (Bloom 1996). For instance, the robot scientist (King et al. 2004) could benefit from the acquisition of biological knowledge from the web in generating functional genomic hypotheses. Agents inspired by the WebCrow architecture could autonomously access and process large amounts of online data in order to stimulate effective interdisciplinary speculations (see, for example, in neuroscience, Voss [2001]).

Interestingly, as chess has been reputed "the

Drosophila of artificial intelligence" (McCarthy 1997), crosswords might be the *Drosophila* of web-based intelligent agents. British cryptic crosswords were used for a first selection of the famous Bletchley Park code-breaking team, which contributed to the birth of computers. Now web-based agents can solve crosswords themselves, and their autonomous reasoning on the huge web repository might significantly help to attack real-world problems.

Acknowledgments

We thank Michael Littman, Ramon López de Mántaràs, Paolo Traverso (all the ECAI-06 team), Michelangelo Diligenti, Joydeep Gosh, Ciro Guariglia, Gino Bartoli, Giovanni Canessa, Stefano Campi, Ennio Peres, and Ernesto Di Iorio. We are especially grateful to Douglas Eck, Clinton Jones, and Marco Lippi for their valuable help during the ECAI-06 competition and to Jon Glick, the author of the *AI Magazine* crossword, who invited us to test WebCrow on his crossword. We would also like to make a special acknowledgment to Clinton Jones who helped improve WebCrow and reviewed the paper. This research was supported by the University of Siena and by Google under the 2006 Google Research Program Award. We also would like to thank Google for its technical support during the design of the web search module and for the provided query service. The last author would like to dedicate this paper to the memory of his father.

Notes

1. The basic idea of WebCrow, to use an agent for online access to the Web for cracking crosswords, was mentioned in a brief news article by Francesca Castellani, "Program Cracks Crosswords," *Nature,* 6 October 2004.

2. A lemmatizer is a software module that when given a word-form inside a sentence, returns its dictionary form (for example, from "playing" to "play.")

4. TF-IDF is a weight, widely used in information retrieval that provides a straightforward measure for the informativity of a term inside a text passage. It is computed by multiplying the term frequency (TF) and its inverse document frequency (IDF).

3. For example, during the ECAI-06 competition the WSM downloaded as many as 15–20 documents per clue.

5. All the competition rankings can be found in the WebCrow site at www.dii.unisi.it/webcrow.

6. The news spread quickly mainly thanks to the article "Crossword Software Thrashes Human Challengers" by Tom Simonite, which appeared the day after the competition in *New Scientist* magazine.

7. For example, with no constraint on the second letter, *fill* and *full* could be two possible answers for the clue "Gas tank."

8. In order to handle this specific topic a minor tuning was applied to the web search module. The query formulation module was quickly modified to enrich the queries with AI-oriented keywords such as *artificial intelligence* and *computer science*.

References

Abney, S. 1997. Part-of-Speech Tagging and Partial Parsing. In *Corpus-Based Methods in Language and Speech Processing,* ed. S. Young and G. Bloothooft. Dortrecht, Holland: Kluwer Academic Publishers.

Aravind, K. J. 1991. Natural Language Processing. *Science* 253(5025): 1242–1249.

Bloom, F. E. 1996. An Internet Review: The Compleat Neuroscientist Scours the World Wide Web. *Science* 274(5290): 1104–1108.

Bouzy, B., and Cazenave, T. 2001. Computer Go: An AI Oriented Survey. *Artificial Intelligence* 132(1): 39–103.

Campbell, M., Jr.; Hoane, A. J.; and Hsu, F. H. 2002. Deep Blue. *Artificial Intelligence* 134(1–2): 57–83.

Chakrabarti, S. 2002. *Mining the Web: Discovering Knowledge from Hypertext Data*. San Mateo, CA: Morgan Kaufmann Publishers.

Cook, S. A. 1983. An Overview of Computational Complexity. *Communications of ACM* 26(6): 400–408.

Ernandes, M.; Angelini, G.; Gori, M. 2005. WebCrow: A WEB-based System for CROssWord Solving. In *Proceedings of the Twentieth National Conference of Artificial Intelligence (AAAI-05),* 1412–1417. Menlo Park, CA: AAAI Press.

Etzioni, E. 2007. Machine Reading: Papers from the 2007 AAAI Spring Symposium. Technical Report SS-07-06, Association for the Advancement of Artificial Intelligence, Menlo Park, CA.

Gelly, S., and Silver, D. 2007. Combining Online and Offline Knowledge in UCT. In *Proceedings of the 24th international Conference on Machine Learning (ICML '07)*, 273–280. New York: Association for Computing Machinery.

Harabagiu, S.; Moldovan, D.; Pasca, M.; Mihalcea, R.; Surdeanu, R.; Bunescu, R.; Girju, R.; Rus, V.; and Morarescu, P. 2000. FALCON: Boosting Knowledge for Answer Engines. In *Proceedings of the Ninth Text Retrieval Conference (TREC-9)*. Washington, DC: National Institute of Standards and Technology.

Hart, P. E.; Nilsson, N. J.; and Raphael, B. 1968. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics* (SSC4)2: 100–107.

Hendler, J. 2005. Knowledge Is Power: A View from the Semantic Web. *AI Magazine* 26(4): 76–84.

Keim, G. A.; Shazeer, N. M.; Littman, M. L.; Agarwal S.; Cheves, C. M.; Fitzgerald, J.; Grosland, J.; Jiang, F.; Pollard, S.; Weinmeister, K. 1999. PROVERB: The Probabilistic Cruciverbalist. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, 710–717. Menlo Park, CA: AAAI Press.

King, R. D.; Whelan, K. E.; Jones, F. M.; Reiser, P. G. K.; Bryant, C. H.; Muggleton, S. H.; Kell, D. B.; and Oliver, S. G. 2004. Functional Genomics Hypothesis Generation and Experimentation by a Robot Scientist. *Nature* 427(6971)(15 January): 247–252.

Kumar, K. 1992. Algorithms for Constraint-Satisfaction Problems: A Survey. *AI Magazine* 13(1): 32–44.

Kwok, C. C. T.; Etzioni, O.; and Weld, D. S. 2001. Scaling Question Answering to the Web. *ACM Transactions on Information Systems (TOIS)* 19(3) (July): 242–262.

Lam, S. K.; Pennock, D. M.; Cosley, D.; Lawrence, S. 2003. 1 Billion Pages = 1 Million Dollars? Mining the Web to

WWW.AAAI.ORG/ICWSM08.PHP

Play "Who Wants to be a Millionaire?" In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, 337–345. Eugene, OR: Association for Uncertainty in Artificial Intelligence.

Littman, M. L. 2000. Review: Computer Language Games. *Computer and Games* 134(1–2): 396–404.

Littman, M. L.; Keim, G. A.; and Shazeer, N. M. 2002. A Probabilistic Approach to Solving Crossword Puzzles. *Artificial Intelligence* 134(1–2): 23–55.

Magnini B., and Cavaglia, G. 2000. Integrating Subject Field Codes into WordNet. Paper presented at the Second International Conference on Language Resources and EvaluationLREC-2000, Athens, Greece, 31 May–2 June.

Manning C., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

McCarthy, J. 1997. AI as Sport. *Science* 276: 1518–1519.

Miller, G. A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38(11): 39–41.

Minsky, M. 1968. *Semantic Information Processing,* 12. Cambridge, MA: The MIT Press.

Mittman, B., and Newborn, M. 1980. Computer Chess at ACM 79: The Tournament and The Man Versus Man and Machine Match. *Communications of the ACM* 23(1): 51–54.

Munakata, T. 1996. Thoughts on Deep Blue Versus Kasparov. *Communications of the ACM* 39(7): 91–92.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann Publishers.

Ross, P. E. 2006. The Expert Mind. *Scientific American* 295(2)(August): 64–71.

Searls, D. B. 2002. The Language of Genes. *Nature* 420(6917): 211–217.

Shannon, C. E. 1950. Programming a Computer for Playing Chess. *Philosophical Magazine* 41(4): 256–275.

Shapiro, S. C. 1992. *Encyclopedia of Artificial Intelligence,* 54–57. New York: John Wiley and Sons, Inc.

Shazeer, N. M.; Littman, M. L.; Keim, G. A. 1999. Solving Crossword Puzzles as Probabilistic Constraint Satisfaction. In *Proceedings of the Sixteenth national Conference on Artificial Intelligence (AAAI-99)*, 156–162. Menlo Park, CA: AAAI Press.

Sheppard, B. 2002. World-Championship-Caliber Scrabble. *Artificial Intelligence* 134(1–2): 241–275.

Simon, H., and Schaeffer, J. 1992. *Handbook of Game Theory with Economic Applications*. Amsterdam: Elsevier North Holland.

Stone, M., and Hirsh, H. 2005. AI—The Next 25 Years. *AI Magazine* 26(4): 85–87.

Tesauro, G. 1994. TD-Gammon, a Self-Teaching Backgammon Program, Achieves Master-Level Play. *Neural Computation* 6(2): 215–219.

Voorhees, E. M. 2003. Overview of the TREC 2003 Question Answering Track. In *Proceedings of the Twelfth Text Retrieval Conference*. Washington, DC: National Institute of Standards and Technology.

Voss, D. 2001. Better Searching Through Science. *Science* 293(5537): 2024–2026.

Marco Ernandes received his master's degree in 2003 from the Università di Siena, Italy, where he is currently a Ph.D. student studying cognitive science. In 2007 he started collaborating with expert systems and cofounded the company QuestIT, focusing on question-answering technology. His main research interests are in heuristic search, machine learning, and text mining. He especially enjoys studying mathematical and linguistic puzzles.

Giovanni Angelini received a master's degree in computer engineering from the Università di Padova (Italy) and a Ph.D. from the Università di Siena with a dissertation on crossword solving by mining the web. In 2007 he started collaborating with expert systems and cofounded the company QuestIT, which operates in the area of question answering. He is also currently undergoing a research collaboration with the Università di Siena. His main research interests are in machine learning and in natural language understanding.

Marco Gori received a Ph.D. degree in 1990 from the Università di Bologna and was a visiting student at the School of Computer Science, McGill University, Montréal. In 1992, he became an associate professor of computer science at the Università di Firenze, and in November 1995, he joined the Università di Siena, where he is currently full professor of computer science. His main interests are in machine learning, with applications to pattern recognition, web mining, and game playing. He is coauthor of *Web Dragons: Inside the Myths of Search Engines Technologies* and has served as chair of the Italian Chapter of the IEEE Computational Intelligence Society and as president of the Italian Association for Artificial Intelligence. He is a fellow of ECCAI and of the IEEE.