Die Rätselrevolution: Automated German Crossword Solving

Andrea Zugarini[1,*,†], Thomas Röthenbacher[2,†], Kai Klede[2], Marco Ernandes[1], Bjoern M. Eskofier[2,3] and Dario Zanca[2]

[1] *expert.ai, Siena, Italy*

[2] *Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany*

[3] *Institute of AI for Health, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany*

Abstract

Crossword puzzles are popular word games played in various languages around the world, with diverse styles across different countries. For this reason, automated crossword solvers designed for a language, may not work well on others. In this paper, we extend Webcrow, an automatic crossword solver, to German, making it the first program for crossword solving in the German language. To address the lack of large clue-answer crossword pairs data, Webcrow combines multiple modules, known as experts, which retrieve potential answers from various resources, including the web, knowledge graphs, and linguistic rules. The system is evaluated on a collection of crosswords from variegate sources, where it is able to solve perfectly 67% of them. Additional analysis reveals that while our solver achieved commendable results, puzzles with poorly constrained schemas and original clues still presented significant hurdles. These findings shed light on the complexity of the crossword-solving problem and emphasize the need for future research to address and overcome these particular challenges effectively.

Keywords

Automated Crosswords Solving, German Crosswords, Question Answering

1. Introduction

Automated crossword solving is a challenging problem in Artificial Intelligence and Natural Language Processing (NLP) research. Solving a puzzle requires multiple skills, ranging from encyclopedic knowledge and linguistics to reasoning and constraint satisfaction. In the past, several automated crossword systems have been proposed for English [1, 2, 3]. Despite the successful results achieved by these approaches, they do not investigate crossword resolution in other languages. All those methods heavily rely on large databases of previously answered clues to retrieve and rank answer candidates, that sometimes are even re-ranked [4, 5, 6, 7]. Berkley Crossword Solver [3] make also use of multiple Language Models that have been fine-tuned to segment answers in words and to correct wrong letters. The need for such resources hinders the application of these solutions to other languages.

WebCrow [8, 9] instead, is a crossword solver that was applied also to Italian puzzles. The architecture which is composed of multiple modules, called experts, facilitates the portability to new languages. In this work, we extend WebCrow to the German language. To the best of our knowledge, we are the first to propose an automatic solver for German crosswords.

The paper is organized as follows. In Section 2, the whole WebCrow architecture is described. Then, in Section 3 we present the data gathered for German and its usage by the WebCrow experts. Experiments are outlined in Section 4. Finally, we summarize our conclusions and directions for future works in Section 5.

2. WebCrow

As showcased in Figure 2, WebCrow, similarly to prior crossword solvers like Proverb [1], works in two stages: candidate answers retrieval and constraints satisfaction. In the first stage, a list of weighted candidate answers is retrieved for each clue. The retrieval is carried out by multiple modules, called *experts*. Candidates' lists are then combined together by the merger module. In the second step, the solver fills the grid given the potential answers with the objective of maximizing the most probable solution given the constraints imposed by the grid.

2.1. The Expert Modules

WebCrow uses multiple modules to retrieve answer candidates. In general, the number of experts can vary, and

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

[*] Corresponding author.

[†] These authors contributed equally.

azugarini@expert.ai (A. Zugarini); thomas.roethenbacher@fau.de (T. Röthenbacher); kai.klede@fau.de (K. Klede); mernandes@expert.ai (M. Ernandes); bjoern@fau.de (B. M. Eskofier); dario.zanca@fau.de (D. Zanca)

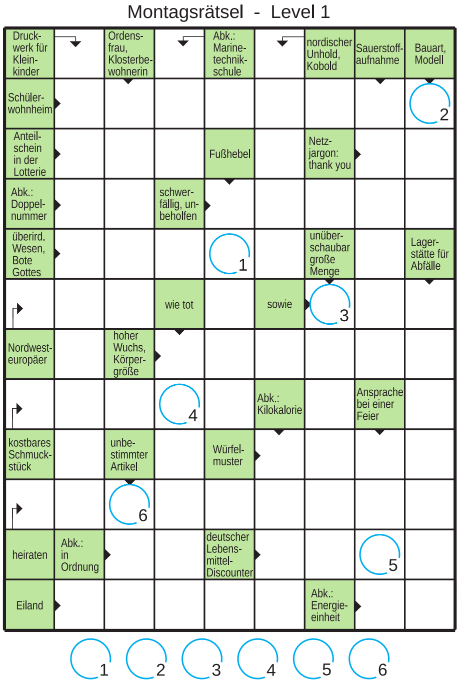

Figure 1: A German crossword puzzle in the "swedish" style.

new experts can always be added. Here, we limit the discussion only to the modules that were then used for the German language.

CWDB. A large collection of previously answered questions is paramount for any crossword solver. The CrossWord DataBase (CWDB) expert retrieves answers from a database of clue-answer pairs. Retrieval and ranking of the answers is based on the semantic similarity between a clue and the clues in the CWDB. We follow the approach in [10]. We used pre-trained encoders [11, 12, 13] to embed both the clues in the database and the ones of the crossword that is being solved.

Knowledge Graph. Ontologies are a rich source of linguistic and encyclopedic knowledge, that is frequently required in crosswords. Knowledge Graphs (KGs) contain structured information about concepts and language. The expert in Webcrow exploits KGs to collect a database analogous to the CWDB in a straightforward way: linguistic concepts in the knowledge graph are paired with their definitions, similarly to the approach followed with the clue-answer pairs from the CWDB. Following the same approach, an answer retrieval step is then applied: the definitions are encoded in a semantic representation and then ranked and retrieved accordingly to a semantic similarity between the clue and the definition in the database.

Web Search. Web Search is the expert that characterizes Webcrow, as the name suggests. This module retrieves answer candidates by searching on the web. Differently from CWDB or Knowledge Graphs, it allows the retrieval of fresh information, that can be crucial to solve some clues. For each clue, the web is queried, making use of Bing APIs.[1] The answer list can be built upon either or both the snippets and the full documents returned by the search. All the frequent words extracted from the documents are ranked according to their TF-IDF, also taking into account the rank of the documents in which they occur [8]. IDF must be pre-computed on a large collection of documents.

Lexicon. The lexicon is a large vocabulary of German words. For each answer to fill in the crossword, the dictionary module returns all the words in the lexicon that fit in length. The returned list is weighted by the n-gram model used in the Implicit Module described below.

2.2. Implicit Module

The implicit module ensures that the crossword solver also works when the correct answer is not present in the candidate list. It generates new candidates on the fly using a collection of character level n-grams, with $n = 4$, together with their relative frequencies in the answers of the CWDB. Whenever the CSP solver reaches a point where none of the answers for a given clue fit into the grid, the implicit module tries to generate the most probable sequence of characters satisfying the constraints.

2.3. Belief Propagation and Grid Filling

The current version of WebCrow looks for a solution that maximizes the expected value of clues answered correctly. To compute the posterior probabilities $q_i(a)$ for an answer a to be in slot i it uses belief propagation. These probabilities then allow us to infer letter probabilities $q_{i,n}(\lambda)$, that quantify the likelihood of a character λ at position n of the answer to clue c_i. It can be computed as in Equation 1, where $a(n) = \lambda$ means the n-th letter of answer a is λ.

$$q_{i,n}(\lambda) = \sum_{a \in D_i, \text{with } a(n)=\lambda} q_i(a). \tag{1}$$

If a cell s in the grid is the n-th cell of an horizontal clue c_i, and the m-th cell in the vertical clue c_j, then the probability $q^s(\lambda)$ becomes:

$$k \cdot q_{i,n}(\lambda) \cdot q_{j,m}(\lambda),$$

[1] https://learn.microsoft.com/en-us/rest/api/cognitiveservices-bingsearch/bing-web-api-v7-reference.

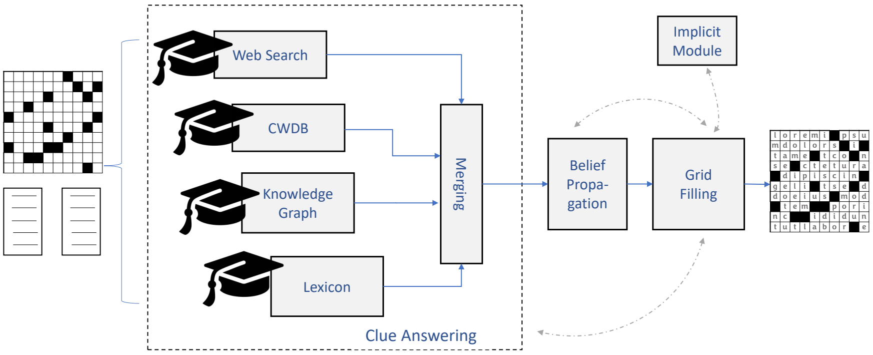

Figure 2: Sketch of the Webcrow architecture for German Crossword Puzzles.

where k is a factor that normalizes q^s to be a probability mass function. If the probability of a letter in a given cell exceeds a certain threshold, the solver "freezes" it in the solution and only allows answers that are consistent with such a letter.

A solution that maximizes the sum of the answer probabilities $q_i(a)$ is then found using Greedy Search, which fills in the missing or incomplete answers by iteratively selecting the most probable from the remaining answers that fit into the missing cells.

Table 1
The full dataset of German crosswords. Listed for each type of crossword are their dimensions, the frequency of publication in their respective newspapers, the publication date of the oldest included crossword, and the number of crosswords included in total.

Source	dims	freq	oldest	# CWs
Bild	7x7	daily	01.11.19	1188
Spiegel	9x9	daily	01.07.20	944
Focus	9x9	daily	02.11.19	1187
HA	16x16	daily	03.01.22	394
FAZ	15x15	weekly	11.03.22	46

3. Adapting WebCrow to German

3.1. Data Gathering

We collected two kinds of data: full crosswords, and a large corpus of clue-answer pairs.

Full Crosswords. Full crosswords were obtained from various newspapers that offer crossword puzzles on their websites. These include "Der Spiegel", "Bild-Zeitung", "Focus", "Frankfurter Algemeine Zeitung" (FAZ) and "Hamburger Abendblatt" (HA). The number of crosswords from each source as well as the crossword dimensions, the frequency of their publication and the earliest retrieved crossword is listed in Table 1. With the exception of FAZ, all the other crosswords schemas are in the so called Swedish style, characterized by brief definitions that goes within the black cells. An example is shown in Figure 1.

The retrieved documents were split into a training, validation, and test set. For each source except for FAZ, the training set included all crosswords published up to and including the Friday 30th September, 2022. For FAZ, an earlier cut-off date (Wednesday 31st August, 2022) was chosen to leave more crosswords for use in validation. The validation set then includes the crosswords that are not in the training set that were published up to the Thursday 24th November, 2022. The crosswords for the test set were published after that date and up to the Wednesday 31st August, 2022, however, the daily publications (every source except FAZ) were sub-sampled to include a similar number of each source. Because some crosswords, like the one from the New York Times, have differing levels of difficulty for each day of the week [1], we sampled every six crosswords, instead of every seven, to make sure all days of the week were present in the test set. This resulted in ten crosswords for FAZ and twelve for each other source.

Clue-answer pairs. We collected a large set of clue-answer pairs crawling two German online crossword web sites: kreuzwortraetsel.de (KWRDE) and kreuzwortraetsel-hilfe.com (KWRH). The download of KWRDE was covered in a period between Friday 9th

Table 2
Crossword Database composition. Clue-answer pairs were collected from different sources.

CWDB	# clue-answer	# unique clue-answer	# unique answers	# unique clues
KWRH	1,312,770	1,312,153	308,680	556,278
KWRDE	1,495,402	1,495,322	357,174	739,146
Spiegel	17247	11251	8783	10711
Bild	13721	8951	6435	8598
Focus	22357	14182	10230	13401
HA	17344	8114	6503	7645
FAZ	1029	1029	941	1029
Total	2,879,870	2,455,604	437,826	1,094,183

September, 2022 and Saturday 1st October, 2022. Similarly, KWRH was downloaded between Thursday 8th September, 2022 and Tuesday 20th September, 2022. As shown in Table 2, a total of 2.4 million unique clue-answer pairs were obtained this way, containing 438 thousand unique answers for German. To supplement the database, the clues from the crosswords in the training set were extracted and added to the database. The individual and total contributions are shown in Table 2.

3.2. Experts Adaptation

CWDB. The 2.4 million clue-answer-pairs crawled were used to build the CWDB expert. After minor preprocessing, clues were embedded with multi-lingual Universal Sentence Encoder (USE) [14] as in [10].

Knowledge Graph. To extend the coverage of our answer retrieval step, we made use of an additional proprietary German ontology[2]. Overall, it contained about 943 thousand lemmas from various web sources like encyclopedias. As for CWDB, we retrieve answer candidates based on semantic similarity [10], thus we treat each lemma as the candidate answer, and its definition is embedded with USE.

Web-Search. To find relevant words, web search experts require a database of document frequency values for the most common words. It was obtained from an online database of words and their frequencies in German movie subtitles [3]. Candidate answers were retrieved only from the web snippets.

Lexicon. The lexicon was constructed from a freely available online German dictionary [4], after romanizing umlauts (e.g. ä to ae) and excluding all non-ASCII and non-alphabetical characters.

Implicit Module. The tetra-grams in the Implicit Module were generated from the answers in the CWDB. They include start and end tokens "$" and "^" to allow for specific n-grams at the beginning and endings of words. They are weighted based on their frequency in the corpus.

4. Experiments

In the evaluation we both measured the end-to-end performance of Webcrow and the answer retrieval capabilities of each single module.

4.1. Answer Retrieval

Although good answer ranking is clearly very important for crossword resolution, it is even more essential to have the target answer present in the candidates' list. Indeed, even poorly ranked target answers can be boosted during belief propagation and grid filling, whereas a missing answer in the list would hardly be recovered, inevitably leading to errors or incomplete solutions. Hence, besides the MRR, we also consider coverage as a performance indicator of our experts.

Results of single experts are summarized in Table 3, where we also report the quality after merging and belief-propagation modules. As we can see, CWDB is the most informative expert, with the highest MRR and coverage. Also, Web Search achieves interesting MRR scores. Lexicon and Knowledge Graph have poorer ranking quality, but they both contribute to increasing the coverage, which is the main purpose of those modules. This can be observed from the coverage after Merging, where all the lists are combined together. Clearly, there is a high overlap in the experts' lists, however, the union of all of them reduces the number of missing target answers by 0.7% absolute, almost a third of all the missing target answers.

From Table 3 we can also observe how belief propagation significantly boosts the ranking quality, enormously facilitating the grid-filling stage.

[2] from Expert.ai.
[3] github.com/hermitdave/FrequencyWords/blob/master/content/2018/de/de_50k.txt
[4] github.com/enz/german-wordlist/blob/master/words

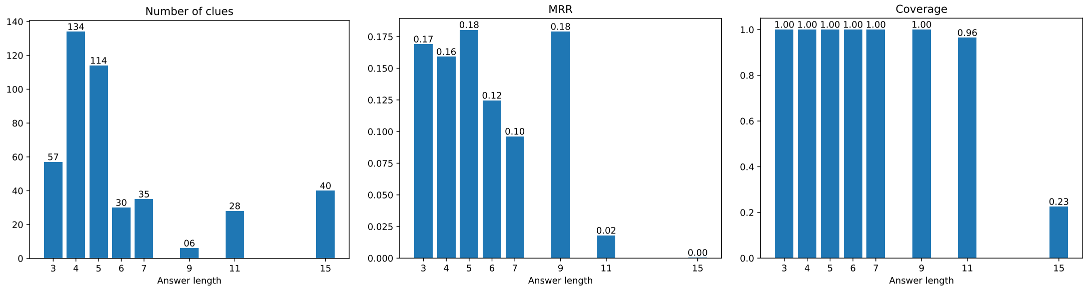

Figure 3: Number of clues, MRR, and Coverage of the merged lists divided by answer length in FAZ crosswords.

Table 3
Results of the individual experts.

Module	MRR	Coverage
Lexicon	0.005	86.8%
Web Search	0.200	75.6%
Knowledge Graph	0.086	91.3%
CWDB	0.634	97.6%
Merging	0.543	98.3%
Belief-Propagation	0.809	98.3%

4.2. Crossword Solving

The whole system was assessed on German full crosswords from the test set described in Section 3. We measured the solving quality with three indicators: the percentage of correctly inserted letters (OK letters), the percentage of correctly inserted words (OK words), and the fraction of perfectly solved crosswords (OK CWs). We report in Table 4 those metrics aggregated per crossword source. Overall 39 out of 58 (about 67% of them) crosswords were perfectly solved by German Webcrow. However, there is a strong variance between different sources. Smaller crossword grids like the ones from “Bild", were always solved without errors. Similar, satisfying performances were obtained in larger grids, also in 16x16 schema from HA. In contrast, Webcrow failed in solving FAZ puzzles, with surprisingly low results. Only 13% of words and 21% of letters were correctly answered in FAZ, far behind near-perfect crosswords from the other sources.

Performance on FAZ Crosswords. Different from other sources, FAZ puzzles are not in Swedish format. Instead, they are characterized by grids populated with many black cells that reduce the number of constraints to impose on the solution. Moreover, the grid layout is disposed in such a way that there is a significant percentage of answers longer than ten characters (see Figure 3), that in our CWDB coming are not present apart from those coming from the few FAZ crosswords in training. Thus, CWDB has little to no coverage of those clues, which are typically very important to constrain a large portion of the grid. Also, the style of the clues is remarkably different. There are many wordplays and linguistic games unusual in the rest of the data, making them challenging even for humans.

Table 4
Results of the end-to-end solving process. For each source, the number of crosswords solved completely correctly over the number of crosswords in the test set are given. Also listed are the average ratio of correct answer words and letters per tested crossword.

Source	OK CWs	OK words	OK letters
Bild	12 / 12	100%	100%
Spiegel	10 / 12	98%	99%
Focus	9 / 12	97%	99%
HA	8 / 12	98%	99%
FAZ	0 / 10	13%	21%

To delve in further, we also analyzed MRR and coverage after merging candidate answers divided by answer length in Figure 3. We can easily notice that both of them are significantly below the scores reported in Table 3 for all the crosswords. In particular, there is a non neglectable portion of long answers with poor coverage and close to zero MRR. Such a modest retrieval quality combined with a less contrained schema inevitably lead the system to fail in solving the crossword.

5. Conclusion

In this work we presented German Webcrow, the first crossword puzzle solver for the German Language. We collected both a dataset of clue-answer pairs and a set of German crosswords from different sources having various formats and styles. Webcrow achieved near-perfect word accuracy in Swedish-type crosswords, that proved to be generally easy to solve, solving overall 39/ 58 perfectly. However, our solver performed poorly on FAZ crosswords. Those puzzles were extremely challenging for multiple reasons, such as the poorly constrained schemas, due to the presence of many black cells, and the rich presence of sophisticated, original clues, involving articulated wordplays that formed words not present in the candidate answers lists.

Challenging puzzles like the ones in FAZ are a clear example of how complex the problem is, and why studying crosswords in multiple languages and formats is important for automated crossword solving. In the future we plan to address the current limitations. In particular, we plan to investigate the use of generative models, to cope with the novel unseen clues.

Acknowledgments

We thank Andreas Weber (raetsel4u.de) for allowing us to reprint one of his crossword puzzles in Figure 1 of this publication.

References

[1] M. L. Littman, G. A. Keim, N. Shazeer, A probabilistic approach to solving crossword puzzles, Artificial Intelligence 134 (2002) 23–55. URL: https://www.sciencedirect.com/science/article/pii/S000437020100114X. doi:https://doi.org/10.1016/S0004-3702(01)00114-X.

[2] M. L. Ginsberg, Dr. fill: Crosswords and an implemented solver for singly weighted csps, Journal of Artificial Intelligence Research 42 (2011) 851–886.

[3] E. Wallace, N. Tomlin, A. Xu, K. Yang, E. Pathak, M. Ginsberg, D. Klein, Automated crossword solving, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3073–3085. URL: https://aclanthology.org/2022.acl-long.219. doi:10.18653/v1/2022.acl-long.219.

[4] G. Barlacchi, M. Nicosia, A. Moschitti, Learning to rank answer candidates for automatic resolution of crossword puzzles, in: Proceedings of the Eighteenth Conference on Computational Natural Language Learning, 2014, pp. 39–48.

[5] G. Barlacchi, M. Nicosia, A. Moschitti, A retrieval model for automatic resolution of crossword puzzles in italian language, in: The First Italian Conference on Computational Linguistics CLiC-it 2014, 2014, p. 33.

[6] M. Nicosia, G. Barlacchi, A. Moschitti, Learning to rank aggregated answers for crossword puzzles, in: European Conference on Information Retrieval, Springer, 2015, pp. 556–561.

[7] A. Severyn, M. Nicosia, G. Barlacchi, A. Moschitti, Distributional neural networks for automatic resolution of crossword puzzles, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2015, pp. 199–204.

[8] M. Ernandes, G. Angelini, M. Gori, Webcrow: A web-based system for crossword solving, in: AAAI, 2005, pp. 1412–1417.

[9] M. Ernandes, G. Angelini, M. Gori, A web-based agent challenges human experts on crosswords, AI Magazine 29 (2008) 77–77.

[10] A. Zugarini, M. Ernandes, A multi-strategy approach to crossword clue answer retrieval and ranking, in: E. Fersini, M. Passarotti, V. Patti (Eds.), Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022, volume 3033 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-3033/paper72.pdf.

[11] M. Iyyer, V. Manjunatha, J. Boyd-Graber, H. Daumé III, Deep unordered composition rivals syntactic methods for text classification, in: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers), 2015, pp. 1681–1691.

[12] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al., Universal sentence encoder, arXiv preprint arXiv:1803.11175 (2018).

[13] T. Mikolov, W.-t. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies, 2013, pp. 746–751.

[14] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung, et al., Multilingual universal sentence encoder for semantic retrieval, arXiv preprint arXiv:1907.04307 (2019).