Softwareproject Topics

Katja Markert: Topics SS2024

Computerlinguistik
Universität Heidelberg

Topic Suggestions

Variations

Most topics can be handled by more than one group via variations of method, language/domains or data. Every group can determine their focus (within reason) themselves. When two groups use the same data, they can also work as if in a "competition".

Topic Suggestions

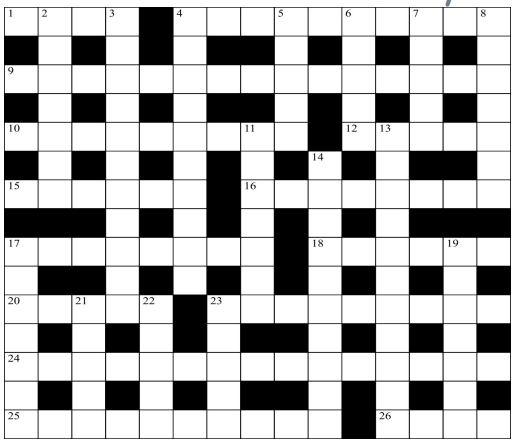
Mostly about word games this year!

1. Word Games
   1. Markert I: Classical crossword puzzles for German and/or with less training data
   2. Markert II: Cryptic crossword puzzles
   3. Markert III: Word games as multi-agent dialogue games
2. Topic Markert IV: **Saints-Memory**: Matching saints in a German historical encylopedia to German Wikipedia

Why are crosswords and word puzzles a fantastic NLP/AI testbed?

- Wide variety of challenges: different reasoning types, ambiguity, puns, knowledge
- Measures whether one can follow instructions: word length, groups of 4 words, given letters in right position
- Underspecification
- Abundant data written by experts and many different constructors
- Reconcile crossing letter and length constraints: search problems
- Large search space: not all answers are words (abbreviations, n-grams etc)

Markert 1: Quick/classical Crosswords



Quick crosswords

Mmostly) clues that need world knowledge or linguistic knowledge but the clue is relatively straightforward. Clue answer often ambiguous. Similar clues repeat in different crosswords.

Example clues and ANSWERS:

- 15D: Youngest of the Brontes (4): ANNE
- 8A: Working at nothing (6): LAZING
- 7D: ___ down the river (4); SOLD
- 8A: A sesame street character (4): ELMO

Markert I: Quick Crosswords Challenges

- What is a good answer for an individual clue? Reward functions and answer generations.
- How to fill in the whole crossword? Search functions. Belief propagation.
- Scoring and Evaluation functions

Markert I: The Berkeley Crossword Solver

Wallace et al (ACL 2022): Automated crossword solving

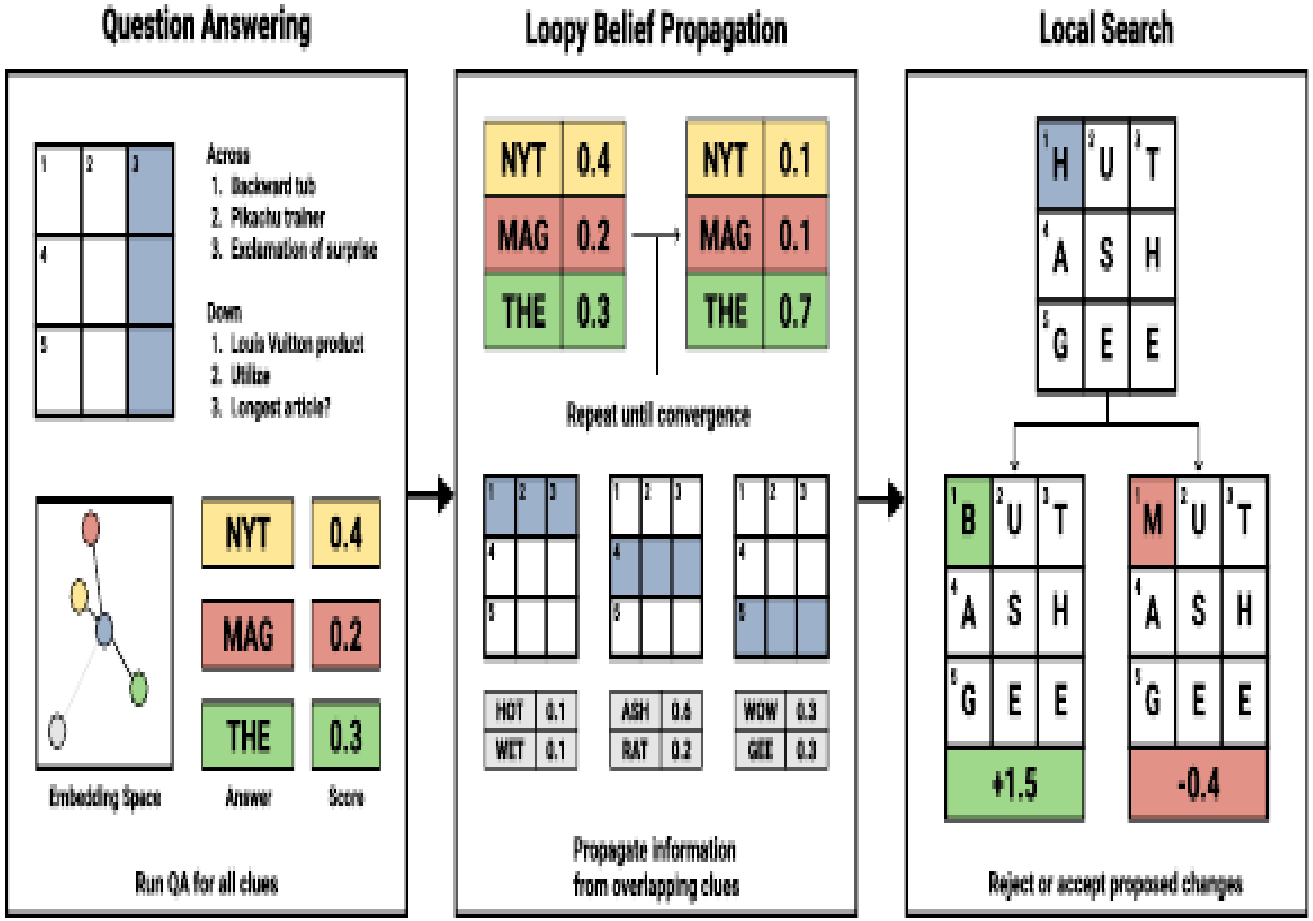


Figure 2: *An overview of the Berkeley Crossword Solver.* We use a neural question answering model to generate answer probabilities for each question, and then refine the probabilities with loopy belief propagation. Finally, we fill the grid with greedy search and iteratively improve uncertain areas of the puzzle using local search.

Markert I: Quick Crosswords Results

Wallace et al (ACL 2022): Automated crossword solving

Source	# Puzzles	Perfect Puzzle (%)		Word Acc. (%)		Letter Acc. (%)	
		Dr. Fill	BCS	Dr. Fill	BCS	Dr. Fill	BCS
The Atlantic	46	82.6	**89.1**	98.5	**99.1**	99.7	**99.8**
Newsday	52	86.2	**94.2**	98.6	**99.6**	99.1	**99.8**
The New Yorker	22	**86.4**	77.2	**99.5**	98.9	**99.9**	99.8
The LA Times	54	81.5	**92.6**	99.4	**99.7**	99.9	99.9
The New York Times	234	70.5	**81.7**	97.9	**98.9**	99.2	**99.7**

Table 3: *Final results of the Berkeley Crossword Solver.* We compare the BCS to Dr. Fill, the previous state-of-the-art crossword solving system, on a range of puzzle sources. The BCS produces significantly more perfect puzzles and achieves better or comparable letter-level and word-level accuracies.

Markert I: Quick Crosswords Open Problems

- Initial answer generation in Berkeley crossword solver needs very large training set of over 6m clue-answer pairs
- Only for English

Markert I: Project Ideas

- Reduce number of training examples needed in Berkeley crossword solver
  - via integrating LLMs in first step
  - Finding best small-scale training subset
  - Extending the WebCrow modular crossword puzzle solver by further experts
- Explore training and test splits; create difficult splits

Markert I: Ressources and Literature

- NYT crossword puzzles clue database at `https:www.github.com/lhlclhl/CP`
- Berkeley crossword solver:
  1. `berkeleycrosswordsolver.com`
  2. 6m clue-answer pairs: `https://github.com/albertkx/berkeley-crossword-solver`
  3. Wallace et al (ACL 2022): *Automated Crossword Solving*
- WebCrow (for French), modular crossword solver `https://arxiv.org/pdf/2311.15626.pdf`
- WebCrow for German: `https://ceur-ws.org/Vol-3596/paper54.pdf`
- Chen et al (ICAPS 2022): *Crossword Puzzle Resolution via Monte Carlo tree Search*

Markert II: Cryptic Crosswords

Quick Crosswords in Berkeley solver use little wordplay and puns (only 8% of clues but 20% of errors) $\longrightarrow$ cryptic crosswords

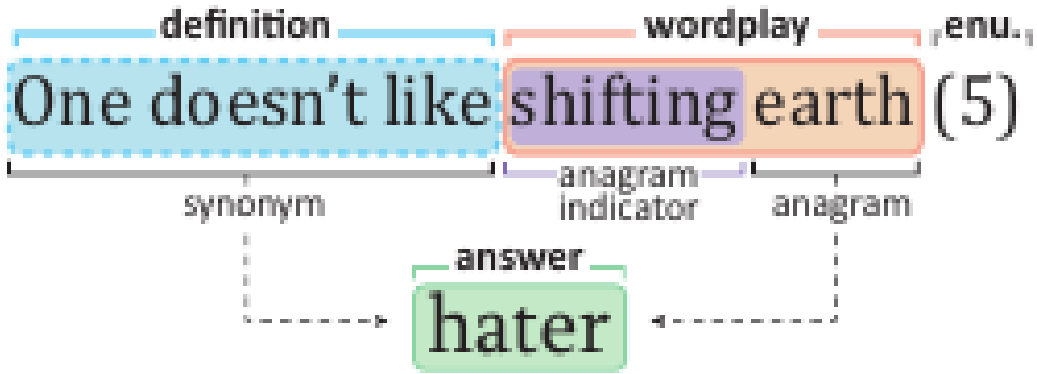
> *"Cryptic crosswords involve a different set of conventions and challenges, e.g. more metalinguistic clues such as anagrams, and likely require different methods from those we propose." (Wallace et al, 2022).*

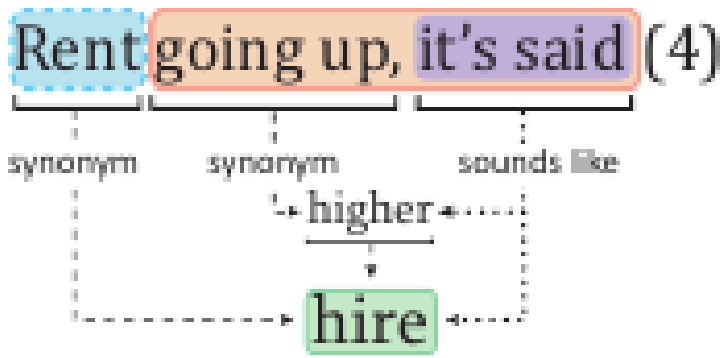
Cryptic Crosswords

- Clue builds on word play, puns (and knowledge)
- often includes a (misleading) surface **definition**
- plus often word-play with **code words or indicators** (anagrams, language type etc)
- word-play can be semantic, syntactic, phonetic or surface-level
- clues are designed to have only one possible answer

Markert II: Cryptic Crosswords English Examples

Graphics from Efrat et al (2021)





Markert II: Cryptic Crosswords German Examples

All from magazine *Stern*, 4.4.2024, Kreuzweise (not that difficult):

- Tubt sich on Tour aus aufs Sauberste, wie aus markantester Waschereilust $\longrightarrow$ Rei
- Wozu ganz unpolitisch Tourist auf Heimweg wird, fehlt'n Siegfried fürn Davonläufer $\longrightarrow$ Ausreiser

Markert II: Cryptic Crosswords German Examples

All from magazine *Stern*, 4.4.2024, Kreuzweise (not that difficult):

- Tubt sich on Tour aus aufs Sauberste, wie aus markantester Waschereilust $\longrightarrow$ Rei
- Wozu ganz unpolitisch Tourist auf Heimweg wird, fehlt'n Siegfried fürn Davonläufer $\longrightarrow$ Ausreiser

Markert II: Cryptic Crosswords German Examples

All from magazine *Stern*, 4.4.2024, Kreuzweise (not that difficult):

- Tubt sich on Tour aus aufs Sauberste, wie aus markantester Waschereilust $\longrightarrow$ Rei
- Wozu ganz unpolitisch Tourist auf Heimweg wird, fehlt'n Siegfried fürn Davonläufer $\longrightarrow$ Ausreiser

Markert II: Cryptic Crosswords Common Word Play Types

Picture from Sadallah et al (2024): Are LLMs good cryptic crossword solvers?

Type	Example Clue	Answer
ANAGRAM is a word (or words) that, when rearranged, form(s) a different word or phrase.	Never upset **a Sci Fi writer** (5)	Verne
HIDDEN CLUES have the answer written in the clue itself, amongst other words.	**Confront them** in the tobacco store (6)	Accost
DOUBLE DEFINITION contains two meanings of the same word.	**In which you'd place the photo** of the **NZ author** (5)	Frame
HOMOPHONE is a word that is pronounced the same as another but spelled differently.	Sounds like a couple *(pair)* **to scale down** (4)	Pare

Markert II: Cryptic Crosswords State-of-the-Art

Efrat et al (2021) *Cryptonite*

- Dataset of around 0.5 million English clues and asnwers
- T5-Large (770m parameters) fine-tuned: 8% accuracy
- On easy beginner level crosswords: 12% accuracy
- No overall crossword solver, only on individual clue level

Markert II: Cryptic Crosswords State-of-the-Art

Rozner et al (Neurips 2021):

- Dataset from *Guardian* with different levels of difficulty, around 55K answers and 142K clues
- **Curriculum learning:** train model first on related tasks such as solving anagrams
- T5 with curriculum learning: 20% top-10 recall
- Also only accuracy on individual clues, not filling in whole grid

Markert II: Cryptic Crosswords State-of-the-Art

Sadallah et al (2024): Language Model prompting and fine-tuning on Rozner et al's dataset

- LlaMa and Mistral (7b), ChatGPT (3.5?)
- Prompting results: accuracy best around 10%
- Finetuning accuracy up to 13% dependent on train and test splits

Markert II: Current Cryptic Crossword Solvers

Why are the results so bad? Why do LLMs perform so badly?

- LLMs struggle with information on the character-level such as anagrams, length of clue etc. due to their tokenization strategies
- Current models only look at individual clues: not combined with search strategies for solving a full puzzle or take advantage of partial solutions
- Identification of code words or indicators?

Markert II: Challenges and Possibilities for Cryptic Crosswords

- Evaluating LLMs for cryptic crossword clues: can we improve their performance with **particular prompting strategies such as chain-of-thought or train-of-thought?**
- Automatic identification of **code words or indicators** for improving cryptic crossword performance: use indicator-specific tools, not just LLMs
- **Search algorithms and strategies for cryptic crosswords:** can we combine Rozner et al (2021) individual clue performance with Berkeley quick crossword puzzle search strategy?
- Cryptic crossword solvers for German cryptic crosswords

Resources and Literature: Cryptic Crosswords

- Cryptonite:
  - Efrat et al (2021): *Cryptonite: A cryptic crossword benchmark for extreme ambiguity in language*
  - `https://github.com/aviaefrat/cryptonite`
- Rozner et al (2021):
  - `https://github.com/jsrozner/decrypt`
  - Rozner et al (Neurips 2021):*Decrypting Cryptic Crosswords: semantically Complex Wordplay Puzzles as a target for NLP*
- Sadallah et al (2024): *Are LLms good cryptic crossword solvers?* `https://arxiv.org/abs/2403.12094`

Markert III: Word Games as Multi-agent or Single-agent Dialogue Games

Chalamalasetti et al (EMNLP 2023): *clembench: Using game play to evaluate chat-optimized language models as conversational agents*

Implementing **dialogue games** as a constructed activity between different LLM players that attempt to reach a goal stage primarily by means of producing and understanding linguistic material

Markert III: Word Games as Multi-agent or Single-agent Dialogue Games

Chalamalasetti et al (EMNLP 2023): *clembench: Using game play to evaluate chat-optimized language models as conversational agents*

#	A	Game Master	B
1		[A|GM] The word you need to describe is "expedition". The taboo words are "expedition", "journey", "discovery", "exploration".	
2	[A>GM] CLUE: A trip taken for a specific purpose.		
3		[GM|GM] [valid]	
4		[GM>B] CLUE: A trip taken for a specific purpose	
5			[GM<B] GUESS: Journey
6		[GM|GM] [valid, wrong]	
7		[A|GM] GUESS: journey	
8	[A>GM] CLUE: A planned and organized trip with a specific goal in mind.		
9		[GM|GM] [valid]	
10		[GM>B] CLUE: A planned and organized trip with a specific goal in mind	
11			[GM<B] GUESS: expedition
12		[GM|GM] [correct]	

Figure 1: An episode of the taboo word game

Markert III: Word Games as Multi-agent or Single-agent Dialogue Games

We can include any other word game into clembench!

For Example NYT Connections Game:

`https://www.nytimes.com/games/connections`

pea	chest	brown	grounds
hazel	head	green	fiction
lawn	chair	coco	director
chief	dogs	field	unchained

Markert III: Word Games as Multi-agent or Single-agent Dialogue Games

We can include any other word game into clembench!

For Example NYT Connections Game:

`https://www.nytimes.com/games/connections`

brown	dogs	fiction	unchained
field	lawn	green	grounds
head	chair	chief	director
coco	chest	pea	hazel

Markert III: Project Idea

- Enhance Clembench with further word games (such as Connections as single-player game but others are possible)
- Develop evaluation scoring for these games, including following of instructions, game completion and success rates
- Evaluate several language models for game performance

Markert III: Ressources and Literature

- Archive for NYT Connections: `https://connections.swellgarfo.com/archive`
- Chalamalasetti et al (EMNLP 2023): *clembench: Using game play to evaluate chat-optimized language models as conversational agents*

Markert IV: **Saints-Memory**

`https://encycnet.github.io/`: aims to create a new semantic resource for historical German in form of a knowledge graph.

- Currently 22 historic German encyclopedias
- Attempt to use DBSpotlight to automatically match entries to DBPedia (and GermaNet)

	Germanet	DBPedia
Brockhaus 1809	32.80	51.15
Eisler Philosophie 1904	46.06	25.40
Wander Sprichtwort 1867	43.75	16.06
Roell Eisenbahnen 1912	24.91	27.88
Heiligenlexikon 1858	2.34	0.35

Use cases:

- Finding gaps in Wikipedia
- Finding mismatched information
- Preserving our cultural heritage

Heiligenlexikon

33,481 entries, 3m tokens

```
<b>S. Bilhildis</b>, (27. Nov.), Wittwe und Stifterin
des Klosters Altmünster (<i>Altum Monasterium B. V. M</i>.)
war die Tochter christlicher Eheleute von vornehmer Abkunft
Namens Iberius und Mechildis (Mechtildis, Mathildis) und
wurde zu Hochheim am Main um das Jahr 625 oder 626 geboren.
Was dieß für ein Hochheim am Main sei, ob der nicht
weit von Wirzburg gelegene Ort, gewöhnlich Veitshöchheim ge
...
Von ihrer Base zu Wirzburg in aller Gottseligkeit erzogen,
ward sie in jungen Jahren, etwa 16 oder 17 Jahre alt, an de
heidnischen Herzog Hettan (in Thüringen) vermählt,
...
Die Zeit, wann sie das Zeitliche
segnete, ist nicht zu ermitteln; Einige jedoch
setzen ihren Tod in das Jahr 630. (<i>El., Buc</i>.)
```

Picture

Bild von Joachim Schäfer - `https://www.heiligenlexikon.de` Ökumenisches Heiligenlexikon, Creative Commons CC BY-NC-SA 4.0

Wikipedia

Bilhildis von Altmünster, auch Bilihild, Bilehild oder Bilihilt und Bilhild (im 7. Jahrhundert in Veitshöchheim; gest. um 734 in Mainz) war eine fränkische Adelige, Klostergründerin und Äbtissin. Der Name Bilhildis ist althochdeutsch und bedeutet „die mit dem Beil Kämpfende“. . . . wurde sie gegen ihren Willen mit dem ungetauften, in Würzburg residierenden Herzog Heden (dux militum gentilis . . . vocabulo Hetan) aus dem Geschlecht der Hedenen vermählt . . . In Veitshöchheim findet jährlich an ihrem Gedenktag, dem 27. November, ein Gottesdienst statt.

That was the simple case...

- Unambiguous and successful name match in Wikipedia
- Matching of feast day ("Gedenktag")
- Even then we can see: name variations *Hettan* - *Heden*, different birth or death dates, uncertain information on very early saints

Normally...

- Obscure and unmatched saints:

```
<b>Boderius</b>, (22. Mai), wird in einigen
Orten als Martyrer verehrt
```

- Several names and several or ambiguous feast days
- Very ambiguous saints where name and day is not enough
  1. Bernardus (1): cistercian, abbot of Clairveaux, approx. 1100
  2. Bernardus (2): arch bishop of Vienne died 842, also named Barcar, or Barnar
  3. Bernardus (3): bishop of Carinola, died 1109
  4. ...
  5. Bernardus (64): simple monk of the Capucines, died 1540
- Normally no infoboxes in Wikipedia

Project Idea

- Define information extraction templates for saints

saint	relation	possible filler
saint	has-name	any string
saint	feast-day	date
saint	has-job	martyr, abbot, bishop, arch-bishop . . . ..
saint	is-born	date
saint	has-died	date
saint	located-in	location

- Algorithms for template filling inspired by IE work or by LLMs for Heiligenlexikon $\rightarrow$ Template 1
- Template filling from Wikidata $\rightarrow$ Template(s) 2
- Template match
- Evaluation

Template filling

- There is no gold training data, so unsupervised and probably no or little standard machine learning
- Possibility:
  1. Preprocessing with Heideltime and German Stanford Core NLP, LLMs
  2. Template 1: Some relations can be filled by (approximate) regular expression match and tag restrictions (names, job titles, feast days); some via LLMs
  3. Template 2: REs or Wikidata `https://www.wikidata.org/wiki/Q477895`
  4. Perform simple, unambiguous saint matches from the templates; seedset 1
  5. Noisy matches from Perplexity API; noisy seedset 2
  6. Semi-supervised learning from the two seedsets

Resources and Literature

- Encyc-Net for the Heiligenlexikon: `https://encycnet.github.io/`
- Hagen et al (2020): Twenty-two historical encyclopedias encoded in TEI: A new resource for the digital humanities. In: LaTeCH-CLfL 2020: 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature.
- Preprocessing of Wikipedia and Heiligenlexikon, small human test set of 200 matches and Perplexity LLM baseline from research module of Johannes Eschbach
- For the TACRED news IE relation dataset: Zhang et al (2017). Position-aware attention and supervised data improve slot filling. In *EMNLP 2017*.
- Stoica et al (2021). Re-tacred: Addressing shortcomings of the tacred dataset. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. No. 15. 2021.
- Han et al (2021): PTR: Prompt Tuning with rules for text classification. `https://arxiv.org/pdf/2105.11259.pdf`