International Journal of Artificial Intelligence,
ISSN 0974-0635; Int. J. Artif. Intell.
2025, Volume 23, Issue Number 1

Methods of Unsupervised Semantic Analysis of Small and Medium-Sized Corpora

S. Dolgikh[1] and **O. Sliusarenko**[2]

[1]Department of Information Technology, National Aviation University,
Lubomyra Huzara Ave, 1, Kyiv (Ukraine)
Email: sdolgikh@nau.edu.ua

[2]Faculty of Computer Science and Technology, National Aviation University,
Lubomyra Huzara Ave, 1, Kyiv (Ukraine)
Email: 3615640@stud.nau.edu.ua

ABSTRACT

Analysis and description of text corpora can present a number of technical challenges, especially in the case of corpora built by automated content extraction that may not allow for readily available annotations and other semantic information about the texts. In this work we describe and test an approach to analysis of the semantic content of corpora based on the methods of unsupervised feature extraction, dimensionality reduction and concept learning. With model corpora represented by texts in English newsgroups, we demonstrate how characteristic semantic types can be identified with methods of unsupervised machine learning and clustering. The results can be an instrumental addition to methods of analysis of semantic context of text corpora where the prior description such as annotations may not be available or is scarce. The approach and methods demonstrated in this work are in no way limited to the English language and can be applied to corpora in any language where the appropriate vectorization and preprocessing methods are available.

Keywords: Natural Language Processing, semantic analysis, unsupervised learning, statistical machine learning, clustering.

Mathematics Subject Classification: 68T50, 68T10

Computing Classification System: I.2

1. INTRODUCTION

Text corpora can be an instrumental source of linguistic information in essential NLP applications not in the least, machine translation. Among other methods, statistical machine translation (Brown et al 1990) (Ahmadnia et al 2019) is based on construction of representative corpora in the source and target languages in the semantic domain of translation. While SMT methods has shown some promise (Och and Ney 2003) (Koehn et al 2007), technical challenges have been noted with construction of representative corpora:

- Effort intensity of creating corpora with good representation of semantic content of the domain;

- Complexity and effort intensity of semantic analysis of large corpora.

With automated methods of construction of corpora recently, including automated corpus sampling (Moreno-Ortiz and Garcia-Gamez 2023) very large and massive corpora of texts can be created. However, with the volume of corpora the complexity of semantic analysis increases as well, and the necessary effort, at least in part manual, rises in accordance. These natural challenges call for development of approaches and methods in unsupervised semantic analysis of large bodies of texts that can be performed in an automated process by machine systems.

2. PRIOR WORK

Approaches in unsupervised analysis of corpora, specifically in the statistical branch of the domain that does not depend on a teacher or significant volume of prior knowledge about the semantic content have been developed over time. The ground-laying works include methods like Latent Semantic Analysis (LSA) developed in late 1990 - early 2000 (Deerwester et al 1990) (Foltz and Dumais 1992) establishing the direction that has been used productively over the subsequent period including in this work: as a first step, describing the corpus in numerical form, ”vectorization” of the corpus in terms of some informative factors (which will be discussed further); next, reducing the dimension of the representation by methods that do not require essential prior knowledge and based on the intrinsic structure of the text data; and analysing the relations between characteristic semantic structures in the informative latent representations or ”embeddings”, commonly of massively reduced dimension of descriptive parameters (latent coordinates).

On the foundation laid by these pioneering works, further methods in unsupervised statistical semantic analysis of corpora were developed, including Probabilistic LSA (Hofmann, T. 2001) that addressed an essential limitation of LSA: the assumption of linearity of the embeddings obtained with Singular Value Decomposition (SVD) that does not necessarily have strong grounding in application to very sparse descriptions of corpora in the term-frequency framework (Sparck Jones 1972). More general non-linear sparse feature extraction methods were developed including Stochastic Neighbor Embedding (SNE) (Hinton and Roweis 2002) and related, that were successfully applied to text data including t-SN (van der Maarten and Hinton 2008), Uniform Manifold Approximation (McInnes, Healy and Melville 2018) and others. Methods of deep learning and generative learning demonstrated an effective ability to learn the structures of characteristic types or concepts with different sets and types of complex data (Welling and Kingma 2019) (Dolgikh 2024).

In a different perspective on semantic analysis of texts, attention was given to the cooccurrence of terms in complete texts or a certain window of text, speech etc. (Distributional semantics, (Lenci 2018), (Boleda 2020)). This approach is based on a much richer and for that reason, significantly more data-intensive and sparser numerical representations of text data that include the context of the terms and can be applied to large bodies of texts. In the state of the art to date in the semantic representation of the language, large contextual distributed semantic models (DSM) based on generative neural networks such as BERT (Devlin et al 2018) among many others and generative pre-trained models have shown the ability to generate text approaching the level of an average human, and a human specialist.

In this work we approached the problem of automated semantic analysis of corpora with methods of unsupervised machine learning that do not depend on massive prior information about the content of the datasets. A specific focus has been on the question, of whether characteristic semantic content of small to medium-sized corpora (in the study, thousands of texts) can be determined with entirely unsupervised methods, without prior semantic analysis or external information on semantic content. From that perspective, the use of large language models, though massively more powerful was deemed above reasonable for the stated problem, including the required resources of computational power, memory and the size of the models. As well, methods discussed here can be readily applied to corpora of specialized texts, which may cause challenges for general language models, in case of insufficient sampling of specialized domains. Still, obtaining matching results with large language models could be an interesting and worthy exercise for a future study.

One can mention a recent study (Choudhari et al 2022) that attempted a similar analysis of methods of unsupervised dimensionality reduction in application to semantic analysis of small corpora. In this work we obtained the following improvements on the reported results:

- We evaluated more methods of unsupervised feature extraction and dimensionality reduction, both linear and non-linear, most of the commonly used with text data, including generative neural networks;
- in the cited study, the latent dimension of the embeddings was fixed (two); that imposed an essential assumption on the numerical descriptions of the corpora data. In this study we consider two and three-dimensional embeddings; moreover, the methods developed can be readily extended to other dimensions without any principal limitations;
- the present study focuses specifically on the ability to determine the characteristic semantic structure of the corpora.

The hypothesis examined in the paper is that a combination of the methods of feature extraction and numerical description of text corpora developed in the NLP domain with methods of unsupervised learning and dimensionality reduction can be instrumental in obtaining a semantic model or intrinsic, natural structure of the corpora in an entirely unsupervised process that does not require prior knowledge about the semantic content of the corpus. If successful, such methods can produce informative geometrical models of distributions of data in the corpora in terms of the informative latent factors, or features computed by the methods in the process of unsupervised learning and dimensionality reduction.

The objectives of this work therefore were:

To test a range of methods of unsupervised feature learning and dimensionality reduction for effectiveness in the construction of informative embeddings of text corpora; to that end, we compared a number of common methods in unsupervised learning and dimensionality reduction, both linear and essentially non-linear, with the model corpora of texts of English newsgroups developed for the study. The second important contribution of this work lies in the development of the methods of unsupervised definition of characteristic semantic types in the corpora, based on methods of unsupervised clustering and learning of characteristic semantic types in the corpora entirely and exclusively from the data itself, without external prior knowledge of its semantic content.

Finally, in this work, small to medium-sized model corpora were used as a demonstration of the approaches in unsupervised semantic analysis. It is believed that term frequency features may not provide sufficiently detailed and accurate descriptions of very large realistic natural language corpora, in addition to resource intensity of working with such features in case of very large bodies of texts. For that task, more sophisticated linguistic models like BERT (Devlin et al. 2018) and more recent ones would be a more natural and effective direction. They were sufficient though for the objective of this work to demonstrate that characteristic semantic content of corpora can be determined with entirely unsupervised methods, without dependency or requirement for prior knowledge of the semantic content; and to test and identify methods of unsupervised machine learning and dimensionality reduction that can be effective in determination of the intrinsic semantic structure of the corpora short of massive investment of training and operational resources.

3. METHODOLOGY

3.1. Model corpora: English newsgroups

To demonstrate developed approach in unsupervised analysis of conceptual structure of corpora, model corpora composed of English texts were constructed on the basis of a dataset of English newsgroups, (20 newsgroups dataset) that contained approximately 20,000 texts, partitioned nearly evenly across 20 different newsgroups.

Two model corpora were constructed from the texts available in the dataset:

- News4, with texts in four topical newsgroups: atheism; Christianity; computer graphics; science–medical;
- News7, comprising the texts from seven newsgroups: atheism; Christianity; sport– hockey; science–medical; commercial–for sale; commercial–auto; politics–Middle East.

The composition of the resulting model English corpora, News4 and News7 are described in Table 1.

Table 1: Model English corpora, 20 newsgroups.

Corpus	Size (text)	Length (mean, tokens)	Team feature
News4	2257	306	35,482
News7	4016	310	50,676

It can be noted that the descriptions of the corpora in the framework of the term frequency (TF) features that were obtained were essentially sparse, meaning the ratio of the informative, non-zero values in the feature vector that corresponds to a text in the corpus to its total length (i.e., the size of the term feature set). For example, the NEWS4 corpus had an average sparsity of approximately 1%, whereas for NEWS7 it was 0.6%.

The labels, that represent the numerical encoding of the type of the newsgroup in the constructed model corpora were not used in the examined methods of unsupervised semantic analysis. However, they were used in the verification of the performance of the methods.

3.2. Unsupervised semantic analysis of model corpora

The process of unsupervised semantic analysis of corpora proposed and investigated in this work included these stages:

- Preprocessing of the original collections of texts represented in the study by the model corpora: processing of the texts with standard NLP methods such as filtering (stop-word removal), reduction to standard form (stemming) and so on;
- Vectorization: obtaining the description of the corpus in terms of numerical feature vectors; in this work, term-frequency methods of vectorization were used;
- Verifiably unsupervised: i.e., not dependent on explicit annotations, reduction of the dimension of the numerical representations of the corpora. Includes computation (extraction) of a small number of informative features relative to the initial dimensionality of the vectorized descriptions;
- Determination of the natural structure of the latent distributions of the data in the informative feature space with methods of unsupervised clustering among others, that can be associated with the underlying semantic differences, types and flavors in the original corpus.

3.3. Preprocessing and feature extraction

Preprocessing of the text data in the corpora was performed with standard methods of processing texts in NLP: tokenization; filtering (stop words removal); standard forming (stemming) as discussed earlier. In the resulting numerical representation or "vectorization", a text in the corpora was represented by a sparse vector of integer token counts.

Next, a standard NLP transformation of token count to token frequency features was applied. In the resulting set, a text *t* in the corpus was represented by a sparse vector *t v* of Term frequency - inverse document frequency (TF-IDF) features (Sparck Jones 1972). In this work, a ready vectorization package in the scikit-learn Python library was used (Scikit-learn Tfidf Vectorizer 2011).
The resulting sparse matrices *corpus_tf* can be considered as numerical representations of the input corpora in the framework of term frequency features, having the form: *corpus_tf* = (*text id, term frequency feature vector*). It was then used as input to the methods of unsupervised analysis of the semantic structure of the corpora. A description in pseudocode of the process of producing numerical description of the model corpora in the framework of term frequency features is provided below.

Tokenized numerical representation of the original (input) corpus, *corpus.data* *corpus_v = Tokenizer(stop_words, lemma).transform(corpus.data)*

Result: *corpus_v*: a sparse matrix of token count vectors in the input corpus; filtration and lemmatization performed.

Transform the token count matrix to token frequency matrix:

corpus_tf = TfidfTransformer.transform(corpus_v)

Result: a sparse matrix of token frequency vectors associated with texts in the input set.

3.4. Unsupervised dimensionality reduction

In following the objectives of this work, we set out to test a range of methods of unsupervised analysis of data distributions (that is, can be used without providing prior information, of any kind, about the distribution of data to be analyzed) in application to vectorized representation of small to medium corpus data that was described in the preceding section 3, and identify methods that can be effective in learning, that is, producing informative low-dimensional embeddings of the corpus data.

Most well-known methods of dimensionality reduction and informative factor extraction were used, including the following:

- Linear feature extraction and embedding: Principal Component Analysis, and more generally, Singular Value Decomposition
- Spectral Embedding
- Manifold learning: t-SN embedding
- Manifold learning: Uniform Manifold Approximation
- Generative learning: informative non-linear embedding with generative neural network model of self-supervised learning.

The selection of methods was guided by demonstrated effectiveness with other types of data and ready availability of packages or libraries. A brief description of the methods of unsupervised feature extraction (dimensionality reduction) is given in the Results section.

In evaluation of the methods, a common process was followed:

- Construction of low-dimensional representations (embeddings) of the numerical representations of model corpora;
- Visualization and evaluation of distributions of overall embedding of the corpora and of the classes;
- Evaluation of a possibility to identify characteristic semantic types (concepts) in the embeddings of the corpora with methods of unsupervised clustering that do not require prior knowledge about the semantic content of the corpora.

The pseudocode of the process of informative feature extraction and dimensionality reduction with vectorized descriptions of the model corpora:

Initialization of the FL,DR method with the dimension of resulting informative embedding:

reducer = Method(num_components = components, parameters)

produce informative embedding of the specified latent dimension:

embedding = reducer.transform(corpus_tf)

Input is the numerical term frequency description of the corpus obtained in the previous step. Result: a numerical low-dimensional embedding of the vectorized corpus *corpus_tff* in the informative feature space.

3.5. Clustering and Inference of the Semantic Structure

The final step in the analysis is to apply methods of clustering in the informative embedding space of a strongly reduced dimension. Again, an essential constraint in this step is that the methods cannot rely on known annotations that may not be available, or any other essential information or hints about the content of the corpora, even the approximate number of distinct semantical types. To satisfy this essential constraint, known methods of density clustering, a broad family of methods that includes DbScan, Orbit, MeanShift and many others (Campello et al 2019). In this work, method MeanShift was used (Fukunaga and Hostetler 1975).

The pseudocode of unsupervised clustering in the informative embedding of the corpora:

```
semantic_types = Clustering.transform(embedding)
```

where Clustering: the method of unsupervised clustering applied to the informative embedding of the corpus obtained in the previous step.

Result: a structure of clusters in the informative embedding space that can be associated with distinct semantic types in the original corpus.

It needs to be noted that the methods of processing corpora described here are not in any essential way limited to English texts. Given the fact that a number of packages have been developed with support for a large set of languages, including preprocessing functions like filtering and stemming used here, the procedure described above can be applied to corpora in languages different from English to obtain numeric representations of the texts. Then, with numerical description (vectorization) of the corpora thus obtained, methods of unsupervised feature learning and dimensionality reduction can be applied with any language of the original corpus for which the utilities of tokenization and term frequency have been implemented.

4. RESULTS

In this section we describe the results in unsupervised modeling of the model corpora following the process described in the preceding sections. In the evaluation of distributions of embedded corpora produced by unsupervised methods, evaluation of characteristics of distributions, including the separation of different semantic groups of texts was done by visual observation. To that end, after embeddings were produced by an entirely unsupervised process as described above, scattering diagrams of embedded corpora were produced with an indication of the labelled class (or external semantic type) of the texts in the corpus. Thus, not only the overall embedding of the corpora but also those of the specific classes / semantic types could be evaluated from the visualizations of latent embeddings of the model corpora.

4.1. Linear embeddings

Truncated Singular Value Decomposition (SVD) is a linear decomposition analysis method related to Principal Component Analysis that can be used with sparse data. This is essential in the analysis of the corpora in the study, because, as mentioned earlier, numeric representations of the corpora in the framework of term frequency features have the form of strongly sparse matrices.

A readily available implementation of truncated SVD in the Scikit-learn Python library was used to obtain linear embeddings of the model corpora (Scikit-learn: TruncatedSVD 2011). It supports the use of sparse data and thus can be used directly with numerical vector representations of the corpora, corpus_v obtained as described in Section 3.3.

The results of low-dimensional embeddings obtained with linear singular value decomposition of the model corpora sparse matrices are shown in Fig 1: NEWS4 corpus, two and three-dimensional embeddings (left); NEWS7, two and three-dimensional embeddings (right).

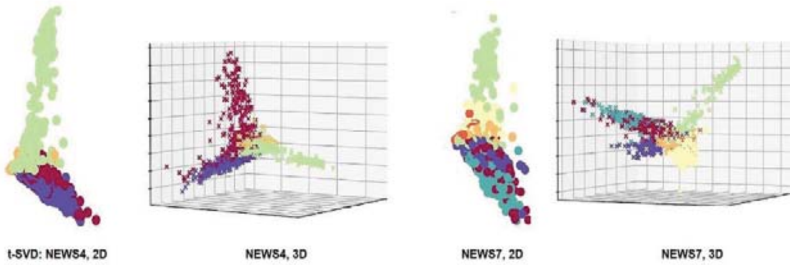

Figure 1. t-SVD embeddings, NEWS4 and NEWS7 corpora.

As can be inferred from visualizations of distributions of the classes in the diagrams above, truncated SVD can achieve separation of some semantic classes in the model corpora, however in quite a limited way. The reasons for the lower effectiveness of linear methods of informative dimensionality reduction with the text data are discussed in the conclusion of this section.

4.2. Spectral embedding

Spectral Embedding is a non-linear dimensionality reduction method based on the calculation of Laplacian eigenmaps produced from the affinity graph of the input data (Belkin and Niyogi 2003). It can be used directly with the sparse matrix descriptions of text corpora.

In the evaluation, an implementation of spectral embedding in the Scikit-learn manifold learning package was used (Scikit-learn: Spectral Embedding 2011).

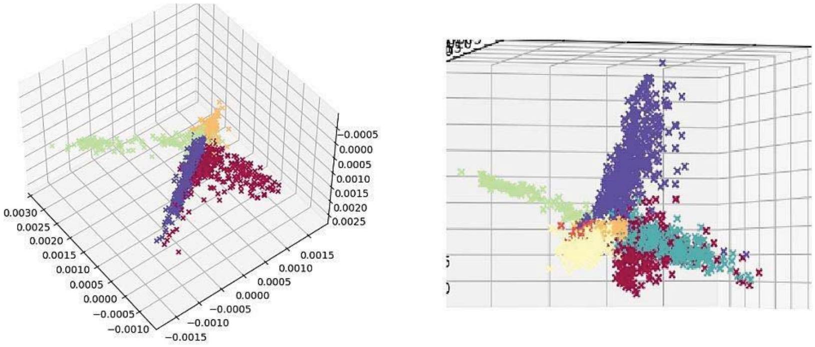

Figure 2. Spectral embedding, NEWS4 (left), NEWS7 (right) corpora.

As can be observed from the resulting visualization (Figure 2), spectral embedding can produce latent embeddings with a good separation of most semantic classes in the model corpora.

4.3. Manifold learning: t-SN Embedding

TSNE (for t-distributed Stochastic Neighbor Embedding) is a non-linear manifold learning technique based on minimization of the Kullback-Leibler divergence between the joint probabilities of the distribution in the low-dimensional embedding and the high-dimensional data (van der Maarten and Hinton 2008). In this work, an implementation of the method in the Scikit-learn manifold learning package was used (Scikit-learn: t-SNE 2011).

The method has been used with text data before (Yellowbrick t-SNE Corpus Visualization 2019) and produced well-structured embeddings of the model corpora in the study, as can be observed in the distribution diagrams Fig 3. However, an essential limitation of this family of methods has to be noted: they do not support embeddings with the latent dimension above four.

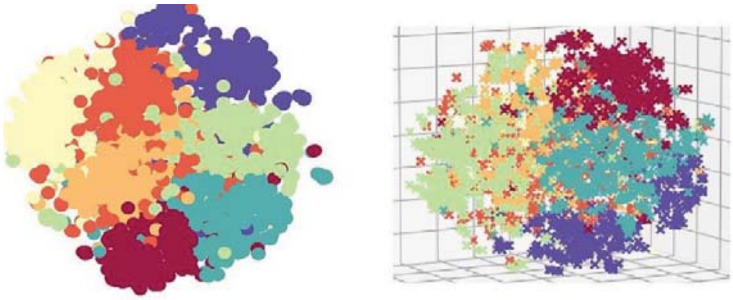

Figure 3. t-SN embedding, NEWS7 corpus 2D (left), 3D (right).

Interesting to note in this case that a lower dimensional embedding Fig 3 (2D, left side of the diagram) produced a more informative distribution of the characteristic semantic types in the model corpora than in higher latent dimensions. Granted, only a limited amount of time and resources was available to investigate the effect of the choice of parameters of the methods, including initialization specifically in the case of this method, that can have a strong influence on the produced embedding.

4.4. Manifold learning: Uniform Manifold Approximation

UMAP (Uniform Manifold Approximation and Projection) is an essentially nonlinear manifold learning technique that can be used for effective dimensionality reduction of high-dimensional sparse data (McInnes, Healy and Melville 2018). The results of its application with the model corpora are shown in Figure 4.

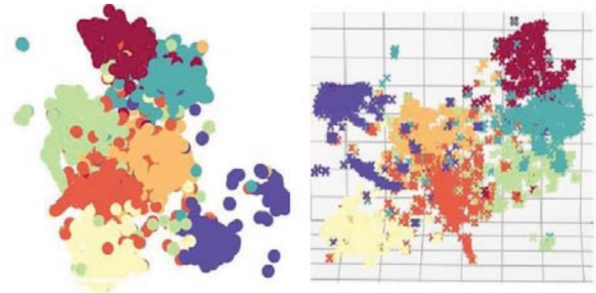

Figure 4. UMAP embedding, NEWS7 corpus; 2D (left), 3D (right).

From the visualizations in Figure 4 it can be concluded that this method can produce a good separation of semantic types in the model corpora in both two and three-dimensional embeddings.

4.5. Generative learning: self-supervised ANN

Generative models of self-supervised learning that do not require labeled sets can be instrumental in producing informative non-linear representations of complex real-world data. In this work we attempted to obtain such representations with simple neural architectures of the type of deep autoencoder (Bengio 2009), (Welling and Kingma 2019) with several deep layers and a latent representation (embedding) layer. To use the available neural network modeling packages with essentially sparse input data produced by the vectorization methods as described, a preprocessing stage was added that performed a preliminary dimensionality reduction with dimensionality reduction, to obtain dense datasets of features that could be used in training of neural network models of self-supervised generative learning. Low-dimensional embeddings of the input data were produced by activations of the neurons in the central (”encoding”) layer of the models after completion of the training phase, thus the dimension of the latent embedding was determined by the architecture of the model. The models of the type described in (Dolgikh 2021) were trained with stochastic gradient for minimization of the distance between input batches and generations produced by the model.

While it was beyond the scope of this study to investigate more complex neural network architectures as well as parameter optimization and tuning, in some of the experiments they produced clearly separated embeddings of the characteristic semantic types in the corpora, as shown in Figure 5.

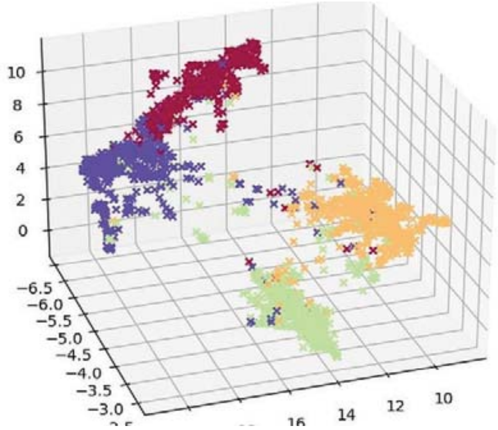

Figure 5. Generative ANN 3D latent embedding, NEWS4 corpus.

4.6. Summary: unsupervised analysis of the semantic structure of corpora

A summary of the results presented in the preceding sections is given in Table 2. We show the overall evaluation of separation of semantic types and usability of the tested method of unsupervised feature learning (dimensionality reduction) for unsupervised analysis of text corpora in the absence of prior information about its semantic content.

Table 2: Unsupervised analysis of model corpora.

Method	Separation, semantic types	Usability
Linear	Low	Not very effective
Spectral	Some	Possible use
TSNE	Most, 2D	Usable
Umap	High	Usable
Generative ANN	Some	Possible use

An initial observation that can be made based on these results (see further in the Discussion section, is that numerical descriptions of text corpora in the framework of term frequency features are essentially non-linear. Whereas most of the non-linear methods of dimensionality reduction were successful in resolving the underlying low-dimensional structure in sparse numerical representations of model corpora to some degree of success, and in certain cases, confidently, the linear methods appeared to be much less effective with text data. This conclusion is of small surprise, given the essential characteristics of the approach chosen for vectorization of corpora, very high dimensionality and sparsity. An effective description of such data in terms of a small number of linearly derived features appears to be unlikely.

In conclusion, it can be noted that though in the presented results we used the latent dimension of the embedding of two or three, there are no principal limitations in the demonstrated approach for specific values of latent dimensions, as long as the methods of dimensionality reduction do not constrain it. Most of the methods considered above do allow the use of higher latent dimensions.

4.7. Clustering and inference of semantic structure

While the results in the preceding sections demonstrated a potential for unsupervised analysis of corpora with methods of unsupervised machine learning and dimensionality reduction, evaluation of the performance of the methods indeed required labeled data.

The next step to avoid this dependency would be to apply methods of unsupervised clustering directly in the informative embeddings of the corpora obtained with the methods described earlier. Such methods indeed exist and have been used widely in the analysis of distributions of general data, where prior information about distributions is not available. In this work, we used MeanShift clustering as it was proven effective in applications in the informative embeddings of reduced dimensionality of complex realistic data (Dolgikh, 2021).

In the plots below, methods of unsupervised clustering were applied to embeddings of the model corpora to identify characteristic semantic types in the corpora without prior knowledge about their semantic content. The positions of red dots identify the centers of density clusters resolved by a density clustering method that has been applied to a low-dimensional embedding of the corpora.

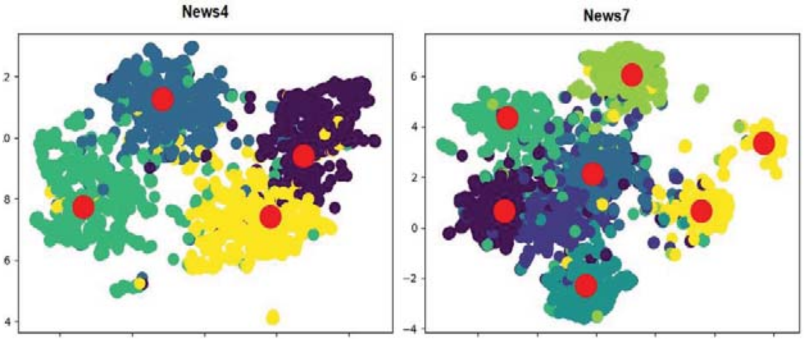

Figure 6. Unsupervised semantic type analysis, UMAP embedding / MeanShift clustering, NEWS4 and NEWS7 corpora**.**

As can be observed in the plots of the distributions of externally known semantic types (that is, the newsgroups) with the characteristic semantic types identified by the proposed unsupervised approach (centers of clusters indicated by red dots), density clustering in combination with unsupervised dimensionality reduction demonstrated good association with distributions of distinct semantic types of texts in the model corpora. Similar results were obtained with embeddings produced by other methods of unsupervised production of informative embeddings of corpora, including t-SN and ANN.

An interesting example of unsupervised semantic analysis can be seen in the plot of the NEWS7 corpus (Figure 6, right). The latent region of the distribution of the newsgroup *”politics - Middle East”* (6) is split between two distinct regions, in a proportion of approximately 3 : 1. It can be conjectured that this separation is not superfluous, and the posts in that newsgroup may have contained at least two essentially different semantic flavours of texts. An analysis of the input texts indeed confirmed this hypothesis: whereas the messages associated with the larger latent region dealt with Arab-Israeli relationships, those in the other subtype were associated with more general issues of history, traditions and religion in the region. Thus, the application of unsupervised semantic analysis in the informative embeddings of corpora revealed an underlying semantic structure information about that was not available in the external annotations.

Once again, this conclusion can be obtained from an analysis of informative low-dimensional embedding of the corpus, without any prior knowledge about its semantic content.

Following the results presented in this section it can be concluded that combining methods of unsupervised production of informative embeddings of corpora data with methods of unsupervised density clustering can produce informative descriptions of the semantic composition of the corpus in the problems and applications without prior information about its semantic content.

5. DISCUSSION AND CONCLUSION

An analysis of the semantic content of corpora produced with means of automatic extraction can be challenging due to the need for manual evaluation and labeling by a human. Methods described in this work allow us to investigate and identify characteristic semantic structure of small to medium-sized corpora with any semantic content with automated and unsupervised methods that do not require significant prior knowledge about the composition and semantic content of the corpora. With a range of methods of unsupervised learning and dimensionality reduction reviewed and tested, a subset of those that can be applied successfully to the analysis of the structure of text corpora has been identified. It is believed that the approach demonstrated here can be an instrumental addition in the toolbox of the methods of analysis of corpora.

An essential finding of this work is that sparse numerical descriptions of realistic natural language corpora in the framework of term frequency features of even relatively simple model corpora studied in this work were essentially, non-linear, in the sense of underlying low-dimensional informative parameter manifolds. Linear methods of dimensionality reductions were noticeably less successful in the separation of semantic types in the corpora, whereas several of the non-linear methods have demonstrated the ability to separate them into distinct regions in the embedding plane. These findings are consistent with the earlier results (McInnes, Healy and Melville 2018), (Choudhari et al 2022) and others.

It can be concluded that in the demonstrated instances, the methods tested in this work were successful in creating a conceptual model of the corpora with geometrically separate regions of natural semantic types in the embedding plane, and an ability to identify the semantic type of a sample, under certain conditions, without any prior information about the content of the corpus, i.e., in such cases, achieve effective zero-shot learning of natural semantic types.

With the examples of informative distributions of model corpora demonstrated here, a natural question can be raised: what are the semantics of the latent parameters identified by the methods, i.e. the meaning of the coordinates in the latent embedding space? While such a study has not been attempted here, the answer can be glimpsed from other results in representation learning (Higgins et al 2012). It was shown, though with the data of a different type, that latent coordinates did not have consistent global semantics, rather having a local, manifold-like structure. With images, the same latent coordinate could translate to the size of an object in one region of the embedding plane and to its type or contrast of the image, in another. It can be conjectured that similar behavior would be found with the text data, with latent coordinates describing different characteristics of texts in different regions of the embedding plane. Still, in the authors view, the semantics of informative latent factors of text data can be an interesting and worthy topic of future studies.

In conclusion, it can be noted that the methods developed and demonstrated in this work can be readily extended to multilingual, parallel corpora and other cases. The only dependency is the support of standard NLP procedures of text data preprocessing and features that are available in many languages with a number of NLP products and packages.

6. REFERENCES

Ahmadnia, B., Haffari, G., Serrano, J., 2019, *Round-trip training approach for bilingually low-resource statistical machine translation systems*, International Journal of Artificial Intelligence, **17(1)**, 167-185.

Belkin, M., Niyogi, P., 2003, *Laplacian Eigenmaps for dimensionality reduction and data representation*, Neural Computation, **15 (3)**, 1373-1396.

Bengio, Y., 2009, *Learning deep architectures for AI*, Foundations and Trends in Machine Learning, **2 (1)**, 1-127.

Boleda, G., 2020, *Distributional semantics and linguistic theory*, Annual Review of Linguistics, **6**, 213-234.

Brown, P.F., Cocke, J., Della Pietra, S.A., et al., 1988, *A statistical approach to language translation*, Association for Computer Linguistics, **1**, 71-76.

Campello, R.J.G.B., Kröger, P., Sander, J., et al., 2019, *Density-based clustering*, WIREs Data Mining and Knowledge Discovery, **10 (2)**, 1343.

Chandar, A. P. S., Lauly, S., et al., 2014, *An autoencoder approach to learning bilingual word representations*, In 27th International Conference on Neural Information Processing Systems (NIPS’14), Canada: Montreal, 1853-1861.

Chodhari, R., Doboli, S., and Minai, A.A., 2022, *A Comparative study of methods for visualizable semantic embedding of small text corpora*, In 2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), 1-8.

Cunningham, J.P., and Ghahramani, Z., 2015, *Linear dimensionality reduction: survey, insights, and generalizations*, JMLR, **16**, 2859-2900.

Deerwester, S., Dumais, G.W., Furnas, G.W., Landauer, T.K., and Harshman, R., 1990, *Indexing by latent semantic analysis*, Journal of the American Society for Information Science, **41**, 391-407.

Devlin, J., Chang, M.-W., et al, 2018, *BERT: pre-training of deep bidirectional transformers for language understanding*, ACL Anthology, **19**, 1423.

Dolgikh, S., 2021, *Categorization in unsupervised generative self-learning systems.* International Journal of Modern Education & Computer Science*,* **13 (3)** 68–78.

Dolgikh, S., 2024, *Biologically feasible generative neural architectures and evolutionary learning in simple visual environments.* International Journal of Artificial Intelligence*,* **22 (1)** 1–19.

Foltz, P.W., and Dumais, G.W., 1992, *An analysis of information filtering methods*, Communications of the ACM, **35 (12)**, 51-60.

Fukunaga, K., and Hostetler, L.D., 1975, *The estimation of the gradient of a density function, with applications in pattern recognition*, IEEE Transactions on Information Theory, **21 (1),** 32-40.

Gondara, L., 2016, *Medical image denoising using convolutional denoising autoencoders*, In IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 241-246.

Higgins, I., Matthey, L., Leawitt, B., Glorot, X., et al., 2016, *Early visual concept learning with unsupervised deep learning*, arXiv, 1606.05579.

Hinton, G., and Roweis, S., 2002, *Stochastic neighbor embedding*, In NIPS'02 Advances in Neural Information Processing Systems, **15**.

Hofmann, T., 2001, *Unsupervised learning by probabilistic Latent Semantic Analysis*, Machine Learning, **42**, 177-196.

Koehn, P., Hoang, H., Birch, A., et al., 2007, *Moses: Open Source toolkit for Statistical Machine Translation*, ACL 2007, Prague, Czech Republic.

Le, Q.V., Ranzato, M.A., Monga, R., et al., 2012, *Building high-level features using large scale unsupervised learning*, In ICML'12 29th International Conference on Machine Learning, 507-514.

Lee, J.A., and Verleysen, M., 2007, *Nonlinear Dimensionality Reduction*, Springer, 2007.

Lenci, A., 2018, *Distributional models of word meaning*, Annual Review of Linguistics, **4**, 151-171.

Manning, C.D., Raghavan, P., and Schütze, H., 2008, *Introduction to Information Retrieval*, Cambridge University Press, Cambridge.

McInnes, L., Healy, J., and Melville, J., 2018, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, arXiv, 1802.03426.

Moreno-Ortiz, A., and García-Gámez, M., 2023, *Strategies for the analysis of large social media corpora: sampling and keyword extraction methods*, Corpus Pragmatics, 2023.

Och, F.J., and Ney, H., 2003, *A systematic comparison of various statistical alignment models*, Computational Linguistics, **29**, 19-51.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al., 2011, *Scikit-learn: Machine Learning in Python*, JMLR, **12**, 2825-2830.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al., 2011, *Scikit-learn Tfidf vectorizer*, Online, https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html.

Sparck Jones, K., 1972, *A statistical interpretation of term specificity and its application in retrieval*, Journal of Documentation, **28 (1),** 11-21.

Van der Maaten, L., and Hinton, G., 2008, *Visualizing data using t-SNE*, JMLR, 9 (86), 2579-2605.
Welling, M., and Kingma, D., 2019, *An introduction to variational autoencoders*, Foundations and Trends in Machine Learning, **12 (4)**, 307-392.

Yellowbrick developers, 2019, *t-SNE Corpus Visualization*, Online, https://www.scikit-yb.org/en/latest/api/text/tsne.html.

20 Newsgroups dataset, Online, http://qwone.com/jason/20Newsgroups/.