Simple Discovery of COVID IS WAR Metaphors Using Word Embeddings

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

Senja Pollak
Jožef Stefan Institute
Ljubljana, Slovenia
senja.pollak@ijs.si

Špela Vintar
University of Ljubljana
Ljubljana, Slovenia
spela.vintar@ff.uni-lj.si

ABSTRACT

In the past year, the discourse on the COVID-19 pandemic has produced a great number of metaphors stemming from the more basic conceptual metaphor ILLNESS IS WAR. In this paper, we present a semi-automatic method to detect linguistic manifestations of the latter in Slovene media. The method consists of assembling a seed vocabulary of war-related words from an existing Slovene metaphor corpus, extending the vocabulary using word embeddings, and refining the extended vocabulary using intersection filtering. Our method offers a quick compilation of corpus data for further analysis, however, we also address issues related to the method's precision and the need for manual filtering.

KEYWORDS

metaphors, covid, word embeddings, media discourse

1 INTRODUCTION

The COVID pandemic has been a ubiquitous topic in the discourse of the past year, featuring in medical, political, public and personal discourse. The emergence of a new virus of yet unknown origin, behaviour and effects has presented itself like a complex and obscure topic. To make sense of it, we have once more resorted to metaphorical language, much like we do when faced with other abstract, obscure concepts. According to Conceptual Metaphor Theory (CMT, [11, 12]), metaphors "are among our principal vehicles for understanding" and "play a central role in the construction of social and political reality" ([12, p. 151]). In CMT, linguistic metaphors such as *"food for thought"* and *"half-baked idea"* are considered manifestations of an established conceptual mapping between a more concrete domain and a more abstract domain, here for example IDEAS ARE FOOD. The domain of DISEASES, on the other hand, is often mapped to the domain of WAR, a more common frame of reference which has taken hold as a fairly conventional way to talk about illnesses and their treatments, as well as several other domains ([8]).

As was already observed in various studies ([19, 2, 5, 7]), the discourse on the current COVID pandemic has also repeatedly used the WAR domain in its metaphors. At the time of our experiment, however, no study has yet addressed the use of such metaphors in Slovene, where they were also adopted for communicating various implications, preventive measures, recommendations and laws to abide by. To investigate the use and pervasiveness of this metaphorical domain in Slovene media, we have conducted a quick analysis of a corpus of COVID-related news articles using an innovative methodological approach. We propose a top-down method to search for expected conceptual metaphors through semi-automatic means employing word embeddings. While most previous corpus-based approaches to identify metaphors either use a small set of candidate words or require manual inspections of large data samples, our approach reduces manual work on assembling linguistic data by combining existing annotated resources and text mining methods.

2 PROPOSED APPROACH

Our method aims to discover linguistic expressions of the conceptual metaphor COVID IS WAR in the corpus by targeting a broader potentially metaphoric vocabulary. Previous related works have relied on either a limited vocabulary set (e.g. [7]) or a list of words laboriously compiled from various sources such as dictionaries, thesauri and other studies on metaphor [19], or have used sophisticated but complex NLP methods and specialized resources (e.g. [6]). In our experiment, we use a simple unsupervised approach using existing resources and language processing technologies.

The main novelty of our approach is using pre-trained word embeddings to extend the vocabulary, used also by e.g. [16] and [18] to extend terminology. As past research has shown [14], word embeddings used for training language models retain linguistic regularities, including syntactic and semantic relationships between words. This means that similar words have similar vectors, and the closer vector representations (word embeddings) are, the higher the chance they share a certain semantic space. We make use of this feature by trying to capture a semantic space that would resemble the conceptual domain of WAR, which represents the source domain of the metaphor.

2.1 Method

First, we start by collecting war-related lexical units from the KOMET corpus [1], the only corpus of metaphors in Slovene which was recently compiled and annotated similarly to the English corpus of metaphors, VUAMC [17]. KOMET contains approximately 200,000 words obtained from journalistic, fiction and online texts and was hand-annotated for metaphoricity on the basis of the MIPVU procedure ([17]). Additionally, the metaphoric expressions are tagged for one of 69 semantic frames, i.e. the source concepts that semantically motivate them. One of these semantic frames is `#met.battle`, which subsumes 105 metaphoric instances with 67 different lemmas, such as *predati, ostrostrelec, orožje, napasti* [surrender, sniper, weapon, attack]. These also form multi-word idioms such as *železna pest* [iron fist] and *boriti se z mlini na veter* [to tilt at windmills] which we exclude from our candidates list because the word embeddings we use only represent tokens, not whole phrases. Moreover, the lemmas within do not themselves necessarily represent the desired domain. We also filter out some words erroneously annotated with the frame such as *številen* [numerous]. This gives a starting vocabulary of

Information Society 2021, 4–8 October 2021, Ljubljana, Slovenia

51 unique seed words. Then, to extend the vocabulary further, we employ Slovene word token embeddings ([13] pre-trained with fastText ([4]) on various large corpora of Slovene (GigaFida, Janes, KAS, slWaC etc.). For each seed word in the list of words extracted from the KOMET corpus, we use the Gensim library ([20]) to find the word's N nearest neighbours in the fastText embeddings' space (using the `most_similar` function).

To increase the robustness of the extended vocabulary, we try to automatically filter out lexis not related to war. To this end, we use the word embeddings intersection method ([18]). The method retains only the candidates that intersect between the sets, meaning they occur in the neighbourhood of at least k input seed words. For our main experiment, presented in this paper, we select the parameters N=50 and k=3. We thus obtain a maximum of 2550 (50 x 51) potential candidates. In the output, there are 2078 unique words, and, after lemmatization, 1539 unique lemmas. After the intersection filtering, the vocabulary extended by word embeddings consists of 184 word lemmas: 44 of them are already included in our initial seed set and 140 are new lemmas. We join the new, extended set with the initial seed set, which yields a total of 191 lemmas to search for.

3 CORPUS

The experiment is carried out on a corpus of Slovene COVID-19-related news articles, automatically crawled from the web by searching for the keyword "covid-19" in article titles (a subset of the Slovene corpus used in the Slav-NER 2021 shared task ([15]). The corpus consists of 233 texts spanning from February 2nd to December 11th, 2020 . To prepare it for analysis, we remove the header of each text (comprised of the article number, locale, date and URL), then parse the text into sentences and tokens using the NLTK library ([3]). We also lemmatize the corpus using the LemmaGen lemmatization module ([9]). The pre-processed corpus contains 7,273 sentences and 151,947 tokens.

3.1 Corpus search

In the next step, we extract all sentences from the corpus containing any of the war-related terms from our expanded vocabulary of 191 lemmas. The results yield 335 instances of potentially metaphorical expressions. Out of the 191 lemmas on the metaphorical candidate list, the COVID corpus contains 49, appearing in 268 sentences. Due to the unsupervised approach these are still only candidate words from the semantic domain of war. A manual analysis shows that in addition to war metaphors, our extracted sentences include the following four cases:

(1) Some of the seed words found in the corpus are used literally;
(2) Some of the seed words found in the corpus are a result of lemmatization errors
(3) Some of the seed words found in the corpus are used metaphorically, but refer to other target domains, such as POLITICS or NATURE (e. g. *boriti se proti podnebnim spremembam* ['fight against climate change'])
(4) Some of the seed words in our initial 191-candidate list are not actually related to the topic of WAR but are more closely related to another topic (e.g. *gol* ['goal'])

On this account we perform a manual analysis of the extracted sentences and categorize them as follows:

(1) falsely extracted instances due to a lemmatization error or literal use, or true metaphorical expressions but with other source or target domain, and
(2) true metaphorical expressions referring to disease as target domain

For example, in the following sentence, the word brigade [brigades] only refers to a name of a street, which we mark as literal usage.

- */.../ odvzem brisov pri pacientih s sumom na Covid-19: ob Cesti proletarskih **brigad** 21 /.../*
  */.../ taking swabs from patients with suspected Covid-19: at 21, Proletarian **Brigades** Road /.../*

In the following example, the word napad [attack] is used to refer to another domain – INTERNET, COMPUTING, which we mark as metaphor for another target domain.

- *Covid-19 je okrepil trend rasti kibernetskih **napadov** [Covid-19 reinforced the growing trend of cyber **attacks**]*

The following three example sentences contain expression that we mark as metaphor for the target domain of DISEASE.

- *Čeprav v **boju** z virusom to nikakor ni hitro.*
  *[Although this is by no means fast in the **fight** against the virus.]*

- *Kako bo jeseni, ko bodo »**udarili**« še drugi virusi?*
  *[What will happen in autumn, when other viruses also "**strike**"?]*

- *Prvi organski sistem v organizmu, ki ga virus **napade**, povzroči pljučnico, ...*
  *[The first system in the organism that the virus **attacks** causes pneumonia ...]*

Results of this analysis are presented in Table 1, whereby we report only lemmas that were metaphorically used for the DISEASE target domain at least once.

As can be derived from Table 1, our proposed method correctly identified 25 different lemmas with a total of 123 occurrences that are used metaphorically to frame the topic of the pandemic. Out of our 233 articles, 68 or 29,18% contained at least one militaristic metaphorical expression. The ostensibly most frequent expression used was *boj* [fight] with 46 metaphorical occurrences, followed by *boriti* [to fight] with 13 metaphorical occurrences and *soočati* [to confront] with 7 metaphorical occurrences. They account for 37.4%, 10.6% and 5.7% of all metaphorical expressions found by our method, respectively, and together, they represent more than 50% of them. This points to the interpretation that the news corpus contains mostly highly conventional and recurrent metaphors. A lot of the war-related vocabulary (potential candidates in our extended war-related lexis) is not used, meaning the corpus does not, at this moment, exhibit very original, novel metaphorical expressions. Using a larger and a more recently compiled corpus would perhaps reveal a more innovative use of COVID IS WAR metaphors. The vocabulary extension method using word embeddings has proven fruitful as it revealed some metaphorical expressions that were not in the initial 51-word list extracted from the KOMET corpus. The 9 newly discovered lemmas are: *soočiti, izbojevati, zmagati, obraniti, uiti, soočanje, spopadati, zoperstaviti, podleči* [to confront, to fight, to win, to defend, to escape, confrontation, to combat, to oppose, to succumb].

The analysis also revealed some additional lemmas that relate the epidemic to the war frame. In the sentences containing the lemmas we searched for, there were other words from the WAR

Table 1: Analysis of metaphoric lemmas from the extended vocabulary

Lemma	Corpus occurences	Literal uses, lemmatization errors or other source/target domain	DISEASE as target domain
Boj [fight]	57	11	46
Boriti [to fight]	16	3	13
Soočati [to confront]	17	10	7
Spopad [to combat]	6		6
Spopadanje [combatting]	6		6
Zoperstaviti [to oppose]	5		5
Bitka [battle]	5	1	4
Napad [attack]	41	37	4
Podleči [succumb]	5	1	4
Spopadati [to combat]	5	1	4
Bojen [combat [ADJ]]	17	15	2
Borba [battle]	3	1	2
Braniti [to defend]	4	2	2
Napasti [to attack]	6	4	2
Obramben [defense [ADJ]]	9	7	2
Soočanje [confronting]	2		2
Soočiti [to confront]	6	4	2
Žrtev [victim]	49	47	2
Borec [fighter]	3	2	1
Izbojevati [to fight]	1		1
Obraniti [to defend]	1		1
Štab [base, headquarters]	3	2	1
Udariti [to hit]	2		2
Uiti [to escape]	2	1	1
Zmagati [to win]	5	4	1
TOTAL	270	147	123

domain forming so called metaphor clusters ([10]). Thus, we managed to capture some metaphorical expressions that appeared in close vicinity (in the same sentence) of the found metaphorical expressions: *fronta, strategija, preboj, akcijski načrt, vojna mentaliteta, sovražnik* [front, strategy, breakthrough, action plan, war mentality, enemy]. For instance, our method found the sentence below which, in addition to the word *bitka* [battle] in our candidate list, contains a metaphorical use of the word *fronta* [front].

- ***Bitka*** *proti virusu na več* ***frontah***
  *[**Battle** against the virus on multiple **fronts**]*

4 ANALYSING DIFFERENT PARAMETER SETTINGS

Some of the expressions mentioned above would have been captured had we modified the parameters of vocabulary extension. Namely, we experimented with using more nearest neighbours (75, 100, 150 and 200). Our initial experiments were carried out on a N of 50 and intersection k of 3. However, by changing the parameters, the results of initial new lemmas could differ. In Figure 1, we analyse how the seed list changes with different parameters: N of 50 and 75 neighbours, each combined with the intersection count k of 2, 3 and 4. Note that these refer only to the list of potentially metaphoric lemmas, and not to the analysis of their use, which can only be analysed in context. We see that the initially selected parameters (50 neighbours and 3 recurrences) are an acceptable middle-ground between precision and size while still maintaining an unsupervised approach, however, had we wanted more examples, we could increase the parameter N or decrease the parameter k.

For the recall, we are not able to carry out a systematic evaluation. Nevertheless, based on metaphor clusters analysis mentioned above, we identified the set of additional words that belong to the military vocabulary: *fronta, strategija, preboj, akcijski, vojen, sovražnik* [front, strategy, breakthrough, action [ADJ], war [ADJ], enemy]. The words *vojen* [war[ADJ]] and *sovražnik* [enemy] would have been included if we lowered the intersection parameter to $k = 2$ at $N = 50$ neighbours or extended the vocabulary by $N = 75$ neighbours while keeping the intersection parameter $k = 3$. Other metaphorical expressions occurring in the corpus *(fronta, preboj, strategija, akcijski)* [front, strategy, breakthrough, action [ADJ]] are not found anywhere in the first 200 neighbours of any of the words, indicating perhaps that the number of neighbours might be further increased. However, we observe that increasing the number of neighbours leads to fuzzier results. The added vocabulary using 75, 100, 150, and 200 nearest neighbours of our initial seed words includes increasingly more words unrelated to the topic of war and some very common words, which would need additional filtering. We assume that the reason for this is that words commonly used metaphorically (conventional or dead metaphors) are “displaced” in the vector space of embeddings, moving away from the words in their original semantic domains and closer to words in other semantic domains – target domains. For example, we observed a lot of sports expressions in our extended vocabulary (e.g. “ball”, “goal”, “goalpost”). This shows how entrenched metaphors are in our language: in the vector space of word embeddings, the semantic domains are already “muddled”. In the present example, this could be a due to the frequent linguistic manifestations of the conceptual metaphor COMPETITION IS WAR.

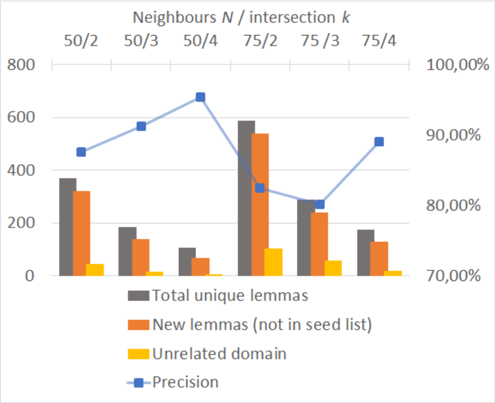

Figure 1: Analysis of vocabulary extension parameters *N* and *k*

5 CONCLUSION

We present an innovative approach using word embeddings as a tool for extending the vocabulary of potentially metaphoric expressions and identify them in corpora. Our approach shows promise in that it correctly identifies numerous such expressions and confirms that intersections of semantic spaces of metaphorical seed words can be used to refine the quest for words pertaining to the military domain. Nevertheless, some metaphoric expressions are missed by our method and the experiment still needs manual analysis. Further research and experiments would be needed for a larger expansion of vocabulary and a finer filtering approach as well as comparing different word embeddings, possibly those trained on more literal language.

ACKNOWLEDGMENTS

This work is supported by the Slovenian Research Agency by the research core funding P6-0215 and P2-0103, as well as by the research project CANDAS (J6-2581). The work has also been supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this paper reflect only the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] Špela Antloga. 2020. Metaphor corpus KOMET 1.0. Slovenian language resource repository CLARIN.SI. (2020). http://hdl.handle.net/11356/1293.

[2] Benjamin R. Bates. 2020. The (in)appropriateness of the war metaphor in response to SARS-CoV-2: a rapid analysis of Donald J. Trump's rhetoric. *Frontiers in Communication*, 5, 50, (June 2020). DOI: 10.3389/fcomm.2020.000505.

[3] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. (1st edition). O'Reilly Media, Inc.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, (June 2017), 135–146. DOI: doi.org10.1162/tacl_a_00051.

[5] Eunice Castro Seixas. 2021. War metaphors in political communication on Covid-19. *Frontiers in Sociology*, 5, 112. DOI: 10.3389/fsoc.2020.583680.

[6] Jane Demmen, Elena Semino, Zsófia Demjén, Veronika Koller, Andrew Hardie, Paul Rayson, and Sheila Payne. 2015. A computer-assisted study of the use of violence metaphors for cancer and end of life by patients, family carers and health professionals. *International Journal of Corpus Linguistics*, 20, 2, 205–231. DOI: 10.1075/ijcl.20.2.03dem.

[7] Damián Fernández-Pedemonte, Felicitas Casillo, and Ana Jorge-Artigau. 2021. Communicating COVID-19: metaphors we "survive" by. *Tripodos*, 2, (February 2021), 145–160. DOI: 10.51698/tripodos.2020.47p145-160.

[8] Stephen J. Flusberg, Teenie Matlock, and Paul H. Thibodeau. 2018. War metaphors in public discourse. *Metaphor and Symbol*, 33, 1, 1–18. DOI: 10.1080/10926488.2018.1407992.

[9] Matjaž Juršič, Igor Mozetič, Tomaž Erjavec, and Nada Lavrač. 2010. Lemmagen: multilingual lemmatisation with induced ripple-down rules. *Journal of Universal Computer Science*, 16, 9, 1190–1214. http://www.jucs.org/jucs_16_9/lemma_gen_multilingual_lemmatisation|.

[10] Veronika Koller. 2003. Metaphor clusters, metaphor chains: analyzing the multifunctionality of metaphor in text. In volume 5, 115–134.

[11] George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago press.

[12] George Lakoff and Mark Johnson. 2003. *Metaphors we live by*. University of Chicago press.

[13] Nikola Ljubešić and Tomaž Erjavec. 2018. Word embeddings CLARIN.SI-embed.sl 1.0. Slovenian language resource repository CLARIN.SI. (2018). http://hdl.handle.net/11356/1204.

[14] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, (June 2013), 746–751. https://aclanthology.org/N13-1090.

[15] Jakub Piskorski, Bogdan Babych, Zara Kancheva, Olga Kanishcheva, Maria Lebedeva, Michał Marcińczuk, Preslav Nakov, Petya Osenova, Lidia Pivovarova, Senja Pollak, Pavel Přibáň, Ivaylo Radev, Marko Robnik-Sikonja, Vasyl Starko, Josef Steinberger, and Roman Yangarber. 2021. Slav-NER: the 3rd cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics, Kiyv, Ukraine, 122–133. https://aclanthology.org/2021.bsnlp-1.15.

[16] Senja Pollak, Andraž Repar, Matej Martinc, and Vid Podpečan. 2019. Karst exploration: extracting terms and definitions from karst domain corpus. In *Proceedings of eLex 2019*, 934–956.

[17] G. Steen. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU. Converging evidence in language and communication research*. John Benjamins Publishing Company. DOI: 10.1075/celcr.14.

[18] Špela Vintar, Larisa Grcic, Matej Martinc, Senja Pollak, and Uroš Stepišnik. 2020. Mining semantic relations from comparable corpora through intersections of word embeddings. In (May 2020). https://aclanthology.org/2020.bucc-1.5.pdf.

[19] Philipp Wicke and Marianna M. Bolognesi. 2020. Framing COVID-19: how we conceptualize and discuss the pandemic on Twitter. *PLOS ONE*, 15, 9, (September 2020), 1–24. DOI: 10.1371/journal.pone.0240010.

[20] Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In (May 2010), 45–50. DOI: 10.13140/2.1.2393.1847.