The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography?

Miloš Jakubíček, Michael Rundell

LEXiCAL COMPUTING

Brno, Czechia

eLex 2023, Brno, Czechia

sketchengine.eu

Can ChatGPT outperform current tools for post-editing lexicography?

ChatGPT

- chat-tuned autoregressive transformer based language model
- released initially in December 2022
- raised general excitement (and worries)
- best promotion for corpora ever?

sketchengine.eu

ChatGPT

language modelling – not language reasoning

- no logical inference
- no knowledge database
- token sequence generation
- beam search through (probabilistic) language model

ChatGPT

no source data reference

- where does it come from?

limited prompt and response length

- thousands of tokens (4K, 16K)

sketchengine.eu

translation through multilingualism

- English = 92 % of data

sketchengine.eu

prompting

- key part to get good responses

sketchengine.eu

ChatGPT

non-deterministic learning and inference

- limited reproducibility

ChatGPT

static model
- up to 2021
- additional training possible
- plugins

Making an English minidictionary

- DANTE sample used for headwords
- very heterogenous
- 99 headwords

sketchengine.eu

3 runs each with 3 questions, 297 answers

1. What does the word *H* mean?
2. Generate a dictionary entry for *H*.
3. Generate a dictionary entry for *H* including possible word forms, word senses, pronunciation, collocations, synonyms, antonyms and examples of usage.

sketchengine.eu

Access

- initially by faking browser access (no API)
- we repeated after API got available with ChatGPT3.5
- published in Lexonomy:
  https://www.lexonomy.eu/chatgpt
  https://www.lexonomy.eu/chatgpt35
- ChatGPT 4 was not available through API by the time of writing the full paper (now it is)

- word sense induction
- definitions
- grammatical and syntactic information
- labelling of 'marked' items
- example sentences

- "false polysemy": variations on the same sense
- missing senses: common meanings not identified
- invented senses: additional senses for which there is no evidence

Climate (noun)

1. The long-term patterns of temperature, humidity, wind, and precipitation in a particular region.

2. The overall weather conditions of a place over a period of time, typically 30 years or more.

3. The typical or average weather conditions of a place.

4. The general set of weather conditions of a planet or region.

5. The state of the atmosphere in a region in terms of temperature, humidity, wind patterns and precipitation.

6. The average of weather conditions over a period of time, typically 30 years or more.

sketchengine.eu

. eneral disputes (6.9).</s><s>In the current judicial **climate** , difficulties in categorisation should not make a de

. ient, and most sellers, at least in today's economic **climate** , desire to retain the flexibility to sell off or secure th

. or because the market is flat due to the economic **climate** .</s><s>The length of the rent free period will be a

. ct.</s><s>In today's changing legal and economic **climate** , business focus should be on market position (not

. igh the Act does not make reference to the current **climate** of opinion about sexual explicitness, juries in obsce

. ind, and to hear expert evidence about the current **climate** of permissiveness in relation to this kind of literatur

sketchengine.eu

Invented sense: *butter* (sense 3)

1. a soft, yellowish-white food made from cream or milk, used as a spread or in cooking and baking.
2. a substance resembling this in consistency or color.
3. Informal- a person or thing that is easy to defeat or manipulate; a pushover.

1. She spread the bread with butter before eating it.
2. He added a big pat of butter to his potatoes.
3. The cake had a buttery flavor that everyone loved.
4. She found the exam to be a piece of butter.

sketchengine.eu

```
careful (adjective):

1. taking or showing care to avoid causing harm or mistakes; being cautious.

"Drive carefully on the icy roads."

2. Mindful and considerate of others or the effects of one's actions.

"She was always very careful not to hurt anyone's feelings."

3. done with attention to detail; meticulous.

"He was careful to proofread his work before submitting it."
```

Guess what this is

```
1. A person who spends a lot of time sitting and watching television or playing video games, and is generally inactive.
2. Someone who is lazy or lacks motivation to engage in physical activity.
```

sketchengine.eu

Haunt (hônt) verb

1. Visit frequently, or reside in as a ghost or spirit.

2. Constantly present in one's mind; an obsession.

3. To frequent a place or places frequently.

1. enjoy + object
   Example: I enjoy watching movies.
2. enjoy + -ing form of a verb
   Example: She enjoys playing the piano.
3. enjoy + noun
   Example: We enjoyed a delicious meal at the restaurant.
4. enjoy + object + as + complement
   Example: He enjoyed the concert as a celebration of his birthday.
5. enjoy + oneself / yourself / themselves
   Example: We enjoyed ourselves at the party.
6. enjoy + preposition + object
   Example: She enjoyed a peaceful afternoon by the lake.
7. enjoy + that-clause
   Example: They enjoyed that they had completed the project on time.

1. enjoy + noun: "I enjoy coffee in the morning."
2. enjoy + gerund (-ing form of a verb): "He enjoys playing soccer with his friends."
3. enjoy + object pronoun: "We enjoy them as our neighbors."
4. enjoy + infinitive: "She enjoys to dance every weekend."
5. enjoy + preposition + noun: "He enjoys going on trips with his family."

```
Betimes (adverb):
1. Early or in good time; before the usual or expected time.
2. Sometimes or on occasion; occasionally.
Example in a sentence: "We must leave betimes in order to make it to the concert on time."


Note: Betimes is archaic word and it's usage is rare in modern english.
```

cookie (noun)

```
1. A small, usually sweet, baked good made from flour, sugar, and other ingredients, such as chocolate chips or nuts.
2. A small text file that a website saves on your computer or mobile device when you visit the site. It enables the v
3. (Computing) A small piece of data stored on the user's computer by a website, containing information about the use
4. (slang) A person, especially a girl or woman, who is attractive in a sexual way.
```

sketchengine.eu

```
Example sentences:
- The echo of the gunshots could be heard for miles.
- The canyon walls created an echo that made it difficult to hear.
- The singer's voice echoed through the large concert hall.
- The children echoed back the teacher's instructions.
- The company's mission statement echoes the values of the founder.
```

Lack of diversity

- Initial "The" in all
- 3rd person subject + simple past pattern in 3 out of 4

Formulaic examples: a persistent problem

Ironically, examples look made-up

- fair (separate experiment)

  The referee made a fair decision by awarding a penalty.

  The garden was filled with fair flowers.

  The price of the item was fair, not too high or too low.

  The fair-skinned woman had to wear a hat and sunscreen to protect her skin from the sun.

- persuade (Lew, forthcoming)

  The salesperson persuaded the customer to buy the product.

  The speaker persuaded the audience to support the cause.

  The friend persuaded the colleague to take a day off

sketchengine.eu

Evaluation summary

- WSI: weak
- grammar and syntax: weak
- labelling: mostly good
- examples: appalling: sackable offences
- definitions: ChatGPT's strongest feature
  - often uses dated formulae ('the act or state of …', 'having or showing'
  - but generally good, both for mainstream vocabulary and technical terms