LLM-SEMREL: Towards a Better Coreference Resolution for Portuguese

Evandro Fonseca[1] **, Joaquim Neto**[1]

[1]Blip

evandro.fonseca, joaquim.neto {@blip.ai}

Abstract. *This paper aims to describe LLM-SEMREL, a new Portuguese semantic database built automatically using currently available large language models. The motivation for this project stems from the lack of rich semantic resources for the Coreference Resolution task in Portuguese. As a result, we provide a new resource that can be used to improve current models and build new ones. LLM-SEMREL is composed of 1,229,399 semantic relations, distributed among 261,731 words and their descriptions.*

1. Introduction

The development of Large Language Models (LLMs) has made it possible to create large volumes of synthetic data automatically, which is particularly useful in contexts where annotated data is scarce. These models, trained on vast text corpora, can be fine-tuned to generate examples that follow specific semantic patterns, replicating the linguistic complexity required for tasks such as coreference resolution. In the context of Brazilian Portuguese, this approach has the potential to overcome the limitation of resources, providing a diverse and enriched dataset that is crucial for training machine learning models. This capability is particularly significant when we consider the complexity and challenges of coreference resolution. According to [Brown et al. 2020], LLMs like GPT-3 can generate text that mimics the intricacies of natural language, which can be leveraged to create high-quality synthetic data for various NLP tasks.

Coreference resolution is a central task in Natural Language Processing (NLP), involving the identification of all expressions in a text that refer to the same entity[Fonseca 2018]. This task is challenging due to the inherent complexity of natural language, where different forms and expressions can refer to the same concept or object. In this context, semantic databases, which categorize and interrelate meanings of words and expressions, can provide the necessary support for machine learning systems to make precise inferences about which expressions are coreferent [Khosla and Rose 2020].

In recent years, the development of LLMs has revolutionized the field of NLP[OpenAI 2023, Dubey et al. 2024, Reid et al. 2024]. Leveraging the capabilities of these advanced models, we present a comprehensive semantic

database specifically tailored for the Portuguese language. This database, created using state-of-the-art LLMs, aims to enhance various NLP applications by providing rich semantic information and nuanced understanding of Portuguese text. This initiative is particularly critical when considering the challenges associated with creating resources in Brazilian Portuguese considering the scarcity of resources for the task in Portuguese.

However, the availability of resources such as semantic databases and annotated corpora in Brazilian Portuguese is limited compared to other languages like English[Hedderich et al. 2021]. While some resources are available, they are often insufficient to cover the linguistic diversity present in Brazilian Portuguese, which hinders the development of robust coreference resolution systems.

The creation of an automated, Portuguese-specific semantic database is, therefore, of utmost importance for advancing the task of Portuguese coreference resolution. This resource would not only facilitate the training of more accurate models but could also be used in a wide range of other NLP tasks, such as machine translation and sentiment analysis. With a robust semantic database, AI systems can achieve a deeper and more contextualized understanding of the Portuguese language, enabling them to solve complex tasks more effectively. Additionally, the availability of such resources would contribute to the democratization of technology, allowing more Brazilian researchers and developers to create innovative solutions for the local market, reinforcing the relevance of initiatives in this direction.

2. Related Work

In the study conducted by [Fonseca et al. 2016], the authors explore the impact of integrating semantic knowledge into coreference resolution systems for the Portuguese language. Their work specifically evaluates the contribution of semantic features derived from the Onto.PT lexical resource, which includes relations such as synonymy, hypernymy, and hyponymy. By incorporating these semantic features into a machine learning model, they were able to enhance the model's ability to correctly identify coreferent entities, particularly in cases where traditional lexical and syntactic features might fall short. The results showed improvements in precision, recall, and F-measure, highlighting the importance of utilizing comprehensive semantic databases to enrich coreference resolution tasks, especially in resource-limited languages like Portuguese.

[Jiang and Cohn 2021] introduced a coreference resolution model that incorporates syntactic and semantic information through a Heterogeneous Graph Attention Network (HGAT). This model constructs a heterogeneous graph that integrates syntactic structures, such as dependency trees, with semantic structures derived from Semantic Role Labeling (SRL). By enhancing word representations with this combined syntactic and semantic context,

the model significantly improves coreference resolution accuracy compared to models relying solely on pre-trained embeddings like SpanBERT. This work underscores the importance of well-structured semantic databases and syntactic inputs in enhancing the performance of coreference resolution systems.

A study conducted by [Lima et al. 2018] analyzed the use of different semantic bases, specifically Onto.PT[Gonçalo Oliveira 2012] and ConceptNet[Speer and Havasi 2012], in the task of coreference resolution for Portuguese texts. The authors demonstrated that integrating semantic information from these bases can improve the performance of coreference tools like CORP, particularly when dealing with relations of hyponymy and synonymy. The research concluded that while both semantic bases contribute to the system's performance, the use of ConceptNet resulted in slightly higher precision, highlighting the importance of semantic bases in enhancing coreference resolution, especially in languages with limited resources.

3. LLM-SEMREL

LLM-SEMREL[1], as its name suggests, was entirely annotated using a large language model. We used GPT-4o [OpenAI 2023] to annotate 261,731 words, considering eight semantic relations. The LLM-SEMREL words were taken from the br.ispell dictionary[2][Ueda 2005].

Regarding our method for developing the semantic database, we used the prompt-tuning technique, specifically zero-shot learning, to develop the resource. Through the Azure API and using the GPT-4o model, we created a Java code that made a request to the API for each word in the br.ispell dictionary and executed the prompt. The response from the LLM request is a JSON in the following format: first, we have the field *word*, which contains the word from the dictionary; then there is the *description* field, which provides a general context of the meaning of the word in the *word* field; finally, we have the *relations* field, which is a list of our semantic relations. Below, we show an example of the JSON for the word "cachorro"

```
{
    "word" : "cachorro",
    "description" : "Mamífero doméstico, conhecido por sua
       lealdade e companheirismo com os seres humanos.",
    "relations" : [ {
      "synonym_of" : [ "cão", "canino" ]
    }, {
      "hypernym_of" : [ "animal de estimação", "mamífero" ]
    }, {
      "hyponym_of" : [ "pastor-alemão", "poodle", "labrador" ]
    }, {
```

[1]the resource is available at: https://github.com/evandrofonsecatake/llm-semrel
[2]https://www.ime.usp.br/ pf/dicios/

```
      "meronym_of" : [ "pata", "cauda", "focinho" ]
    }, {
      "holonym_of" : [ "matilha" ]
    }, {
      "paronym_of" : [ "cachorrinho", "cachorrão" ]
    }, {
      "troponym_of" : [ ]
    }, {
      "antonym_of" : [ "gato" ]
    } ]
  }
```

Below, we present our prompt, which is applied to each word in the dictionary to generate semantic relations. We begin by establishing that the agent is responsible for generating semantic annotations when they exist for a given word. Following this, we provide detailed instructions on how the output should be formatted.

```
Você é um assitente que realiza anotação linguistica de relações
semânticas.  Você precisa anotar as relações quando existentes
para cada palavra recebida, seguindo o json exemplo:
```

"word":
"description":
"relations": [
"synonym_of": []
"hypernym_of": []
"hyponym_of": []
"meronym_of": []
"holonym_of": []
"paronym_of": []
"troponym_of": []
"antonym_of": []
]

Figure 1. Prompt used to collect the semantic relations of a word.

In Table 1, we show the relations, their definitions, and examples of each. The generated database comprises a total of 1,229,399 semantic relations.

In Table 2, we compare the number of relations between LLM-SEMREL and Onto.PT. We can see that Onto.PT has more types of relations. However, our resource was built with a focus on semantic relations to solve coreferences. In Table 3, we show the number of tokens consumed by the LLM model

Table 1.

Relation	Definition	Example	Number of Relations
synonym_of	Synonymy refers to the relationship between words that have similar or identical meanings.	**feliz** and **alegre**	447,773
antonym_of	Antonymy refers to the relationship between words that have opposite meanings.	**quente** and **frio**	335,558
hyponym_of	Hyponymy refers to the relationship where a word has a more specific meaning than a general or superordinate term.	**pardal** is a hyponym of **pássaro**	68,360
hypernym_of	Hypernymy refers to the relationship where a word has a broader meaning that encompasses more specific words.	**veículo** is a hypernym of **carro**	60,310
meronym_of	Meronymy denotes a part-whole relationship where a word represents a part of something larger.	**roda** is a meronym of **carro**	15,554
holonym_of	Holonymy is the relationship where a word represents the whole to which parts belong.	**árvore** is a holonym of **galho**	10,461
paronym_of	Paronymy refers to words that are similar in form or derivation but have different meanings.	**cavalheiro** and **cavaleiro**	274,994
troponym_of	Troponymy is the relationship where a verb denotes a specific manner of doing something that another verb denotes.	**sussurrar** is a troponym of **falar**	16,389

to generate the whole database. It is possible to see that 71,129,390 tokens were expended.

Table 2.

Relation	**LLM-SEMREL**	**Onto.PT**
Synonym_of	447,773	168,858
Antonym_of	335,558	92,598
Hyponym_of	68,360	91,466 (combined)
Hypernym_of	60,310	91,466 (combined)
Meronym_of	15,554	9,436
Holonym_of	10,461	7,431
Paronym_of	274,994	-
Troponym_of	16,389	-
Contained_In	-	644
Material_of	-	873
Cause_of	-	12,369
Producer_of	-	2,303
Purpose_of	-	16,271
Has_Quality	-	2,256
Has_State	-	561
Property_of	-	38,048
Place_of	-	1,393
Manner_of	-	3,966
Manner_Without	-	265
Total	**1,229,399**	**448,738**

4. Error Analysis

Regarding error analysis, we have examined several instances to understand the primary errors in our presented semantic relations. According to our anal-

Table 3.

Tokens			Words	Semantic Relations
Prompt	**Completion**	**Total**		
39,360,451	31,768,939	71,129,390	261,731	1,229,399

ysis, the main errors found in the database are due to the directionality between the semantic relations of hypernym, hyponym, meronym, and holonym. It is known that hyponyms are more specific than hypernyms (subclass relation), just as meronyms represent parts of something and holonyms represent the whole. To illustrate that we present two instances: “cachorro”*(dog)* and “Terra”*(Earth)*:

Table 4. pathways of semantic relations

Word	Relations				
	synonym_of	**hypernym_of**	**hyponym_of**	**meronym_of**	**holonym_of**
Cachorro	cão, canino	animal de estimação, mamífero	pastor-alemão, poodle, labrador	pata, cauda, focinho	matilha
Terra	mundo, planeta	continente, país, cidade	planeta do sistema solar	crosta, manto, núcleo	sistema solar

Analyzing Table 4, specifically for the word "cachorro," we can trace the following semantic relations and their terms: each term in the array of each relation has a connection with the word. For example, "pastor-alemão" is a hyponym of "cachorro"; "poodle" and "labrador" are also hyponyms of "cachorro." Similarly, "pata" and "cauda" are parts of "cachorro," making them meronyms of "cachorro." The same behavior is noted for "matilha" (a holonym of "cachorro").

However, when we look at the word "Terra" and its relations, this direction sometimes changes. For example, "continente" is not a hypernym of "Terra"; rather, "Terra" is a hypernym of "continente" or "país." On the other hand, "crosta," "manta," and "núcleo" are indeed meronyms of "Terra." This error appears in several instances of LLM-SEMREL. We believe it is due to prompt interpretation by LLM.

5. Conclusion

In this paper, we introduced LLM-SEMREL, a comprehensive semantic database for the Portuguese language, created using the capabilities of large language models (LLMs). This resource addresses the significant gap in semantic resources available for the Coreference Resolution task in Portuguese, providing a rich dataset of 1,229,399 semantic relations. By leveraging state-of-the-art LLMs, we have been able to generate a diverse and nuanced set of semantic relations that can enhance various NLP applications, including but not limited to coreference resolution, machine translation, and sentiment analysis.

Our error analysis revealed that while LLM-SEMREL is a robust resource, there are areas for improvement, particularly in the directionality of certain semantic relations. These errors highlight the challenges inherent in automatic annotation using LLMs and suggest avenues for refining our approach.

Looking forward, several directions for future work emerge. Firstly, refining the prompt design and interpretation mechanisms used by LLMs could mitigate the identified errors, enhancing the accuracy of the semantic relations. Additionally, integrating human-in-the-loop approaches for validation and correction could further improve the quality of the database.

Another promising direction is the expansion of LLM-SEMREL to include more semantic relations and a broader vocabulary. This could involve incorporating additional linguistic resources and leveraging advancements in LLMs to generate even richer datasets. Furthermore, applying LLM-SEMREL to other NLP tasks beyond coreference resolution could demonstrate its versatility and utility across different applications.

In conclusion, LLM-SEMREL represents a significant step forward in the creation of semantic resources for the Portuguese language. While there are challenges to address, the potential benefits for NLP applications are substantial, paving the way for more accurate and contextually aware language models. We look forward to the continued evolution of this resource and its impact on the field of NLP.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Rozière, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I. M., Misra, I., Evtimov, I., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Upasani, K.,

Plawiak, K., Li, K., Heafield, K., Stone, K., and et al. (2024). The llama 3 herd of models. *CoRR*, abs/2407.21783.

Fonseca, E., Vieira, R., and Vanin, A. A. (2016). Improving coreference resolution with semantic knowledge. In *Computational Processing of the Portuguese Language - 12th International Conference, PROPOR 2016, Tomar, Portugal, July 13-15, 2016, Proceedings*, volume 9727 of *Lecture Notes in Computer Science*, pages 213–224. Springer.

Fonseca, E. B. (2018). *Resolução de correferência nominal usando semântica em língua portuguesa*. PhD thesis. Escola Politécnica.

Gonçalo Oliveira, H. (2012). *Onto. PT: Towards the Automatic Construction of a Lexical Ontology for Portuguese*. PhD thesis, Ph. D. thesis, University of Coimbra.

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., and Klakow, D. (2021). A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568. Association for Computational Linguistics.

Jiang, F. and Cohn, T. (2021). Incorporating syntax and semantics in coreference resolution with heterogeneous graph attention network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1584–1591. Association for Computational Linguistics.

Khosla, S. and Rose, C. (2020). Using type information to improve entity coreference resolution. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 20–31. Association for Computational Linguistics.

Lima, T., Collovini, S., Leal, A. L., Fonseca, E., Han, X., Huang, S., and Vieira, R. (2018). Analysing semantic resources for coreference resolution. In *Computational Processing of the Portuguese Language - 13th International Conference, PROPOR 2018, Canela, Brazil, September 24-26, 2018, Proceedings*, volume 11122 of *Lecture Notes in Computer Science*, pages 284–293. Springer.

OpenAI (2023). GPT-4 technical report. *CoRR*, abs/2303.08774.

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T. P., Alayrac, J., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., Antonoglou, I., Anil, R., Borgeaud, S., Dai, A. M., Millican, K., Dyer, E., Glaese, M., Sottiaux, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Molloy, J., Chen, J., Isard, M., Barham, P., Hennigan, T., McIlroy, R., Johnson, M., Schalkwyk, J., Collins, E., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Meyer, C., Thornton, G., Yang, Z., Michalewski, H., Abbas, Z., Schucher, N., Anand, A., Ives,

R., Keeling, J., Lenc, K., Haykal, S., Shakeri, S., Shyam, P., Chowdhery, A., Ring, R., Spencer, S., Sezener, E., and et al. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Speer, R. and Havasi, C. (2012). Representing general relational knowledge in conceptnet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 3679–3686.

Ueda, R. (2005). Dicionário br.ispell. `https://github.com/fititnt/br.ispell-dicionario-portugues-brasileiro?tab=readme-ov-file`.