# Detecting Idiomatic Multiword Expressions in Clinical Terminology using Definition-Based Representation Learning

**François Remy**
Ghent University - Imec
francois.remy@ugent.be

**Alfiya Khabibullina**
University of Malaga
0611289993@uma.es

**Thomas Demeester**
Ghent University - Imec
thomas.demeester@ugent.be

## Abstract

This paper shines a light on the potential of definition-based semantic models for detecting idiomatic and semi-idiomatic multiword expressions (MWEs) in clinical terminology. Our study focuses on biomedical entities defined in the UMLS ontology and aims to help prioritize the translation efforts of these entities. In particular, we develop an effective tool for scoring the idiomaticity of biomedical MWEs based on the degree of similarity between the semantic representations of those MWEs and a weighted average of the representation of their constituents. We achieve this using a biomedical language model trained to produce similar representations for entity names and their definitions, called BioLORD. The importance of this definition-based approach is highlighted by comparing the BioLORD model to two other state-of-the-art biomedical language models based on Transformer: SapBERT and CODER. Our results show that the BioLORD model has a strong ability to identify idiomatic MWEs, not replicated in other models. Our corpus-free idiomaticity estimation helps ontology translators to focus on more challenging MWEs.

## 1 Introduction

Translation in the biomedical domain remains particularly challenging due to the large number of specific and ad-hoc usage of terminology (Neves et al., 2018, 2022). Some medical ontologies such as UMLS (Bodenreider, 2004) contain more than 4 million entities. Out of these, only a fraction has already been labelled in languages other than English. While large efforts to translate some medical ontologies such as SnomedCT (Schulz and Klein, 2008) can be noted, few if any of these efforts have yet to yield full coverage of the ontology in their target language (Macary, 2020; Auwers, 2020).

Popularity is of course one factor motivating the prioritization of the expert translation of some entity names over others, as translating popular entities makes the ontology usable to a large number of practitioners at a lower cost. But, with the rise of automatic translation tools, another factor worth considering in the prioritization is the translation difficulty of the entities being passed on to medical translation experts. Their efforts should indeed better be directed to cases where automatic translation does not provide good results.

In this context, idiomaticity has a key role to play. Indeed, the automatic translation of idiomatic[1] MWEs poses a significant challenge, as juxtaposing the translation of each individual constituent often results in a loss of meaning that can, in some cases, be catastrophic. This difficulty has been noted by prominent researchers such as Koehn and Knowles (2017) and Evjen (2018). As a result, identifying such idiomatic MWEs would therefore immensely benefit the prioritization of translation efforts of medical ontologies.

While many strategies for identifying MWEs have been presented in the past (Ramisch et al., 2010; Kafando et al., 2021; Zeng and Bhat, 2021), we found that applying them to the medical domain (and especially its clinical counterpart) was challenging due to the extreme corpus size that would be required to produce statistically significant results for the long tail of medical entities.

In this paper, we investigate another approach relying on an ontological representation learning strategy based on definitions, and the empirical properties of semantic latent spaces, described by Nandakumar et al. (2019) and Garcia et al. (2021). In particular, we investigate whether semantic models trained from ontological definitions perform better than other semantic models for the task of identifying idiomatic MWEs without relying on their usage in context, using a novel self-explainability score which will be introduced in Section 2.

[1] MWEs are referred to as idiomatic if their meaning cannot be deduced from the interpretation of their constituents, in line with the definition of "Multiword Terms" presented by Ramisch et al. (2010); examples in the biomedical domain include "Gray Matter" or "Morning Sickness".



*Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 73–80

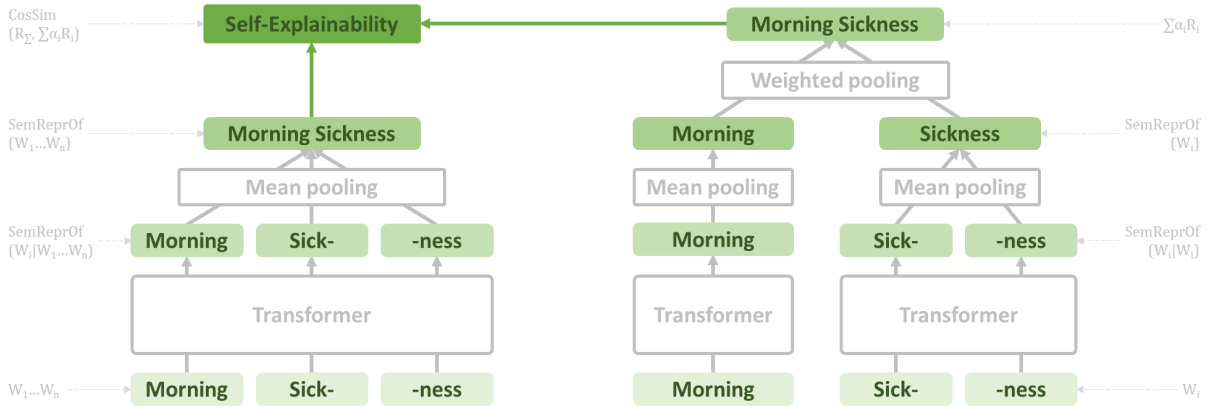


Figure 1: In this paper, we use a cosine similarity metric to compare the representation of a MWE with the weighted average of the representations of its two constituents, after embedding each of these with the same semantic model which is based on a Transformer pipeline. Any difference in representation between these must come from interactions between the constituents within the Transformer when these constituents are combined in the MWE.

## 2 Methodology

After collecting multiword entity names, a chosen semantic model is used to map the obtained MWEs $(W_1...W_n)$ to their latent representations, either as a whole $(\overline{\mathrm{R}_\Sigma})$ or word-per-word $(\overline{\mathrm{R}_i})$.

$$\overline{\mathrm{R}_\Sigma} := \mathrm{SemReprOf}(W_1...W_n)$$
$$\overline{\mathrm{R}_i} := \mathrm{SemReprOf}(W_i)$$

Our semantic model, being based on a Transformer + Mean Pooling pipeline (see Figure 1), produces its representations by averaging the representation of the tokens it is provided as an input (after taking their interactions into account):

$$\overline{\mathrm{R}_\Sigma} = \frac{1}{n}\sum \mathrm{SemReprOf}(W_i|W_1...W_n)$$

To isolate the effect of these interactions, we compute a weighted average of the independent representations of the constituents of the MWE (with weights $\alpha_i$) as a generalization of the above:

$$\sum \alpha_i \overline{\mathrm{R}_i} = \sum \alpha_i \mathrm{SemReprOf}(W_i)$$

Our novel self-explainability score for MWEs corresponds to the degree of similarity between their latent semantic representation $(\overline{\mathrm{R}_\Sigma})$ and the best[2] weighted average of the independent representations of their constituents $(\sum \alpha_i \overline{\mathrm{R}_i})$.

$$\mathcal{S} := \max_{\alpha_i} \left[\mathrm{CosSim}(\sum \alpha_i \overline{\mathrm{R}_i}, \overline{\mathrm{R}_\Sigma})\right]$$

Only strong inter-constituent interactions should be able to explain low self-explainability scores.

Based on this insight, we hypothesize that low self-explainability scores identify the MWEs that the semantic model treats as idiomatic. To validate our hypothesis, we will demonstrate that there is indeed a statistically significant difference in self-explainability scores between idiomatic and non-idiomatic MWEs, among a chosen population.

For our analysis, we construct a set of two-words MWEs obtained from UMLS[3], which were then subsequently divided into two groups by our annotators[4]: those which were "perceived as idiomatic or semi-idiomatic" and those which were "perceived as self-explanatory".

We also hypothesize that a definition-based pretraining is essential for this analysis to produce good results. However, as the proposed analysis could be applied to any contextual text representation model, we set out to evaluate the benefits of the definition-based pretraining of the BioLORD model (Remy et al., 2022) by comparing its results with two strong alternatives: SapBERT (Liu et al., 2021) and CODER (Yuan et al., 2022). These two state-of-the-art biomedical language models were also trained using contrastive learning and UMLS, but not using definitions as a semantic anchor.

[2]We determine the optimal weights $\alpha_i$ in Appendix A.

[3]All two-words entity names from UMLS were included, after filtering out pairs containing words which are either too frequent (>10000 occurences) or too rare (<10 occurences) in the UMLS ontology. This amounts to about 100 thousand two-words MWEs (98.307 to be precise).

[4]The labelling was performed by two annotators: a trained linguist specialized in MWEs who is currently following a course on medical translation, and a NLP practitioner with multiple years of experience in clinical NLP (with an inter-annotator agreement of 82.5% and a kappa score of 0.54).

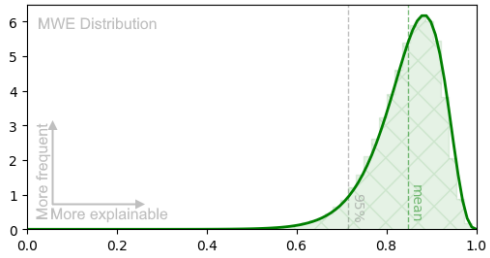


Figure 2: Density of self-explainability scores produced by BioLORD for all the MWEs of our dataset.

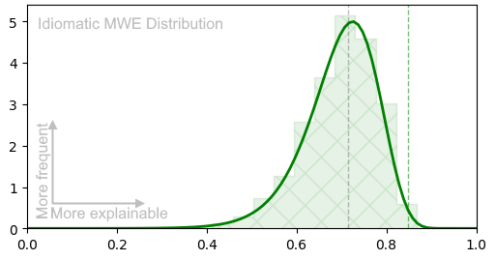


Figure 3: Density of self-explainability scores produced by BioLORD for the idiomatic MWEs of our dataset.

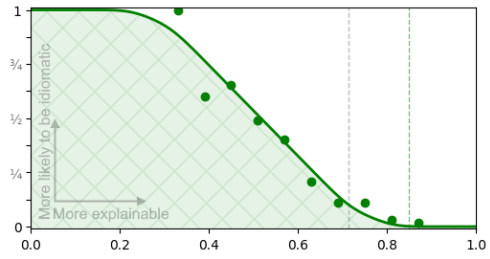


Figure 4: Proportion of MWEs preceived as idiomatic, in function of the self-explainability score produced by BioLORD (bullets represent our annotations).

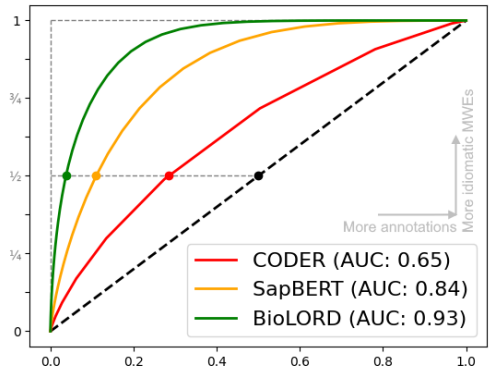


Figure 5: Comparison between the ROC curves of various biomedical models, which shows that BioLORD has a much large area under curve than the other models. The green dot represents the 95th percentile operating point described in the paper; this is the point where about half of the idiomatic MWEs are recalled; achieving the same result with the other models (orange and red dots) or through chance (black dot) requires processing multiple times more MWEs than with BioLORD.

## 3 Experimental Results

We start our analysis by plotting the empirical distribution of self-explainability scores for all considered UMLS entities. We report this empirical distribution as a histogram in Figure 2.

Interestingly, this distribution is unimodal, which seems to give weight to the hypothesis that MWEs exist on a spectrum of idiomaticity, as described by Cowie (1981), and do not form clearly distinct idiomaticity classes.

Based on our annotations, we evaluate the proportion of idiomatic MWEs present in a subset of 10 bins of self-explainability scores (see Figure 4).

This enables us to estimate the full distribution of idiomatic MWEs by multiplying these ratios with the population counts (see Figure 3).

These two distributions have very different means (0.850 vs 0.697), indicating that our self-explainability score is indeed significantly lower for idiomatic MWEs than for non-idiomatic ones.

We determined based on our annotations that about 2.6% of the MWEs in our dataset appeared idiomatic or semi-idiomatic in nature. To evaluate how effectively our self-explainability score can help identifying idiomatic MWEs, we determined the threshold score enabling a recall of about 50% of idiomatic MWEs in our dataset. This corresponds to about 4000 MWEs featuring a similarity below 0.714, consisting of the outliers at or below the 95% percentile of our self-explainability scores.

To confirm this, we annotated more extensively the MWEs of our dataset falling into these 5 outlier percentiles. We find that about 23% of these MWEs appear idiomatic to our annotators, which is in line with our population-based estimates of 26% (2.6% of idiomatic MWEs * 50% recall = 1.3% of idiomatic MWEs out of these 5% of outliers, yielding an expected precision of 26%).

Of course, a threshold of 0.714 represents only one of the possible operating points of our model. By varying this threshold, we compute the receiver operating characteristic (ROC) of our classifier, and plot it in Figure 5 (green curve). We find that our model shows an area under curve (AUC) of 93%.

Repeating this analysis for other semantic biomedical models demonstrates the importance of BioLORD's definition-based training. Indeed, both SapBERT (orange curve) and CODER (red curve) fail to provide a classifier that is as effective as BioLORD for this task, with AUC scores of 0.84 and 0.65 respectively. See also Figure 6.

To enable a more qualitative appreciation of the results, we also report the MWEs featuring the lowest self-explainability scores, for each of the considered models (see Table 1). Based on this, we note that the outliers of the BioLORD model are not only of higher quality, but also feature a significantly lower self-explainability scores. We interpret this as an indication that, to produce definition-grounded representations for MWEs, the BioLORD model has to devote more of its weights to memorize and specialize idiomatic MWEs than the other models.

We can further this impression by looking at Figure 6. While SapBERT has a distribution of scores similar to BioLORD, the difference between the idiomatic and self-explanatory MWEs is less pronounced, leading to more mixups. Looking further, we also notice that the CODER model seems to feature almost no score variation between MWEs in general, and appears to treat few MWEs as idiomatic (besides a few general-purpose hold-outs from its original pre-training). These findings again comfort the idea that a definition-based pre-training is important to achieve good results.

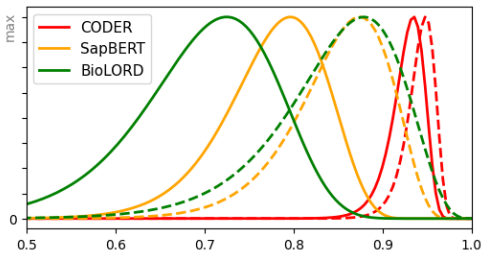


Figure 6: Density of self-explainability scores produced by the compared models for the idiomatic (solid) and self-explainable (dotted) MWEs of our dataset.

| Model | MWE | $\mathcal{S}$-score |
|---|---|---|
| BioLORD | Gray Matter | 0.30 |
| | Neprogenic rest | 0.32 |
| | Heyman operation | 0.33 |
| SapBERT | Ibuprofen dose | 0.49 |
| | Anal Lymphoma | 0.53 |
| | Hemoglobin Wood | 0.54 |
| CODER | United Kingdom | 0.75 |
| | Small Molecule | 0.77 |
| | United States | 0.78 |

Table 1: Most extreme self-explainability outliers for the models compared in this study. An extended version of this table can be found in Appendix A.

## 4 Conclusion

In this paper, we investigated the suitability of definition-based semantic models for detecting idiomatic MWEs in the terminology of a domain. We were able to demonstrate that our proposed self-explainability score can indeed serves as a proxy for idiomaticity, and observed that the BioLORD model indeed displays strong ability to perform this evaluation in the biomedical domain.

The corpus-free idiomaticity estimation thereby developed is powerful enough to help ontology translators to focus on more challenging MWEs, with about half of the idiomatic MWEs contained in the 5% of self-explainability score outliers.

Finally, we were also able to show that biomedical models which were not trained using a definition-based strategy perform significantly worse than our chosen definition-based model, showing the importance of a definition-based pre-training strategy in the development of reliable semantic representations for idiomatic MWEs.

## Limitations

It is worth noting that the approach described in this paper can only be expected to operate reliably on entities which can be accurately represented in the latent space by the chosen semantic model (either through its exposure to textual definitions or ontological relationships about the entity during pre-training, or through its generalization abilities).

Unlike past approaches for detecting idiomatic MWEs, our strategy cannot make use of context to recognize idiomatic MWEs from their usage in a corpus. It would be an interesting future work to investigate how to combine examples of uses and ontological knowledge to develop a better in-context idiomaticity evaluation for MWEs.

An additional limitation of our work, is that we limited our analysis to UMLS entities consisting of exactly two words. This is not a limitation of our proposed approach per se, but we acknowledge that further work should probably be carried out to investigate how to best handle longer sequences.

## Ethics Statement

The authors of this paper do not report any particular ethical concern regarding its content.

## References


Tom Auwers. 2020. Snomed ct translated into dutch and french by belgian national release centre.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, 32(Database issue):D267–70.

A. P. Cowie. 1981. The Treatment of Collocations and Idioms in Learners' Dictionaries. *Applied Linguistics*, II(3):223–235.

John Mervyn Evjen. 2018. Highlighting difficulties in idiomatic translation. *Spectrum*, 2.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

Rodrique Kafando, Rémy Decoupes, Sarah Valentin, Lucile Sautot, Maguelonne Teisseire, and Mathieu Roche. 2021. ITEXT-BIO: Intelligent term EXTraction for BIOmedical analysis. *Health Inf. Sci. Syst.*, 9(1):29.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Francois Macary. 2020. An exemplar of collaboration: The first release of the snomed ct common french translation.

Navnita Nandakumar, Timothy Baldwin, and Bahar Salehi. 2019. How well do embedding models capture non-compositionality? a view from multiword expressions. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 27–34, Minneapolis, USA. Association for Computational Linguistics.

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Cristian Grozea, Amy Siu, Madeleine Kittner, and Karin Verspoor. 2018. Findings of the WMT 2018 biomedical translation shared task: Evaluation on Medline test sets. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 324–339, Belgium, Brussels. Association for Computational Linguistics.

Mariana Neves, Antonio Jimeno Yepes, Amy Siu, Roland Roller, Philippe Thomas, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, Giorgio Maria Di Nunzio, Federica Vezzani, Christel Gerardin, Rachel Bawden, Darryl Johan Estrada, Salvador Lima-lopez, Eulalia Farre-maduel, Martin Krallinger, Cristian Grozea, and Aurelie Neveol. 2022. Findings of the WMT 2022 biomedical translation shared task: Monolingual clinical case reports. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 694–723, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. mwetoolkit: a framework for multiword expression identification. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

François Remy, Kris Demuynck, and Thomas Demeester. 2022. BioLORD: Learning ontological representations from definitions for biomedical concepts and their textual descriptions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1454–1465, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Stefan Schulz and Gunnar O. Klein. 2008. Snomed ct – advances in concept mapping, retrieval, and ontological foundations. selected contributions to the semantic mining conference on snomed ct (smcs 2006). *BMC Medical Informatics and Decision Making*, 8(1):S1.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022. Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of Biomedical Informatics*, 126:103983.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

## A An analytical solution for the optimal vector averaging problem

In this appendix, we derive the analytical solution for the problem of finding the optimal weighted average (of the representation of the constituents of a MWE) given the task of maximizing the cosine similarity between their weighted average and the representation of the MWE itself.

Let $\overline{\mathrm{R}_1}$ and $\overline{\mathrm{R}_2}$ be two vectors (the representation of the words $W_1$ and $W_2$ through the BioLORD model). Let $\overline{\mathrm{R}_\Sigma}$ be a vector (the representation of the MWE through the BioLORD model).

... see Figure A.1 ...

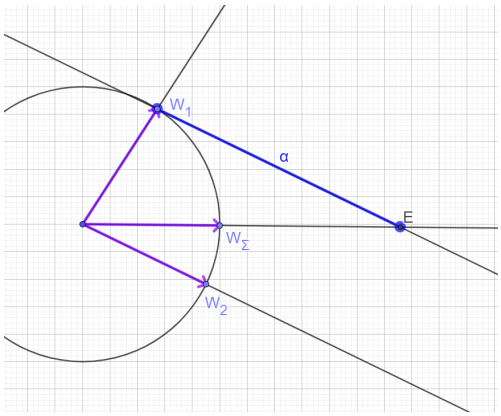


Figure A.1: Representation of the problem

Our objective is to maximize the cosine similarity between $\overline{\mathrm{R}_\Sigma}$ and a weighted average of the vectors $\overline{\mathrm{R}_i}$ (with weights $\alpha_i$). Because the cosine similarity between two vectors does not depend on their respective lengths, we can without loss of generality try to maximize the following expression for the mixing parameter $\alpha = \alpha_2/\alpha_1$.

$$\mathit{CosSim}(\overline{\mathrm{R}_1} + \alpha\overline{\mathrm{R}_2}, \overline{\mathrm{R}_\Sigma}) := \frac{(\overline{\mathrm{R}_1} + \alpha\overline{\mathrm{R}_2}) \cdot (\overline{\mathrm{R}_\Sigma})}{|\overline{\mathrm{R}_1} + \alpha\overline{\mathrm{R}_2}|.|\overline{\mathrm{R}_\Sigma}|}$$

Because the maximum cosine similarity will necessarily be positive, we can look for the maximum of its square instead. We will find our optimum by looking at the points where the derivative is equal to 0:

$$\frac{d}{d\alpha}\left[\mathit{CosSim}^2(\overline{\mathrm{R}_1} + \alpha\overline{\mathrm{R}_2}, \overline{\mathrm{R}_\Sigma})\right] = 0$$

... recalling $\frac{d}{dx}\left[\frac{f}{g}\right] = \left[g\frac{df}{dx} - f\frac{dg}{dx}\right]/\left[g^2\right]$ ...

$$(\overline{\mathrm{R}_1} + \alpha\overline{\mathrm{R}_2})^2 \frac{d}{d\alpha}\left[((\overline{\mathrm{R}_1} + \alpha\overline{\mathrm{R}_2}) \cdot (\overline{\mathrm{R}_\Sigma}))^2\right]$$
$$= ((\overline{\mathrm{R}_1} + \alpha\overline{\mathrm{R}_2}) \cdot (\overline{\mathrm{R}_\Sigma}))^2 \frac{d}{d\alpha}\left[(\overline{\mathrm{R}_1} + \alpha\overline{\mathrm{R}_2})^2\right]$$

... computing the inner derivatives ...

$$(\overline{\mathrm{R}_1} + \alpha\overline{\mathrm{R}_2})^2 (2((\overline{\mathrm{R}_1} + \alpha\overline{\mathrm{R}_2}) \cdot (\overline{\mathrm{R}_\Sigma}))(\overline{\mathrm{R}_2} \cdot \overline{\mathrm{R}_\Sigma}))$$
$$= ((\overline{\mathrm{R}_1} + \alpha\overline{\mathrm{R}_2}) \cdot (\overline{\mathrm{R}_\Sigma}))^2 (2(\overline{\mathrm{R}_1} + \alpha\overline{\mathrm{R}_2})(\overline{\mathrm{R}_2}))$$

... dividing both sides by 2 and $(\overline{\mathrm{R}_1} + \alpha\overline{\mathrm{R}_2})(\overline{\mathrm{R}_\Sigma})$ ...

$$(\overline{\mathrm{R}_1} + \alpha\overline{\mathrm{R}_2})^2 (\overline{\mathrm{R}_2} \cdot \overline{\mathrm{R}_\Sigma})$$
$$= ((\overline{\mathrm{R}_1} + \alpha\overline{\mathrm{R}_2}) \cdot (\overline{\mathrm{R}_\Sigma}))((\overline{\mathrm{R}_1} + \alpha\overline{\mathrm{R}_2}) \cdot (\overline{\mathrm{R}_2}))$$

Let's introduce a more convenient notation for the scalar products ($\mathrm{R}_{xy} = \overline{\mathrm{R}_x} \cdot \overline{\mathrm{R}_y}$). Given we are trying to find scaling coefficients for $\overline{\mathrm{R}_i}$ vectors, we can first normalize them to make their norm is equal to one, without loss of generality, such that $\mathrm{R}_{11} = \mathrm{R}_{22} = \mathrm{R}_{\Sigma\Sigma} = 1$.

... expanding the products ...

$$(\mathrm{R}_{11} + 2\alpha\mathrm{R}_{12} + \alpha^2\mathrm{R}_{22})(\mathrm{R}_{2\Sigma})$$
$$= (\mathrm{R}_{1\Sigma}\mathrm{R}_{12} + \alpha\mathrm{R}_{2\Sigma}\mathrm{R}_{12} + \alpha\mathrm{R}_{1\Sigma}\mathrm{R}_{22} + \alpha^2\mathrm{R}_{2\Sigma}\mathrm{R}_{22})$$

... isolating $\alpha$ on the left side ...

$$\alpha(\mathrm{R}_{12}\mathrm{R}_{2\Sigma} - \mathrm{R}_{1\Sigma}\mathrm{R}_{22}) = (\mathrm{R}_{1\Sigma}\mathrm{R}_{12} - \mathrm{R}_{11}\mathrm{R}_{2\Sigma})$$

... giving us the formula of $\alpha$ ...

$$\alpha = \frac{\mathrm{R}_{1\Sigma}\mathrm{R}_{12} - \mathrm{R}_{2\Sigma}\mathrm{R}_{11}}{\mathrm{R}_{2\Sigma}\mathrm{R}_{12} - \mathrm{R}_{1\Sigma}\mathrm{R}_{22}} = \frac{\mathrm{R}_{1\Sigma}\mathrm{R}_{12} - \mathrm{R}_{2\Sigma}}{\mathrm{R}_{2\Sigma}\mathrm{R}_{12} - \mathrm{R}_{1\Sigma}}$$

... giving us the formula of $\alpha_i > 0$ ...

$$\alpha_1 = \mathrm{R}_{1\Sigma} - \mathrm{R}_{12}\mathrm{R}_{2\Sigma}$$
$$\alpha_2 = \mathrm{R}_{2\Sigma} - \mathrm{R}_{21}\mathrm{R}_{1\Sigma}$$

**Intuition:** If we assume that the constituents of the entity have orthogonal meanings ($\overline{\mathrm{R}_1} \cdot \overline{\mathrm{R}_2} = 0$), this gives $\alpha_1 = \mathrm{R}_{1\Sigma}$ and $\alpha_2 = \mathrm{R}_{2\Sigma}$ which are the cosine similarities of each constituent with respect to the entire MWE.

## B Examples of similarity outliers for the considered models

| Word1 | Word2 | Score |
|---|---|---|
| Gray | Matter | 0.303302 |
| Nephrogenic | rest | 0.317366 |
| Heyman | operation | 0.328952 |
| Chemical | procedure | 0.331814 |
| Morning | sickness | 0.359685 |
| Morning | Sickness | 0.359685 |
| Green | Card | 0.364002 |
| Yellow | Fever | 0.365865 |
| Nitrogen | retention | 0.372655 |
| molecular | function | 0.374572 |
| osseous | survey | 0.384946 |
| Refsum | Disease | 0.38831 |
| Monteggia's | Fracture | 0.392137 |
| Silver | operation | 0.393802 |
| Worth | disease | 0.395263 |
| Diseases | Component | 0.398678 |
| Root | stunting | 0.402461 |
| McBride | operation | 0.403504 |
| Air | hunger | 0.405719 |
| Storage | disease | 0.414184 |
| Border | Disease | 0.415117 |
| Intersection | syndrome | 0.417804 |
| Retinal | correspondence | 0.420826 |
| Patch | Testing | 0.423289 |
| Dot | haemorrhages | 0.423748 |
| Coordination | Complexes | 0.4248 |
| White | matter | 0.426788 |
| Molar | concentration | 0.432153 |
| Book | Syndrome | 0.432465 |
| Circulatory | depression | 0.4349 |
| German | Syndrome | 0.436444 |
| Nissen | Operation | 0.438874 |
| Physical | shape | 0.440117 |
| External | features | 0.442601 |
| Anoxic | neuropathy | 0.443183 |
| Compartment | syndromes | 0.445978 |
| Visceral | Myopathy | 0.447205 |
| Tumour | haemorrhage | 0.447391 |
| Mountain | Sickness | 0.44767 |
| Growth | Factor | 0.451592 |

Table B.1: Self-explainability outliers for BioLORD

| Word1 | Word2 | Score |
|---|---|---|
| ibuprofen | dose | 0.488790 |
| Anal | Lymphoma | 0.531192 |
| Hemoglobin | Wood | 0.542635 |
| Ovarian | injury | 0.548922 |
| Ovarian | perforation | 0.557121 |
| Ibuprofen | overdose | 0.569412 |
| hemoglobin | Aurora | 0.575010 |
| miconazole | injection | 0.575241 |
| diphenhydramine | Cartridge | 0.580044 |
| phenylephrine | Injection | 0.584401 |
| Hemoglobin | Mexico | 0.585959 |
| Dexamethasone | Powder | 0.589987 |
| Hydrocortisone | phosphate | 0.592702 |
| Guaifenesin | poisoning | 0.592808 |
| hydrocortisone | receptor | 0.594878 |
| Vaginal | adenocarcinoma | 0.595991 |
| iv | lidocaine | 0.598489 |
| Gonadal | Thrombosis | 0.598919 |
| Rectal | artery | 0.603538 |
| hemoglobin | Cook | 0.606404 |
| Ibuprofen | Powder | 0.606984 |
| hemoglobin | Thailand | 0.608336 |
| Ovarian | vessels | 0.609299 |
| Intestinal | hematoma | 0.610457 |
| diphenhydramine | Injection | 0.611432 |
| hemoglobin | Chicago | 0.611646 |
| Ornithine | Ql | 0.612263 |
| Aspirin | dose | 0.613269 |
| Hydrocortisone | Injection | 0.613701 |
| Ovarian | hematoma | 0.613911 |
| hemoglobin | Oita | 0.614288 |
| Wrist | injection | 0.614621 |
| Hemoglobin | Ohio | 0.614865 |
| Aspirin | overdose | 0.615012 |
| Oral | hemangioma | 0.615188 |
| Hemoglobin | Shanghai | 0.618727 |
| Sodium | retention | 0.619068 |
| Diphenhydramine | overdose | 0.619255 |
| hemoglobin | Bristol | 0.619368 |
| Gonadal | artery | 0.620956 |

Table B.2: Self-explainability outliers for SapBERT

| Word1 | Word2 | Score |
|---|---|---|
| United | Kingdom | 0.754104 |
| Small | Molecule | 0.772967 |
| United | States | 0.775555 |
| Dependent | Variable | 0.796870 |
| patch | clamp | 0.799848 |
| Index | finger | 0.809509 |
| Eggshell | nail | 0.810650 |
| single | molecule | 0.812445 |
| Data | Administration | 0.818826 |
| Alkaline | Phosphatase | 0.818921 |
| Brush | Border | 0.820135 |
| Czech | Republic | 0.821894 |
| CrAsH | compound | 0.822972 |
| Nuclear | medicine | 0.823420 |
| Nuclear | Medicine | 0.823420 |
| Hydrogen | Bonds | 0.823888 |
| Replication | Origin | 0.825065 |
| Wild | Type | 0.825602 |
| Antigen | Presentation | 0.826336 |
| outer | membrane | 0.827730 |
| Inclusion | Bodies | 0.829212 |
| Health | administration | 0.829440 |
| Active | Site | 0.829467 |
| Focus | Groups | 0.830125 |
| Natural | killer | 0.830615 |
| Click | Chemistry | 0.831714 |
| Strand | breaks | 0.832437 |
| proc | gene | 0.832669 |
| Lewis | antigen | 0.833199 |
| lucifer | yellow | 0.833356 |
| Mass | Spectrometry | 0.833356 |
| Foreign | Bodies | 0.833412 |
| Foreign | body | 0.833504 |
| Uvea | language | 0.836055 |
| Williams | Syndrome | 0.836802 |
| pyridoxine | clofibrate | 0.837463 |
| Precision | Medicine | 0.838389 |
| Antigen | Switching | 0.838619 |
| Public | Domain | 0.838712 |
| Data | Acquisition | 0.838931 |

Table B.3: Self-explainability outliers for CODER

## Acknowledgements

This work would not have been possible without the joint financial support of the Vlaams Agentschap Innoveren & Ondernemen (VLAIO) and the RADar innovation center of the AZ Delta hospital group.

We also want to thank the SIGLEX-MWE 2023 (19th Workshop on Multiword Expressions) and the ClinicalNLP 2023 Workshop for organizing the joint session to which this paper was submitted, which enabled the blooming of this collaboration between machine learning engineers and linguists thanks to its existence.