Semiotic analysis of narrative legal texts using asymmetric
document divergency

Vladimir Krylov, Sergei Krylov and Grigory Zhigalov
National Research University Higher School of Economics, Russia

E-mail: vkrylov@hse.ru

Abstract. In the paper the case is studied then semiotic signs can be represented as language constructs in the same language as the text for the interpretation. The goal is to obtain estimates of the depth of interpretability with the respect to each of the signs by finding the projections of the narrative on these language constructs. That is, we find the probability of the presence of each of the signs. By use the embedding of linguistic constructions into a multidimensional numerical vector space and the grid layout, which is the image of the system of signs. At the stage of the inference, our model computes the nodes of the grid, which are the most adequate to the input text by the method of new proposed asymmetric similarity. The constructed model was studied using Russian criminal, civil and labour legislation. We suggest using the models called the constellation as for the grid of normative acts and the narrative together with the corresponding divergence. In this paper, examined the pre-trained models known in NLP as doc2vec and Fast Text. The evaluation of the quality of models was carried out using open databases of court decisions.

1 Introduction
Among the tasks of natural language processing (NLP) [1] the often task is the interpretation of narrative texts in terms of codes, rules and taxonomies. The task looks like we need to find the hidden predefined signs in the presented story. The example is the diagnosis of the disease according to anamnesis (e.g. using Classification of Disease ICD-10) [2]. The text of anamnesis contains recollection of signs of illness, medical history of patient that to be converted into the diagnosis. The other example is the human behavior interpretation in terms of legislative acts (e.g. the criminal code). This is the frequent problem for courts of law to convert plain description of the human behavior into the list of the law violations [3]. According to NLP the interpretation is the mapping of narrative texts set into subset of finite set of articles, classes, acts, taxes etc. In formal terms, this is the task of classification. In some cases, the solution by supervised learning classification is acceptable. For example, text sentiment analysis (classification by positive or negative key) can be accomplished by model training on large set of short texts [4]. However, the medical diagnosis and the legislative tasks have huge number of classes to choose and they operate with the long narratives. In this case, the classifier gives the result with unacceptable error level.

Another example is LSA – Latent Semantic Analysis of texts [5]. The mapping of the texts to the small dimension numerical space using thematic clustering is the approach. This is interpretation in terms of internal entities of the text called themes. This is the analysis applied to themes in the texts but not external themes. In our case the external themes have to used to interpretation. If external and
internal texts can be combined, then the task can be solved by LSA. However, the external themes extraction from external signs description but not narrative is almost mission impossible.

Concerning the described problems, there can state that there is need to create new algorithmic pipeline to solve tasks of the following kind:

1. The events and the supporting facts are narrated in natural language (narrative text [6]).
2. The finite set of signs is defined. The meaning of a sign is a text in natural language. E.g., it could be facts, conditions, and definitions, anything else in the form of meaningful text.
3. The problem is to find the set of semiotic signs that corresponds the narrative text. The term “corresponds” is task dependable and it says “narrative contains of the same meaning as sign” in general case.

We are using term “sign” as defined in semiotics [7] to enhance the task definition extremely wide. However, we are going to limit usage of the signs by word defined signs only. Anyway, we call the task in the aforementioned form (1-2-3) the task of semiotic text analysis.

2 The Juridical Narrative Semiotic Analysis Pipeline
Our research concerns the practical aspects of semiotic analysis [8] – the interpretation of legislative texts in the law terms. Below we use the terms corresponding to the application area. However, the pipeline can be used in other practical semiotic tasks too.

2.1 Semiotic analysis commutative diagram
Many other mathematical concepts and properties, especially in algebraic topology, homological algebra, and category theory, can be formulated in terms of commutative diagrams. Let’s describe our task by the diagram of factorization of mapping f (the set G to the set H).

![Factorization Diagram](image)

**Figure 1.** The common commutative factorization diagram

In the task, the set G is the collection of the narratives of event flow and of human behavior descriptions. The set H is the set of subsets of finite set of semiotic signs. For legislative case it is set of legislation act IDs. The finding of mapping f is the building of solution of the task of narrative interpretation in terms of law. The implementation of the mapping is the algorithmic solution of semiotic analysis task. However, the direct finding of the mapping for any set of narratives is the insoluble task. Let’s use factorization shown in the above diagram. Instead of the finding f we can find pair of mappings π (epimorphism) and  \( \tilde{f} \) (isomorphism) and corresponding factor set G/ker(f) for equivalence of mapping f. Let’s make first mapping of the pair by the following composition:

\[
\pi = \text{emb} \cdot \text{kNN}
\]

The first mapping is the embedding [9] of narrative into n-dimensional real space \( \mathbb{R}^n \), second – the operation of k-nearest neighbors selection using some distance in \( \mathbb{R}^n \). As \( \tilde{f} \) we can use bijection of embedding indices and linguistic representations. The factor set G/ker(f) is the set of points in \( \mathbb{R}^n \).
where these points correspond to selected mapping \( \text{emb} \). Let’s define the models are going to use to implement mapping \( \hat{f} \).

### 2.2 Embedding of linguistic constructions

As defined above the embedding operation must be applied to the narrative as well as textual description of semiotic signs – legislation act set. The embedding of legislation act is executed at the stage of training. The embedding of the narratives is executed at the stage of inference of the interpretation. It requires some limitations of the calculation complexity. In our research we pay attention to highly effective known methods of embedding of linguistic entities only. We construct needed implementation of mapping \( \text{emb} \) having needed properties. The legislation act as well as the narratives are the long texts. This is why at the preparation phase of embedding we take the shortened representation using methods of text summarization. This transformation allows to build textual objects used at the embedding as ordered by importance sequences. The models of extraction-based summarization are more precise for text meaning preservation than abstraction-based summarization models. However, the developed pipeline works fine with any types of summarization. The sentences and parts of sentences of the size less than 30 words selected as minimal semantic units. Every such text fragment having number \( i \) of text \( m \) is mapped to a point of \( n \)-dimensional real space \( R^n \). The fragment represented as vector \( x_i^m \) of size 300. We used the embedding of sentences by doc2vec and FastText methods for the start. Next step will be to use multilanguage Universal Sentence Encoder. The implementation of the mapping \( \text{emb} \) requires not only the embedding of minimal semantic units-sentences. It is necessary the integration of the embeddings of all sentences of the narrative and legislation act as well. If \( m \) (number of sentences) is equal to \( M \) then the full embedding of the text is the set of \( m \) points or constellation into \( R^n \).

The text of your paper should be formatted as follows:

\[
X^m = \{x_i^m\} i = 1, \ldots, M
\]  

So the mapping \( \text{emb}: G \rightarrow (R^n)^M \) transforms every narrative or legislation act to constellation of \( R^n \). The set of such constellations forms factor set \( X=G/\text{ker}(f) \).

### 2.3 k-neighbor task in the constellation space

Every point of constellation is the element of metric vector space \( R^n \) having distance \( d(x_i, x_j) \), then we can define \( X \) to have constellation divergence \( \partial(X^a, X^b) \) from \( X^a \) to \( X^b \) as the following:

\[
\partial(X^a, X^a) = 0, \forall a; \quad \partial(X^a, X^b) \geq 0, \forall a, b
\]  

The equality \( \partial(X^a, X^b) = \partial(X^b, X^a) \), \( \forall a, b \) may be violated. Moreover, the distance from \( X^b \) to \( X^a \) can be other. This case called anisotropy or asymmetric of distance. We define such divergence \( X^a \) to \( X^b \) as

\[
\partial(X^a, X^b) = \sum_{i=1}^{M} \min_{k} d(x_i^a, x_k^b)
\]  

In Fig. 2 you can see the constellation divergence from 2 points constellation to 3 points constellation. Here \( d(x_i^a, x_k^b) = d_{ik}^{ab} \) can be cosine distance.
The above divergence is used to implement the mapping \( kNN \). Let’s introduce the integer \( k>0 \) – number of signs to interpret the text represented by constellation \( X^c = \{x_i\} \); \( i = 1, \ldots, M \). We can calculate divergences from \( X^c \) to all constellation of signs \( \{X^1, X^2, \ldots, X^S\} \). Then we can define order them ascending.

\[
\partial(X^c, X^{i1}) \leq \partial(X^c, X^{i2}) \leq \cdots \leq \partial(X^c, X^S)
\]  

(5)

Let’s call first \( k \) constellations in that order as \( X^c \) for the mapping \( kNN \). Having the implementations of all mappings of decomposition of mapping \( f \) defined we can specify Narrative Semiotic Analysis Pipeline

### 2.4 Pipeline

We used two pipelines: the first for main models training and the second for real time inference – The creation of the list of the descriptions of legislation acts semiotic signs that interpret input narrative during the inference. Both pipelines use several common models and processing modules. They are the following.

- **PP1** – preprocessing module to prepare text for embedding.
- **EW1, EW2** – modules of tokenized texts embedding to numeric multidimensional space.
  - **EW1** – uses FastText model to embed words and to build vector of sentence as average vector of all included words.
  - **EW2** – uses doc2vec model
- **DW** – module to calculate divergency from a constellation of given narrative to constellations of the signs with identifiers from input list. The output is the list of specified length. It contains the identifiers of constellations that have minimal divergency.

There are auxiliary modules in pipelines also. They provide parameters adjustment, variables conversion, the lists control.

### 3 Experiments

Now compare the results by different embedding methods on a sign interpretation task and let’s show that the approach is effective enough to build the computer-aided system where law acts match the narrative. Our main findings are:

- semiotic analysis pipeline can find articles of legislative acts corresponding to sequences of actions, events, behaviors with the correctness defined by embedding method;
• the representation of narratives and semiotic signs descriptions by constellations of multi-dimensional vectors and constellation divergency to measure the proximity of signs and narratives give best results.

3.1 Datasets
The research based on the following text data in Russian language:
• Crime Code of Russian Federation;
• Civil Code of Russian Federation;
• Criminal courts orders
• Civil court orders
• Russian Language corpus.

3.2 FastText embedding
EW1 module uses FastText model. At preprocessing we used library functions of genism package [10].
At first, we have separated texts by sentences having less than 500 words in each one. We have formed Pandas DataFrame and we named it s. This way has been created corpus of sentences to train the model. At the second stage has been done embedding of the words of the corpus to space of 300-dimensional real vectors.

```
fasttext = gensim.models.FastText(corpus_file='all/tmp_sent.txt', size=300)
```

The averaging of normalized vectors of words of the sentence is the core of this step. We got 300-dimensional vector for every sentence. We stored results in special column of DataFrame.

```
s['fasttext'] = s['sent'].map(sent_emb)
```

The following picture (Fig.3) shows visualization of the space of the embeddings for five main sources which are descriptions of semiotic signs: criminal code, civil code, juridical narratives criminal court orders, civil court orders and the corpus of Russian language (opencorpora).

![Figure 3. The FastText space sentence based PCA 2D representation](image)

3.3 doc2vec embedding
EW2 module uses model doc2vec [11] which returns multi-dimensional numeric vector of the sentence. We have selected the same size of embedding space as earlier equal to 300.
At the preprocessing stage we have prepared corpus of lemma based documents.

```
gensim.utils.save_as_line_sentence(df['lemmas'].str.split(), 'all/tmp_lemmas.txt')
```

The training of the model is the same as the described in previous section. The only difference is that the model works with the sentence at the beginning.

```
doc2vec = gensim.models.doc2vec.Doc2Vec(corpus_file='all/tmp_lemmas.txt', epochs=300, vector_size=300, workers=8)
```
The space of embeddings can be visualized for EW2 also (Fig.4).

![Figure 4. The doc2vec space sentence based PCA 2D representation](image)

3.4 Hidden semiotic signs finding
If we look at the pictures above, we can say that the structure of the spaces is significant different. But we cannot compare these embeddings models before the end goal is not reached. If we choose the more detailed (two classes) view for example for criminal code space and the criminal court orders only, the difference between the spaces according FastText and doc2vec models is significant as well. The fig.5 shows how the FastText based spaces look like.

![Figure 5. The FastText criminal code and criminal court orders spaces PCA 2D representation](image)

The next step in our study was to study the possibility of finding the most appropriate semiotic signs for the meaning that correspond to a given narrative. To determine the quality of the solution to the problem of finding signs, we used as a narrative the dataset of court decisions. It contains as a label a set of articles of the criminal code that were assigned to the description by the court. Thus, we have the opportunity to compare the result of the work of our pipeline and the real court order. To assess the quality, the Weak Accuracy metric was used, which shows the share of narratives for which the pipeline generated at the output at least one article of the criminal code the same as the court. The pipeline module generates a predetermined number of characters closest according constellation divergency for each narrative. It is obvious that the greater the number of signs specified, the greater the likelihood of at least one falling into the number of correct ones. In the limit, this probability tends to 1. The pipeline works the better, the more correct signs fall into a smaller neighborhood. We investigated how Weak Accuracy depends on the size of a given neighborhood. Figure 7 shows the dependence of Weak Accuracy on the size of the neighborhood for the pipeline with modules EW1 and EW2, that is, using embeddings with FastText and doc2vec methods. As shown by experiments, embedding by FastText gives a significantly better result. However, the obtained correctness of finding the signs remains unsatisfactory for practical purposes of the criminal court.
To find out the possible reasons for such a low probability of getting the description of the correct sign in a close vicinity of a narrative, it makes sense to study in more detail the structure of the space of signs, built on their textual descriptions. To this end, we built threshold graphs of space with the applied constellation divergence. The Fig. 6 on the right shows one of such graphs for signs corresponding to the articles of the criminal code. A discussion of the features of the resulting graph and the corresponding legal space is beyond the scope of this article. Here we only note that the results show one of the possible reasons for the low quality of the semiotic analysis of the narratives in order to find the signs of the criminal code. This reason lies in the vagueness of the definitions of the signs in the criminal code text in Russian. This conclusion serves as a basis for choosing the directions of future work – the use of other systems of semiotic signs, on which the results of the analysis will be more practically significant. Another direction is the improvement of the pipeline. We intend to explore other methods of embedding and space metrification.

4 Conclusions
In our work, we showed that the FastText methods make it possible to obtain almost twice as accurate results as the use of the doc2vec technique. At the stage of the nearest neighbor search, significant improvements were achieved by applying a new space of finite sets of vectors (constellations) of a multidimensional real space, equipped with an asymmetric similarity of distance – the divergence of constellations. Next step we plan to do is using the Universal Sentence Encoder to embed the signs and narrative sentences.

We have also made significant progress in understanding the analysis of the link structure of semiotic signs by their descriptions. We constructed graph representations of several sign systems based on the constellation divergence we introduced. For example, it turned out that the sign system of the national Criminal Code of the Russian Federation, the Civil Code and the Code of Labor laws of the Russian Federation are qualitatively different.

References
[1] Goldberg Y 2017 Neural Networks Methods for Natural Language Processing (Morgan&Claypool Publishers)
[2] American Medical Association ICD-10-CM 2019: The Complete Official Codebook
[3] Lanier M, Henry S, Desire’ J M Anastasia 2018 Essential Criminology (4th edn., Routledge)
[4] Liu B 2012 Sentiment Analysis and Opinion Mining (Morgan & Claypool Publishers)
[5] Landauer T, McNamara D, Dennis S and Kintsch W 2013 Handbook of Latent Semantic Analysis (Psychology Press)
[6] Spencer L 2005 A Step-by-step Guide to Narrative Writing (The Rosen Publishing Group)
[7] Chandler D 2017 Semiotics: The Basics (Taylor & Francis)
[8] Deledalle G and Charles S 2001 Peirce’s Philosophy of Signs: Essays in Comparative Semiotics (Indiana University Press)
[9] Ma E 3 silver bullets of word embeddings in NLP [Online] https://towardsdatascience.com/3-silver-bullets-of-word-embedding-in-nlp-10fa8f50cc5a
[10] Gensim Tutorial – A Complete Beginners Guide [Online] https://www.machinelearning-plus.com/nlp/gensim-tutorial/
[11] Shperber G A gentle introduction to Doc2Vec [Online] https://medium.com/scaleabout/a-gentle-introduction-to-doc2vec-db3e8c0ce5e