Dispersing the *clouds of doubt*: can cosine similarity of word embeddings help identify relation-level metaphors in Slovene?

Mojca Brglez
Faculty of Arts, University of Ljubljana
mojca.brglez@ff.uni-lj.si

Abstract

Word embeddings and pre-trained language models have achieved great performance in many tasks due to their ability to capture both syntactic and semantic information in their representations. The vector space representations have also been used to identify figurative language shifts such as metaphors, however, the more recent contextualized models have mostly been evaluated via their performance on downstream tasks. In this article, we evaluate static and contextualized word embeddings in terms of their representation and unsupervised identification of relation-level (ADJ-NOUN, NOUN-NOUN) metaphors in Slovene on a set of 24 literal and 24 metaphorical phrases. Our experiments show very promising results for both embedding methods, however, the performance in contextual embeddings notably depends on the layer involved and the input provided to the model.

1 Introduction

In recent decades, metaphors have been recognized as a ubiquitous phenomenon in all types of discourse (Reijnierse et al., 2019; Cameron, 2003; Semino, 2008), and because of their central role in both language, thought and communication (Lakoff and Johnson, 1980; Steen, 2017; Burgers et al., 2016), they have been addressed by various fields and disciplines, from linguistics, neurolinguistics, psycholinguistics, cognitive linguistics, social science, and computer science. The main underlying mechanism of metaphor involves representing one domain in the terms of another (Lakoff and Johnson, 1980, 2003; Kövecses, 2020). The represented domain, usually more abstract, is called the target domain, and the domain it is represented by is called the source domain, which is usually more concrete and based on physical experience. For example, in the expression *political storm*, we represent the target domain of POLITICS in terms of the source domain of WEATHER.

For a metaphor to be apt (Tourangeau and Sternberg, 1981), the domains have to share certain features or relations, but otherwise be sufficiently different from one another. On the one hand, this semantic difference can be observed between the metaphorically used word and its context. Wilks (1978) put forward the idea of metaphors as "selectional preference violations", that is, the context of the metaphorically used word is not the context this word would normally select. On the other hand, metaphorically used expressions also exhibit some form of polysemy in themselves. The contextual meaning of the metaphorically used word is different from its most basic meaning which is expressed in literal contexts. The latter is also the defining factor of the most frequently used procedure for manual metaphor identification in texts (MIPVU, Steen, 2010).

These two facets of metaphors have often been used and explored in automatic metaphor identification approaches. Various methods have been proposed that model language and meaning on the basis of the distributional hypothesis (Harris, 1954), according to which similar words have similar contexts. In these models, the meanings of words are determined by their relationships to other words in that same space, and similar words tend to have similar vectors and similar neighbourhoods. Older approaches to metaphor modelling use distributional vectors created with the help of e.g. latent semantic analysis (Kintsch, 2000; Utsumi, 2011), while more recent ones use distributed word embeddings obtained through deep-learning (Mao et al., 2018; Su et al., 2017). An important distinction can also be drawn depending on the level of metaphor processing: word-level, relation-level, or sentence-level. On the word-level, the task is to determine the metaphoricity of a (or each) word. On the sentence-level, the whole sentence is classified for containing metaphor(s) or not. On the relation-level, which we are concerned with in this experiment, the expressions under question are pairs of words that have a syntactic relation between a source and target term, for instance verb-object (***break*** *a promise*) or adjective-noun constructions (***deep*** *thought*). Related to and sometimes overlapping with relation-level metaphors is the wider class of multi-word expressions (MWEs), which include phraseological units such as idioms and proverbs, and other fixed expressions such as compounds and collocations (Gantar et al., 2018). Especially idioms can overlap with metaphoric expressions by their meaning non-compositionality by which the meaning of the whole cannot be directly derived from the meaning of its parts. Some idioms may in fact even stem from metaphorical conceptualizations, e.g. *to throw dust in someone's eyes*. Another shared characteristic can be lexicalization, that is, both MWEs and conventional metaphors can be included in dictionaries, for example *parent company*. However, MWEs mostly require some extent of syntactic fixedness, and, more importantly, they always require at least 2 constituent words, while metaphors can take form of a word, a phrase, or even a whole paragraph.

In Slovene, automatic figurative language processing is still in its early stages, with only a few semi-supervised (Brglez et al., 2021) and supervised automatic models proposed (Škvorc et al., 2022; Zwitter Vitez et al., 2022). The direct use of cosine similarity between the source and target word for metaphor identification in Slovene has not yet been explored and could possibly allow unsupervised extraction of metaphorical candidates from text, avoiding the need for manually annotated data.

The aim of the experiments presented here is two-fold: 1) to investigate the representation of metaphorical expressions in both static and dynamic embeddings and evaluate their use for metaphor identification, 2) to establish a baseline by which to distinguish between metaphor and non-metaphor.

2 Related Work

Metaphor identification has been approached from various perspectives, using or combining several tools and resources. State-of-the-art approaches for English and other more resourced languages use deep learning methods to train metaphor identification models on large annotated corpora in a supervised manner. Because the focus of our work is on unsupervised classification and evaluation of word embeddings for this purpose, here we only report on some previous work in this same direction.

One of the first unsupervised approaches is by (Shutova et al., 2010) to identify verbal metaphors in the BNC corpus. Starting with a seed set of 62 metaphorical verb-object and verb-subject pairs, they apply unsupervised noun and verb clustering on vectors obtained from corpus frequencies in order to extend the range of target and source concepts. Then, they search the BNC for metaphorical expressions using these two expanded lexicons and achieve a precision of 0.71.

Agres et al. (2016) evaluate both static Word2vec and distributional vectors on data from a behavioural study to test if they encapsulate metaphoricity, familiarity and meaningfulness. They test these features with a multiple regression analysis, to see if they are correlated with cosine similarity. For both vector types, their results show that low values of metaphoricity were predictors of high cosine similarity.

Su et al. (2017) also use word2vec embeddings trained on reference corpora for Chinese and English to investigate their use for the identification of nominal metaphors (X is Y). They devise a method that combines calculating the relatedness of words (X,Y) via cosine similarity with checking for hyper-/hyponymy relation in WordNet. If the similarity is lower than a predefined threshold and the concepts have no taxonomic relationship in WordNet, the candidate is classified as a metaphor. They establish the threshold value of cosine similarity as the best overlap (convergence) between literal recall, metaphor recall and accuracy, and determine it to be at 0.235 for English and 0.575 for Chinese. This also shows that the threshold varies greatly on the language involved and that language-specific baselines need to be determined.

Mao et al. (2018) use CBOW and SkipGram embeddings and WordNet to predict the metaphoricity of a verb in a sentence. For each target word, they find the best-fit synonym, hypernym or hyponym in WordNet that matches the context by having the highest cosine similarity to the context vector of the sentence. Then, they compute the cosine similarity between the best-fit word and the target word, and classify the target word as metaphor if the similarity is lower than a threshold of 0.6, which was pre-established on the basis of a development set.

Shutova et al. (2016) experiment with both visual and linguistic embeddings in predicting phrase-level metaphors. They obtain both individual word embedddings and joint phrase embeddings based on the SkipGram method, and investigate various combinations of measuring similarity via cosine distance. They obtain best results with their multimodal approach, while in linguistic embeddings-only setting, computing the similarity between the words in the phrase outperforms computing similarities of phrase embeddings.

In a semi-supervised manner, Zayed et al. (2018) use a seed set of verb-noun phrases to determine the metaphoricity of the candidate verbs on the phrase level. First, they find the most similar verb in the seed set using cosine distance and Word Mover's Distance, and compare the similarity of the candidate noun to the nouns associated with the most similar verbs in the seed set. They also compare GloVe and Word2Vec static embeddings methods, and achieve the best results using GloVE embeddings and cosine distance.

More recently, Pedinotti et al. (2021) tested the knowledge instilled in BERT models by applying the "landmark method" introduced in (Kintsch, 2000), which tries to determine which properties are transferred from the source to the target domain. Namely, metaphoricity relies on some common ground between the two domains which makes the comparison plausible. In their experiment, Pedinotti et al. check whether the representations of metaphors are closer to these common ground 'landmarks' or to the literal properties of source domain words that are not relevant to the metaphor mapping. They conclude that metaphorically used words are consistently more similar to literal landmarks in the first few layers of BERT embeddings. Moreover, they observe a difference comparing conventional and creative expressions: while models achieve steadily better accuracy (in terms of wrong answers) for conventional metaphors, the accuracy actually drops in the later layers for creative metaphors.

Among unsupervised approaches to MWEs, which are somewhat similar to relation-level metaphors, we can mention Cordeiro et al. (2019) and Garcia et al. (2021). Cordeiro et al. (2019) investigate English nominal compounds, where the head of the phrase is a noun (adjective-noun and noun-noun), and their syntactic counterparts in French and Portuguese. To distinguish compositional from non-compositional (idiomatic) MWEs, they measure the cosine similarity between the combined vectors of the parts and the vector of the compound. Moreover, they investigate the influence of various variables: different distributional models, preprocessing methods, dimension sizes, and context sizes. They find that the models can successfully capture idiomaticity, with word2vec as the best performing model for English, while for French and Portuguese, the PPMI-based models fared better. In addition, they find that models for the morphologically richer French and Portuguese benefit from preprocessing steps such as lemmatization and stopword removal. In a more recent approach, Garcia et al. (2021) investigate various contextual models for their representation of potentially idiomatic expressions, i.e. expressions that can be literal or idiomatic depending on the context, in English and Portuguese. They measure the cosine similarity of the embeddings of idiomatic compounds with 1) the embeddings of their meaning-preserving compounds and 2) literal synonyms of the components. Their experiments show the idiomatic phrases are closer to the literal synonyms than to their meaning-preserving paraphrases, leading to the conclusion that idiomaticity is not yet adequately captured by contextual models.

3 Methods

3.1 Dataset

To test our hypotheses, we create a small dataset consisting of metaphorical and non-metaphorical examples of use for 24 Slovene words (8 adjectives and 16 nouns). The examples include three types of constructions: adjective-noun with a potentially metaphorical adjective; adjective-noun with a potentially metaphorical noun; and noun-noun, where the first noun can be metaphorical. All the literal pairs are by default, in the absence of additional context to the contrary[1], considered literal. To provide a sentential context for later use with contextualized embeddings, we concordance one example sentence from the Slovenian reference corpus Gigafida 2.0 (Krek et al., 2019). For each metaphorical-literal pair, we take heed of acquiring syntactically equivalent pairs, thus matching in grammatical gender, case, and number in their

[1] It is possible to use a phrase considered literal on its own in a metaphorical manner. For instance, *dark clouds* is used literally in *The dark clouds spread over the city.*, and metaphorically in *I am plagued by the dark clouds of depression.*

Phrase type	Phrase	Frequency	Example sentence
$NOUN_m-NOUN_l$	**oblaki** dvoma *[**clouds** of doubt]*	11	*Politiki včasih izgubijo zaupanje, četudi se jim laganja izrecno ne dokaže; dovolj je že, da njihovo podobo zastrejo **oblaki dvoma**.* Politicians sometimes lose trust even if their lying is not explicitly proven; it suffices if their image is shrouded by **clouds of doubt.**
$NOUN_l-NOUN_l$	**oblaki** metana *[**clouds** of methane]*	9	*Temperatura na Titanu je ravno prava, da v spodnjih plasteh atmosfere nastajajo **oblaki metana**, iz katerih le ta občasno dežuje.* The temperature on Titan is just right for the **clouds of methane** to form in the lower layers of the atmosphere, where they occasionally rain.
ADJ_m-NOUN_l	**prežvečena** fraza *[**chewed-up** phrase]*	18	*Njegove besede so z dnevi postale **prežvečena fraza**, a so bile prispodoba vsega, kar se je sprehajalo skozi glave številnih, ki so lovili misli, da bi dojeli resničnost.* His words eventually became a **chewed-up phrase** but were a metaphor for everything that went through the heads of many who were hunting for thoughts to understand reality.
ADJ_l-NOUN_l	**prežvečena** hrana *[**chewed-up** food]*	39	*Neredko je vzrok za povečano dejavnost bakterij v črevesu tudi premalo **prežvečena hrana**.* Oftentimes the reason for the increased activity of gut bacteria is insufficiently **chewed-up food**.
ADJ_l-NOUN_m	moralni **steber** *[moral **pillar**]*	12	*Na vasi učitelja dojemajo kot **moralni steber** in pričakujejo, da je vseh pogledih trden in pošten.* In the countryside, people perceive the teacher as a **moral pillar** and expect them to be firm and fair in all aspects.
ADJ_l-NOUN_l	sredinski **steber** *[central **pillar**]*	9	*Ob **sredinski steber** vgradimo leseno pomično steno, s katero ohranimo krožni prehod med prostori, hkrati pa omogoča ločevanje kuhinjskega ali jedilnega dela od dnevne sobe.* By the **central pillar** we build a wooden sliding wall, which maintains the circular passage through the rooms while also allowing to separate the kitchen or dining area from the living room.

Table 1: Examples from the dataset: type of construction, phrase, frequency of the phrase in the reference corpus and an example sentence from the corpus. The subscripted letters $_l$ and $_m$ indicate literal or metaphorical use, respectively.

phrasal and sentential form. Moreover, in order to obtain comparable phrases and to alleviate the potential frequency bias in the embedding space, we avoid overly conventional, common phrases and only choose phrases with less than 65 occurrences in the corpus.

Examples of the three types of phrases are shown in Table 1. For example, for the word *oblak*[cloud], we find a literal word pair *oblaki metana*[clouds of methane] and a metaphorical word pair *oblaki dvoma*[clouds of doubt], and one sentence per pair where the phrases match in grammatical number, gender and case, while also having a similar (low) frequency in the corpus.

3.2 Word embedding models

We compare word embeddings obtained by two methods: static and dynamic. For static embeddings, we use the 100-dimensional CLARIN.SI-embed.sl fastText embeddings (Ljubešić and Erjavec, 2018). For dynamic/contextual embeddings, we obtain 768-dimensional embeddings from SloBERTa 2.0 (Ulčar and Robnik-Šikonja, 2021) a Slovene pre-trained RoBERTA model. Among the contextualized embedding models for Slovene, this architecture has performed best in most monolingual tasks (Ulčar et al., 2021).

In the static embeddings setting, we obtain the same FastText vector regardless of the context. To obtain contextual embeddings from SloBERTA, we test providing the model with different contexts:

- no-context (IND). The input to the model is just the individual word.
- phrase (PHR). The input to the model is the phrase only.
- sentence context (SENT). We present the model with the complete example sentence.

According to Wang and Zhang (2022) who explore word embedding similarity for word-sense disambiguation in different layers of contextual models, BERT-based models exhibit "first word position bias". In their experiments, the cosine similarity of two words that appeared at the start of the input sentences was considerably higher than the similarity of words that appeared in later positions. However, when simply prefixing and suffixing the input with quotation marks ("), the similarity dropped and lead to higher accuracy. For this reason, we also decide to prepend each of the inputs with a simple prompt *"Primer: "* ["Example: "]. Secondly, we experiment with embeddings obtained separately from each layer (input layer and all subsequent 12 layers). Ethayarajh (2019) has showed that BERT embeddings become increasingly more contextualized, i.e. context specific in the upper layers. Thus, we would expect to observe most relevant semantic differences between the constituent words of metaphorical phrases in the lower layers of the model.

3.3 Similarity metric

The first basic assumption driving our method is that because words participating in a metaphoric phrase originate in different conceptual / semantic domains, they should exhibit less similarity than words participating in a literal phrase that originate in the same or similar conceptual domain. This means the former should be represented further apart in the vector space than words participating in a literal phrase. To measure semantic similarity, we thus apply the frequently used cosine similarity metric that estimates the similarity of the words through the cosine of the angle between the words' vectors:

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|}$$

For words that are split into subword tokens during tokenization with SloBERTa, we calculate the vector of the word from the element-wise mean of all its subword tokens. Secondly, from the perspective of word polysemy which is underscored in the MIPVU procedure, the contextual sense of a word that is used metaphorically is sufficiently different from its non-metaphorical, basic sense. Thus, in contextual word embedding models, we would we expect to observe a substantial self-dissimilarity of the word's embeddings if used metaphorically. However, we do not directly compare a word's embedding in a literal and a metaphorical sentence, because this would not enable unsupervised detection (one would always have to compare a metaphorical and a literal sentence). Instead, we compare the self-similarity of a candidate word between three contexts (individual word, phrase, or full sentence input to the model) outlined in the previous section. In other words, we hypothesise that without additional context, the model would retain and represent the most basic meaning if it sees the word individually, and, conversely, assign a more contextual, shifted meaning of the word in the context of a metaphorical phrase or a full sentence.

4 Results and discussion

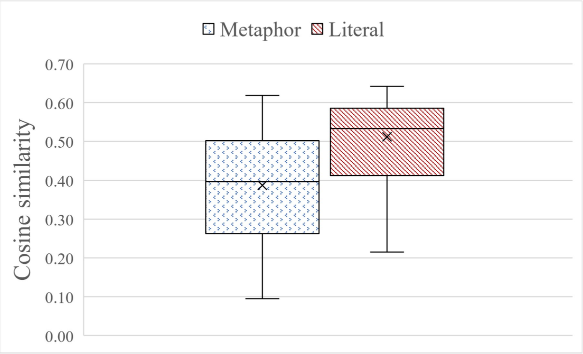

Figure 1: Average cosine similarity between fastText embeddings of words in literal and metaphorical phrases

Figure 1 shows that on average, words participating in metaphorical phrases tend to be more dissimilar than words in literal phrases, i.e. their cosine similarity is lower. This holds for most, but

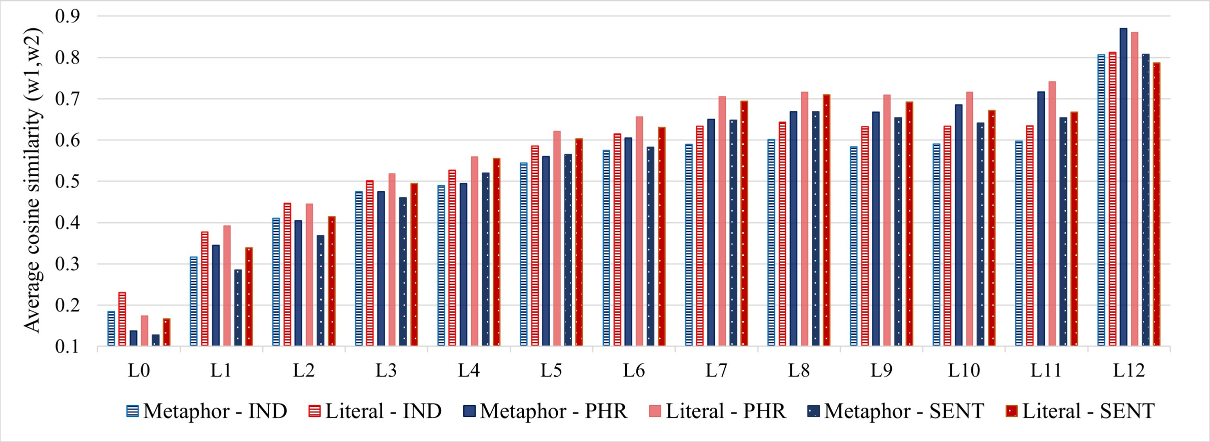

Figure 2: Average cosine similarity of words in metaphorical and literal phrases in different inputs (IND = individual word inputs, PHR = phrase input, SENT = full sentence input), by SloBERTa layer.

not all of the FastText embeddings (18 out of 24 pairs). As for contextualized SloBERTA embeddings (Figure 2), the same trend is observed for all three types of inputs (individual words, phrase inputs or sentences). Indicative are the differences between the blue (metaphorical) and the red (literal) column. The differences are larger in the first few layers, however, in the last layer, the metaphorical word pairs achieve even average higher cosine similarity than those in literal phrases. This would indicate that the initially present semantic distance is neutralized in the later layers of the model.

For the static embeddings, we first analyze the relationship and balance between recall and precision for the literal and metaphorical classes at different cosine similarity thresholds. As shown in Figure 3, the values converge and balance out with cosine similarity values between 0.42 and 0.49.

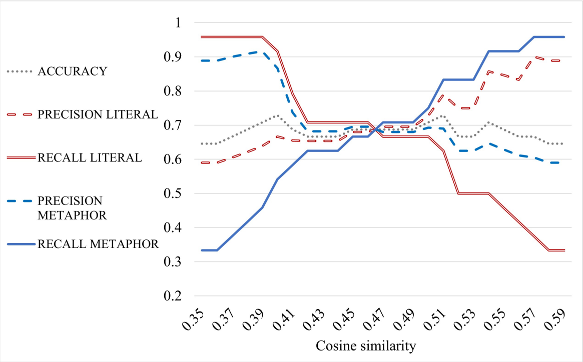

Figure 3: Balance of precision and recall in predicting metaphoricity by cosine similarity of fastText embeddings.

Embedding	**Significance F**	**Cosine similarity threshold**
fastText	<**0.001**	0.4495
SloBERTA IND		
Layer 0	<**0.01**	0.2076
Layer 1, 4	<0.05	
Other layers	>0.05	
SloBERTA PHR		
Layer 0-4, 9	<0.05	
Layer 4-8	<**0.01**	(0.5267, 0.5905, 0.6309, 0.6777, 0.6924, 0.6887)
Other layers	>0.05	
SloBERTA SENT		
Layer 0-2	<**0.01**	(0.1473, 0.3119, 0.3914)
Layer 3, 6-9	<0.05	
Other layers	>0.05	

Table 2: Linear regression results for different embedding methods with cosine similarity as the predictor and metaphoricity as the dependent variable.

Then, for each of the scenarios (static, contextual, different layers of contextual embeddings, different inputs for contextualized embeddings, word similarity and self-similarity), we try to fit a linear regression model to the cosine similarity values to test the relevance for metaphor identification, and to determine the best threshold for unsupervised classification. The results in Table 2 show significance levels for cosine similarities between the first and second word in the phrase in different settings, and the cosine similarity threshold calculated with linear regression.

In the next step, we computed the self-similarity of the word in different contexts. We only focused on the words that are used both literally and metaphorically in our dataset. The average self-similarities are depicted in Figure 4. Not surprisingly, embeddings from the individual- and phrase-inputs are very similar throughout the model, as the context is practically identical. The least similar, as expected, are embeddings from the individual inputs compared to those from sentence inputs. There seem to be observable differences in the average word self-similarity, especially when comparing the individual word embedding to its sentence embedding and the word's embeddings in the phrase and sentence contexts. However, the linear regression and ANOVA tests show no significant relationship between the word's self-similarity in any of the layers and any of the settings: the absolute highest $R^2 = 0.2431$ ($f<0.1$) was achieved when comparing the embedding from the individual word to the embedding of the word in the sentence on the 4th layer. We assume that this is due to the design of contextualized models, which are intended to represent word meaning in wider contexts and fail to produce sensible representations when presented with narrower contexts.

Embedding	**A**	**P**	**R**	**F1**
FastText	0.69	0.70	0.67	0.68
SloBERTA IND				
Layer 0	0.69	**0.68**	0.71	0.69
SloBERTA PHR				
Layer 4	0.66	0.65	0.71	0.68
Layer 5	0.66	0.67	0.67	0.67
Layer 6	0.60	0.61	0.58	0.60
Layer 7	0.64	0.65	0.63	0.64
Layer 8	0.68	**0.68**	0.71	0.69
SloBERTA SENT				
Layer 0	**0.71**	0.68	**0.79**	**0.73**
Layer 1	0.69	**0.68**	0.71	0.69
Layer 2	0.58	0.59	0.54	0.57

Table 3: Prediction results in terms of accuracy (A), metaphor precision (P), metaphor recall (R), and F1 score.

To further evaluate cosine similarity as a predictor of metaphoricity, we classify our examples according to the thresholds obtained from linear regression models with significance levels $f<0.01$. We report the results in Table 3. The results are very comparable across models. The highest overall scores are achieved by predicting metaphoricity from the cosine similarities of words on the input layer (0) when the model receives the whole sentence as input. However, the differences in performance obtained from the embeddings from the 0th layer from different inputs must be purely incidental, as the embeddings there are not contextualized yet. The difference is only due to the additional positional embeddings that encode the position of the word in the sequence.

5 Conclusion

In this paper, we presented the first experiment on unsupervised identification of metaphors on the phrase level in Slovene with word embeddings. Based on a dataset of 24 comparable pairs of metaphorical and literal phrases, we investigated the use of cosine similarity in both static and contextual embeddings. The results show that both methods achieve comparable results in terms of precision, recall and accuracy when comparing cosine similarities between the phrase's constituent words. In line with previous research, we also intuit that lower layers exhibit less contextualized information and are generally more suited to the task. However, in our experiments with self-similarity, where we compared the candidate word's embeddings in different contexts, the results show no statistical significance and cannot be used to determine a metaphorical shift in meaning. In conclusion, this preliminary experiment showed promising results for unsupervised metaphor identification, but will have to be evaluated on more data which may contain less clear-cut examples of metaphorical and literal language. Future work includes testing the method on more examples and other embedding models. We also plan to investigate the use of psycholinguistic measures such as abstractness for relation-level metaphor identification, and evaluate the methods with respect to the syntactic type of construction used. Another interesting avenue for further research could be investigating other methods for combining subword embeddings, which could potentially provide a better word representation for the purposes of metaphor identification.

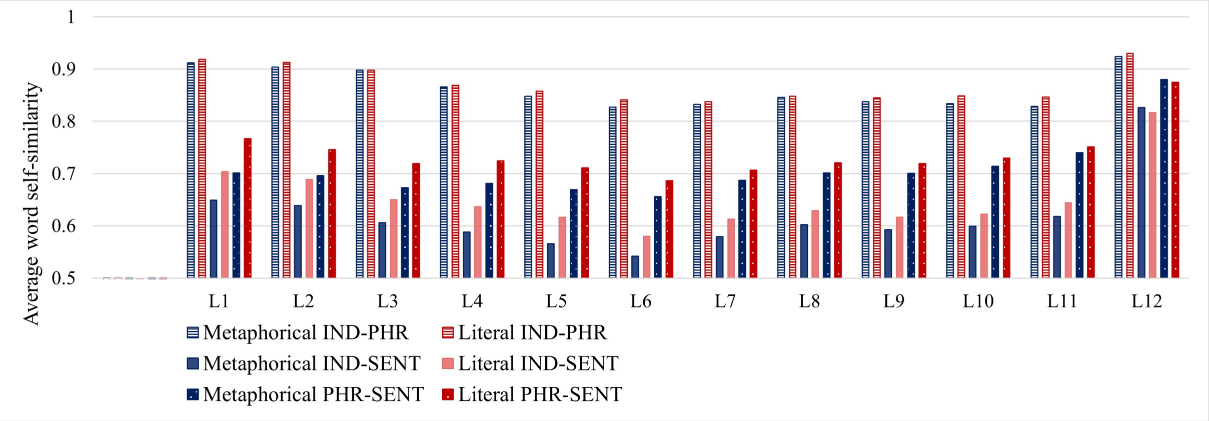

Figure 4: Average self-similarity of a candidate word in different inputs (IND=individual word, PHR=phrase input, SENT=full sentence input), by SloBERTa layer.

Limitations

Although the paper shows promising results, the findings can only be applied to the small set of data we used in our experiment. To validate them further, the approach would have to be tested on a much larger dataset containing less clear-cut examples. Secondly, our unsupervised metaphor identification approach was limited to adjective-noun and noun-noun phrases, meaning we cannot draw definite conclusions for the usefulness of this approach for identification of metaphors in other constructions. Thirdly, there is a plethora of language models available for Slovene. In this work, we only experimented with fastText and SloBERTa embeddings because of their good performance on other linguistic tasks. Other models, such as GPT, T5, BERT, or ELMo, could turn out to be more suitable for metaphor processing.

Acknowledgements

This project has received funding by the Slovenian Research Agency (ARRS) under the research core funding P6-0215.

References

Kat R. Agres, Stephen McGregor, Karolina Rataj, Matthew Purver, and Geraint A. Wiggins. 2016. Modeling metaphor perception with distributional semantics vector space models. In *Proceedings of the ESSLLI Workshop on Computational Creativity, Concept Invention, and General Intelligence (C3GI)*, page 1–14.

Mojca Brglez, Senja Pollak, and Špela Vintar. 2021. Simple discovery of COVID IS WAR metaphors using word embeddings. In *Odkrivanje znanja in podatkovna skladišča - SiKDD: 4 October 2021, Ljubljana, Slovenia*, page 37–40. Institut "Jožef Stefan".

Christian Burgers, Elly Konijn, and Gerard Steen. 2016. Figurative framing: Shaping public discourse through metaphor, hyperbole, and irony. *Communication Theory*, 26:410–430.

Lynne Cameron. 2003. *Metaphor in Educational Discourse*. Advances in Applied Linguistics. Bloomsbury Publishing.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Polona Gantar, Lut Colman, Carla Parra Escartín, and Héctor Martínez Alonso. 2018. Multiword Expressions: Between Lexicography and NLP. *International Journal of Lexicography*, 32(2):138–162.

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.

Zellig S. Harris. 1954. Distributional structure. *WORD*, 10(2-3):146–162.

Walter Kintsch. 2000. Metaphor comprehension: A computational theory. *Psychonomic bulletin & review*, 7:257–66.

Zoltán Kövecses. 2020. *Extended Conceptual Metaphor Theory*. Cambridge University Press.

Simon Krek, Tomaž Erjavec, Andraž Repar, Jaka Čibej, Špela Arhar Holdt, Polona Gantar, Iztok Kosem, Marko Robnik-Šikonja, Nikola Ljubešić, Kaja Dobrovoljc, Cyprian Laskowski, Miha Grčar, Peter Holozan, Simon Šuster, Vojko Gorjanc, Marko Stabej, and Nataša Logar. 2019. Corpus of written standard Slovene Gigafida 2.0. Slovenian language resource repository CLARIN.SI.

George Lakoff and Mark Johnson. 1980. *Metaphors we Live by*. University of Chicago Press.

George Lakoff and Mark Johnson. 2003. *Metaphors we Live by*. University of Chicago Press.

Nikola Ljubešić and Tomaž Erjavec. 2018. Word embeddings CLARIN.SI-embed.sl 1.0. Slovenian language resource repository CLARIN.SI.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and WordNet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1222–1231, Melbourne, Australia. Association for Computational Linguistics.

Paolo Pedinotti, Eliana Di Palma, Ludovica Cerini, and Alessandro Lenci. 2021. A howling success or a working sea? testing what BERT knows about metaphors. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 192–204, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gudrun Reijnierse, Christian Burgers, Tina Krennmayr, and Gerard Steen. 2019. Metaphor in communication: the distribution of potentially deliberate metaphor across register and word class. *Corpora*, 14(3):301–326.

Elena Semino. 2008. *Metaphor in Discourse*. Cambridge University Press.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, San Diego, California. Association for Computational Linguistics.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, volume 2, pages 1002–1010.

Gerard Steen. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Converging evidence in language and communication research. John Benjamins Publishing Company.

Gerard Steen. 2017. Deliberate metaphor theory: Basic assumptions, main tenets, urgent issues. *Intercultural Pragmatics*, 14:1–24.

Chang Su, Shuman Huang, and Yijiang Chen. 2017. Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*, 219:300–311.

Roger Tourangeau and Robert J. Sternberg. 1981. Aptness in metaphor. *Cognitive Psychology*, 13(1):27–55.

Matej Ulčar and Marko Robnik-Šikonja. 2021. Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0. Slovenian language resource repository CLARIN.SI.

Matej Ulčar, Aleš Žagar, Carlos S. Armendariz, Andraž Repar, Senja Pollak, Matthew Purver, and Marko Robnik-Šikonja. 2021. Evaluation of contextual embeddings on less-resourced languages. *Computer Research Repository, https://arxiv.org/abs/2107.10614. Version 1.*

Akira Utsumi. 2011. Computational exploration of metaphor comprehension processes using a semantic space model. *Cognitive Science*, 35(2):251–296.

Yile Wang and Yue Zhang. 2022. Lost in context? on the sense-wise variance of contextualized word embeddings. *Computer Research Repository, https://arxiv.org/abs/2208.09669. Version 1.*

Yorick Wilks. 1978. Making preferences more active. *Artificial Intelligence*, 11(3):197–223.

Omnia Zayed, John Philip McCrae, and Paul Buitelaar. 2018. Phrase-level metaphor identification using distributed representations of word meaning. In *Proceedings of the Workshop on Figurative Language Processing*, pages 81–90, New Orleans, Louisiana. Association for Computational Linguistics.

Ana Zwitter Vitez, Mojca Brglez, Marko Robnik Šikonja, Tadej Škvorc, Andreja Vezovnik, and Senja Pollak. 2022. Extracting and analysing metaphors in migration media discourse: towards a metaphor annotation scheme. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2430–2439, Marseille, France. European Language Resources Association.

Tadej Škvorc, Polona Gantar, and Marko Robnik-Šikonja. 2022. MICE: Mining idioms with contextual embeddings. *Knowledge-Based Systems*, 235:107606.