ISSN 2499-4553

IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 11, Number 1
june 2025

aAccademia university press

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

Associazione Italiana di Linguistica Computazionale

direttore responsabile
Michele Arnese

isbn 9791255001492

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_11_1

IJCoL

Volume 11, Number 1
june 2025

CONTENTS

Italian Crossword Generator: An In-depth Linguistic Analysis in Educational Word Puzzles

Kamyar Zeinalipour*
Università degli Studi di Siena

Tommaso Iaquinta**
Università degli Studi di Siena

Asya Zanollo†
Università degli Studi di Siena

Giovanni Angelini‡
Expert.AI

Leonardo Rigutini§
Expert.AI

Marco Maggini‖
Università degli Studi di Siena

Marco Gori#
Università degli Studi di Siena

Educational crosswords offer numerous benefits for students, including increased engagement, improved understanding, critical thinking, and memory retention. Creating high-quality educational crosswords can be challenging, but recent advances in natural language processing and machine learning have made it possible to use language models to generate nice wordplays. The exploitation of cutting-edge language models like GPT3-DaVinci, GPT3-Curie, GPT3-Babbage, GPT3-Ada, and BERT-uncased has led to the development of a comprehensive system for generating and verifying crossword clues. A large dataset of clue-answer pairs was compiled to fine-tune the models in a supervised manner to generate original and challenging clues from a given keyword. On the other hand, for generating crossword clues from a given text, Zero/Few-shot learning techniques were used to extract clues from the input text, adding variety and creativity to the puzzles. We employed the fine-tuned model to generate data and labeled the acceptability of clue-answer parts with human supervision. To ensure quality, we developed a classifier by fine-tuning existing language models on the labeled dataset. Conversely, to assess the quality of clues generated from the given text using zero/few-shot learning, we employed a zero-shot learning approach to check the quality of generated clues. The results of the evaluation have been very promising, demonstrating the effectiveness of the approach in creating high-standard

* Università degli Studi di Siena (UNISI), Via Roma 56, 53100 Siena, Italy.
E-mail: kamyar.zeinalipour2@student.unisi.it
** Università degli Studi di Siena (UNISI), Via Roma 56, 53100 Siena, Italy.
E-mail: tommaso.iaquinta@student.unisi.it
† Università degli Studi di Siena (UNISI), Via Roma 56, 53100 Siena, Italy.
E-mail: a.zanollo@student.unisi.it
‡ expert.ai, Via Virgilio, 48/H – Scala 5 41123, Modena, Italy. E-mail: gangelini@expert.ai
§ expert.ai, Via Virgilio, 48/H – Scala 5 41123, Modena, Italy. E-mail: lrigutini@expert.ai
‖ Università degli Studi di Siena (UNISI), Via Roma 56, 53100 Siena, Italy.
E-mail: marco.maggini@unisi.it
Università degli Studi di Siena (UNISI), Via Roma 56, 53100 Siena, Italy. E-mail: marco.gori@unisi.it

educational crosswords that offer students engaging and rewarding learning experiences. In this new extended version of (Zeinalipour et al. 2023b) we also propose a linguistic analysis of crossword clues developed from a syntactic perspective. This preliminary analysis provides insights into how clues are derived from complete sentences and what structures, if any, are preferred. The present linguistic study represents a useful support, in future research, for the generation of personalized puzzles in educational and medical environments.

1. Introduction

Crossword puzzles serve as a highly effective educational tool for numerous reasons. Firstly, they play a crucial role in enhancing children's vocabulary and spelling abilities, as solving the puzzles requires accurate word spelling (Orawiwatnakul 2013), (Dzulfikri 2016), (Bella and Rahayu 2023). Moreover, crossword puzzles are particularly beneficial for acquiring new lexicons in language classes and subjects that involve specialized technical terms (Nickerson 1977), (Sandiuc and Balagiu 2020), (Yuriev, Capuano, and Short 2016). Secondly, these puzzles foster problem-solving skills since students must engage in critical thinking to match clues with appropriate phrases (Kaynak, Ergün, and Karadaş 2023), (Dol 2017). Additionally, crossword puzzles contribute to memory retention, as students need to recollect previously learned material to complete the puzzles (Mueller and Veinott 2018), (Dzulfikri 2016). Lastly, they create an enjoyable and engaging learning experience, motivating students to continuously practice and improve their skills (Zirawaga, Olusanya, and Maduku 2017), (Bella and Rahayu 2023). In summary, crossword puzzles offer an enjoyable and effective approach to practice and enhance essential educational abilities (Zamani, Haghighi, and Ravanbakhsh 2021), (Yuriev, Capuano, and Short 2016).
Creating educational crosswords requires skill, but this process can be time-consuming and limited by human resources. Recent advancements in natural language processing and machine learning offer an alternative solution: training Large Language Models (LLMs) on vast amounts of data to generate diverse and engaging crossword clues and reduce creation time.
This paper makes several contributions to the field. Our initial contribution involves the utilization of this paper to introduce an extensive dataset comprising Italian crossword clue-answer pairs, on the other hand, contributions to the field by proposing a system that uses LLMs to generate high-quality educational crosswords. Our approach includes fine-tuning, zero/few-shot learning, and prompt engineering to generate clues from text and keywords. To ensure quality, we developed a set of models to filter out undesirable clues. We additionally employ an algorithm to create educational crossword schema. The resulting system can generate and filter crossword clues, creating educational crosswords with the generated clue-answer pairs.
Additionally, an important improvement in the educational characterization of crossword puzzles is given by a linguistic study of crossword language. Different linguistic skills come into play when solving crosswords, so designing clues to specifically activate certain linguistic abilities is intriguing from an educational standpoint. To pursue this aim, a preliminary study is carried out to highlight the properties of crossword language and which, if any, linguistic structures are the most widespread. The proposed linguistic study also represents a useful contribution for automatically generating crosswords. The results obtained from the categorization of crossword clues can be also viewed as guidelines to create AI-generated crosswords that are increasingly similar to human-generated puzzles and to evaluate LLMs' capabilities of exploiting the linguistic

richness shown by human-designed crosswords to possibly create ad-hoc clue-answer pairs depending on the solver's needs.
The paper's organization is as follows: Section Two provides a comprehensive review of relevant work, and Section Three outlines the dataset used in this study. In Section Four, we detail our investigation's approach, followed by the presentation of our test findings in Section Five. Finally, Section Six concludes this study, highlighting its implications and potential future directions.

2. Related works

The art of crafting crossword puzzle clues has been a puzzle in itself, prompting diverse strategies to tackle the challenge. Traditional methods often lean on well-established dictionaries, thesauri, or language analysis of web-retrieved texts to define clues (Rigutini et al. 2008), (Rigutini et al. 2012). However, in a groundbreaking leap forward, Rigutini and colleagues unveiled the world's first fully automated crossword generator in 2008. Embracing the realm of natural language processing and machine learning, their innovative system autonomously generated crossword puzzle clues. The approach involved web crawling for documents, extracting word meanings, and utilizing techniques like part-of-speech tagging, dependency parsing, WordNet-based similarity measures, and classification models to rank clues by relevance, uniqueness, and readability.
Taking another path, (Ranaivo-Malançon et al. 2013) proposed an NLP-driven method for constructing crossword puzzles. They commenced by assembling a collection of texts related to the puzzle's theme. Subsequently, four critical components were built: pre-processing, candidate generation, clue production, and answer selection, altogether orchestrating a comprehensive and captivating crossword puzzle.
Venturing into the realm of Spanish language puzzles, (Esteche et al. 2017) explored extracting definitions from news articles to craft crossword puzzles. They employed a two-stage process: first, identifying crucial words and phrases and extracting their meanings from a trustworthy online dictionary, followed by utilizing those definitions as clues to construct engaging crosswords.
In another linguistic context, (Arora and Kumar 2019) presented SEEKH, a software application employing natural language processing to extract keywords and craft crossword puzzles in a multitude of Indian languages. Combining statistical and linguistic tools, SEEKH adeptly pinpointed essential keywords, bringing to life a medley of crosswords across linguistic landscapes.
Recently, the works of Zeinalipour *et al.* (Zeinalipour et al. 2023c, 2023b, 2023a) demonstrated a transition from manually crafted crossword puzzles to the use of pre-trained large language models (LLMs). These LLMs were employed to generate puzzles in English, Arabic, and Italian, underscoring the significant impact of computational linguistics in producing culturally diverse puzzles. On a related note, Zugarini *et al.* (Zugarini et al. 2024) proposed a method for developing an educational crossword clues dataset specifically for the English language. However, the task of creating educational crossword puzzles in Turkish remains unexplored.
Despite extensive research efforts, effectively producing comprehensive and distinctive sets of clues and answers from linguistic corpora remains a formidable challenge, especially when dealing with the nuanced intricacies of the Italian language. To tackle these challenges head-on, we present an innovative methodology utilizing Language Models (LLMs) to craft sophisticated educational clues. Representing a pioneering endeavor, our approach successfully generates Italian educational crossword puzzles, addressing

a void that previous methods have left unattended. By creating intellectually stimulating and original crossword puzzles, this novel technique enriches learners' profound comprehension of the subjects through detailed and encompassing answers. Therefore, our proposed work not only introduces novelty to the realm of Italian crossword generation but also provides a groundbreaking solution within the domain of educational tools.

3. Dataset

To fine-tune the LLMs, we leveraged a comprehensive collection of Italian crossword clues and answers. The sources of the clues-answer pairs are both internet sites that release solutions for crossword clues as `https://www.dizy.com/` and `https://www.cruciverba.it/` that we scraped through apposite scripts. And also *pdf* versions of famous Italian crossword papers like *Settimana Enigmistica* and *Repubblica*, that we suitably converted to clue-answer pairs. The various sources were then cleaned and merged and the duplicates were removed. We intend to release this dataset with the support of this paper. This dataset consists of 125,600 entries that correspond to unique clue-answer pairs. It included clues related to different domains, such as history, geography, literature, and pop culture. The dataset under investigation contains a diverse array of linguistic features, including grammatical structures, syntactic patterns, and lexical elements.

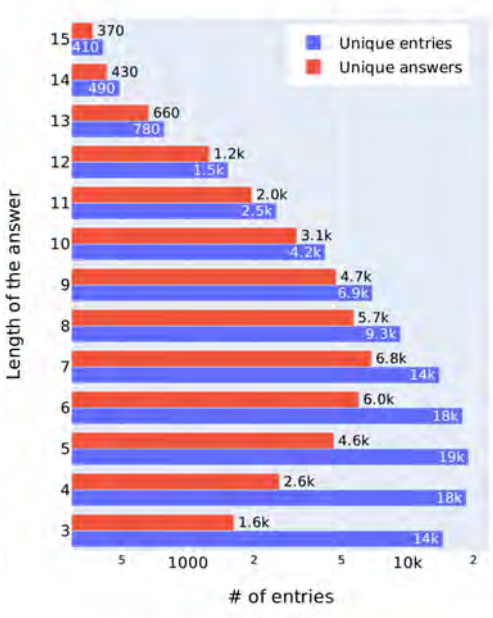


Figure 1
Distribution of the database entries by answer length, in blue the unique answer-clue pairs and in red the unique answers.

A recurring structural pattern in the dataset is the usage of the phrase "known for" or "used for" to define a particular place or object. For example, the definition of a certain location might be "a place known for its historical significance" or "an

object used for a specific purpose." In both cases, the answer is a specific instance of the category described in the definition. Moreover, the dataset includes instances where the definition employs clever wordplay or exploits general category definitions to arrive at a specific answer. For example, “In the middle of the Lake" might elicit the response “AK", while “An exotic legume" could be answered with “SOY" by virtue of its membership in the broader category of legumes. In figure 1 you can further go into detail regarding the distribution of the data divided by the length of the answers. Shorter answers tend to have more clues associated while as the answer gets longer the number of clues diminishes in proportion. One of the primary goals of this study was to establish the groundwork for future research by making the processed dataset publicly accessible, with the aim of encouraging other scholars to contribute to this field."[1]

3.1 Linguistic Analysis

A realistic idea of the clue’s linguistic structure emerges from an analysis of the Italian clue-answer pairs dataset introduced in 3, which can be carried on from different perspectives. We will propose a syntactic categorization of clue-answer pairs which allows us to investigate what kinds of structural movements and transformations can be applied to derive crossword clues. This could represent an interesting chance also for educational crosswords. As we’ve seen, the cryptic register is not required here, but playing with the syntax of clues could be useful to test psycholinguistic hypotheses on groups of participants, which can be chosen on the basis of their age, abilities, native speakers or L2 speakers, and even patients with precise cognitive pathologies or impairments.
A first general differentiation, which is quite trivial, regards non-clausal and clausal clues, based on the presence or absence of an inflected verb. Following this first distinction, it is possible to identify specific patterns, categorize clues depending on them, and analyze their distribution in a dataset of Italian crosswords. This would give us an idea of which structures characterize clues.
Non-clausal clues can be articulated in different structures, distinguished by the nature of their root node: noun phrases (NP), determiner phrases (DP), prepositional phrases (PP), adjectival phrases (AdjP), and adverbial phrases (AdvP).
Clausal clues represent syntactically relevant items in virtue of the presence of an inflected verb and they can be categorized on the basis of their matrix verb, indeed clausal clues include copular sentences, clauses with verbal predicates, and relative clauses. These main categories differentiate internally, and some subcategories can be extracted. We will see that, in our dataset, relative clauses represent a feature and not a category by themselves.
Starting from this first characterization an essential difference stands out: clausal clues are usually truncated sentences, with the omission of the answer, whereas non-clausal clues just define the answer in many different ways. There could be also cases of non-clausal clues in which some material is omitted, for instance in fill-the-blanks clues. In particular, clausal clues can be traced back to the original complete structure, which undergoes varying degrees of elaboration. In all cases, the determiner (if present) of the answer word is deleted. The amount of modification in the derivation of the elliptical construction is signaled by the presence of clitics, pronouns and inversions.

1 The dataset is available at `https://huggingface.co/datasets/Kamyar-zeinalipour/ITA_CW`

To shed light on the possible derivations we must refer to some examples:

- <è simbolo di fedeltà, edera>, complete sentence <l'edera è simbolo di fedeltà> which undergoes DP/subject omission;
- <lo è il salto del circense, acrobatico>, complete sentence <Il salto del circense è acrobatico> which undergoes predicate "acrobatico" omission, inversion <predicate copula subject> and clitic "lo" insertion in predicate position.

With respect to relatives, they usually can have a non-clausal counterpart. Indeed, subordinate clauses are frequently used to define the answer by specifying its hypernym. The same could be done in non-clausal clues with modifiers or adjuncts (adjectives, prepositional phrases) but also with a simple bare noun phrase, in a synonymy relation with the answer:

- <La scienza che utilizza il calcolo delle probabilità, statistica> can be rephrased as a nominal clue like <La scienza probabilistica> or <La scienza delle probabilità>.

A more detailed study results from the observation of the reference dataset. First of all a qualitative data analysis has been carried out following a syntactic approach to define what structures are used in crossword clues. For this first general investigation, Regular Expressions (RegEx) and Part-of-Speech (PoS) tagging have been employed to extract examples of different syntactic formulations and see whether their distribution was significant or not. RegEx represents the most superficial approach to our aim. Indeed, proceeding by selecting clues that contain explicitly specified characters is useful for a first observation of the dataset and of the possible combinations between elements like (copular + prepositional phrase or determiner phrase + relative clause). Improving this approach with PoS tagging results in a pretty accurate distinction between typologies which was used as a reference to define the interesting categories and to refine the classification on a syntactic level. The quality of PoS tagging in extracting clues, however, depends a lot on the tagger's accuracy and requires a detailed specification of the sequence of items that we want to select, given that the hierarchical syntactic structure is not considered here. These first two methods represent a useful way of approaching a categorization of an unlabelled dataset, a starting point for further fine-grained classifications depending on the analysis goal. However, the proposed analysis focuses on the clue syntax from the perspective of generative grammar, therefore, not only the clue PoS tagging is relevant but, more than this, the hierarchical structure needs to be taken into account, which is frequently (perhaps willingly) counterintuitive in the clue formulation.

Once some relevant syntactic structures have been outlined, the extraction has been improved using the Python library spaCy (Honnibal and Montani 2017), which, given a sentence, returns the corresponding syntactic structure in terms of dependency grammar. It is important to highlight that the proposed categorization is based on the generative grammar approach thus we proceed by thinking of specific rules that work around the difference between the parser and our theoretical categorization. This allowed us to extract categories based on the clue hierarchical structure and to define and extract mutually exclusive categories. Indeed, macro-categories have been

identified by looking at the root node of each sentence (considering the entire structure with both clue and answer) and, by gradually subdividing data, additional features have been observed and classified as micro-categories. In other words, bareNP and DP represent nominal macro-categories, which are then further defined by looking at their root node NP or DP and among DP, at the article type: definite or indefinite. The `nlp` function has been applied to all the clues in our dataset to obtain parsed clues. After the parsing, it was possible to extract the root node of each clue and to proceed by classifying clues as 'clausal', 'nominal' etc. For example, a clue is classified as clausal if it contains a verb in finite form in its root node.

Code 1
This code will select the clue if it contains a verb in finite form.

```
if token.pos_ -- 'VERB'
   and token.dep_ -- 'ROOT'
   and "VerbForm-Fin" in token.morph:
        return True
```

Further specifications can be made on the kind of verb like copula, auxiliary, active or passive, reflexive, by adding another if statement with the condition that we want to categorize, e.g. 'passive'.

Code 2
This code will select the clues having an auxiliary in passive voice as root node.

```
def passive(doc):
    for token in doc:
        if token.dep_ -- 'aux:pass'
           and token.head.dep_ -- 'ROOT':
                return True
```

Keeping the first distinction between non-clausal and clausal clues, specific macro categories were created.
To distinguish the different structures, a number of features have been defined and associated with a value 'true' or 'false' depending on whether they were present or not in the clue. In other words, a nominal clue can be of type NP or DP. The determiner in DP can be morphologically defined as definite or indefinite, which represents an interesting distinction in defining words for logical reason: a definite article identifies a unique entity, while an indefinite DP indicates one entity among a set or a set itself, without further restriction for a specific item. Therefore, the features useful for this categorization are:

- article: 'true' in DP and 'false' in NP;
- definite: 'true' in defDP and 'false' in indDP.

A nominal clue of type defDP is selected through this rule:

- clausal=False;
- nominal=True;
- art=True;
- definite=True.

where 'nominal=True' selects all the clues having a noun as root node, while 'art=True' and 'definite=True' are the clues in which child nodes are morphologically defined as article and definite. Another category was derived by adding additional features to the previous rule:

- clausal=False;
- nominal=True;
- art=True;
- definite=True;
- relCl=True.

which extracts all the DP followed by a relative clause. The feature 'relCl' corresponds to the presence of a relative pronoun in the structure, as a child node of type 'relative pronoun'.

To build efficient rules, one must first know how the nlp function of spaCy understands and displays syntactic dependencies. Consequently, the meaning of such rules is to be understood internally in this language. The parser used in spaCy is based on dependency grammar; thus, specific functions have been created with the aim of adapting it to our approach based on generative grammar. Thanks to this approach only 1,58% of the pairs remain uncategorized (category 'other'). Specific distribution percentages are shown in Figure 3.

	clausal	cop	act	pass	imp_refl	inf	subj	cl	pron	nom	art	def	relCl	prep	adj	adv	pronom
cop:missSubj	✓	✓					✗	✗									
cop:clitic	✓	✓					✓	✓									
cop:pron	✓	✓					✓	✗									
act:missSubj	✓	✗	✓			✗		✗	✗								
act:clitic	✓	✗	✓			✗		✓									
act:pron	✓	✗	✓			✗	✓	✗	✓								
pass:missSubj	✓	✗		✓		✗	✗	✗									
pass:other	✓	✗		✓		✗	✓										
imp_refl:missSubj	✓	✗			✓	✗	✗	✗									
imp_refl:other	✓	✗			✓	✗	✓										
inf_VP	✓					✓											
bare_NP	✗									✓	✗		✗	✗	✗	✗	
bare_NP:rel	✗									✓	✗		✗	✗	✗	✗	
def_DP	✗									✓	✓	✓	✗	✗	✗	✗	
def_DP:rel	✗									✓	✓	✓	✓	✗	✗	✗	
ind_DP	✗									✓	✓	✗		✗	✗	✗	
PP	✗													✓			
adjP	✗							✗	✗						✓		
adjP:pron	✗							✗	✓						✓		
advP	✗															✓	
bare_Det	✗																✓

Figure 2
Feature combinations used for clues categorization

As we can see in Figure 3, the most used structure is the bare noun phrase (bare_NP).

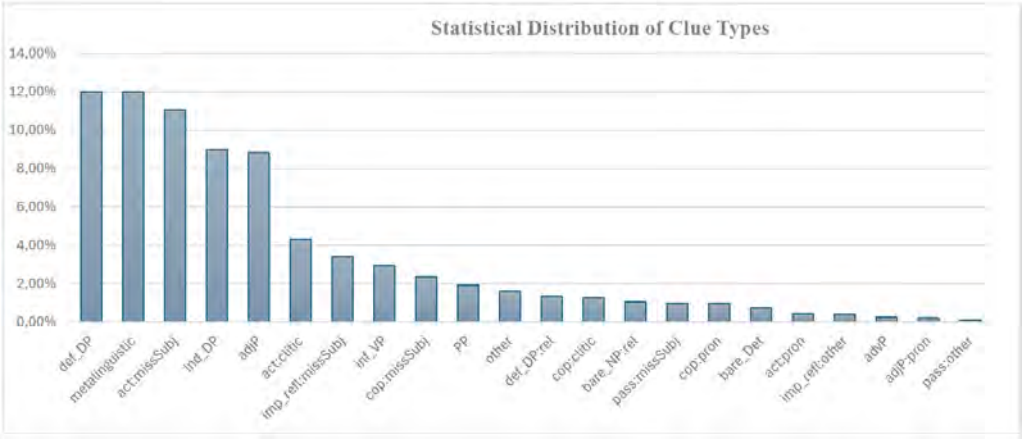


Figure 3
Statistical Distribution of Clue Types in the dataset.

A short description with examples is provided for each extracted category.

bare_NP Bare noun phrases without determiner, represent 23,05 % of the clues. The category includes simple noun phrases, lists of nouns, proper nouns, NP with one or more PP, or adjectives in child nodes. The wide use of bare noun phrases in crossword clues can be explained by the fact that they can be used to define from very common to more specific nouns depending on the number of complements added in the structures. The variety of additional elements constitutes a way of increasing complexity and creating ambiguity.

1. Donne da sole = abbronzate
2. Infuso paglierino = tè.
3. Bevanda ambrata ottenuta per infusione = tè.

In (1) we have a $[_{NP}$ N$[_{PP}$ P N/Adj]] with a lexical ambiguity on 'sole' (it could either be an adjective or a noun). For the very same answer, "tè", we can find different NP structures whose meaning is very similar: (2) has a very simple structure but is very general; (3) has a more complex structure, with adjectives and a prepositional phrase which help narrow the search. The relation between the clue and the answer changes depending on the structure of the clue, i.e. a bareNP with a preposition can highlight a trait of the answer while a list of nouns is usually in a synonymy relation as Stoino, zerbino = tappetino.

def_DP and ind_DP DefiniteDP and indefiniteDP represent respectively 12,01% and 8,98% of the clues. The main difference between these two types is the article meaning, as explained above.

4. Il conto delle spese da farsi = preventivo

5. Una brutta abitudine perdonabile = vizietto

Examples of the kind [$_{DP}$ D N] are rarely found, while more complex DP with adjectives or prepositional phrases is preferred. This could be expected given the role of the clue in guiding the search. PP or adjP function as complements added in the structure to specify the noun in the DP, which alone could be too general and useless to guess the answer. Even considering this aspect, the average word count for these two types of clues is of 4,42 words for defDP and 4,13 for indDP, which means that complements are present but in a reasonable amount, to avoid trivial definitions.

bare_NP:rel and def_DP:rel The semantic role of complements like PP or adjP is sometimes fulfilled by relative clauses, which are found in bareNP or defDP. These structures represent the 1,04% for bareNP and 1,33% for defDP.

6. Cilindri commestibili che vengono affettati = polpettoni
7. Lo Stato di cui fanno parte le Isole Azzorre = Portogallo

metalinguistic Metalinguistic clues (that can be both nominal or clausal) constitute 11,99% of the clues. These clues refer to answers, composed of 2 letters, which are already given in the clue, usually inside a word or as a common trait between two words. Such clues are used to fill empty spaces in the grid, resulting from other answers, but their style and meaning certainly contribute to rendering the puzzles more engaging. The formulation is characterized by a certain amount of ambiguity, the correct interpretation is not straightforward, and the intended meaning derives from the crossword context. Indeed, they require the solver to reason on the orthographic form of a certain word, rather than its meaning, and to extract two letters based on their position inside the word. For example, in 8, the solution consists of the two letters located in the center of the word "Matera".

8. Il centro di Matera = TE

act:missSubj and pass:missSubj The largest clausal typology represents 11,05% and consists of clauses with an active verbal predicate, missing the subject, which is used as the answer (through ellipsis). The corresponding typology in passive form represents only 0,98% of the data.

9. Risiede in uno spazio geografico determinato = abitante (active)
10. È detta Il Continente Bianco = Antartide (passive)

pass:other Passive sentences, where the missing element is not the subject, constitute only 0,09% of the entire dataset.

11. Vi furono ritrovati noti bronzi = Riace

act:clitic The second most numerous clausal category is represented by clauses with an active verbal predicate and a clitic pronoun replacing a nominal constituent different from the subject. In these clues, the solution is the referent of the clitic pronoun.

12. La segue il medico = ammalata
13. Ne fanno parte battesimo e cresima = sacramenti

This category represents 4,32% of the clues, thus less than half of 'active missing subject', but given the whole distribution, 'active missing object' still represents a relevant kind of clue.

imp_refl: missSubj 3,41% represents a clausal structure with a reflexive verb or an impersonal pronoun, missing the subject:

14. Si reca spesso al catasto = geometra
15. Si lascia al telefono dopo il bip = messaggio

imp_refl:other Other structures with impersonal pronouns or reflexive verbs constitute only 0,38% of the clues. Some examples are (16) relative clauses without a NP or DP as the root node and (17) with the pronoun in subject position:

16. Che si riferisce all'Università = accademico;
17. Alcune si scatenano a primavera = allergie or Quello Flavio si trova a Roma = anfiteatro.

act:pron Among active clauses with verbal predicates, we find 0,41% of structures with a full pronoun inside the subject or object phrase:

18. Quelli d'America hanno per capitale Washington = Stati uniti
19. Wagner compose quella delle Valchirie = cavalcata

This category includes also relative clauses, characterized by a relative pronoun as subject:

20. Chi la subisce ha perso = sconfitta
21. Chi la dà prende parte = adesione.

Clausal clues also include copular sentences:

cop:missSubj Copular structures without the subject are the most widespread, representing 2,35% of the entire dataset, and they can be characterized by different features like adverbs (23), a modal verb (24) or a conditional clause (25).

22. Fu Cancelliere della Germania dal 1949 al 1963 = Adenauer
23. Di solito è reciproca = antipatia or A volte sono insanabili = contrasti or Non è analogo all' analogico = digitale
24. Possono essere fumogeni = candelotti
25. Se è rapida è brillante = carriera

cop:clitic The presence of a clitic pronoun in copular structures signals that the solution is either the predicate or some other constituent different from the subject. 1,26% of the dataset are copular structures of this kind.

26. Venere ne era la dea = bellezza (indirect object della bellezza is replaced by the clitic ne)
27. Lo è la falsa bionda = ossigenata (the clitic replaces the predicate)

cop:pron 0,97% of the clues are copular sentences containing a full pronoun. This typically happens when the pronoun, substituting the solution, targets a position inside the subject or the predicate phrase, but not the entire phrase. Cases like (28) are apparent counterexamples, but a closer inspection suggests that the pronoun may be heading some silent partitive structure (*Una (anidride/delle anidridi) è carbonica*).

28. È celebre quella di Trinità dei Monti = scalinata
29. Una è carbonica = anidride
30. Il cielo è il suo mondo = astronomo (Pronoun suo replaces dell'astronomo)
31. In Italia la più lunga è quella del sole = autostrada (Pronoun quella refers to the elliptical answer in subject position l'autostrada più lunga)

As in all clausal categories, finding the answer means reconstructing the complete sentence, thus the answer always represents the omitted argument of the matrix verb.

adjP 8,82% of the clues are adjectival phrases, which stands in a similar relation with the answer to bareNP: lists of answers, definitions of the answers at different levels depending on which and how many complements are added.

inf_VP and advP The same function is performed also by infinitive verbs and adverbial phrases, which are way less frequent than adjP and NP, constituting respectively the 2,93% and only the 0,23% of the clues.

32. Puniti o morigerati = castigati (adjP)
33. Investire di un grado = nominare (infVP)
34. Lentamente = adagio (advP)

In these three typologies, clue and answer are mainly related through synonymy. Indeed, we can observe a direct correspondence between the PoS of the answer and the clue root node. In non-clausal clues the answer PoS depends on the relation holding with the clue: if the answer is a synonym of the clue, then the clue and answer belong to the same PoS; if the clue speaks of some trait of the answer or some characteristic of it, then there will be more variability in the answer type.

PP This is the case of prepositional clues that represent 1,92% of the data.

bare_Det The unique instance of pronominal clues, 'bare_Det', constitutes only 0,70% of the dataset and includes all the clues having a root node morphologically defined as a pronoun. Different structures can be encountered in this category, such as demonstrative pronoun followed by a relative clause (35); indefinite article and relative clause (36); or a relative clause, i.e. in which the subject is a relative pronoun (37).

35. Quel che si può fare davanti alle evidenze = arrendersi
36. Uno che vuole salire in alto = arrivista
37. Chi abiura la propria religione = apostata

adjP:pron As we've seen, pronouns are found very frequently in clues, clearly referring to the answer. 0,19% of clues have an adjectival phrase as a root and contain a pronoun in the structure, like Pittoresco quello siciliano = carretto. Interestingly, these clues differ from simple adjectival phrases in that the solution is not a synonym of the clue, but it is an entity to which the adjective applies. In this respect, they are more similar to clues in the cop:pron category.

In general, as Figure 3 displays, non-clausal clues are preferred to clauses, and among verbal clues, verbal predicates are preferred to copular verbs. Active sentences are preferred to passives. However, typologies distribute homogeneously. The highest peak is 23,05% so still less than half of the dataset. From a syntactic analysis, there is ample scope for creativity from a structural perspective.

4. Methodology

The system extracts clue-answer pairs from provided texts (path (a) of Figure 4), or generates clues based on given keywords (path (b) of Figure 4). As input texts, we use paragraphs selected from Wikipedia pages on educational topics like science, geography, and economics. Using this type of text allows us to create direct clues like definitions, appropriate for educational usage. The system evaluates the quality of the generated clue-answer pairs using various validators. Following the generation process, users are granted the opportunity to review all the produced clue-answer pairs and select their preferred combinations. These selected pairs are then utilized by the final component of the system to generate the crossword puzzle schema.

In this segment, we will delve into the system's fundamental aspects, encompassing three essential components: the generation and validation of clue-answer pairs from provided text, the creation of clues based on given keywords, the validation of the result, and lastly, the generation of the crossword puzzle layout or schema.

4.1 Path (a)

In this section, we analyze the path (a) of Figure 4. We used a multi-step process to apply zero-shot and few-shot learning techniques to text. First, we divided the text into paragraphs and extracted precise keywords. Then, we created personalized clues inspired by the original text using those keywords. To ensure high quality, we

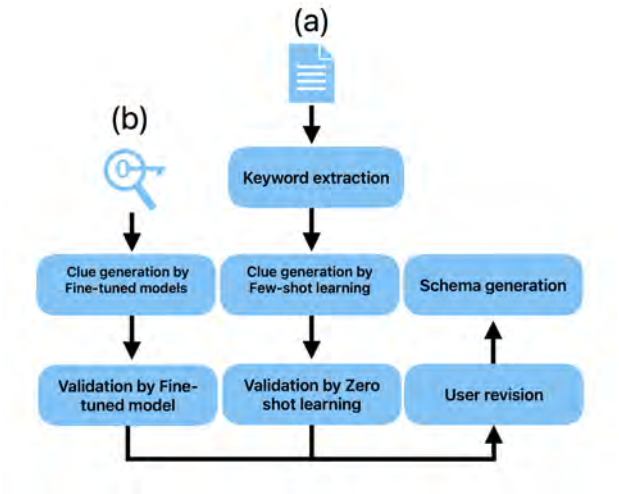


Figure 4
Overall System Architecture

thoroughly validated the generated clue-answer pairs. Our primary tool was the GPT-3 DaVinci base model (Brown et al. 2020). We'll explore each step in detail in the following.

Keyword extraction: Our innovative strategy harnesses the power of zero-shot learning for an approach to our task. We meticulously craft two prompts in both Italian and English, ensuring they are well-structured with clear objectives and detailed steps to achieve them. You can access it in the appendix under the section labeled Prompts 1 and 4. This thoughtful design empowers the Language Model (LLM) to precisely extract the most relevant keywords, capitalizing on its robust zero-shot learning capabilities. By providing guidance through our prompts, we optimize the model's ability to understand and respond to the intricacies of the task at hand.

Clue generation: We use a few-shot learning approach to create compelling crossword clues for each identified keyword in the paragraph. By leveraging an example educational text, crossword keywords, and valid clue examples, we empower the Language Model (LLM) to craft meaningful clues. We presented the paragraph and extracted clues as prompts to the LLM, allowing it to generate clues based on the provided text and keywords. This technique ensures precise and contextually relevant clues. We crafted prompts in both Italian and English, similar to the previous section. Two distinct types of prompts were developed, and all of them are accessible in the Appendix under Prompts 2 and 5.

Validation: We improved the quality of generated keywords and clues by implementing a multi-stage filtering process. First, we filtered out long keywords (over 3 words) as they were less suitable for crossword puzzle answers. Some generated clues inaccurately described their corresponding keywords, and some were hallucinations from the provided text. To address this, we used zero-shot learning to identify and

filter out unwanted clues, resulting in a significant improvement in the final output. We created Italian and English prompts, akin to the previous section. Both prompt types can be found in the Appendix under Prompts 3 and 6.

4.2 Path (b)

Referring to pipeline (b) of Figure 4; addressing situations where users lack access to the original text and wish to generate crossword clues solely from given answers, we devised an approach to cater to this scenario. Our strategy encompassed multiple stages, each contributing to the overall effectiveness of the solution. Initially, we focused on fine-tuning various language models specifically tailored for this unique task. Leveraging the data generated from these fine-tuned models, we then proceeded to create diverse classifiers. These classifiers were carefully designed with the primary objective of distinguishing high-quality clue-answer pairs from those that were deemed less suitable.

Fine-tuned models: In the pursuit of generating crossword clues from given answers, we undertook various fine-tuning processes of language models, using data collected from Section 3. Our selection of models comprised GPT3-DaVinci (175B parameters) and GPT3-Curie (13B parameters). GPT3-DaVinci, with its vast parameter count, demonstrated unmatched depth, enabling it to uncover intricate patterns and craft nuanced clues. On the other hand, GPT3-Curie, while slightly smaller, proved remarkable in grasping language subtleties, further enhancing the fine-tuning process (Brown et al. 2020). In our fine-tuning process, we employ a distinctive approach by inputting the answer and tasking the model to generate the corresponding crossword clue. This iterative method not only refines the model's ability to comprehend context but also hones its skill in crafting clues that are both challenging and contextually fitting. By continually providing the answer as input during fine-tuning, we guide the model toward a nuanced understanding of how to construct clues that align seamlessly with the given solution. This tailored training methodology further enhances the model's proficiency in delivering accurate and engaging crossword clues, solidifying its role as a versatile and effective tool in the clue-generation process.

Validation: We developed different strong classifiers using fine-tuned language models to distinguish good crossword clues from poorly crafted ones since not all generated clues fit the given answers perfectly. In pursuit of this goal, we fine-tuned several models, each boasting unique capacities: GPT3-DaVinci (175B parameters), GPT3-Curie (13B parameters), GPT3-Babbage (1.3B parameters), GPT3-Ada (350M parameters) (Brown et al. 2020), and BERT-uncased-base (110M parameters) (Raffel et al. 2020). By harnessing the collective power of these models, each with varying parameter counts, we gained a comprehensive perspective on their effectiveness in filtering and validating the generated clues. Through this approach, our goal was to ensure that only high-quality and contextually relevant crossword clues remained, thereby elevating the overall accuracy and usability of our system.

4.3 Educational Crossword Schema Generator

Our algorithm for creating educational crosswords takes input such as answer lists, work area dimensions, and stopping criteria. It starts by randomly placing a central answer, and then adds other answers nearby. The algorithm iteratively adds answers, sometimes removing recent ones or restarting. The best solution is selected based on a global score of the generated schemes. Each solution produced is evaluated using the following formula:

$$\text{Score} = (\text{FW} + 0.5 \cdot \text{LL}) \cdot \text{FR} \cdot \text{LR}$$

where FW (Filled Words) is the number of words added; LL (Linked Letters) is the number of letters that belong to two crossing words; FR (Filled Ratio) is the number of total letters divided by the minimum rectangle area used; and LR (Linked Letters Ratio) is the Linked Letters (LL) divided by the number of total letters. The algorithm incorporates various stopping criteria, including the minimum number of answers added to the grid; reaching the threshold of minimum Filled Ratio; the limit on the number of times the grid is rebuilt from scratch, and the maximum time duration. The solution with the highest score is deemed the best. These stopping criteria play a crucial role in guiding the algorithm's decision-making process, determining when to conclude the crossword construction. Through the establishment of thresholds and limitations, we successfully ensure the efficient and effective generation of crosswords. Within the filling process, we have the option to designate a list of "preferred answers." The algorithm places a higher priority on selecting answers from this list, increasing the probability of their incorporation into the grid.

5. Experiments

The experimental evaluation of the designed system is presented in this section, focusing on the individual components and their roles in the overall framework. The system's performance is thoroughly analyzed to assess its effectiveness and efficiency, providing insights into its strengths and weaknesses.

5.1 Experimental Evaluation: Path (a)

In our experiments, we observed variations in model output quality when altering the language of the prompts. To demonstrate this, we conducted two sets of experiments using two types of prompts: one in English and the other in Italian. Our system underwent a rigorous evaluation process using 50 paragraphs sourced from Wikipedia to assess the performance of each component using Italian and English. Human supervision was employed, with the annotation performed by a native Italian speaker holding a degree in linguistics and guidelines for evaluation can be found in Appendix A. The results of these evaluations are summarized in Table.
Initially, our focus was on keyword extraction, and we achieved promising results in our experiments. Specifically, employing the zero-shot learning approach, we obtained 79.73% and 75.60% accuracy in generating suitable keywords for crossword clues using Italian and English prompts, respectively. Subsequently, we subjected the clue-generation process to human evaluation and found that, with Italian and English prompts, 68.34% and 76.70% of the generated clues were considered acceptable, respectively. To ensure the validity of our results, we employed various approaches

outlined in Section 4.1. Through this validation, we were able to identify 56.76% and 69.72% of the unacceptable clue-answer pairs generated using the Italian and English prompts, respectively. These results clearly demonstrate the effectiveness of our system in producing satisfactory crossword clues based on the evaluated text.

Table 1
Assessment outcomes of the clue-answer pairs generated from the provided Text.

System part	*Italian Prompt*	*English Prompt*
Acceptable keywords	79.73 %	75.60%
Acceptable clues	68.34 %	76.70 %
Validator performance	56.76 %	69.72 %

Figure 5 demonstrates the step-by-step process of generating crossword clue-answer pairs from input text. The image shows the various stages, such as keyword extraction, clue creation, and pair validation, and illustrates how our system converts input text into pertinent crossword clues. The results with the Italian data revealed that, when the prompt is in English, the performance of the model is better than when the prompt is in Italian.

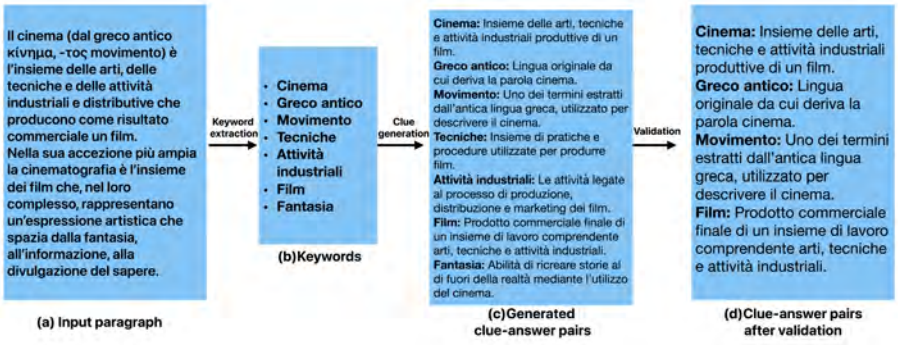


Figure 5
A concrete example of the path (a)

5.2 Experimental Evaluation: Path (b)

This section delves into our experimental endeavors on generating and validating clues from keywords. Building upon the insights presented in Section 4.2, we developed and fine-tuned two distinct models, GPT-3 DaVinci and GPT-3 Curie, specifically focused on generating clues from given keywords. The models were trained using a batch size of 16, with a learning rate of 0.01, over three epochs, without using any additional prompts for fine-tuning. Due to cost constraints, we utilized a subset of our dataset for training, specifically encompassing 50,000 unique clue-answer pairs, rather than the entire dataset. The output format followed the structure: `"ANSWER:[ANSWER] \n CLUE:[CLUE]"`.
Once the fine-tuning phase concluded, we generated 4,000 clues from each of the fine-

tuned models and subjected them to human evaluation using the guidelines provided in Appendix A. The outcomes of this evaluation are summarized in Table 2. Remarkably, GPT-3 DaVinci outperformed GPT-3 Curie, yielding an impressive 60.1% of acceptable clues compared to Curie's 34.9%. To gain deeper insights into the quality of the gen-

Table 2
Assessment outcomes of the clues generated from the provided keyword.

Model	*% of acceptable clues*
GPT3-DaVinci	60.1
GPT3-Curie	34.9

erated clues, we meticulously assembled a collection of acceptable and unacceptable clues. These were randomly sampled from the human-supervised label dataset, offering a diverse clue for each answer. Please consult Table 3 (refer to table 1 in the Appendix for translation). This detailed analysis helps us evaluate the quality and suitability of the clues for creating engaging crossword puzzles.

We developed multiple classifiers that integrate different language models to differen-

Table 3
Acceptable and unacceptable clues from given keywords using various models.

Clue-Answer pair	*Model*	*Accepted*
Mitologia: La conosce chi conosce i miti	DaVinci	Yes
Elettricità: Uno dei segni zodiacali	DaVinci	No
Curiosità: Il desiderio di sapere	Curie	Yes
Collaborazione: Lo si raggiunge con chiunque	Curie	No

tiate between acceptable and unacceptable clue-answer pairs. The result of the analysis on the test set is shown in Table 4. We utilized a dataset of 6,000 human evaluations from the previous step to construct various classifiers. This is the data which we tried to evaluate GPT-3-Davinci and GPT-3-Curie by human supervision. For training and evaluation, we employed 80% of this data for fine-tuning the classifiers and reserving the remaining 20% for testing the classifiers. Within the dataset, 51% comprised acceptable clues, while the remaining 49% consisted of unacceptable clues.

The evaluation results reveal significant distinctions among the classifiers in their

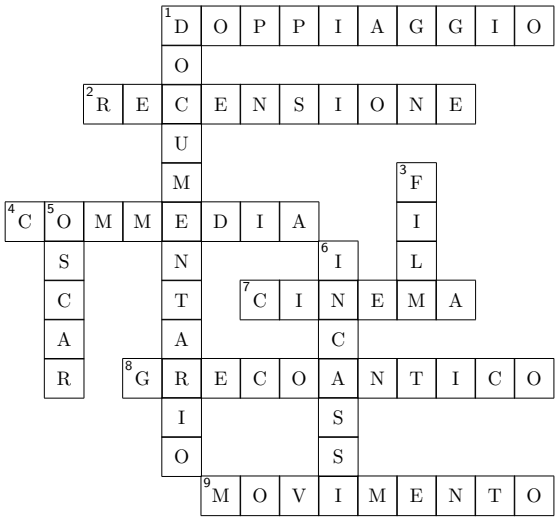
Table 4
Classifier performance on distinguishing acceptable Clue-Answer pairs

Model	*accuracy %*	*precision %*	*recall %*	*F1 Score*
GPT3-DaVinci	79.88	80.16	76.67	0.7838
GPT3-Curie	77.82	78.80	72.99	0.7578
GPT3-Babbage	74.12	72.58	73.25	0.7291
GPT3-Ada	69.17	67.77	67.06	0.6741
BERT-uncased-base	65.62	63.71	64.47	0.6409

ability to differentiate between acceptable and unacceptable clue-answer pairs. Earning the top position, the GPT3-DaVinci model achieved an accuracy of 79.88%, solidifying its role as the most effective classifier in this task. Following closely, the GPT3-Curie base model attained a commendable 77.82% accuracy. The GPT3-Babbage model demonstrated respectable performance with 74.12% accuracy, while GPT3-Ada and BERT-uncased achieved accuracies of 69.17% and 65.62%, respectively.

5.3 Schema Generation

Our schema generation algorithm creates educational crosswords with diverse layouts using a single batch of words. Below is an illustration, check the Figure 6 of a comprehensive Italian educational crossword about movies produced with our system. The clue-answer pairs are both extracted from a text (path (a), see Figure 5) and generated directly from a keyword (path (b), contr-assigned with a ⋆ below). Since our focus is on creating educational crosswords on specific topics, the resulting puzzles may not be fully interconnected like traditional crosswords. This is because constructing a classic crossword requires a large number of intersecting words, which may not be strictly related to the specific topic of interest. To maintain adherence to the selected topic, we build crosswords in an educational style, which requires fewer constraints and allows us to create puzzles where each clue-answer pair is strictly related to the theme.



Orizzontali 1 ⋆ Traduzione Simultanea 2 ⋆ Una valutazione critica 4 ⋆ Lo è uno spassoso racconto 7 Insieme delle arti, tecniche e attività industriali produttive di un film 8 Lingua originale da cui deriva la parola cinema 9 Uno dei termini estratti dall'antica lingua greca, utilizzato per descrivere il cinema **Verticali** 1 ⋆ Un film... come documento 3 Prodotto commerciale finale di un insieme di lavoro compredente arti, tecniche e attività industriali 5 ⋆ Un premio assai ambito 6 ⋆ Entrano nelle casse del botteghino

Figure 6
An illustrative crossword created using the newly introduced system.

6. Conclusions

In this paper, we present various contributions, including the introduction of a substantial dataset for Italian clue-answer pairs, we developed an innovative system using Large Language Models to generate educational crossword puzzles from given texts or answers. Our approach combines human supervision and specific guidelines to ensure high-quality and relevant clues.
Our system includes a keyword extraction component (79.73% high-quality keywords) and a crossword clue generation component (76.6% relevant and acceptable clues). A validation component filters out unacceptable pairs, achieving a 69.72% detection rate. We conducted an in-depth investigation of fine-tuned generators and classifiers to enhance the quality of clues. Among the models tested, GPT3-Davinci demonstrated exceptional performance in generating clues based on given keywords, producing a remarkable 60.1% of acceptable clues. Moreover, GPT3-Davinci proved to be the most proficient classifier, accurately distinguishing between good clue-answer pairs and unacceptable ones with an impressive 79.88% accuracy.
Our algorithm for generating educational crossword schemes is efficient and produces diverse layouts. This study aims to enhance student skills and promote interactive learning. Educators can integrate our system into their instruction for more effective teaching practices.
Our future works comprise the application of the presented linguistic analyses to the generation of personalized puzzles, varying in terms of the clue's syntactic structure, using various prompts including specification on the clue's structure. Future research involves developing advanced models for direct clue-answer pair generation and exploring specialized models for different clue types. Our vision is to revolutionize educational crossword generation and unlock new innovations in teaching practice.

Acknowledgments
We extend our heartfelt gratitude to Achille Fusco for his invaluable contribution to the linguistic analyses on crossword clues, which have significantly enriched this project. This work was partially supported by the Provincia Autonoma di Trento through the project MAESTRO (CUP: C79J23001170001)[2] and by the FAIR programme (PNRR MUR) through the project ReSpiRA (CUP: B43D22000900004)[3].

Appendix A: Guidelines for Validating Clue-Answer Pairs

In the course of our study, we embraced an enthralling challenge: constructing a classifier capable of discerning between acceptable and non-acceptable crossword clue-answer pairs. Crossword puzzles have held a cherished place as a beloved pastime, demanding a harmonious fusion of linguistic prowess, creative acumen, and adherence to intricate puzzle construction rules to fashion top-tier clue-answer pairs. Our pursuit of creating an automatic evaluator for generated crossword clues and their

2 "MAESTRO - Mitigare le Allucinazioni dei Large Language Models: ESTRazione di informazioni Ottimizzate" a project funded by Provincia Autonoma di Trento with the Lp 6/99 Art. 5:ricerca e sviluppo, PAT/RFS067-05/06/2024-0428372, CUP: C79J23001170001 - https://www.opencup.gov.it/portale/web/opencup/home/progetto/-/cup/C79J23001170001

3 "ReSpiRA - REplicabilità, SPIegabilità e Ragionamento", a project financed by FAIR, Affiliated to spoke no. 2, falling within the PNRR MUR programme, Mission 4, Component 2, Investment 1.3, D.D. No. 341 of 03/15/2022, Project PE0000013, CUP B43D22000900004 - https://www.opencup.gov.it/portale/web/opencup/home/progetto/-/cup/B43D22000900004

corresponding answers holds tremendous potential. This advancement promises to aid puzzle creators, enrich puzzle-solving experiences, and unlock profound insights into the subtle nuances of language and puzzle design. Ultimately, this endeavor not only elevates the world of crossword puzzles but also kindles a deeper appreciation for their linguistic artistry and cognitive allure.

Table 1
Translation of Table 3

Clue-Answer pair	*Model*	*Acc.*
Mythology: It is known by anyone who knows myths	DV	Yes
Electricity: One of the zodiac signs	DV	No
Curiosity: The desire to know	Curie	Yes
Collaboration: One reaches it with anyone	Curie	No

To create a powerful classifier for crossword clue-answer pairs, we must establish a strong and comprehensive guideline that clearly delineates the attributes of acceptable and non-acceptable pairs. This guideline will be the cornerstone for training our classifier, enabling it to discern the defining characteristics that set apart high-quality clues from irrelevant or inappropriate ones. With strict adherence to this guideline, we can guarantee the accuracy of our classifier in assessing the quality of clue-answer pairs, ultimately leading to the creation of more captivating and enjoyable crossword puzzles.
Let us now explore the pivotal components of the guideline, essential for evaluating crossword clue-answer pairs.

Relevance and Cohesion: A top-notch crossword clue-answer pair thrives on a profound and meaningful connection between the clue and the answer. The clue should provide ample context or clever hints that smoothly lead solvers to the intended solution. Simultaneously, the answer must be directly tied to the clue, fitting flawlessly within the puzzle's theme or topic.

Wordplay and Inventiveness: Elevate your crossword clues with ingenuity and wordplay that challenge and delight solvers. Seek clues that encourage lateral thinking, incorporate witty twists, or conceal intriguing meanings. A well-crafted clue-answer pair captures the solver's imagination, transforming the puzzle into an exhilarating journey of discovery.

Clarity and Precision: Precision is key in creating crossword clues. Ensure your clues are crystal clear and unambiguous, presenting solvers with a distinct and precise solution. Avoid any ambiguity that might lead to multiple interpretations or numerous possible answers. The goal is to deliver a single correct solution that aligns perfectly with the clue's intended meaning.

Grammar and Language: Pay meticulous attention to grammar, syntax, and linguistic conventions in both the clue and the answer. Maintain grammatical correctness, coherence, and an appropriate level of complexity for a crossword puzzle.

General Knowledge and Fairness: Strike a balance between challenge and accessibility by grounding your clues in general knowledge or commonly known facts. Avoid overly obscure or specialized references that could alienate solvers. A great clue-answer pair caters to a diverse range of puzzle enthusiasts, offering a fair and engaging experience for all.

Through the adoption of this framework, a robust dataset can be generated, facilitating the development of a dependable classifier that discerns commendable crossword clue-answer pairs from incongruous or inappropriate ones. This transformative classifier holds the promise of revolutionizing crossword puzzle creation, assessment, and solving, offering invaluable revelations into the craft of constructing captivating and mentally stimulating puzzles.

Appendix B: Italian Prompts

Prompt 1
Italian prompt for keyword extraction.

```
prompt = f"""
    Obiettivo: Il tuo compito è estrarre delle parole chiave, descritte nel
    testo proposto. Le parole chiave estratte saranno utilizzate per creare
    brevi definizioni di cruciverba riguardanti il testo da cui sono estratte
    le parole chiave. Le definizioni saranno d'aiuto per trovare la soluzione
    corrispondente e completare il cruciverba.

    Completa l'obiettivo attraverso i seguenti passaggi:

    1- Estrai le parole chiave più importanti del testo.
    2- Controlla le parole chiave: controlla se le parole chiave sono
    descritte e definite nel testo o non sono descritte e definite nel testo.
    3- Parole chiave finali : sulla base del passaggio precedente, rimuovi
    tutte le parole chiave che non sono definite nel testo.

    Utilizza il seguente formato di output:

    Parole chiave: <Parole chiave finali>

    Text: ```{text}```
    """
```

Prompt 2
Italian prompt for clue generation.

```
prompt = f"""
    Genera brevi definizioni di cruciverba per ciascuna delle parole chiave
    fornite: {keywords} sulla base del seguente testo: {text}.

    Completa l'obiettivo attraverso i seguenti passaggi:
    1- Per ciascuna delle parole chiave fornite, trova il passaggio del testo
    contentente l'informazione riguardante la parola chiave.
    2- Genera brevi definizioni: per tutte le parole chiave genera brevi
    definizioni riguardanti il testo. Nella definizione non deve essere
    presente la parola chiave.
    3- Non usare virgolette e apostrofi nell'output.

    Segui questo esempio per completare l'obiettivo:
    "Testo: La scienza è un sistema di conoscenze ottenute attraverso una
    attività di ricerca prevalentemente organizzata con procedimenti metodici
    e rigorosi, coniugando la sperimentazione con ragionamenti logici condotti
    a partire da un insieme di assiomi, tipici delle discipline formali.
    Uno dei primi esempi del loro utilizzo lo si può trovare negli Elementi
    di Euclide, mentre il metodo sperimentale, tipico della scienza moderna,
    venne introdotto da Galileo Galilei, e prevede di controllare
    continuamente che le osservazioni sperimentali siano coerenti con le
    ipotesi e i ragionamenti svolti.

    Parole chiave: conoscenze, ricerca, rigorosi, assiomi, ipotesi, Galileo

    Definizioni:
    Conoscenze: informazioni acquisite tramite ricerca organizzata con
    procedimenti metodici e rigorosi.

    Ricerca: attività organizzata prevalentemente con procedimenti metodici
    e rigorosi finalizzata allottenimento di conoscenze.

    Rigorosi: esatti e precisi nello svolgimento delle azioni.

    Assiomi: un insieme di verità accettate come base dei ragionamenti
    logici.

    Ipotesi: assunte per comprendere le osservazioni sperimentali e testare
    le conoscenze

    Galileo : egli introdusse il metodo sperimentale nel processo di scienza
    moderna.
    "

    """
```

Prompt 3
Italian prompt for auto check.

```
prompt = f"""
    Obiettivo: il tuo obiettivo è controllare se il contenuto di ogni
    definizione è presente o no nel testo proposto Per ciascuna definizione
    scrivi "True" se il contenuto è presente nel testo e "False" se il
    contenuto non è contenuto nel testo.

    Sentences: ```{clue}```

    Text: ```{text}```
    """
```

Appendix C: English Prompts

Prompt 4
English prompt for keyword extraction.

```
prompt = f"""
    Objective: Your task is to extract described keywords in Italian from a
    given Italian text. These keywords will be used to create Italian crossword
    short definitions based on the extracted text. The clues will help Italian
    solvers to find the corresponding answers and complete the puzzle grid.

    Please follow these steps to achieve the objective:

    1- Extract the most important Italian keywords in the Italian text.

    2- Check keywords: check if the Italian keywords are well Explained in
    the given Italian text or not.

    3- Final keywords : Remove all the Italian keywords which are not well
    defined in the Italian text based on the last step.

    Use the following output format:

    Keywords: <Final keywords>

    Text: ```{text}```
    """
```

Prompt 5
English prompt for clue generation.

```
prompt = f"""
    Generate short crossword definitions in Italian for each provided Italian
    keyword: {keywords} based on the following Italian text: {text}.

    Follow these steps to achieve the objective:

    1- For each provided Italian keyword detect the part of the Italian text
    which contains the keyword information.
    2- Generate short definitions in Italian: For all the Italian keywords
    generate short definitions in Italian based on the Italian text, and
    place the correspondent keyword after each generated definition.
    Make sure that the Italian keyword is not present in the correspondent
    definition.
    3- Do not use quotation marks and apostrophes in the output.

    Follow this example to complete the task:
    "Text: La scienza è un sistema di conoscenze ottenute attraverso una
    attività di ricerca prevalentemente organizzata con procedimenti
    metodici e rigorosi, coniugando la sperimentazione con ragionamenti
    logici condotti a partire da un insieme di assiomi, tipici delle
    discipline formali. Uno dei primi esempi del loro utilizzo lo si
    può trovare negli Elementi di Euclide, mentre il metodo sperimentale,
    tipico della scienza moderna, venne introdotto da Galileo Galilei,
    e prevede di controllare continuamente che le osservazioni sperimentali
    siano coerenti con le ipotesi e i ragionamenti svolti.

    Keywords: conoscenze, ricerca, rigorosi, assiomi, ipotesi, Galileo

    Clues:

    Conoscenze: informazioni acquisite tramite ricerca organizzata con
    procedimenti metodici e rigorosi.

    Ricerca: attività organizzata prevalentemente con procedimenti metodici
    e rigorosi finalizzata allottenimento di conoscenze.

    Rigorosi: esatti e precisi nello svolgimento delle azioni.

    Assiomi: un insieme di verità accettate come base dei ragionamenti
    logici.

    Ipotesi: assunte per comprendere le osservazioni sperimentali e testare
    le conoscenze

    Galileo : egli introdusse il metodo sperimentale nel processo di scienza
    moderna.
    "

    """
```

Prompt 6
English prompt for auto check.

```
prompt = f"""
    Objective: Your objective is to check whether each given Italian sentence
    content is present in the provided Italian text or not. Print "True" if
    it is present in the provided Italian text and "False" if it is not
    present in the provided Italian text.

    Sentences: ```{clue}```

    Text: ```{text}```
    """
```

References

Arora, Bhavna and N.S. Kumar. 2019. Automatic keyword extraction and crossword generation tool for indian languages: Seekh. In *2019 IEEE Tenth International Conference on Technology for Education (T4E)*, pages 272–273, Goa, India, December. IEEE.

Bella, Yolanda Dita and Endang Mastuti Rahayu. 2023. The improving of the student's vocabulary achievement through crossword game in the new normal era. *Edunesia: Jurnal Ilmiah Pendidikan*, 4(2):830–842.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Dol, Sunita M. 2017. Gpbl: An effective way to improve critical thinking and problem solving skills in engineering education. *Journal of Engineering Education Transformations*, 30(3):103–13.

Dzulfikri, Dzulfikri. 2016. Application-based crossword puzzles: Players' perception and vocabulary retention. *Studies in English Language and Education*, 3(2):122–133.

Esteche, Jennifer, Romina Romero, Luis Chiruzzo, and Aiala Rosá. 2017. Automatic definition extraction and crossword generation from spanish news text. *CLEI Electronic Journal*, 20(2).

Honnibal, Matthew and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Kaynak, Serap, Sibel Ergün, and Ayşe Karadaş. 2023. The effect of crossword puzzle activity used in distance education on nursing students' problem-solving and clinical decision-making skills: A comparative study. *Nurse Education in Practice*, 69:103618.

Mueller, Shane T. and Elizabeth S. Veinott. 2018. Testing the effectiveness of crossword games on immediate and delayed memory for scientific vocabulary and concepts. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society (CogSci 2018)*, Madison, Wisconsin, USA, July.

Nickerson, RS. 1977. Crossword puzzles and lexical memory. In *Attention and performance VI*. Routledge, pages 699–718.

Orawiwatnakul, Wiwat. 2013. Crossword puzzles as a learning tool for vocabulary development. *Electronic Journal of Research in Education Psychology*, 11(30):413–428.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ranaivo-Malançon, Bali, Terrin Lim, Jacey-Lynn Minoi, and Amelia Jati Robert Jupit. 2013. Automatic generation of fill-in clues and answers from raw texts for crosswords. In *2013 8th International Conference on Information Technology in Asia (CITA)*, pages 1–5, Kuching, Sarawak, Malaysia, July. IEEE.

Rigutini, Leonardo, Michelangelo Diligenti, Marco Maggini, and Marco Gori. 2008. A fully automatic crossword generator. In *2008 Seventh International Conference on Machine Learning and Applications*, pages 362–367. IEEE.

Rigutini, Leonardo, Michelangelo Diligenti, Marco Maggini, and Marco Gori. 2012. Automatic generation of crossword puzzles. *International Journal on Artificial Intelligence Tools*, 21(03):1250014.

Sandiuc, Corina and Alina Balagiu. 2020. The use of crossword puzzles as a strategy to teach maritime english vocabulary. *Scientific Bulletin" Mircea cel Batran" Naval Academy*, 23(1):236A–242.

Yuriev, Elizabeth, Ben Capuano, and Jennifer L. Short. 2016. Crossword puzzles for chemistry education: learning goals beyond vocabulary. *Chemistry education research and practice*, 17(3):532–554.

Zamani, Peyman, Somayeh Biparva Haghighi, and Majid Ravanbakhsh. 2021. The use of crossword puzzles as an educational tool. *Journal of Advances in Medical Education & Professionalism*, 9(2):102.

Zeinalipour, Kamyar, Tommaso Iaquinta, Giovanni Angelini, Leonardo Rigutini, Marco Maggini, and Marco Gori. 2023a. Building bridges of knowledge: Innovating education with automated crossword generation. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1228–1236, Jacksonville, Florida, USA, December. IEEE.

Zeinalipour, Kamyar, Tommaso Iaquinta, Asya Zanollo, Giovanni Angelini, Leonardo Rigutini, Marco Maggini, and Marco Gori. 2023b. Italian crossword generator: Enhancing education through interactive word puzzles. *arXiv preprint arXiv:2311.15723*.

Zeinalipour, Kamyar, Mohamed Saad, Marco Maggini, and Marco Gori. 2023c. Arabicros: Ai-powered arabic crossword puzzle generation for educational applications. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, pages 288–301, Singapore, December.

Zirawaga, Victor Samuel, Adeleye Idowu Olusanya, and Tinovimbanashe Maduku. 2017. Gaming in education: Using games as a support tool to teach history. *Journal of Education and Practice*, 8(15):55–64.

Zugarini, Andrea, Kamyar Zeinalipour, Surya Sai Kadali, Marco Maggini, Marco Gori, and Leonardo Rigutini. 2024. Clue-instruct: Text-based clue generation for educational crossword puzzles. *arXiv preprint arXiv:2404.06186*.