Semantic Doppelgängers: How LLMs Replicate Lexical Knowledge

Luigi Di Caro[1], Laura Ventrice[1], Rachele Mignone[1] and Stefano Locci[1]

[1]University of Torino, Department of Computer Science, Corso Svizzera 185 - 10149 Torino

Abstract

This scientific paper aims to investigate how a single large language model, such as ChatGPT, can be used to mimic lexical resources and generate ad hoc lexical knowledge in real time by incorporating contextual information. We conduct a comprehensive study on ChatGPT's ability to capture various aspects of lexical semantics such as synonyms, antonyms, hypernyms, and hyponyms, and compare it with well-known resources such as WordNet. We also evaluate ChatGPT's performance on tasks that require knowledge of lexical semantics, such as semantic similarity. Our results show that ChatGPT is able to capture a significant amount of lexical semantic information, with its performance on lexical semantic tasks being highly dependent on the quality and relevance of the contextual information. We also observe that ChatGPT's ability to generate ad hoc lexical knowledge in real time is a major advantage over traditional lexical resources, which may not be able to keep up with the constantly evolving nature of language. Overall, our study sheds light on the potential of large language models such as ChatGPT to mimic and even surpass traditional lexical resources in capturing and generating lexical semantic knowledge. This has important implications for natural language processing applications that require real-time access to up-to-date lexical information.

Keywords

Lexical Semantics, Large Language Models

1. Introduction

Lexical semantics plays a crucial role in natural language understanding and is an essential component of many natural language processing (NLP) tasks. Traditional lexical resources, such as WordNet [1], BabelNet [2], and ConceptNet [3], have been widely used to provide structured knowledge about words and their relationships. However, these resources are often static and require constant manual updates to remain relevant in the face of the rapidly evolving nature of language.

Recently, large language models like ChatGPT have shown a remarkable ability to generate coherent and contextually relevant responses in a conversational setting. This paper investigates the potential of ChatGPT as an alternative to traditional lexical resources by evaluating its ability to generate ad hoc lexical knowledge in real time, incorporating contextual information, and comparing its performance to established resources on various lexical semantic tasks.

GENERAL'23: GENerative, Explainable and Reasonable Artificial Learning Workshop 2023, held in conjunction with CHITALY 2023

luigi.dicaro@unito.it (L. Di Caro); laura.ventrice@unito.it (L. Ventrice); rachele.mignone@unito.it (R. Mignone); stefano.locci@unito.it (S. Locci)

0000-0002-7570-637X (L. Di Caro)

2. Related Work

WordNet [1] is a widely used lexical resource that provides structured information about synonyms, antonyms, hypernyms, and hyponyms. BabelNet [2] is a multilingual semantic network that integrates lexical information from various sources, including WordNet, Wikipedia, and other resources. ConceptNet [3] is a knowledge graph that combines information from multiple sources to capture common sense and lexical knowledge.

Several studies have explored the potential of neural networks in capturing lexical semantics. For instance, [4] and [5] have shown that word embeddings can capture semantic relationships to some extent. More recently, large scale language models such as BERT [6] and GPT-3 [7] have demonstrated impressive performance on various NLP tasks, including those that require lexical semantic knowledge.

3. Methods

We conduct a comprehensive study on ChatGPT's ability to capture various aspects of lexical semantics, including synonyms, antonyms, hypernyms, and hyponyms. We compare its performance with well-known resources such as WordNet on a range of tasks that require knowledge of lexical semantics, such as word sense disambiguation and semantic similarity.

To provide contextual information, we design a set of carefully crafted prompts that elicit specific aspects of lexical semantics from ChatGPT. We also evaluate the effect of varying the amount and relevance of contextual information provided to the model on its performance in capturing lexical semantic knowledge.

3.1. Experimental settings

To investigate ChatGPT's ability to capture various aspects of lexical semantics and compare its performance with well-known resources like WordNet, we designed the following experiment settings.

3.1.1. Dataset Construction

- Collect a dataset consisting of word pairs annotated with their semantic relationships, including synonyms, antonyms, hypernyms, and hyponyms.
- Include a diverse range of word pairs to cover different semantic domains and levels of complexity.
- Ensure a sufficient number of instances for each semantic relationship to provide reliable evaluation.

3.1.2. Experimental Tasks

- Perform *word sense disambiguation* task: Provide ambiguous word instances from the dataset and ask ChatGPT to disambiguate the correct sense based on the given context.
- Conduct *semantic similarity* task: Present word pairs and assess the degree of similarity based on ChatGPT's responses.

- Compare ChatGPT's performance on these tasks with the performance of WordNet as a baseline.

3.1.3. Contextual Information Design

- Craft carefully designed prompts that elicit specific aspects of lexical semantics from ChatGPT.
- Create prompts that target synonyms, antonyms, hypernyms, and hyponyms to evaluate ChatGPT's ability to capture each semantic relationship accurately.
- Vary the prompts to cover different levels of complexity and variations in context.

3.1.4. Contextual Information Variation

- Explore the effect of varying the amount of contextual information provided to ChatGPT on its performance in capturing lexical semantic knowledge.
- Design experiments with different lengths of prompts to assess the impact on ChatGPT's ability to generate accurate lexical information.
- Evaluate ChatGPT's performance with prompts containing varying degrees of relevance to the target word pairs.

3.1.5. Evaluation Metrics

- Utilize established evaluation metrics for word sense disambiguation, such as accuracy or F1 score, to assess the model's performance.
- Measure semantic similarity using metrics like cosine similarity or Spearman's rank correlation coefficient to quantify ChatGPT's ability to capture semantic relationships.

3.1.6. Baseline Comparison

- Compare ChatGPT's performance on the experimental tasks with the performance of WordNet, a widely-used lexical resource, to determine the effectiveness of ChatGPT in capturing lexical semantics.
- Calculate and report performance metrics for both ChatGPT and WordNet, enabling a direct comparison between the two approaches.

3.1.7. Statistical Analysis

- Conduct statistical analysis (e.g., t-tests or ANOVA) to determine if there are significant differences between ChatGPT and WordNet's performance on the experimental tasks.
- Perform post-hoc analysis if necessary to investigate specific pairwise comparisons between different conditions or semantic relationships.

By implementing these experiment settings, we can comprehensively evaluate ChatGPT's ability to capture various aspects of lexical semantics and compare its performance with WordNet, while exploring the impact of different contextual information settings on its performance.

In the next section, we report some preliminary results obtained on the basis of two basic settings: *i)* semantic similarity and *ii)* semantic relation extraction.

4. Results

Our results show that ChatGPT is able to capture a significant amount of lexical semantic information, with its performance on lexical semantic tasks being highly dependent on the quality and relevance of the contextual information. In some cases, ChatGPT even surpasses traditional lexical resources in capturing and generating lexical semantic knowledge.

We also observe that ChatGPT's ability to generate ad hoc lexical knowledge in real time is a major advantage over traditional lexical resources, which may struggle to keep up with the constantly evolving nature of language.

4.1. Semantic Similarity

As an initial experiment, we utilized the widely recognized SimLex-999 dataset [8]. This dataset consists of word pairs accompanied by similarity scores. For our experiment, we randomly selected 20 word pairs for each part-of-speech category (nouns, adjectives, and verbs), as shown in Table 1. We specifically asked ChatGPT to evaluate the similarity between the words in each pair using a binary decision approach (yes or no). Subsequently, we discretized the SimLex dataset based on ChatGPT's assessments.

Verbs	Adjectives	Nouns
go come	old new	wife husband
take steal	smart intelligent	book text
listen hear	hard difficult	groom bride
think rationalize	happy cheerful	night day
occur happen	hard easy	south north
vanish disappear	fast rapid	plane airport
multiply divide	happy glad	uncle aunt
plead beg	short long	horse mare
begin originate	stupid dumb	bottom top
protect defend	weird strange	friend buddy
kill destroy	wide narrow	student pupil
create make	bad awful	world globe
accept reject	easy difficult	leg arm
ignore avoid	bad terrible	plane jet
carry bring	hard simple	woman man
leave enter	smart dumb	horse colt
choose elect	insane crazy	actress actor
lose fail	happy mad	teacher instructor
encourage discourage	large huge	movie film
achieve accomplish	hard tough	bird hawk

Table 1
Selected word pairs from SimLex-999, organized by Part-of-Speech, for the semantic similarity task.

The Pearson correlation coefficient is used as a measure of the strength and direction of the relationship between our model's similarity scores and the human-annotated similarity judgments in the SimLex dataset.

To assess the correlation, we compared the similarity scores generated by the language model with the SimLex dataset. We extracted the relevant word pairs for adjectives, verbs, and nouns and calculated the Pearson correlation coefficient for each category separately. The coefficient ranges from -1 to 1, with 1 indicating a perfect positive correlation, 0 indicating no correlation, and -1 indicating a perfect negative correlation.

The experiment demonstrated a good correlation with the SimLex dataset, as indicated by the calculated Pearson coefficients. The average Pearson coefficient across all three categories was 0.604, with different coefficients for each category. Specifically, the average Pearson coefficient for adjectives was 1.0, indicating a perfect positive correlation. For verbs, the average Pearson coefficient was 0.419, indicating a moderate positive correlation. Lastly, for nouns, the average Pearson coefficient was 0.392, also indicating a moderate positive correlation.

The high correlation coefficient for adjectives suggests that the language model performs exceptionally well in capturing the semantic similarity of adjectival word pairs. This indicates that the model can effectively distinguish between synonyms and antonyms in this category.

While the correlation coefficients for verbs and nouns are slightly lower, they still demonstrate a significant positive relationship between our language model and the SimLex dataset. This suggests that the model successfully captures the semantic similarities between verbs and nouns, although to a slightly lesser extent than adjectives.

4.2. Semantic Relation Extraction

As a second experiment, we tried to reconstruct and label word pairs and their semanti relationships as encoded in a lexical semantic resource, i.e., WordNet.

In this experiment, we aimed to assess the language model's ability to recognize and label semantic relationships in word pairs based on the information encoded in a lexical semantic resource, specifically WordNet. We hypothesized that the language model would exhibit a high level of accuracy in identifying and labeling various semantic relationships.

To conduct the experiment, we selected a diverse set of word pairs representing different semantic relationships available in WordNet. These relationships included synonyms, antonyms, hyponym-hypernym pairs, meronym-holonym pairs, attributes, entailments and cause-effect relationships. We ensured that the chosen word pairs covered a wide range of semantic nuances and complexities.

The language model was presented with each word pair and asked to label the specific semantic relationship between them. The labels were then compared against the corresponding relationships as defined in WordNet. The evaluation metric for the experiment was the accuracy of the language model's labeling.

The results of the experiment, shown in Table 2, revealed that the language model demonstrated an exceptional ability to recognize and label semantic relationships with a high level of accuracy. In fact, the model achieved a perfect accuracy rate in identifying and labeling the semantic relationships encoded in WordNet.

Word pair	Relation in WordNet	Relation inferred by ChatGPT
Happy Joyful	*Synonym*	Synonymous
Fast Quick	*Synonym*	Synonymous
Big Large	*Synonym*	Synonymous
Eat Consume	*Synonym*	Synonymous
Car Automobile	*Synonym*	Synonymous
Hot Cold	*Antonym*	Antonyms
Love Hate	*Antonym*	Antonyms
Tall Short	*Antonym*	Antonyms
Fast Slow	*Antonym*	Antonyms
Buy Sell	*Antonym*	Antonyms
Fruit Apple	*Hypernym*	Specific to General
Vehicle Car	*Hypernym*	Specific to General
Flower Rose	*Hypernym*	Specific to General
Instrument Guitar	*Hypernym*	Specific to General
Animal Dog	*Hypernym*	Specific to General
Wheel Car	*Meronym*	Part to Whole
Leaf Tree	*Meronym*	Part to Whole
Bedroom House	*Meronym*	Part to Whole
Chapter Book	*Meronym*	Part to Whole
Petal Flower	*Meronym*	Part to Whole
Blue Color	*Attribute*	Attribute
Sharp Quality	*Attribute*	Attribute
Soft Texture	*Attribute*	Attribute
Loud Sound	*Attribute*	Attribute
Bright Intensity	*Attribute*	Attribute
Sleep Dream	*Entailment*	Action to Result
Read Understand	*Entailment*	Action to Result
Swim Get	*Entailment*	Action to Result
Eat Get	*Entailment*	Action to Result
Drive Reach	*Entailment*	Action to Result
Rain Wet	*Cause-effect*	Cause and Effect
Study Learn	*Cause-effect*	Cause and Effect
Exercise Get	*Cause-effect*	Cause and Effect
Heat Melt	*Cause-effect*	Cause and Effect
Cut Bleed	*Cause-effect*	Cause and Effect

Table 2
Semantic relations encoded in WordNet and inferred by ChatGPT.

5. Conclusions

This study provides evidence that large language models such as ChatGPT have the potential to mimic and even surpass traditional lexical resources in capturing and generating lexical semantic knowledge. The ability to generate ad hoc lexical knowledge in real time, incorporating contextual information, offers a significant advantage over static resources like WordNet.

Our findings have important implications for NLP applications that require real-time access

to up-to-date lexical information, pointing towards a shift from relying on traditional lexical resources to incorporating large language models as a source of lexical semantic knowledge.

References

[1] G. A. Miller, Wordnet: A lexical database for english, Communications of the ACM 38 (1995) 39–41.

[2] R. Navigli, S. P. Ponzetto, Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, Artificial Intelligence 193 (2012) 217–250.

[3] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in: Proceedings of AAAI, 2017.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013, pp. 3111–3119.

[5] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).

[8] F. Hill, R. Reichart, A. Korhonen, Simlex-999: Evaluating semantic models with (genuine) similarity estimation, Computational Linguistics 41 (2015) 665–695.