Probabilistic Topic Modeling of the Russian Text Corpus on Musicology*

Olga Mitrofanova[1]

[1] St.-Petersburg State University, Faculty of Philology,
Department of Mathematical Linguistics,
Universitetskaya emb., St.-Petersburg, 199034, Russia
oa-mitrofanova@yandex.ru

Abstract. The paper describes the results of experiments on the development of a statistical model of the Russian text corpus on musicology. We construct a topic model which is based on Latent Dirichlet Allocation and process corpus data with the help of GenSim statistical toolkit. Results achieved in course of experiments allow to distinguish general and special topics which describe conceptual structure of the corpus in question and to analyse paradigmatic and syntagmatic relations between lemmata within topics.

Keywords: musicology, corpus linguistics, text corpora, probabilistic topic modeling

1 Introduction

It has been proved that language and music reveal similar structure which admits reasonable treatment in symbolic and statistical models: phenomena occurring in natural language and in musical texts can be properly explained in terms of statistically enriched grammatical formalisms [Bod 2002; Scha, Bod 1993].

Evidently there is a considerable overlap between natural language and music which appears in the study of prosody, tonal languages, verse in vocal pieces, etc. This overlap also manifests itself in musicology.

Musicology is a peculiar field of knowledge which attracts much attention of linguists as it has regular and well-structured terminological system [Koryhalova 2006], although conceptual core of musicology is extremely complex, it springs up from heterogeneous sources and is extremely rich in emotionally coloured terms [Mitrofanova 2002a,b]. At the same time it has seldom been an object of quantitative

* The research discussed in the paper is supported by the grant of St.-Petersburg State University № 30.38.305.2014 «Quantitative linguistic parameters for defining stylistic characteristics and subject area of texts».

research. The present study fills a gap in our knowledge about statistical features of texts on musicology.

We support corpus-based approach in linguistic research: it has generated an original philosophical trend and developed a novel methodology which facilitates the transition from frequency data and distributional analysis to explanation of discourse phenomena, and builds a bridge between text corpora and cultures [Gries 2009]. Hereinafter we treat corpus as a digital collection of natural language data, large in size, unified, well-structured, linguistically annotated and intended for various experimental procedures [Zakharov 2005].

Thus, the aim of our research is to construct the Russian text corpus on musicology and to develop a probabilistic topic model for the given corpus, therefore to describe distribution of words over topics and topics over texts in our text collection. To meet the need we choose a version of topic models which is based on Latent Dirichlet Allocation and process corpus data with the help of GenSim statistical toolkit.

The structure of the paper is as follows:

section 1 gives an outline of our research,

section 2 discusses the notion of a topic model,

section 3 is devoted to the software used in the study,

section 4 describes the Russian text corpus on musicology,

section 5 discusses experimental settings and results achieved in course of experiments,

section 6 provides a brief conclusion,

section 7 deals with further development of research.

2 Topic modeling of text corpora

Topic modeling is a research procedure which allows to expose implicit factors constituting conceptual structure of text corpora. A topic model reflects the transition of a set of texts in a given corpus and a set of words in texts to a set of topics describing the content of a corpus. In fact, the process of topic modeling results in extracting a set of topics from a corpus, i.e. a set of clusters containing words revealing similarity in meaning which is based on distributional similarity within a corpus.

Each word or document within a topic model is related to a number of topics with certain probability. A topic model provides a description of a text corpus in terms of a family of probability distributions over a set of topics. Probabilistic topic model is constituted by

(1) $p(w|d) = \Sigma p(t|d)\ p(w|t)$, where $p(w|d)$ is a certain frequency of occurrence of a word w in a text d,

(2) $p(w|t)$ is an uncertain probability of occurrence of a word w in a topic t,

(3) $p(t|d)$ is an uncertain probability of occurrence of a topic t in a text d.

In order to construct a topic model for a text corpus D it is necessary to find a set of topics T, a distribution $p(w|t)$ for all the topics and a distribution $p(t|d)$ for all the texts. The desired distributions provide a condensed description of text corpora applicable in Natural Language Processing tasks (especially in semantic compression procedures like automatic text indexing and categorization, word and document clustering and classification, extraction of distributionally similar words, etc.).

There is a great variety of topic models, the basic types being algebraic models (Vector Space Model VSM, Latent Semantic Analysis LSA, etc.) and probabilistic models (Probabilistic Latent Semantic Analysis pLSA, Latent Dirichlet Allocation LDA, etc.). LDA model proved to be one of the most productive in processing text collections of both large and moderate size.

Results on topic models for English text corpora are widely discussed in numerous publications, cf. [Blei et al. 2003; Daud et al. 2010; TMB], etc. In most cases the procedure is performed for scientific and news texts. The paper [Rhody 2012] describes a remarkable study in the field of topic modeling of poetic texts. At the same time, Russian corpora are seldom involved in such research, with few exceptions: positive results have been described in [Bodrunova et al. 2013; Vorontsov, Potapenko 2014; Mitrofanova 2015].

3 Probabilistic topic modeling toolkit

There are certain computer implementations of various topic models including LDA, e.g. MALLET [http://mallet.cs.umass.edu/topics.php], Stanford Topic Modeling Tool [http://nlp.stanford.edu/software/tmt/tmt-0.4/], etc. A list of available software is given in [http://www.cs.columbia.edu/~blei/topicmodeling.html].

In our study the task of topic modeling is fulfilled with the help of a set of Python libraries GenSim [http://radimrehurek.com/gensim/] developed by R. Řehůřek. We have chosen GenSim as includes a powerful statistical module for developing probabilistic topic models for text corpora. GenSim is also known by its flexibility and operability. We composed a script activating GenSim components and performing automatic text processing based on LDA.

The script provides the following procedures:

(1) extraction of a dictionary from the input corpus (plain text format, UTF8 encoding);

(2) elimination of stop-words and low frequency words;

(3) transformation of a dictionary into a matrix;

(4) construction of LDA model with changeable parameters, the number of iterations, topics, and topic size being assigned by a user;

(5) output of results: topics extracted from a corpus are represented as lists of lemmata with weights and values of perplexity.

4 The Russian text corpus on musicology

The subject area of music seems to be of particular interest in corpus studies, cf. papers on the Corpus of Russian romances [Martynenko 2006, 2013]. Our purpose is to construct a resource which gives a general overview of various types of texts dealing with music. In order to perform experiments on probabilistic topic modeling we have developed the Russian text corpus on musicology covering such areas as theory of music, history of music, performing art, musical instruments, biographies of composers and musicians, etc. Our corpus includes texts of encyclopaedias and reference books as well as texts of monographs and collected works on certain subjects.

Pre-processing of texts in our corpus includes the following stages:

(1) elimination of non-text elements (tables, images, diagrams, hyperlinks, etc.);

(2) construction of stop-list from the Corpus Dictionary of Multi-Word Lexical Units extracted from the Russian National Corpus (RNC) [http://www.ruscorpora.ru/obgrams.html] and frequency lists including lemmata of linkwords, pronouns, numerals, abbreviations from the Frequency Dictionary of Contemporary Russian (RNC) [http://dict.ruslang.ru/freq.php];

(3) lemmatization and automatic morphological disambiguation with the help of Yandex morphological parser mystem 3.0 [https://tech.yandex.ru/mystem/];

(4) splitting texts into documents in accordance with their initial logical structure (chapters, sections, etc.).

Pilot experiments were performed on a subcorpus of over 300 000 tokens in size which included encyclopaedic texts [Samin 2002, 2006]. To prove representativeness of the subcorpus we performed its statistical analysis using the functions embedded into AntConc corpus manager [http://laurenceanthony.net/software/antconc/]. We examined the top 100 frequent lemmata, among which we found words *концерт (concerto), музыкант (musician), музыка (music), играть (play* [v]*), игра (play* [n]*), музыкальный (musical), писать (write), скрипач (violinist), оркестр (orchestra), искусство (art), исполнение (performance), пианист (pianist), выступать (perform), скрипка (violin), композитор (composer), произведение (composition), консерватория (concervatoire), джаз (jazz), концертный (concert* [adj]*), выступление (performance), артист (artist), стиль (style), инструмент (instrument), звук (sound),* etc. which also occur in the Russian Associative Dictionary [RAS 2002]. Lemmata of the upper zone of the frequency list, as will be shown further, constitute the principal part of the topics generated by the LDA model.

5 Design of experiments and analysis of results

Series of tests on probabilistic topic modeling of subcorpus from our collection of Russian texts on musicology were performed with the following parameters:

(1) number of iterations –10;

(2) number of topics – 10, 20, 50, 100;

(3) topic size – 10 lemmata.

Below we list some fragments of output which illustrate our results. Lemmata constituting topics are arranged in accordance with values of the association measure which is automatically calculated but removed from the output.

The model produces topics of general character which partly overlap: such topics include widespread words dealing with music, e.g.:

опера (opera), музыка (music), композитор (composer), произведение (composition), время (time), жизнь (life), музыкальный (musical), новый (new), театр (thearte)…;

музыка (music), композитор (composer), опера (opera), жизнь (life), произведение (composition), симфония (symphony), песня (song), время (time), новый (new)…;

концерт (concerto), музыка (music), музыкант (musician), время (time), музыкальный (musical), игра (play), писать (write), большой (large), искусство (art)….

Alongside with general topics we have managed to form topics of more definite content which refer to particular composers, performers, styles and genres of music, e.g.:

«*Шопен (Chopin)*»: *Шопен (Chopin), польский (polish), Варшава (Warsaw), Польша (Poland), друг (friend), Жорж (George), Санд (Sand), родина (native land), предчувствие (presentiment), композитор (composer)…*;

«*Вена в музыке композиторов (Vienna in the life of composers)*»: *Людвиг (Ludwig), Бетховен (Beethoven), Моцарт (Mozart), Гайдн (Haydn), Вена (Vienna), Барток (Bartok), жизнь (life), композитор (composer)*…;

«*Скрипачи (violinists)*»: *скрипач (violinist), игра (play), музыкант (musician), Крейслер (Kreisler), Байо (Baillot), скрипка (violin), Тартини (Tartini), концерт (concerto), Ромберг (Romberg)*…,

«*Джаз (Jazz)*»: *группа (group), альбом (album), песня (song), джаз (jazz), Иоахим (Joachim), пластинка (disc), компания (company), записывать (record), Джон (John)…*, etc.

Probably lemma *Ромберг (Romberg)* within the topic «*Скрипачи (violinists)*» implies the family of musicians famous for its cellists, composers, singers, pianists,

clarinettists, and violinists. It seems that the topic «*Джаз (Jazz)*» refers to the jazz musician John Coltrane and probably to vinyl discs with jazz records of Joachim Kühn.

We also found out several topics of mixed character, e.g.

музыка (music), музыкант (musician), играть (play), гитарист (guitarist), фестиваль (festival), концерт (concert), говорить (talk), Виктор (Viktor), Цой (Tsoi)…, etc.

The given topic combines general terms and the mention of the Russian rock-musician and songwriter Viktor Tsoi.

As regards the inner structure of the topics generated by LDA model, it is possible to distinguish certain paradigmatic and syntagmatic relations between words within topics. Most of the relations extracted from the topics can be described in terms of lexical functions in «Sense<=>Text» linguistic model (e.g. Syn, Gener, Der, $Oper_{1,2}$, S_i, S_c, Mult, etc.) [Melchuk 1999].

A. Paradigmatic relations, e.g.

– synonymy (Syn): «*произведение – сочинение (composition)*»,

– hyponymy (Gener): «*музыкант (musician) – композитор (composer), пианист (pianist), скрипач (violinist), гитарист (guitarist)*, …», «*произведение (composition) – опера (opera), симфония (symphony), песня (song),* …», etc.,

– meronymy (partitive relations): «*произведение (composition) – фраза (phrase), такт (bar)*», etc.

– derivational relations (Der) *музыка (music) – музыкант (musician), музыкальный (musical); Польша (Poland) – польский (polish); опера (opera* [n]*) – оперный (opera* [adj]*), оперетта (operette); скрипка (violin* [n]*) – скрипач (violinist), скрипичный (violin* [adj]*); гитара (guitar) – гитарист (guitarist); выступать (play on the stage) – выступление (performance); играть (play* [v]*) – игра (play* [n]*); джаз (jazz* [n]*) – джазовый (jazz* [adj]*)*, etc.,

Further analysis of paradigmatic relations within topics allows us to extract lexical semantic-groups like «*Музыкальные формы и жанры (musical forms and genres)*»*: симфония (symphony), соната (sonata), концерт (concerto), опера (opera), оперетта (operetta), оратория (oratorio), песня (song), пастораль (pastorale)…,* etc.

B. Syntagmatic relations, e.g.

– verb-object relations ($Oper_{1,2}$) : *музыка (music) – писать (write)*; *пластинка (disc) – записывать (record) ;*

– noun-modifier relations (S_i, S_c): *играть на скрипке (play the violin)*; *играть в оркестре (play in an orchestra)*;

– noun-attribute relations: *искусство игры*; *джазовый музыкант (jazz musician)*; *джазовый оркестр (jazz band);*

– item-set relations (Mult): *музыкант (musician): оркестр (orchestra), консерватория (concervatoire), капелла (choir)*…

Thorough treatment of syntagmatic relations between separate lemmata included in topics provides sets of constructions, e.g.

– full names: e.g. *Людвиг ван Бетховен (Ludwig van Beethoven)*; *Жорж Санд (George Sand);*

– appositive constructions, e.g. *композитор (composer) X (X = Бах (Bach), Бетховен (Beethoven), Гайдн (Haydn), Моцарт (Mozart), Мендельсон (Mendelssohn), Гуно (Gounod), Дворжак (Dvořák), Брамс (Brahms), Шопен (Chopin), Гендель (Handel), Сен-Санс (Saint-Saëns), Глазунов (Glazunov), Рахманинов (Rachmaninov)…); музыкальный критик и композитор Серов (musical critic and composer Serov); пианист Оборин (pianist Oborin); Концерт Виотти (Viotti Concerto)*; *опера Россини (Rossini's opera); опера Вагнера (Vagner's opera); Концерт Венявского для скрипки с оркестром (Wieniawski Concerto for Violin and Orchestra); опера "Царь Давид" (opera "King David"); французский композитор Гуно (the French composer Gounod*); etc.

6 Conclusion

Thus, we performed linguistic analysis of the data generated by LDA model:

1) we described three types of topics as regards their content: general topics, special topics as well as mixed topics, and

2) we analysed the inner structure of topics in terms of paradigmatic and syntagmatic relations between lexical items.

Evidence obtained in course of our experiments proves consistency of the statistical model and compliance of linguistic results generated by the model with common knowledge on musical terminology.

7 Further development of research

We hope to continue our research in the following directions:

1) enlargement of the Russian text corpus on musicology: addition of texts on musical criticism as well as educational texts (e.g. children's literature on music), development of parallel subcorpus of comparable texts;

2) refinement of pre-processing procedure and purification of morphological analysis (lemmatization in particular);

3) improvement of the topic modeling toolkit, addition of several topic models besides LDA (LSA and pLSA), and investigating optimal parameters for topic modeling;

4) application of results achieved in course of topic modeling in further studies of musical texts (musical terminology extraction, construction identification, ontology development);

5) comparison of topical structure of musicology corpus with other specialized corpora.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I. (2003): Latent Dirichlet Allocation. In: Journal of Machine Learning Research. Vol. 3 (4–5). January 2003. URL: http://jmlr.csail.mit.edu/papers/volume3/blei03a/blei03a.pdf
2. Bod, R. (2002): A Unified Model of Structural Organization in Language and Music. In: Journal of Artificial Intelligence Research. Vol. 17. Pp. 289–308.
3. Bodrunova, S., Koltsov, S., Koltsova, O., Nikolenko, S.I., Shimorina, A. (2013): Interval Semi-Supervised LDA: Classifying Needles in a Haystack. In: MICAI–2013. URL: http://www.hse.ru/data/2013/10/03/1277898420/micai2013-182-final-easychair.pdf
4. Daud, A., Li, J., Zhou, L., Muhammad, F. (2010): Knowledge Discovery through Directed Probabilistic Topic Models: a Survey. In: Proceedings of Frontiers of Computer Science in China. Vol. 4(2). URL: http://www.researchgate.net/publication/215904200_Knowledge_Discovery_through_Directed_Probabilistic_Topic_Models._a_Survey
5. Gries, St. (2009): What is Corpus Linguistics. In: Language and Linguistics Compass. Vol. 3. Pp. 1–17.
6. Koryhalova, N.P. (2006): Muzykal'no-ispolnitel'skije Terminy. [Terminology of Musical Performance]. St.-Petersburg, 2006.
7. Martynenko, G.Ja. (2006): Semantika Korpusa Russkogo Romansa. [Semantics of the Russian Romance Corpus]. In: Trudy Mezhdunarodnoj Konferencii «Korpusnaja Lingvistika – 2006». [Proceedings of the International Conference «Corpus Linguistics – 2006»]. St.-Petersburg.
8. Martynenko, G.Ja. (2013): Korpus Russkogo Romansa kak Osnova Issledovanija Verbal'no-muzykal'nyh Tekstov. [The Corpus of Russian Romances for Studying Poetry and Music]. In: Trudy Mezhdunarodnoj Konferencii «Korpusnaja Lingvistika – 2013». [Proceedings of the International Conference «Corpus Linguistics – 2013»]. St.-Petersburg.

9. Melchuk, I.A. (1999): Opyt Teorii Lingvisticheskih Modelej «Smysl<=>Tekst» [«Experience of the Theory of Linguistic Models «Sense<=>Text»»]. Moscow.
10. Mitrofanova, O.A. (2015): Verojatnostnoje Modelirovanije Tematiki Russkojazychnyh Korpusov Tekstov s Ispol'zovanijem Kompjuternogo Instrumenta GenSim. [Probabilistic Topic Modelling of the Russian Text Corpora by means of GenSim Toolkit]. In: Trudy Mezhdunarodnoj Konferencii «Korpusnaja Lingvistika – 2015». [Proceedings of the International Conference «Corpus Linguistics – 2015»]. St.-Petersburg.
11. Mitrofanova, O.A. (2002a): Regulyarnoje i Irregulyarnoje v Terminologii Muzyki: o Jazykovyh Sposobah Zadanija Risunka Muzykal'nogo Proizvedenija. [Regular and Irregular Items in Terminology of Music: on Linguistic Means of Defining the Contour of the Musical Composition]. In: Materialy XXXI Nauchno-Prakticheskoj Konferencii Filologicheskogo Fakul'teta SPbGU. Vyp. 4. Sekcija Prikladnoj i Matematicheskoj Lingvistiki. [Proceedings of the XXXI Research Conference of the Philological Faculty, St.-Petersburg State University. Issue 4. Section of Applied and Mathematical Linguistics]. St.-Petersburg.
12. Mitrofanova, O.A. (2002b): Jazykovyje Sposoby Zadanija Risunka Muzykal'nogo Proizvedenija: Shtrihi k Lingvisticheskomu Portretu A.N. Skryabina. [Language Means of Defining the Contour of the Musical Composition: the Features of A.N. Skryabin's Linguistic Portrait]. In: Avtor. Tekst. Auditorija. [Author. Text. Audience]. Saratov.
13. RAS (2002): Russkij Assiciativnyj Slovar'. [The Russian Associative Dictionary]. Ed. by Ju.N. Karaulov et al. Vol. 1–2. Moscow.
14. Rhody, L.M. (2012): Topic Modeling and Figurative Language. In: Journal of Digital Humanities. Vol. 2(1). Winter 2012. URL: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/
15. Samin, D. (2002): 100 Velikih Muzykantov. [100 Great Musicians]. Moscow, 2002.
16. Samin, D. (2006): 100 velikih kompozitorov. [100 Great Composers]. Moscow, 2006.
17. Scha R., Bod, R. (1993): Computational Aesthetics. In: Informatie en Informatiebeleid. Vol. 11(1). Pp. 54–63.
18. TMB – Topic Modelling Bibliography. URL: http://www.cs.princeton.edu/~mimno/topics.html
19. Vorontsov, K.V., Potapenko, A. (2014): Additive Regularization of Topic Models. In: Machine Learning. URL: http://link.springer.com/article/10.1007/s10994-014-5476-6?sa_campaign=email/event/articleAuthor/onlineFirst
20. Zakharov, V.P. (2002): Korpusnaja Lingvistika [Corpus Linguistics]. St.-Petersburg.