John Kausch, Western University

Cosine Similarity Indexing of Word Embeddings using Knowledge Organization Systems

Abstract

This paper proposes a new technique for cosine similarity indexing in the era of large language models (LLMs). It investigates how knowledge organization systems (KOS) can be used to index the latent spaces which LLMs produce. A latent space is a multidimensional feature space used by a model to encode the context of data items. In the case of an LLM, a typical latent space is a word embedding, which gives every word a "position" in a multidimensional feature space, where the features are opaque, and not human-readable. This work asks: can indexing such latent spaces with KOSs help make LLMs more explainable? It builds on previous work in latent semantic indexing for information retrieval models to see if similar techniques can be used to bridge KOSs and LLMs. It also investigates how this method can be applied to improving the performance of multilingual information retrieval. A cross-lingual ontology (called Horapollo) is used to index two latent spaces containing Wikipedia articles written in English and Arabic. Then, the distance between equivalent articles in both spaces are taken, raising questions about the use of KOSs for multilingual and transdisciplinary information retrieval tasks in the era of semantic search.

1. Background

There are similarities between word embeddings in LLMs and latent semantic indexing techniques for information retrieval (Golub, 2019). Both contain lists of words followed by a large matrix of many dimensions, where each dimension contains a number for each word. In some count-based information retrieval spaces, each dimension is a document, and each number is a count of the times a term occurs in a document (Qi et al., 2024). In the case of a word embedding, each dimension is a feature determined by the LLM (Mikolov et al., 2013). Human researchers cannot interpret the significance of these features, although some have tried (Olah et al., 2018). In this case, the value for each feature represents an opaque quantity which has meaning only to a machine. This makes determining bias difficult. For both count-based information retrieval indices and word embeddings the similarity between words can be determined using common methods like the Euclidian distance, or especially cosine similarity (Toshevska e tal., 2020).

In classical problems of latent semantic indexing studies sought to demonstrate that it was possible to represent a very large count-based matrix as a smaller space with opaque features, but where relationships between terms would be preserved (Deerwester et al., 1988). This saved memory and computational costs for performing retrieval tasks. The trade-off is that the features of the latent space which such methods produce are typically not human readable. This study investigates a technique which would work in the opposite direction: is it possible to index a non-human readable word embedding space of an LLM with a KOS to make it human readable? In this case the dimensions of the feature space are the classes or terms of the KOS, and the values for each word or document are its cosine similarity to that term in the original space. Another question this raises is: can KOS produced latent spaces achieve similar results on word embedding benchmarks for information retrieval tasks as the original, LLM-produced word embeddings? It also provokes questions about bias, e.g., will a word embedding space indexed by a KOS biased towards a particular domain perform better on benchmark tasks related to that domain rather than others?

The value of this is not that it eliminates bias; rather, techniques like latent semantic indexing using a KOS can serve to make bias explicit (Hjorland, 2007). All KOSs are biased, but unlike the opaque bias of a LLM, the bias of a KOS can be critiqued, because the classes or terms in a KOS have explicit labels (Feinberg, 2007). It is easier to debate the ethics of the use of specific terms in certain conceptual contexts than it is to debate the ethics of abstract feature spaces which no one researcher can understand. Thus, this method could potentially provide a

way of investigating the bias contained in specific layers of an LLM beyond interrogating the training data.

Another value of this technique is that it could create new forms of interoperability, although that prospect is very tentative, and outside the scope of this study. Recently the machine learning literature has put forward the idea of knowledge distillation (Guo et al. 2021) where the parameters of a larger, source model are distilled into the parameters of a smaller, target model; models such as DeepSeek (Guo et al., 2024) had a large impact on the press and the industry using this approach. Cosine-similarity indexing with a KOS might yield similar results to knowledge distillation in that it could potentially take knowledge stored in the weights of a neural network and transfer them to another context. The potential value of this is that knowledge from different LLMs trained on either related or different domains could be indexed using a KOS, and the "distance" between the classes and various words, documents and data items compared.

The main application where this technique may be highly useful however is for multilingual information retrieval (Sujata, 2011). In this use-case, a set of documents indexed by a KOS in one language could be related to documents indexed by the same KOS in another, leading to an alignment between the latent spaces which could produce a single, higher-level latent space, which could be used to support multilingual queries. It is this application that is briefly explored in this paper.

Cosine similarity methods for information retrieval are well established. However, the aim of this study is not to produce a list of recommendations when a user enters a query; rather, it is to produce a secondary latent space, indexed by a KOS, which can aid in multilingual and transdisciplinary information retrieval when a user enters a query. It is beyond the scope of this paper to test whether a KOS-indexed latent space can perform similarly on benchmark tasks to either a count-based document embedding or a semantic-search word embedding model. Instead, this paper aims to accomplish an exploratory study of the cross-linguistic distance between concepts, to prepare the way for future studies which will investigate questions of how KOS-indexed latent spaces can be used. The fundamental question this paper asks is: is there a noticeable difference in the distance between documents on the same topics written in different languages when indexed by the same KOS? This exploratory step will prepare the way for later studies.

2. Methods

The exploratory method uses the cosine similarity between terms in a KOS and a latent space to index the latent space. In order to generate a KOS-indexed latent space from word embeddings, exemplar objects for each class in the KOS are chosen from the latent space. Then, the cosine similarity from these exemplars to every other object in the latent space is taken, to produce a new latent space whose dimensions are the classes and where each word is categorized according to its cosine similarity to the exemplar of that class.

For instance, for a class "Bird," the Wikipedia article for 'Bird' could be selected as an exemplar for that dataset. Then, the cosine similarity for that document to every other document in the dataset is taken to create a feature vector. For the purposes of calculating, the arccosine of the cosine similarity is taken to give each document a position on the unit circle. This permits the conceptualization of the difference between documents in different languages as the angle of rotation. The arccosine is represented in degrees rather than radians.

For this study, a word embedding space made using Word2Vec (Mikolov et al., 2013) known as Wikipedia2Vec (Yamada et al., 2020) was selected. Wikipedia2Vec was used as the data source because it contains both words and entities (which are documents) in the same embedding space, while also meeting size constraints. It also supports pre-computed datasets in several languages, including both English and Arabic Wikipedia articles.

For the experiment, the 26 bases classes of the Horapollo Ontology (namespace located at https://realcharacterlanguage.world/horapollo) were used to index the embedding space in both languages. For the English embedding space words were selected using the English labels for the Horapollo ontology; for the Arabic embedding space they were selected by manually retrieving the headings of Arabic pages which were translations of English topics. Because the Arabic precomputed embeddings came from a 2016 dump retrievable from the Wikipedia2Vec website, only 18 of the 26 classes for Arabic had exemplars in the language; thus only 18 classes were used to compare documents cross-linguistically.

Initially, a random sample of 100 English and 100 Arabic Wikipedia articles were chosen. Equivalent articles were then selected by manually crawling to the Wikipedia pages in both languages and selecting the desired translation into a target language using Wikipedia's language selection function. However, it was found that only roughly 20% of English articles had Arabic translations, as opposed to 80% of Arabic articles having English translations; thus, in order not to bias the dataset, only Arabic to English translation were used. Of the 80 Arab articles with English equivalents, only 35 of these English equivalents were in the pre-computed embedding dataset of Wikipedia2Vec; thus, the final sample was very small, containing only 35 articles, which is a definite limitation of this study. With a sample size that small, only a few descriptive statistics are presented in the next section, along with exploratory remarks for further studies.

Following the indexing process, the arccosine was taken of each cosine similarity to give each document a position on the unit circle for each class. These positions could then be averaged for each of the 18 classes to give an average distance from the indexing class for each document. Following this the difference between the average angles of distance were taken between a document in L1 (English) and L2 (Arabic) – by convention, the angle of distance from the Arabic document was subtracted from the angle of distance of the English.

3. Results

In a presentation on Abstract Wikipedia and the Wikifunctions project, ontologist Denny Vrandecic discusses the problem of un-even distribution between Wikipedia articles in different languages, necessitating the existence of Wikidata identifiers and its sister project Wikifunctions (Sigalov, 2021). The first result of this experiment confirms what is already well-established: there was a proportional misdistribution of overlap between articles in Arabic and English. An English article had a 20% distribution of having an Arabic equivalent, as opposed to an Arabic article having an 80% distribution. Future work in this area could potentially make use of Wikidata identifiers as a mediation tool to eliminate potential biases in sampling exemplars, and permitting greater authority control between selected translations of documents.

Wikipedia Article	Average Difference Per Document
https://en.wikipedia.org/wiki/Lake_Moeris	6.214369792
https://en.wikipedia.org/wiki/Hemothorax	8.902158687
https://en.wikipedia.org/wiki/Emmeline_Pankhurst	-3.05991257

https://en.wikipedia.org/wiki/Mahmoud_El_Nokrashy_Pasha	6.06601974
https://en.wikipedia.org/wiki/Mauritian_Creole	6.101634875
https://en.wikipedia.org/wiki/Brahma	3.375648818
https://en.wikipedia.org/wiki/Azilal	1.537106561
https://en.wikipedia.org/wiki/Rugby_football	3.837840794
https://en.wikipedia.org/wiki/Shrike	-6.08495878
https://en.wikipedia.org/wiki/Top-level_domain	22.01524386
https://en.wikipedia.org/wiki/Mailing_list	0.23913082
https://en.wikipedia.org/wiki/Consumer_protection	9.303928509
https://en.wikipedia.org/wiki/Battle_of_Plassey	-0.58807958
https://en.wikipedia.org/wiki/Hisban	-2.11604202
https://en.wikipedia.org/wiki/Mir_Osman_Ali_Khan	14.26797266
https://en.wikipedia.org/wiki/Talkalakh	-5.56306966
https://en.wikipedia.org/wiki/Abu_Qir_Bay	8.004979869
https://en.wikipedia.org/wiki/Convergent_series	-2.88444509
https://en.wikipedia.org/wiki/Treaty_of_Verdun	1.102277847
https://en.wikipedia.org/wiki/Maeda_clan	4.494729057
https://en.wikipedia.org/wiki/Minority_group	0.755303445
https://en.wikipedia.org/wiki/Reye_syndrome	6.656590883
https://en.wikipedia.org/wiki/Kingston_University	7.716768376
https://en.wikipedia.org/wiki/Frosinone_Calcio	3.31832679
https://en.wikipedia.org/wiki/Lichen	1.104749211
https://en.wikipedia.org/wiki/Banu_Qasi	-0.76116405
https://en.wikipedia.org/wiki/Woman	0.976662377
https://en.wikipedia.org/wiki/Nodal_analysis	-2.91380938
https://en.wikipedia.org/wiki/History_of_Baghdad	2.583410747
https://en.wikipedia.org/wiki/Glazov	8.876932784
https://en.wikipedia.org/wiki/Hot_spring	7.54298593
https://en.wikipedia.org/wiki/Human_Rights_Campaign	4.384513787
https://en.wikipedia.org/wiki/Khalil_al-Wazir	-1.84065524
https://en.wikipedia.org/wiki/Persepolis	1.701910863
https://en.wikipedia.org/wiki/Public_policy	0.850052484
AVERAGE DIFFERENCE ACROSS ARTICLES:	3.317688948

Table 1: Difference Between Average Distances

Table 1 shows the difference between the average distances as calculated for English and Arabic. The maximum difference between documents was the Wikipedia article “Top Level Domain,” whose English position was on average 22 degrees more distant than its Arabic equivalent. The minimum difference article was “Mailing List” whose average English position was only 0.2

degrees more distant than the Arabic. English articles tend to be more distant from the 18 source concepts than Arabic articles; 25 articles have differences where the English article is on average more distant, as opposed to 10 articles in Arabic. The mean difference is slight, with an average across articles of English articles being 3.3 degrees more distant; the standard deviation from the mean is 5.5 degrees. While the sample size is too small to reasonably draw conclusions from inferential statistics, the correlation between the feature spaces for each of the 18 classes was too small in each case to suggest any relationship even were the results significant. This is not surprising, as the distances between classes and documents has no absolute meaning, and instead signifies the relative similarity within the latent space.

4. Discussion

There are several heuristic problems in how to interpret these results. There is no absolute meaning to an angle of distance between a class and similar documents; rather, the angles represent relative difference between contexts. This raises the problem of the incommensurability of contexts across domains and across languages. A cursory look at the results suggests that the difference between the positions of Arabic and English documents on the same topics is relatively similar, supporting the idea of KOS-indexing as a method for creating mappings across different latent spaces. However, without deeper metrics which can look at the shape of the data, it is difficult to determine the significance of the distances. For instance, the shape of the Arabic data may be such that the articles cluster within certain bands of difference, meaning that a minute difference in rotation between Arabic articles could correspond to a much greater angle of rotation for English articles. Even if such measurements of the general shape of the data could be taken, there is no way of establishing the subjective or qualitative difference in meaning without doing studies with human participants. As such, although the results look promising, there are several potential confounds, and little can be stated with certainty at this time. Future research is necessary to examine questions of incommensurability, translation, and this indexing method.

5. Conclusion

This experiment demonstrated that it is possible to cross-linguistically index documents in different latent spaces with a pre-defined KOS, exploring new ways in which multilingual information retrieval could be practiced in the age of semantic search. While there are several confounds in interpreting the results, initial results look promising, suggesting viability for future studies with more qualitative methods and human participants. This work demonstrates that KOSs can still be applied to indexing tasks for latent spaces in a new era, but more work is necessary to test the utility of such an index for specific information retrieval tasks, benchmarks, and user intuitions.

6. References

Brickley, D., & Miller, L. (2007). FOAF vocabulary specification 0.91. [Most recent version available at: http://xmlns.com/foaf/spec/]

Ceccarelli, D., Lucchese, C., Orlando, S., Perego, R., & Trani, S. (2013). Learning relatedness measures for entity linking. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management (pp. 139-148).

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., & Beck, L. (1988). Improving information-retrieval with latent semantic indexing. In Proceedings of the ASIS annual meeting (Vol. 25, pp. 36-40). 143 OLD MARLTON PIKE, MEDFORD, NJ 08055-8750: INFORMATION TODAY INC.

Feinberg, M. (2007). "Hidden bias to responsible bias: an approach to information systems based on Haraway's situated knowledges" Information Research, 12(4) paper colis07. [Available at http://InformationR.net/ir/12-4/colis/colis07.html]

Golub, Koraljka. (2019). “Automatic subject indexing of text”. Knowledge Organization 46, no. 2: 104-121. Also available in ISKO Encyclopedia of Knowledge Organization, eds. Birger Hjørland and Claudio Gnoli, https://www.isko.org/cyclo/automatic

Guo, D., Zhu, Q., Yang, D., Xie, Z., Dong, K., Zhang, W., ... & Liang, W. (2024). DeepSeek-Coder: When the Large Language Model Meets Programming--The Rise of Code Intelligence. arXiv preprint arXiv:2401.14196.

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, *129*(6), 1789-1819.

Hjørland, B. (2008). Deliberate bias in knowledge organization. Advances in Knowledge Organization, 11, 256-261.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems (Vol. 26). Retrieved from https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The building blocks of interpretability. Distill, 3(3), e10.
Qi, Q., Hessen, D. J., & Van Der Heijden, P. G. (2024). Improving information retrieval through correspondence analysis instead of latent semantic analysis. Journal of Intelligent Information Systems, 62(1), 209-230.

Sigalov , S. (Director). (2021, May 31). *Abstract Wikipedia, WikiFunctions & WikiLambda—A talk with Denny Vrandečić* [Video recording]. https://www.youtube.com/watch?v=0GjNkrvT7Yc

Sujatha, P., & Dhavachelvan, P. (2011). A review on the cross and multilingual information retrieval. *International Journal of Web & Semantic Technology*, *2*(4), 115.

Toshevska, M., Stojanovska, F., & Kalajdjieski, J. (2020). Comparative Analysis of Word Embeddings for Capturing Word Similarities. CoRR.

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., & Matsumoto, Y. (2020). Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 23–30. Association for Computational Linguistics.