Do topics make a metaphor? Topic modeling for metaphor identification and analysis in Russian.

Yulia Badryzlova[1][0000-0003-3901-4816], Anastasia Nikiforova[1] [0000-0002-7800-9430], and Olga Lyashevskaya[1, 2][0000-0001-8374-423X]

[1] National Research University Higher School of Economics, Moscow, Russia
[2] Vinogradov Russian Language Institute RAS, Moscow, Russia
yuliya.badryzlova@gmail.com, steysie@gmail.com, olesar@yandex.ru

Abstract. The paper examines the efficiency of topic models as features for computational identification and conceptual analysis of linguistic metaphor on Russian data. We train topic models using three algorithms (LDA and ARTM – sparse and dense) and evaluate their quality. We compute topic vectors for sentences of a metaphor-annotated Russian corpus and train several classifiers to identify metaphor with these vectors. We compare the performance of the topic modeling classifiers with other state-of-the-art features (lexical, morphosyntactic, semantic coherence, and concreteness-abstractness) and their different combinations to see how topics contribute to metaphor identification. We show that some of the topics are more frequent in metaphoric contexts while others are more characteristic of non-metaphoric sentences, thus constituting topic predictors of metaphoricity, and discuss whether these predictors align with the conceptual mappings attested in literature. We also compare the topical heterogeneity of metaphoric and non-metaphoric contexts in order to test the hypothesis that metaphoric discourse should display greater topical variability due to the presence of Source and Target domains.

Keywords: Metaphor Identification, Topic Modelling, LDA, ARTM, Topical Predictors of Metaphoricity, Topical Profiles, Topical Heterogeneity.

1 Introduction

1.1 The task of computational metaphor identification

Contemporary cognitive theory states that human reasoning is intrinsically metaphorical and imaginative, based on various kinds of prototypes, framings, and metaphors [1, 2]. Our abstract conceptual representations are grounded in sensorimotor systems, and conceptual metaphor connects these two realms by mapping the domain of familiar, concrete and distinct experiences (the Source Domain) onto the domain of predominantly abstract and complex concepts (the Target Domain), thus enabling us to conceptualize the rich fabric of the reality that surrounds us. The Source-Target mappings are systematic, i.e. they reproduce themselves across similar situations; some of them are claimed to be universal, while others may be culture-specific.

Conceptual metaphors manifest themselves in language and discourse as linguistic metaphors, that is, the lexical units and constructions which express the Target, the Source, and the relations between them. An example of a conceptual metaphoric mapping is CORRUPTION IS A DISEASE which may be linguistically conveyed, for example, by the following English sentences (with *T* and *S* indicating the Source and the Target terms, respectively): "Corrupt *(T)* officials are infecting *(S)* our government at every level." or "Our government is afflicted *(S)* with the cancer *(S)* of corruption *(T)*." [3].

Evidence from psycholinguistic research demonstrates that metaphor guides reasoning and decision-making in societal, economic, health-related, educational, and environmental issues [see, for example, 4–7]. As deeply as conceptual metaphor is engrained in the mind, as much linguistic metaphor is integrated into the language and its usage, forming an organic part of them. According to various estimates, up to nearly one third of words in a corpus may be used metaphorically [8, 9]. Such pervasiveness of metaphor in language and thought – as well as the ambiguity it creates – make metaphor a challenge to various NLP applications, such as machine translation, information retrieval and extraction, question answering, opinion mining, etc. The interest of the NLP community to computational metaphor research expressed itself in the series of dedicated workshops in 2013-2016 [10–13] and the two metaphor detection shared tasks in 2018 and 2020 [14, 15].

How can the underlying conceptual properties of metaphoric utterances be captured in order to train machine learning algorithms to tell them apart from non-metaphoric ones? Different types of features have been explored in the state-of-the-art research:

- Lexical features [16];
- Morphological and syntactic features [17, 18];
- Distributional semantic features [19, 20];
- Features from lexical thesauri and ontologies: e.g., WordNet [21], FrameNet [22], VerbNet [23], ConceptNet [18], and the SUMO ontology [24, 25];
- Psycholinguistic features: concreteness and abstractness, imageability, affect, and force [26–29];
- Topic modelling [16, 30, 31] – the feature which is explored in the present paper.

1.2 Topic modelling in metaphor identification: previous work

Application of topic modelling to identification of metaphor relies on the assumption that metaphoric contexts should contain terms from both the Source and the Target domains, whereas non-metaphoric sentences should be more homogeneous in terms of topical composition; the topics are regarded as proxies for the conceptual domains.

Heintz et al. [30] use topic models to identify linguistic metaphors belonging to the Target domain of Governance in English and Spanish. They train LDA models on the full text of Wikipedia in these languages and automatically align the topics with the manually collected lists of seed words representing the Target and the Source domains. A sentence is judged to contain a linguistic metaphor on the account of the strength of association between topics and the sentence, between the annotated words and the topics, and between the topics and their aligned concepts. The authors carry out two evaluations of their system. In the first, the predictions of the algorithm on the English data are compared to the judgements of two annotators, with the reported F1-scores of 0.66 and 0.5, respectively; however, the agreement between the annotators (κ) was rather low (0.48). In the second evaluation, the annotation task was crowdsourced, and the metaphoricity of a sentence was defined as the fraction of the subjects who annotated it as being metaphoric. Sixty-five per cent of the English sentences that were judged metaphoric by the algorithm had human-generated metaphoricity scores greater than 0.25, and 73% greater than 0.2; on the Spanish data, the respective results were 60 and 73%.

Ghavidel et al. [31] train an LDA model to detect linguistic metaphors in Persian. They generate a topic vector of each sentence in the corpus, and run the rule-based classifier to check whether there is any word which does not belong to the overall topic of the sentence. If the topic of a word is recognized as deviant, the sentence is marked as metaphoric, and non-metaphoric if otherwise. The system is reported to yield the F1-score of 0.68 when evaluated on a random sample of 100 sentences.

Klebanov et al. [16] use LDA topic modelling in combination with other features (lexical unigrams, part of speech tags, and concreteness indexes) to identify metaphor on the word level (i.e. to tag each content word in running text as either metaphoric or non-metaphoric). The F1-score ranges between 0.21 to 0.67 depending on the dataset and the genre. The authors investigate the relative contribution of each feature and report that topics is the second most effective feature (after lexical unigrams).

Besides, topic models, along with the other types of features, were suggested for use to the participants of the First and the Second shared tasks on metaphor detection [14, 15].

In this paper, we apply topic modelling to sentence-level metaphor identification in Russian on a representative metaphor-annotated corpus [32]. We also compare the performance of the topic models in metaphor classification to other state-of-the-art features, and estimate the contribution of topics to the most efficient classifier. Moreover, we take an in-depth look into the topic models of metaphoric and non-metaphoric discourse in order to identify the topical cues of metaphoricity. We also examine the topical heterogeneity of metaphoric and non-metaphoric contexts in order to explore the hypothesis that metaphoric contexts should feature a greater variety of topics due to the presence of two conceptual domains (the Source and the Target).

To the best of our knowledge, this is the first research to apply topic modelling to the problem of metaphor identification in Russian.

2 Topic modelling for metaphor identification in Russian

2.1 Training the topic models: LDA and ARTM

For our experiments, we train two types of topic models: LDA and ARTM.

LDA (latent Dirichlet allocation) [33] is the topic modelling method which is most widely used in NLP tasks. In LDA, the parameters Φ (the matrix of term probabilities for the topics) and *Θ* (the matrix of topic probabilities for the documents) are constrained by an assumption that vectors *φt* and *θd* are drawn from Dirichlet distributions with hyperparameters *β = (βw)w*∈*W* and *α = (αt)t*∈*T* respectively (where *T* is a set of topics, *W* is a set of all terms in a collection of texts).

Two major problems arise when training topic models with LDA – noise from stop-words and other high-frequency words, and assigning words to multiple topics, which negatively affects the overall interpretability of the topics. This issue is addressed by Additive Regularization of Topic Models (ARTM) [34]; in this study, we used the following regularizers available in the BigARTM library[1]:

1. The smoothing / sparsing regularization of terms over topics, where the smoothing regularizer sends high-frequency words into dedicated background topics, and the sparsing regularizer highlights the lexical nuclei of domain-specific topics covering a relatively small proportion of the vocabulary;
2. The smoothing / sparsing regularization of topics over documents, in which the smoothing regularizer indicates the background words in each document of the collection, while the sparsing regularizer pinpoints the domain-specific words in each document.

As a result of such regularization, zero probability is assigned to words that do not describe domain-specific topics, as well as to high-frequency and general vocabulary; each term is assigned to a relatively small number of topics, so that the resulting topics become more interpretable. The smoothing and the sparsing regularizations of matrices Φ and Θ are presented in the equation:

$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{w \in W} \beta_{wt} ln\phi_{wt} + \sum_{d \in D} \sum_{t \in T} \alpha_{td} ln\theta_{td},$$

where D is a collection (set) of texts, $\beta_0 > 0$, $\alpha_0 > 0$ are regularization coefficients, and β_{wt}, α_{td} are user-defined hyperparameters, so that

- $\beta_{wt} > 0, \alpha_{td} > 0$ results in smoothing,
- $\beta_{wt} < 0, \alpha_{td} < 0$ results in sparsing
- $\beta_{wt} > -1, \alpha_{td} > -1$ results in an LDA model.

For our study we trained two types of ARTM models: in the sparse models, the Θ matrix was regularized using the sparsing coefficient $\tau = -0.1$; in the dense models, the smoothing coefficient $\tau = 0.1$ was applied. In both types of models, the *Φ* matrix was regularized using the sparsing coefficient $\tau = 0.25$ and the topic decorrelation coefficient $\tau - 10^4$ (so that words with high frequency throughout the collection received lower weights in each document).

All the models (LDA, sparse ARTM, and dense ARTM) were trained on ≈ 600,000 randomly sampled entries from Russian Wikipedia (the dump of 1 March 2020[2]). The

[1] https://bigartm.readthedocs.io/en/stable/intro.html

[2] https://dumps.wikimedia.org/ruwiki

data was cleaned with the corpuscula[3] tool, and lemmatized and POS-tagged using the pymorphy2 parser[4]; bigram collocations (e.g. *чемпионат_мир* ‘world_cup’) were identified using gensim[5]. The Wikipedia corpus was chosen on the assumption that it is likely to represent a large variety of common topics. The Wiki data was vectorized using count vectorization; the topic models were incorporated with BERT word embeddings [35] by concatenating topic vectors with averaged BERT vectors.

All the resources related to this project (the preprocessed Wikipedia dump, the trained topic models, the Russian metaphor-annotated corpus, and the scripts) are available in a github repository[6].

2.2 Experimental setup

The metaphor identification experiment was run on the Russian corpus of metaphor-annotated sentences [32]. The corpus consists of 7,020 sentences; each of them contains one of the 20 polysemous target verbs (e.g. *бомбардировать* ‘to bombard’, *нападать* ‘to attack’, *утюжить* ‘to iron (about clothes)’, *взвешивать* ‘to weigh’, etc.) which is used either metaphorically or non-metaphorically. The number of sentences per target verb ranges from 225 to 693; each of these subsets is balanced by class. Below are examples of metaphoric and non-metaphoric sentences with the target verb *взрывать* ‘to explode (smth)’; the first metaphoric sentence contains an unconventional metaphor, while the second metaphoric sentence demonstrates a conventionalized metaphor:

- Example 1: (Metaphoric) *Ксенофобия – это то, что, возможно, станет бомбой замедленного действия, которая < взорвет > наше общество.* ‘Xenophobia is what may become a ticking bomb which will < explode > our society.’
- Example 2: (Metaphoric) *Для нее было необходимо < взорвать > ситуацию любым способом…* ‘It was necessary for her to < explode > the situation by any means.’
- Example 3: (Non-metaphoric) *Главнокомандующий князь Горчиков приказал < взорвать > уцелевшие укрепления и оставить город.* ‘The commander-in-chief, Prince Gorchikov, gave orders to < explode > the remaining fortifications and to flee the town.’

The metaphor identification task was formulated as sentence-level binary classification. We experimented with several conventional ML algorithms (logistic regression, SVM, Naïve Bayes, Random Forest, etc., including a simple neural network – multilayer perceptron); no deep learning methods (such as LSTM or CNN) were applied to the task, firstly, due to the relatively small size of the experimental corpus and, secondly, due to the fact that in topic modelling documents are represented as bags of words. For each of the three types of topic models (LDA, ARTM dense, and ARTM

[3] https://github.com/fostroll/corpuscula

[4] https://pymorphy2.readthedocs.io/en/latest/

[5] https://radimrehurek.com/gensim/models/phrases.html

[6] https://github.com/steysie/topic-modelling-metaphor

sparse), we took vectors varying between 30 and 130 topics in size. The experiments were run using 5-fold cross-validation.

2.3 Results

The best classification results (in terms of accuracy) – 0.7 – were yielded by the logistic regression (LogReg) and the multilayer perceptron (NN) models with 40, 50, 80, and 90 topics vectors, as summarized in Table 1 (models with 60 and 70 topics are not shown since they produced slightly lower results). It can be seen that somewhat higher results are obtained with the non-regularized LDA-based models.

At the same time, most of the highest results in terms of F1-score are delivered by the SVM classifier on the ARTM sparse and the ARTM dense models.

Table 1. Classification results with topic modelling.

		LDA				**ARTM sparse**				**ARTM dense**			
number of topics	**classifier**	**acc**	**prec**	**rec**	**f1**	**acc**	**prec**	**rec**	**f1**	**acc**	**prec**	**rec**	**f1**
40	LogReg	**0.70**	0.70	0.72	0.71	0.69	0.69	0.72	0.70	0.69	0.69	0.71	0.70
	SVM	0.67	0.64	0.80	0.71	0.67	0.63	0.81	0.71	0.67	0.63	0.82	0.71
	NN	**0.70**	0.70	0.71	0.71	0.69	0.68	0.71	0.70	0.69	0.69	0.71	0.70
50	LogReg	**0.70**	0.69	0.70	0.70	0.69	0.68	0.71	0.69	0.69	0.69	0.71	0.70
	SVM	0.67	0.63	0.79	0.70	0.66	0.63	0.80	0.70	0.66	0.63	0.80	0.70
	NN	0.69	0.68	0.72	0.70	0.69	0.68	0.70	0.69	0.69	0.68	0.72	0.70
80	LogReg	**0.70**	0.72	0.67	0.69	0.69	0.68	0.70	0.69	0.69	0.69	0.70	0.69
	SVM	0.67	0.66	0.73	0.69	0.65	0.61	0.86	0.71	0.67	0.62	0.83	0.71
	NN	**0.70**	0.72	0.67	0.69	0.68	0.68	0.71	0.69	0.69	0.68	0.71	0.70
90	LogReg	**0.70**	0.70	0.67	0.69	0.68	0.67	0.69	0.68	0.68	0.68	0.70	0.69
	SVM	0.67	0.65	0.77	0.70	0.65	0.61	0.84	0.70	0.66	0.61	0.84	0.71
	NN	0.69	0.69	0.69	0.69	0.68	0.67	0.70	0.69	0.68	0.67	0.72	0.69

To compare the results of the topic-based classifiers to other state-of-the-art features, we replicated the features proposed by Badryzlova [36]: lexical (LEX), morphosyntactic (POS), Concreteness-Abstractness (CONC), and semantic coherence (SEM) features. In order to assess the contribution of the topic-based classifier to metaphor identification, we conducted an ablation experiment in which the performance of each feature, as well as their combinations, was evaluated with the topic-based feature

Table 2. Feature ablation experiment (accuracy). Asterisk denotes statistically significant differences between combinations with and without the topic-based model.

feature / classifier	SVM -TM	LogReg -TM	NN -TM	SVM +TM	LogReg +TM	NN +TM
LEX	0.8164	0.8173	0.8179	0.8318	0.8287	0.8301
POS	0.6757	0.6749	0.6702	0.6032	0.5668*	0.5958*
CONC	0.7173	0.7158	0.7178	0.7595*	0.7473*	0.7603*
SEM	0.7195	0.7310	0.7359	0.7319	0.7430	0.7372
TM	0.6715	0.7033	0.7018	---	---	---
LEX+SEM	0.8074	0.8484	0.8340	0.8094	0.8517	0.8382
LEX+POS	0.8204	0.8201	0.8204	0.8353	0.8294	0.8320
LEX+CONC	0.8327	0.8327	0.8323	0.8384	0.8331	0.8359
LEX+POS+CONC	0.8352	0.8350	0.8337	0.8418	0.8339	0.8377
LEX+POS+SEM+CONC	0.8176	0.8544	0.8542	0.8121	0.8537	0.8377

(+TM) and without it (-TM) – see Table 2 (there we show the results for the LDA model with 80 features).

When comparing the performance of the topic-based classifier to the other uni-feature classifiers, we see that the accuracy of TM surpasses the result of the classifier informed with morphosyntactic features (POS); TM is slightly outperformed by the classifiers operating on Concreteness-Abstractness (CONC) and semantic coherence indexes (SEM). In comparison to the lexical classifier, the topic-based classifier falls behind by a tangible margin – similarly to the other three types of features.

When analyzing the contribution of the topic-based model to the other uni-feature models, we observe that addition of TM improves the performance of LEX, CONC, and SEM; however, only the CONC + TM increase proves statistically significant[7]. At the same time, addition of TM to the POS model considerably worsens the result.

Adding topicality to multi-feature models increases the efficiency of classification in at least one of the classifiers in almost all combinations of features (the exception is the last, most complex model); however, this increase of accuracy is rather narrow and does not prove to be statistically significant.

Overall, the highest results are attained with combinations of three to five features, one of which is lexical (LEX). The importance of this feature for metaphor classification is consistent with previous findings [16] and is closely examined by Badryzlova [36]. Lexical cues seem to be the most potent predictors of metaphoricity; therefore, adding other features does not dramatically affect the performance of the classifier. Five of the features implemented in present study bear on the lexico-distributional

[7] Wilcoxon signed-rank test [37] (SciPy implementation) is used in this study to evaluate the statistical significance of results.

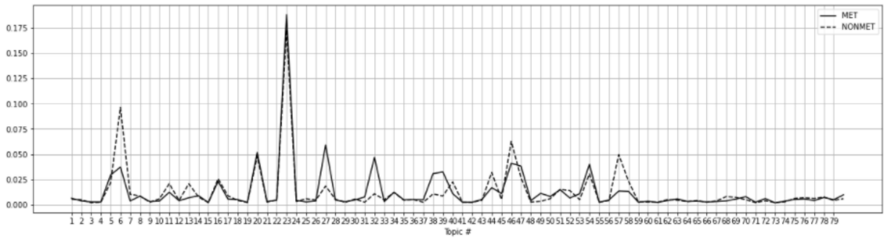

Fig. 1. Distribution of topics in metaphoric (MET) and non-metaphoric (NONMET) contexts (LDA, 80 topics).

properties of words: LEX, SEM, CONC, and TM – and thus they complement each other in this regard. In contrast, the POS feature is based on patterns of words' morphosyntactic combinability which are highly idiosyncratic and thus less generalizable and reliable in metaphor prediction [36]. The substantial drop in classification accuracy in the POS+TM model most likely occurs because the POS predictor, rather weak as it is, is collapsed with the topic-based model, which is intended to capture a different type of distributions, and, moreover, is not the strongest predictor in itself.

3 Analysis of topic distribution in metaphoric and non-metaphoric contexts

In order to test whether metaphoric and non-metaphoric contexts contain different sets of topics, we applied the Kolmogorov-Smirnov statistic [38] which tests the hypothesis that two sets belong to the same distribution. On all our matrices of topics the *p-value* proved above the significance level – therefore, we cannot claim that the distributions of topics in metaphoric vs. non-metaphoric contexts are statistically different. However, this does not mean that the topics are distributed uniformly across these two types of discourse. Analysis of distribution revealed that there are topics that are indicative of either metaphoric or non-metaphoric utterances.

Fig. 1 shows the distribution of topics in the metaphoric (MET) and the non-metaphoric (NONMET) subcorpora, as generated by the LDA model with vector dimensionality of 80 topics. It is easy to notice that topics 27, 32, 38, and 39 prevail in metaphoric contexts, while topics 6, 11, 44, 46, 57, and 58 are more salient in non-metaphoric sentences. Remarkably, topics 16, 20, and 23 are equally frequent in both subcorpora.

Analyses of the topic matrixes generated with the LDA, the ARTM dense and the ARTM sparse models indicated the following topical cues of metaphoricity (the names of the topics were assigned manually): Literature and Writing, Economy, Judicial System, Corporate Management, and Railway. While the metaphoricity of the first four topics is quite expected, explained by the high frequency of analogies and comparisons in their metaphoric contexts, the metaphoricity of the Railway topic arises from the conventionalized indirect meanings of some of the target verbs, for example, *пилить* ‘to travel a long distance’ (lit. ‘to saw’):

- Example 4: (Metaphoric) *Поезд подошел и оказалось, что до нашего вагона еще < пилить > и < пилить >*. The train pulled in, and we discovered that we had to < do a great deal of sawing > (lit. ‘to walk a long distance’) to reach our carriage

The topics that prevail in the non-metaphoric subcorpus are: Biology, Language, Cars, Chemistry, Aviation, Peoples and Traditions, and Religion, e.g.:

- Example 5: (Non-metaphoric) *В верхней части карты Таро находится божественная фигура, обычно представленная крылатым ангелом, [который] смотрит из облаков и < трубит > в трубу.* The upper part of the Tarot card depicts a divine figure which is usually represented by a winged angel who is looking down from the clouds, < trumpeting >.

The topics that have high frequency in both metaphoric and non-metaphoric contexts are: Military and Warfare, Cinema, TV Series and Computer Games, and Architecture and Construction. The following sentences demonstrate examples of metaphoric and non-metaphoric occurrences of the Military and Warfare topic:

- Example 6: (Metaphoric) *Когда немцы с земли и воздуха < утюжили > снарядами и бомбами наши армейские позиции, только воля божья спасла их на дне окопа и в землянке.* When Germans were < ironing > (lit. ‘bombing out’) our army’s positions with shells and bombs, it was but for the grace of God that they survived at the bottom of a trench and in a dugout.
- Example 7: (Non-metaphoric) *Враги рыли под землей галереи, чтобы, заложив мины, < взорвать > русские укрепления.* The enemies were digging underground galleries in order to plant mines and < explode > the Russian fortifications.

The identified topical cues seem to reflect certain broadly defined realms of reality rather than the more fine-grained conceptual structures suggested by the cognitive metaphor theory and attested in empirical linguistic research [e.g. 3]. Thus, the present implementation of topic modeling for metaphor analysis falls short of capturing the expected conceptual mappings. Yet, it demonstrates that differences exist in the topical profiles of metaphoric and non-metaphoric discourse, calling for further investigation. Besides, it should be borne in mind that the inventory of topical predictors of metaphoricity / non-metaphoricity in the present study is by no means exhaustive: it is limited by the scope and the size of the experimental corpus, and is likely to alter with expansion of the corpus.

4 Heterogeneity of topic distribution in metaphoric and non-metaphoric discourse

According to the conceptual metaphor theory, metaphoric contexts may represent at least two topics associated with the Target and Source Domains (see Politics and Military/Warfare in Example (1) above) while non-metaphoric contexts can be limited to one topic space (see Military/Warfare in Example (3)). Therefore, we can expect more salient topics per sentence in the MET class than in the NONMET class. Besides that, the Source Domain can be mapped to different Target Domains in different sentences, which assumes the topic space to be potentially more variable in MET than in NONMET.

We used several probability thresholds to empirically define the number of salient topics per sentence for the LDA matrix with k=80 topics. The average number of topics is significantly larger in the MET class as compared to the NONMET class for thresholds below 0.1 (t-test at the threshold 0.05: $t = 5.718$, $p = 1.122e\text{-}08$). As for the individual verbs comprising the metaphor-annotated corpus (see Section 2.2), this trend holds for 11-15 out of the 20 verbs. However, the verb *уколоть* ‘prick’ follows a different pattern, with metaphoric contexts having in general fewer topics per sentence as compared to non-metaphoric ones. We can explain this by the specifics of the Wikipedia-based topic modeling as both everyday physical events (Source Domain) and emotional reactions (Target Domain) are underrepresented in the training Wiki data and therefore in the topic clusters. This is in line with another observation that the verbs of everyday activity such as *утюжить* ‘to iron (about clothes)’ and *причесать* ‘comb’ form a subgroup that shows fewer topics per sentence in non-metaphoric discourse than other verbs.

We run latent profile analysis [39] to visualize most common topical profiles in each verb in MET and NONMET. We conclude that there is not enough evidence to prove the heterogeneity hypothesis from the point of view of the variability of topics in metaphoric and non-metaphoric discourse since verbs are inconsistent in their behaviour in the current settings. All this suggests that other pre-trained topic models, with a greater number of domains covered, could be used to further test the hypothesis of topic heterogeneity. For example, topic models trained on a corpus of fiction could be expected to reveal the currently underrepresented topics (such as everyday activity or emotional reactions) and, besides, to capture the topics formed by indirect, figurative usages of words.

Conclusions

We applied topic-based features to the task of sentence-level metaphor identification in Russian. In doing so, we compared three types of topic models – a conventional LDA model and two models with additive regularization – ARTM dense and ARTM sparse. When taken alone, the topic-based classifier yields the accuracy of 0.7; in comparison to other state-of-the-art features, topic-based classifier performs on the par with Concreteness-Abstractness and semantic coherence indexes, yet it underperforms in comparison to the lexical baseline. Combining the topic-based model with the other features resulted in statistically significant improvement only in the combination with the Concreteness-Abstractness model; integrating the topic-based model into the morphosyntactic one led to a sharp decrease in performance, which is likely due to the weak predictive power of both features and the differences in patterns (morphosyntactic vs. lexico-distributional combinability) captured by them.

However, application of topic modelling to metaphor analysis allowed us to test two hypotheses about the conceptual nature of metaphor suggested in linguistic literature and practice of metaphor studies.

Firstly, we analyzed the topical profiles (i.e. the distribution of topics) in metaphoric and non-metaphoric discourse, and identified the topical cues, that is, the topics that are indicative of metaphoric and non-metaphoric contexts. These cues do not resemble the Source and Target domains attested in linguistic studies; yet, the existence of these cues suggests a promising direction for further research.

The second hypothesis concerned topic heterogeneity of metaphoric and non-metaphoric discourse. According to the conceptual metaphor theory, metaphoric contexts should be more topically heterogeneous (due to the presence of two conceptual domains, the Source and the Target) than non-metaphoric ones. We found some evidence that metaphoric uses are associated with a larger number of topics than those identified in non-metaphoric uses. However, larger studies are needed to support our findings; for example, applying topic models trained on corpora other than Wikipedia (e.g. fiction) might be able to capture the topics that are currently underrepresented in our models.

References

1. Lakoff, G., Johnson, M.: Philosophy in the flesh. New York: Basic books (1999).
2. Lakoff G., Johnson M.: Metaphors we live by. Chicago: University of Chicago Press (1980).
3. Honga, J., Stickles, E., Dodge, E.: The MetaNet metaphor repository: Formalized representation and analysis of conceptual metaphor networks. In: 12th International Cognitive Linguistics Conference. Edmonton, Canada (2013).
4. Flusberg, S.J., Matlock, T., Thibodeau, P.H.: Metaphors for the war (or race) against climate change. Environmental Communication. 11, 769–783 (2017).
5. Hauser, D.J., Schwarz, N.: The war on prevention: Bellicose cancer metaphors hurt (some) prevention intentions. Personality and Social Psychology Bulletin. 41, 66–77 (2015).
6. Landau, M.J., Oyserman, D., Keefer, L.A., Smith, G.C.: The college journey and academic engagement: How metaphor use enhances identity-based motivation. Journal of Personality and Social Psychology. 106, 679 (2014).
7. Thibodeau, P.H., Boroditsky, L.: Metaphors We Think With: The Role of Metaphor in Reasoning. PLoS ONE. 6, e16782 (2011). https://doi.org/10.1371/journal.pone.0016782.

8. Shutova, E., Teufel, S.: Metaphor Corpus Annotated for Source-Target Domain Mappings. In: LREC. pp. 2–2 (2010).
9. Steen, G., Herrmann, B., Kaal, A., Krennmayr, T., Pasma, T.: A Method for Linguistic Metaphor Identification: From MIP to MIPVU. John Benjamins Publishing Company, Amsterdam; Philadelphia, PA (2010).
10. Klebanov, B.B., Shutova, E., Lichtenstein, P.: Proceedings of the Second Workshop on Metaphor in NLP. In: Proceedings of the Second Workshop on Metaphor in NLP. Association for Computational Linguistics, Baltimore, MD (2014).
11. Klebanov, B.B., Shutova, E., Lichtenstein, P.: Proceedings of the Fourth Workshop on Metaphor in NLP. In: Proceedings of the Fourth Workshop on Metaphor in NLP. Association for Computational Linguistics, San Diego, California (2016).
12. Shutova, E., Klebanov, B.B., Tetreault, J., Kozareva, Z.: Proceedings of the First Workshop on Metaphor in NLP. In: Proceedings of the First Workshop on Metaphor in NLP. Association for Computational Linguistics, Atlanta, Georgia (2013).
13. Shutova, E., Klebanov, B.B., Lichtenstein, P. eds: Proceedings of the Third Workshop on Metaphor in NLP. Association for Computational Linguistics, Denver, Colorado (2015).
14. Leong, C.W.B., Klebanov, B.B., Shutova, E.: A report on the 2018 VUA metaphor detection shared task. In: Proceedings of the Workshop on Figurative Language Processing. pp. 56–66 (2018).
15. Leong, C.W. (Ben), Beigman Klebanov, B., Hamill, C., Stemle, E., Ubale, R., Chen, X.: A Report on the 2020 VUA and TOEFL Metaphor Detection Shared Task. In: Proceedings of the Second Workshop on Figurative Language Processing. pp. 18–29. Association for Computational Linguistics, Online (2020).
16. Klebanov, B.B., Leong, B., Heilman, M., Flor, M.: Different texts, same metaphors: Unigrams and beyond. In: Proceedings of the Second Workshop on Metaphor in NLP. pp. 11–17 (2014).
17. Hovy, D., Srivastava, S., Jauhar, S.K., Sachan, M., Goyal, K., Li, H., Sanders, W., Hovy, E.: Identifying metaphorical word use with tree kernels. In: Proceedings of the First Workshop on Metaphor in NLP. pp. 52–57. Citeseer (2013).
18. Ovchinnikova, E., Israel, R., Wertheim, S., Zaytsev, V., Montazeri, N., Hobbs, J.: Abductive inference for interpretation of metaphors. In: Proceedings of the Second Workshop on Metaphor in NLP. pp. 33–41 (2014).
19. Shutova, E., Kiela, D., Maillard, J.: Black holes and white rabbits: Metaphor identification with visual features. In: Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 160–170 (2016).
20. Panicheva, P.: Analiz parametrov semantičeskoj svjaznosti s pomošč’ju distributivny xsemantičeskix modelej (na materiale russkogo jazyka) [Analysis of parameters of semantic coherence by means of distributional semantic models (on Russian data)], (2019).

21. Gandy, L., Allan, N., Atallah, M., Frieder, O., Howard, N., Kanareykin, S., Koppel, M., Last, M., Neuman, Y., Argamon, S.: Automatic Identification of Conceptual Metaphors With Limited Knowledge. In: AAAI (2013).
22. Gedigian, M., Bryant, J., Narayanan, S., Ciric, B.: Catching metaphors. In: Proceedings of the Third Workshop on Scalable Natural Language Understanding. pp. 41–48. Association for Computational Linguistics (2006).
23. Klebanov, B.B., Leong, C.W., Gutierrez, E.D., Shutova, E., Flor, M.: Semantic classifications for detection of verb metaphors. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 101–106 (2016).
24. Dunn, J.: Evaluating the premises and results of four metaphor identification systems. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 471–486. Springer (2013).
25. Dunn, J.: What metaphor identification systems can tell us about metaphor-in-language. In: Proceedings of the First Workshop on Metaphor in NLP. pp. 1–10 (2013).
26. Neuman, Y., Assaf, D., Cohen, Y., Last, M., Argamon, S., Howard, N., Frieder, O.: Metaphor Identification in Large Texts Corpora. PLOS ONE. 8, e62343 (2013). https://doi.org/10.1371/journal.pone.0062343.
27. Strzalkowski, T., Broadwell, G.A., Taylor, S., Feldman, L., Yamrom, B., Shaikh, S., Liu, T., Cho, K., Boz, U., Cases, I., others: Robust extraction of metaphors from novel data. In: Proceedings of the First Workshop on Metaphor in NLP. pp. 67–76 (2013).
28. Turney, P.D., Neuman, Y., Assaf, D., Cohen, Y.: Literal and metaphorical sense identification through concrete and abstract context. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 680–690. Association for Computational Linguistics (2011).
29. Beigman Klebanov, B., Leong, C.W., Flor, M.: Supervised Word-Level Metaphor Detection: Experiments with Concreteness and Reweighting of Examples. In: Proceedings of the Third Workshop on Metaphor in NLP. pp. 11–20. Association for Computational Linguistics, Denver, Colorado (2015). https://doi.org/10.3115/v1/W15-1402.
30. Heintz, I., Gabbard, R., Srivastava, M., Barner, D., Black, D., Friedman, M., Weischedel, R.: Automatic extraction of linguistic metaphors with lda topic modeling. In: Proceedings of the First Workshop on Metaphor in NLP. pp. 58–66 (2013).
31. Abdi Ghavidel, H., Khosravizadeh, P., Rahimi, A.: Impact of Topic Modeling on Rule-Based Persian Metaphor Classification and its Frequency Estimation. International Journal of Information and Communication Technology Research. 7, 33–40 (2015).
32. Badryzlova, Y., Panicheva, P.: A Multi-feature Classifier for Verbal Metaphor Identification in Russian Texts. In: Conference on Artificial Intelligence and Natural Language. pp. 23–34. Springer (2018).
33. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of machine Learning research. 3, 993–1022 (2003).

34. Vorontsov, K., Potapenko, A.: Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. In: International Conference on Analysis of Images, Social Networks and Texts. pp. 29–46. Springer (2014).
35. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M.: HuggingFace’s Transformers: State-of-the-art Natural Language Processing. ArXiv. arXiv–1910 (2019).
36. Badryzlova, Y.: Automated metaphor identification in Russian texts, (2019).
37. Wilcoxon, F.: Individual comparisons by ranking methods. In: Breakthroughs in statistics. pp. 196–202. Springer (1992).
38. Massey Jr, F.J.: The Kolmogorov-Smirnov test for goodness of fit. Journal of the American statistical Association. 46, 68–78 (1951).
39. Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E.: mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R journal. 8, 289 (2016).