Isolating conceptual semantics by averaging natural language translations using Sparse Autoencoders

Anonymous ACL submission

Abstract

We describe an experiment that isolates conceptual semantics by averaging concept activations derived via Sparse Autoencoders. By translating between natural languages and averaging the concept vectors we can mechanistically interpret more accurate meaning from internal states. We apply the experiment to the domain of Ontology Alignment, which seeks to align concepts across different representations of domains. Our results show that improvements occur when averaging the concept activations of English texts and their French and Chinese translations. The trend of improvement correlates to the reduction in symbolic representation from French to Chinese, indicating that the overall process is isolating conceptual semantics by averaging out language specific symbolic representations.

1 Introduction

We parse OWL ontology classes into text representations and use these as prompts to input into Sparse Autoencoders (SAEs). The output concept set of activations is initially noisy (low correlation from final analysis) and so we perform a natural language translation and compare the translated text's concept activations. The resulting average of different natural language text representations is more semantically relevant to the annotated ground truth that accompanies the corpus (from the Ontology Alignment Evaluation Initiative 2024 [1]). We compare the concept activations for each ontology class and correlate with the ground truth class mappings. We find that the difference between the multi-language averages and the single language analyses is significant and that the effect of translating and averaging appears to isolate conceptual semantics.

2 Related Work

An *ontology* in general terms can be thought of as a formal representation of a domain of knowledge. Any representation has to be subjective and, in practice, ontologies tend to be bespoke to a domain or application.

2.1 Ontology Parsing

OWL[2] ontologies are a standard, machine-readable and flexible format for representing any domain. Due to the subjective nature of semantic representation, our goal of creating a text prompt from an OWL ontology is also subjective. The extraction of OWL classes, properties and relationships can be performed with libraries for various programming languages (we used OWLAPI (Horridge and Bechhofer, 2011)), and tools have been created to generalise text extraction, e.g. NaturalOWL (Androutsopoulos et al., 2013) and OWL Verbalizer (Kaljurand and Fuchs, 2007). The *recursive concept verbaliser* approach for ontology subsumption inference (He et al., 2023) presents a toolbox for OWL ontology analysis (OntoLAMA).

2.2 Ontology Alignment

The challenge of matching concepts between ontology representations is as old as the representations themselves. Since 2004, the Ontology Alignment Evaluation Initiative[3] has provided a framework for evaluating various approaches. From straightforward lexical approaches, through structural and semantic techniques to more recent innovations with machine learning (Qiang et al., 2023) (and a multitude of hybrid methods (Euzenat et al., 2004; Codescu et al., 2014; Jiménez-Ruiz and Cuenca Grau, 2011)), we believe our research is novel in approaching the problem with analyses of LLM internal concept states.

[1] https://oaei.ontologymatching.org/2024/conference/index.html

[2] https://www.w3.org/OWL/

[3] https://oaei.ontologymatching.org/

2.3 Mechanistic Interpretability and Sparse Autoencoders

Mechanistic Interpretability (MI) is a domain which aims to interpret the internal activation states of neural networks for various purposes such as AI safety, neural network decision-making and improving network design (Sharkey et al., 2025). Our interest in MI is for the learned concept activations — the correlation between conceptual semantics and node activations of LLMs.

When applied to language models, Sparse Autoencoders are unsupervised algorithms that learn to map from latent representations to interpretable concepts (also called features). An SAE is a pair of encoder and decoder functions that compresses an input into a hidden representation and tries to reconstruct the input from the hidden representation — thereby learning a set of activation features which can be correlated via techniques such as Dictionary Learning (Bricken et al., 2023) to a vocabulary of human understandable concepts. The sparsity controls that are applied during training result in a reduced set of activations that are more easily computed (compared with billions of activations in a full LLM).

Gemma Scope (Lieberum et al., 2024) is an open suite of SAEs trained on Google's Gemma 2 LLM — at every layer and sublayer.

3 Method

The corpus used in this experiment comes from the conference track of the Ontology Alignment Evaluation Initiative 2024. Across the 16 ontologies, there are 867 class definitions and a set of 174 reference class mappings as ground truth alignments.

The method we apply can be broken down into stages, which are:

1. We use the Java library *OWLAPI*[4] to parse each owl ontology file. Due to the nature of owl representations, each ontology can take a different format and so the Java script has bespoke logic that extracts classes, any related subclasses, superclasses, object properties and data properties. Some manipulation of the representation is needed. There are two styles of output we create: a summary and a verbose version — summary and verbose examples for the class Author are shown in quotes (A) and (B), below. The verbose output is a text string which encapsulates a description of the class and includes connecting and descriptive words, but the summary version is simply a concatenation of the target class name and any associated class names.

> (A) Author is a SuperClassOf Presenter and hasRelatedPaper Paper

> (B) Author is a SubClassOf some writes Contribution and is a SubClassOf Person and is a SubClassOf only writes Contribution and writes Contribution

2. We use the *googletrans* Python library[5] to perform a natural language translation from English to French. There are no parameters supplied to this process - it's a straightforward translation service. Examples of French and Chinese translations of summary and verbose representations are shown below:

> (C) Personne auteur uniquement Contribution Certaines écritures Contribution Contribution

> (D) L'auteur est une sous-classe de contribution des écritures et est une personne sous-classe et est une sous-classe unique en rédaction de contribution et écrit la contribution

> (E) 作者有些人写贡献只写贡献

> (F) 作者是一个子类人，是一个仅写作贡献的子阶级，并且是某些撰写贡献的子类别，并写下了贡献

3. Using the *huggingface* library[6] (Wolf et al., 2020) to access the Gemma Scope open suite of sparse autoencoders, we process each text representation as a prompt to a PyTorch neural network (using a Jump ReLu activation function). The particular SAE set used here is the 2 billion parameter model based on Google's Gemma 2 Large Language Model. We take every layer (0 to 25) of the 16.4k width model and we take the L0 Norm variant for model regularisation (the number of non-zero elements in the activation vector) where the average is between 13 and 23 active features as possible (e.g. 13 out of 16.4k on average). If L0 is set too high then features overlap and interpretability breaks down. Set L0 too low and the network underfits and misses important structures. The output is a set of concepts and an activation weighting. An example tensor with 8 concept identifiers and activation values is shown in (G).

[4] https://github.com/owlcs/owlapi

[5] https://pypi.org/project/googletrans/

[6] https://huggingface.co/google/gemma-scope

(G) [[9664.0000, 50.2211], [3923.0000, 25.5779], [4819.0000, 19.8034], [1072.0000, 28.4082], [4819.0000, 20.4854], [15978.0000, 18.5160], [9271.0000, 29.4970], [8433.0000, 20.8372]]

4. The same process as in (3) is repeated for the translated (French and Chinese) texts.

5. For each class representation, the English and translated concept activation sets are averaged, using a weighting average function that preferences the same concept activations taking the relative weights into account. The average is of the same concept identifier and weighting value format.

6. Every class average is compared with every other class average using a Cosine Similarity function. For control purposes, the similarity is also computed for the same language concept sets - all English values are compared with all English, and the same for the French and Chinese equivalents. Where the class comparison has a pre-defined mapping in the ground truth dataset, we align the similarity score with a target variable value of *1*, else it is set to *0*. An example record, showing the similarity score of the *Author* classes from the *emt* and *edas* ontologies, with a target of *1* is shown below, at (H).

(H) *cmt-Author,edas-Author,0.8362799,1*

7. The resulting output of the previous steps, is a set of differences between representations of classes from the source ontologies. The ground truth class reference mappings are used to create a correlation between the correct relationship and the conceptual difference. The correlation algorithm used is the Point-Biserial Correlation, which is ideal for correlations between binary and continuous variables.

Due to the nature of the corpus, the layer-by-layer analysis only has 174 ground truth mappings (from a total of 370,000 class comparisons) and hence there is a large class imbalance for each layer. We reduce the imbalance by using a random re-sampling to reduce the *false* target variable size to be the same as the *true* size.

A further analysis was undertaken using the same approach, but instead of translating to French, simplified Chinese was used.

Example code is available for validation[7].

[7]TBC

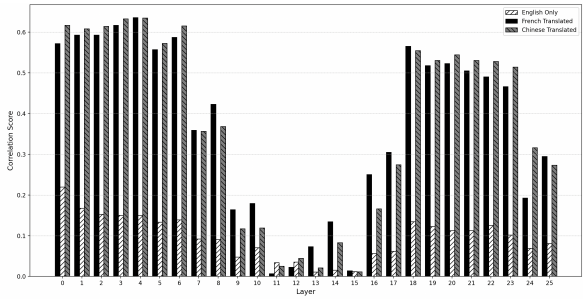

Figure 1: Summary prompt - translated correlation vs English-only

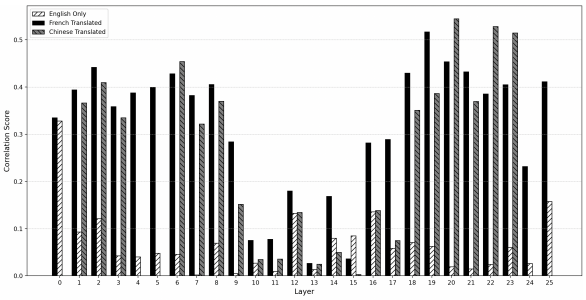

Figure 2: Verbose prompt - translated correlation vs English-only

4 Results

The outputs consist of a correlation between ontology classes and over a number of variations such as the nature of the text representation (either a summary or verbose representation) and the layer of concept activation (0-25). These results are then compared at the level of language, e.g. are the English and French concept sets different from the concept sets generated after averaging the different language representations.

Results are shown in Figure 1, for the summary text representation, and Figure 2 for the verbose prompt. Both results show a clear improvement in correlation after averaging the activations between English and the translations (either French or Chinese). There is a small increase in correlation when we look at averages of translations between English and Simplified Chinese for summary texts, but a reduction for the verbose texts.

Table 1 shows the correlations compared.

NB: some results are incomplete at time of writing.

5 Discussion

In the same way that "Conceptual Semantics takes the meanings of words and sentences to be structures in the minds of language users" (Jackendoff,

Text Version	Language	Correlation
Summary	English only	0.093
Summary	Avg Eng/French	0.371
Summary	Avg Eng/Chinese	0.372
Verbose	English only	0.004
Verbose	Avg Eng/French	0.302
Verbose	Avg Eng/Chinese	0.249

Table 1: Average correlations for each text version and the translated language

2006), we might assume that LLMs have structures which represent the meaning of words that are processed through their neural network states. We might also assume that LLMs don't have concepts of meaning in themselves, but instead are learning and storing correspondences between symbolic and linguistics structures upon which LLMs are trained. Arguments are emerging, however, which show that LLMs do represent real world concepts (Gurnee and Tegmark, 2023 and Kim et al., 2025) beyond the purely linguistic. When we peek into the network internals (via Sparse Autoencoders) we often see that concepts are activated which relate to surface cues such as syntactic and linguistic semantics, but our results show that we can reduce the symbolic concept space for a set of activations and isolate concepts that reflect a purer semantic representation.

After we extract an English text version of an ontology class, we put that string through a set of SAE neural networks and compare the concept activations between other classes from related ontologies. We have a (small) set of ground truth correspondences between classes and we see a fairly weak correlation emerge. This is potentially due to the small size of the corpus and also the relatively subjective extraction from OWL representation to a string of words. We notice that there is a difference in overall correlations between the summary extracted text and the more verbose version.

We take the same English text and translate it to French and Simplified Chinese and put these two prompts through the same SAEs. The resulting concept activation sets are averaged for each ontology class between the English and French and between the English and Chinese versions. This represents a different concept activation set for each ontology class. When we calculate the same correlations as with the English only activations, we see a significant improvement in correspondence for the French translation and an even stronger correlation for the Chinese translations. The difference in average correlations between the French and Chinese (from Table 1) is small, however the average percentage difference is 9% (summary).

We suggest that the translation and averaging process is removing linguistic specific concepts and leaving concepts that are a purer representation of the core semantics of the original prompt. The trend for this pattern to be stronger for the Chinese translation adds credence to the argument since Chinese language tokens contains fewer syntactic elements, "e.g. more frequent functional words in English texts" (Wang and Jiang, 2024). Given that the dataset is small and the representations and translations relatively subjective, this result should be validated.

This slightly unexpected result hints at a new technique for improving conceptual analyses of LLMs, especially via SAEs. We expect future research to confirm and extend this result.

6 Future Work

We highlight some problems which we hope to address in future versions of the research.

The accuracy of extraction of class representations is exposed to problems of subjectivity. Both the conceptual model used to create an OWL ontology and the extraction process to create a text string version are prone to idiosyncrasies in design.

The corpus used is relatively small, having a low number of OWL classes. There is also a class imbalance because there are missing ground truth mappings for many OWL classes. The ground truth is also a manual annotation and liable to potential bias.

The use of SAEs for interpretability is a relatively novel approach and there are known challenges e.g. feature splitting (Chanin et al., 2024), terse concept dictionaries and reconstruction errors (Shu et al., 2025).

More explicitly, these areas are in scope for next steps: (*i*) Extend the corpora to confirm and explore this result, (*ii*) Explore a generalised ontology class extraction process, (*iii*) Analyse concept features for a common sense analysis, (*iv*) Apply improvements in interpreting conceptual semantics to Ontology Alignment tasks.

7 Limitations

At the time of writing, the results presented are not fully complete across all layers of the model (some Chinese results were not completed). Other limitations are described in the main text above, related to the subjectivity of various elements in the experiment.

References

Ion Androutsopoulos, Gerasimos Lampouras, and Dimitrios Galanis. 2013. Generating natural language descriptions from owl ontologies: the naturalowl system. *Journal of Artificial Intelligence Research*, 48:671–715.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, and 1 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, and Joseph Bloom. 2024. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*.

Mihai Codescu, Till Mossakowski, and Oliver Kutz. 2014. A categorical approach to ontology alignment. In *Proceedings of the 9th International Workshop on Ontology Matching collocated with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Trentino, Italy, October 20, 2014*, volume 1317. CEUR-WS. org.

Jérôme Euzenat, David Loup, Mohamed Touzani, and Petko Valtchev. 2004. Ontology alignment with ola. In *Proc. 3rd ISWC2004 workshop on Evaluation of Ontology-based tools (EON)*, pages 59–68. No commercial editor.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Yuan He, Jiaoyan Chen, Ernesto Jimenez-Ruiz, Hang Dong, and Ian Horrocks. 2023. Language model analysis for ontology subsumption inference. *arXiv preprint arXiv:2302.06761*.

Matthew Horridge and Sean Bechhofer. 2011. The owl api: A java api for owl ontologies. *Semantic web*, 2(1):11–21.

Ray Jackendoff. 2006. On conceptual semantics.

Ernesto Jiménez-Ruiz and Bernardo Cuenca Grau. 2011. Logmap: Logic-based and scalable ontology matching. In *International Semantic Web Conference*, pages 273–288. Springer.

Kaarel Kaljurand and Norbert E Fuchs. 2007. Verbalizing owl in attempto controlled english.

Junsol Kim, James Evans, and Aaron Schein. 2025. Linear representations of political perspective emerge in large language models. *arXiv preprint arXiv:2503.02080*.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*.

Zhangcheng Qiang, Weiqing Wang, and Kerry Taylor. 2023. Agent-om: Leveraging llm agents for ontology matching. *arXiv preprint arXiv:2312.00326*.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, and 1 others. 2025. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. 2025. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. *arXiv preprint arXiv:2503.05613*.

Letao Wang and Yue Jiang. 2024. Do translation universals exist at the syntactic-semantic level? a study using semantic role labeling and textual entailment analysis of english-chinese translations. *Humanities and Social Sciences Communications*, 11(1):1–12.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.