E-hypertext Media Topic Model with Automatic Label Assignment

Olga Mitrofanova¹(ES), Anna Kriukova², Valery Shulginov³, and Vadim Shulginov⁴

¹ Saint-Petersburg State University, Universitetskaya emb., 11, 199034 Saint-Petersburg, Russian Federation
o.mitrofanova@spbu.ru
² Charles University, Ovocný trh 5, Prague 1 116 36, Czech Republic
krukova.ann@gmail.com
³ Higher School of Economics National Research University, Myasnitskaya Street, 20, 101000 Moscow, Russian Federation
shulginov.val@yandex.ru
⁴ Rostelecom, Svetlanskaya Street, 57, 690091 Vladivostok, Russian Federation
vadim.shulginov@yandex.ru

Abstract. This article deals with the principles of automatic label assignment for e-hypertext markup. We’ve identified 40 topics that are characteristic of hypertext media, after that, we used an ensemble of two graph-based methods using outer sources for candidate labels generation: candidate labels extraction from Yandex search engine (Labels-Yandex); candidate labels extraction from Wikipedia by operations on word vector representations in Explicit Semantic Analysis (ESA). The results of the algorithms are label’s triplets for each topic, after which we carried out a two-step evaluation procedure of the algorithms’ results: at the first stage, two experts assessed the triplet’s relevance to the topic on a 3-value scale (non-conformity to the topic/partial compliance to the topic/full compliance to the topic), second, we carried out evaluation of single labels by 10 assessors who were asked to mark each label by weights «0» – a label doesn’t match a topic; «1» – a label matches a topic. Our experiments show that in most cases Labels-Yandex algorithm predicts correct labels but frequently relates the topic to a label that is relevant to the current moment, but not to a set of keywords, while Labels-ESA works out labels with generalized content. Thus, a combination of these methods will make it possible to markup e-hypertext topics and create a semantic network theory of e-hypertext.

Keywords: E-hypertext · Topic modelling · Label assignment · Media discourse

1 Introduction

The invention of e-hypertext is determined by the cultural context of the 20th century: Theodor Nelson, who coined this term, defined it as an implementation of the postmodernist concept in which the text exists only in relation to reading. He defined the hypertext as «a series of text chunks connected by links, which offer a reader different...
pathway» [1]. By having a direct connection from one position in a text to another, the reader is actively involved in the construction of the hypertext, which is very different from reading traditional linear text.

The structure of non-linear interconnections in Internet communications also modifies the author’s interaction with e-hypertext. Following the principle of cooperation, the author creates a hypertext’s structure according to two potential strategies of the reader: «selecting the text semantically related to the previously read section (coherence strategy) or choosing the most interesting text, delaying reading of less interesting sections (interest strategy)» [2]. So, every link becomes the «ephemeral contract» [3] between author and reader, which suggests that the semantics of hypertext transition is clear to each of the communication participants.

Another important factor determining the topic structure of e-hypertext is the type of discourse in which hypertext is created. We study the functioning of e-hypertext in media discourse, which provides communication among groups so large, heterogeneous, and widely dispersed that they could never interact face-to-face or through any other but mass-produced and technologically mediated message systems [4]. This defines with the standardization of e-hypertext connections between source and target text, which are used to extend the information and confirm the identity of the source text. Thus, the topical structure of e-hypertext becomes a lens for studying linguistic (discursive characteristics) and anthropological (users strategies) features of Internet communication.

This research describes the principle of automatic label assignment of e-hypertext components: source and target texts, which is a necessary stage in describing the topic structure of e-hypertext and simplifies the linguistic interpretation of hypertext transition semantics. Labels are considered as general terms or multiword expressions exhibiting the content of a topic in a precise and brief manner. As a result, this approach allows you to automate the creation of hypertext media systems through topic clustering of texts. This research is based on the e-hypertext media corpus, which includes two databases: the texts of a media discourse with paratextual elements (title, announcement, date, author), and a database of links’ nominations taking into account the context of use, POS-tagging and vector representation.

2 E-hypertext Media Topic Model

The e-hypertext media corpus involved in our experiments includes texts and hypertext elements extracted from the Russian e-media «Kommersant», «Izvestia», «RBC», «Novaya Gazeta», «TASS», «Dozhd», «Vedomosti», «Interfax», etc. Alongside with the text of the article the corpus includes such metadata as media title, article title, subtitle, name of author, tags, date of publication. Experiments were carried out in the test set of 53 000 articles (total size 12 million tokens). The texts are linked by 70 000 tripartite hypertext elements. Development of the e-hypertext media corpus was performed by means of Python libraries BeautifulSoup [5], NLTK [6] and re [7]. The process of corpus development is discussed in detail in [4]. Corpus preprocessing implied tokenization, normalization, lowercase transformation, stop-word and punctuation deletion. Further we performed collocation analysis by means of phrases module in gensim library for Python [8]). Collocations were extracted and ranked according to the TF × IDF scheme.
Collocation extraction was necessary as it allowed to preserve multiword expressions in the topic model of the corpus. Topic modelling was fulfilled by the ensemble of algorithms including multiple learning method t-Distributed Stochastic Neighbor Embedding [9], DBSCAN clustering [10] and non-negative matrix factorization [11], these algorithms being applicable in processing high-dimensional datasets and clustering textual data as regards topic content. Thus, we generated a topic model for the e-hypertext media corpus which contained 40 topics of 10 topic words. In choosing such parameters of topic modelling we followed [12] where the given topic number and size were proposed for evaluation procedure.

A fragment of the output is given below:

**Topic 2:** следствия, адвокат, уголовный, дело, обвинять, следователь, сыр, арест, УК, фигурант, обвинение (investigation, lawyer, criminal, case, accuse, investigator, investigation committee, arrest, criminal code, person of interest, prosecution).

**Topic 4:** производство, предприятие, продукция, сеть, завод, технология, оператор, производитель, продукт, оборудование (manufacture, enterprise, production, network, plant, technology, operator, manufacturer, product, equipment).

**Topic 9:** матч, команда, сборная, чемпионат, клуб, футболист, игра, сборная, турнир, тренер (match, group, team (attr.), championship, club, football player, game, national team, tournament, coach).

The next step of our research deals with topic interpretation in course of label assignment.

## 3 Approaches to Topic Labelling

As a rule, topic modelling does not imply assignment of topic labels as an obligatory intrinsic operation. However, labelling may evidently enhance linguistic interpretability of generated topics. There are various solutions to the problem of label assignment. Topic labelling requires automated or manual choice of candidate labels which are expected to be unigrams or multiword expressions with general meaning correlating with the meanings of topic words. The most evident ways of label assignment are the choice of the first word (or $n$ first words) of a topic as a label or manual label assignment (e.g., [12, 13]). Those methods can hardly be used to form a solid baseline procedure for the following reasons. The first word (or $n$ first words) of a topic may be inconsistent as topic labels: the ordering of words within a topic does not necessarily reflect hierarchical structure of the lexicon. On the one hand, topic words are expected to establish numerous paradigmatic (synonymous, hyponymic and similar relations, e.g., завод – предприятие (plant – enterprise), команда – сборная (group – national team), матч – чемпионат – турнир (match – championship – tournament), топливо – нефть, бензин (fuel – oil, gasoline), etc.), syntagmatic (e.g. ядерный – комплекс, испытание (nuclear – complex, test), цена – вырасти (price – grow), etc.), epidigmatic (derivational) relations (e.g., обвинять – обвинение (accuse – prosecution), следствие – следователь (investigation – investigator), производство – производитель (manufacture – manufacturer), нефть – нефтяной (oil – oil (attr.)), etc.). On the other hand, their ordering within a topic can not be strictly defined in terms of abstractness-concreteness or
taxonomic relations (hypernym vs. Hyponyms, holonym vs. Meronyms, etc.). Moreover, the first words in a topic may be polysemous (e.g., акция, следствие (campain, investigation) etc.), that depreciates them as possible topic labels. As for manual label assignment, this procedure relies upon on the intuition of researchers and annotators, it requires involvement of experts as annotators in case of domain-specific corpora processing (e.g. medical, legislative, etc. texts) thus it may bring subjectivity into empirical results. At the same time, human expertise remains a significant step in evaluation of topic labelling results.

Computational linguistics provides several automatic techniques of label assignment, these techniques are distributed into classes as regards:

1) the sources of labels: inner sources (the labels are extracted from the corpus involved in topic modelling) or outer sources (the labels are extracted from reference corpora, output of search engines, knowledge databases (Wikipedia, WordNet, etc.);
2) type of algorithms involved: supervised or unsupervised;
3) type of labels: single words, phrases or both.

The works on label assignment using inner sources describe algorithms based on a) definition of Kullbach-Leibler distance between word distributions and maximization of mutual information between candidate labels and topics [14]; b) proper reranking of topical words as regards their attraction to the topic [12]; c) ranking candidate labels by means of summarization algorithms [15]; d) candidate label collocations extraction from documents most relevant to topics, mapping candidates to word vectors and letter trigrams, ranking candidates according to similarity between topics and label vectors [16]; e) detection of documents closest to topics, extracting single terms and multiword expressions and ranking them by information measures [17], etc.

There are various approaches to label assignment using outer sources, the most significant are as follows: a) using terms from Google Directory (gDir) hierarchy [18]; b) extracting article titles from Wikipedia or DBpedia and further ranking them as candidate labels [19, 20]; c) using web as a corpus for extracting candidate labels by searching Google and ranking candidates by PageRank [13]; d) using Wikipedia titles as candidate labels, ranking candidates by operations with neural embeddings for words and documents [21], e) incorporating a formal ontology into topic model for knowledge extraction (KB LDA) [22], f) using k-nearest neighbor clustering and similarity-preserved hashing for fast assignment of labels for newly emerging topics [23], etc.

In the study we proceed from two assumptions. First, topics consisting of sentences are more accessible for interpretation than topics consisting of 2–3 keywords (this is where we see the disadvantage of method b)). Secondly, the source of topics interpretation should have a genre similar to the dataset. Some of these methods are based on hard-to-reach data: a) Google Directory is no longer an actual source and e) shows stable results compared to the usual LDA but requires ontology. As a result, methods c) and d) are close to our approach and are analogous to English.
4 Our Approach to Topic Labelling

Procedures of automatic label assignment are widely discussed and evaluated for English corpora and knowledge resources, the Russian data being poorly represented in research projects. Our recent investigations are aimed to fill in the gap. In this study we used an ensemble of two graph-based methods using outer sources for candidate labels generation [24]: a) candidate labels extraction from Yandex search engine with their further ranking by TextRank (this is a graph-based model that takes into account the value of the each graph’s vertex depending on how many links it forms) [25, 26]; b) candidate labels extraction from Wikipedia by operations on word vector representations in Explicit Semantic Analysis (ESA) model [27, 28]. These procedures comprise two stages: candidate extraction and ranking. Both methods are compatible with any topic models and were tested on LDA from scikit-learn [29] for the corpus of Russian encyclopedic texts on linguistics. Experimental results and their evaluation proved that both methods are applicable. Candidate labels extraction from Yandex provides more consistent labels, although this algorithm turned out to be time-consuming (search engine imposes restrictions to the number of queries from an IP per minute) and hard to reproduce (the output may be adjusted to informational preferences of particular users). Losses in the quality of labels generated by Yandex-based algorithm are explained by the unstability of the external source. Candidate labels provided by ESA seem to be less general (in most cases they correspond to hyponyms of topic words), however the given approach is not time-consuming and provides reproducible results due to the stability of Wikipedia dump. Losses in the quality of labels in case of ESA labelling procedure deal with the specific content of the external source. Thus, the ensemble of two labelling methods could counterbalance and mitigate their drawbacks.

4.1 Topic Labelling Using Yandex

The algorithm for candidate labels extraction from Yandex search engine (Labels-Yandex) [25, 26] is an elaboration of the procedure originally designed for visual information processing [13]. At the stage of candidate extraction the first 10 topical words for each topic form a separate query to Yandex. The output of the query consists of top 30 titles for documents retrieved by the search engine. The list of titles is transformed into a continuous text which is subjected to preprocessing, namely, stop-words removal and lemmatization which is performed by pymorphy2 [30]. At the stage of candidate ranking a text is converted to an oriented graph $G = \{V, E\}$, where $V$ is a set of nodes corresponding to lemmata, and $E$ is a set of weighted edges marked by values of co-occurrence frequency within the text. Two nodes are connected by an edge if they occur together in a context window $[-1, +1]$. Further TextRank values are calculated for all nodes of the graph. Lemmata obtaining higher scores are considered as more significant, while edges getting larger weights imply strong semantic relations in word pairs. The algorithm was adjusted so that it could generate not only single word labels, but also multiword expressions. We introduced morphosyntactic patterns $Adj+N, N+N, N+prep+N$, etc., which were used to extract key phrases, each of them was assigned a TextRank value calculated as a sum of weights for each component.
4.2 Topic Labelling by Means of ESA

Explicit Semantic Analysis (ESA) is a variety of distributional semantic models projecting words from Wikipedia dump to a high-dimensional vector space. The original ESA algorithm was developed to improve monolingual and crosslingual search [31], the paper [24] discusses the first experience in using it for label assignment, the papers [27, 28] show its applications in detecting Russian text similarity/relatedness. In ESA model each article is treated as a separate «concept» represented by a vector generalizing all words co-occurring in an article. Wikipedia is converted to a term-document matrix, its cells containing TF-IDF values showing association strength in word – concept pairs. For each word ESA creates an inverted index showing its relations with the concepts (i.e. articles in which a word occurs). Concepts with low weights for a given word are removed from the model, this allows to reduce irrelevant links in ESA.

The algorithm for candidate labels extraction from Wikipedia by operations on word vector representations in ESA model (Labels-ESA) is based on the assumption that candidate labels considered as Wikipedia article titles should have vectors close to the topic words vectors. At the stage of candidate extraction the first 10 topical words for each topic form an averaged query vector, the output of ESA contains lists of article titles ordered according to cosine values indicating closeness of concept vectors and a query vector. The step of candidate ranking is performed in the same way as for topic labelling using Yandex.

5 Experiments on Topic Labelling for the E-hypertext Media Corpus

5.1 Experimental Procedure

Each topic was submitted as an input to the algorithms of topic labelling Labels-Yandex and Labels-ESA. The first three candidate labels ranked by the algorithms were chosen for further analysis (cf. Table 1).

5.2 Evaluation of Results

Experimental results require verification. As regards topic labelling, verification poses certain problems, the main obstacle being the absence of the golden standard for this procedure for any language (none for English, not to mention Russian). In our study we carried out a two-step evaluation procedure which is based on [13, 21].

First, we performed annotation of label triplets generated by Yandex-based algorithm and ESA. The annotation was based on a questionnaire modified from [24, 26]. Two assessors were asked to mark label triplets according to 3-value scale of weights:

«0» – a triplet doesn’t cover the content of a topic and can’t be used as a topic label;
«1» – a triplet somehow reflects the content of a topic and to more or less extent can be used as a topic label;
«2» – a triplet covers the content of a topic and can be used as a topic label.
Table 1. Examples of candidate labels generated by Labels-Yandex and Labels-ESA.

| Labels-Yandex | Labels-ESA |
|---------------|------------|
| **Topic 2:** следствие, адвокат, уголовный, дело, обвинять, следователь, СКР, арест, УК, фигурант, обвинение (investigation, lawyer, criminal, case, accuse, investigator, investigation committee, arrest, criminal code, person of interest, prosecution) | **Labels-ESA** |
| дело в отношении адвокатов РФ (case against lawyers of the Russian Federation) | дело в отношении (case against) | дело о взяточничестве (case of bribery) | уголовный кодекс (criminal code) | уголовное обвинение (criminal charge) | проект уголовного уложения (criminal code draft) |
| **Topic 21:** акция, митинг, задержать, Навальный, полиция, мероприятие, проведение, оппозиционер, площадь, участник (campaign, rally, detain, Navalny, police, event, holding, oppositionist, square, participant) | задержание на акциях в поддержку (detention at support events) | акции в поддержку (actions in support) | организаторы несогласованных митингов в Москве (organizers of uncoordinated rallies in Moscow) | список маршей (list of marches) | полиция (police) | сплит акций (stock split) |

We calculated averaged weights for label triplets: Yandex label triplets get high weight 1,4 and ESA triplets get medium weight 0,98 (maximum threshold 2), preliminary results being satisfactory.

Inter-annotator agreement was calculated by Kendall’s coefficient of concordance (Kendall’s W), its values being very high both for Yandex label triplets (W = 0,75) and for ESA label triplets (W = 0,87). The agreement between the raters proves to be quite strong, that testifies in favour of data and algorithms consistency.

Second, we carried out evaluation of single labels by 10 independent assessors who were asked to mark each label by weights «0» – a label doesn’t match a topic; «1» – a label matches a topic, cf. Table 2 (the main purpose of the second stage is to verify the received data, so we reduced the choice to two options). In this case we used a more general questionnaire as we addressed the intuition of the native speakers of Russian, not necessarily specialists in NLP and Data Science. The assessors were unaware of the label sources and were supposed to evaluate 6 labels per topic.

We calculated averaged weights for Labels-Yandex and for Labels-ESA: \( w(\text{Labels-Yandex}) = 0,68, \) \( w(\text{Labels-ESA}) = 0,39, \) medium weight 0,54 (cf. in experiments on the corpus of Russian encyclopedic texts on linguistics [24] \( w(\text{Labels-Yandex}) = 0,54, \) \( w(\text{Labels-ESA}) = 0,47, \) medium weight 0,51). Labels-Yandex performs better than Labels-ESA as it addresses to a wide range of web-documents of various subject areas, styles and genres, while ESA uses Wikipedia texts which reveal the traits of scientific style. On the one hand, in previous tests on linguistic texts the performance of two algorithms in question was almost equal due to the choice of the subject area which is well-represented both in the web and in wiki-articles. On the other hand, in experiments with the hypercorpus which contains news articles Labels-Yandex algorithm provides
much more precise labels than Labels-ESA, this sharp difference is explained by semantic and stylistic homogeneity of the corpus and search engine output and proves the divergence of news texts and encyclopedic articles.

Table 2. Examples of weighting candidate labels generated by Labels-Yandex and Labels-ESA.

| Topic 22: цена, нефть, рост, топливо, баррель, нефтяной, бензин, вырасти, снижение, fас (price, oil, growth, fuel, barrel, oil (attr.), gasoline, grow, decline, FAS) |
|---------------------------------------------------------------|
| Labels-Yandex | Labels-ESA |
| --- | --- |
| цена на бензин и нефть (price of gasoline and oil) | бензин и нефть (gasoline and oil) | цена на фьючерс на нефть (price of oil futures) | нефть (oil) | страны по добыче (countries in mining) | лицензия на добычу (production licence) |
| Assessor 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Assessor 2 | 1 | 1 | 0 | 1 | 0 | 0 |
| Assessor 3 | 1 | 1 | 1 | 1 | 0 | 0 |
| Assessor 4 | 1 | 1 | 1 | 1 | 1 | 1 |
| Assessor 5 | 1 | 1 | 1 | 1 | 1 | 1 |
| Assessor 6 | 1 | 1 | 1 | 1 | 0 | 0 |
| Assessor 7 | 1 | 1 | 1 | 1 | 0 | 0 |
| Assessor 8 | 1 | 1 | 1 | 1 | 0 | 0 |
| Assessor 9 | 1 | 1 | 0 | 0 | 0 | 0 |
| Assessor 10 | 1 | 1 | 1 | 1 | 1 | 0 |
| Total weight | 10 | 10 | 8 | 9 | 4 | 3 |

| Topic 38: Роснефть, нефть, месторождение, нефтяной, иск, соглашение, добыча, акция, дивиденд, сделка (Rosneft, oil, field, oil (attr.), lawsuit, agreement, mining, share, dividend, deal) |
|---------------------------------------------------------------|
| Labels-Yandex | Labels-ESA |
| --- | --- |
| нефть в России (oil in Russia) | тонны в год (tons per year) | Тоталь в России (Total in Russia) | месторождение (field) | лицензия на добычу (production licence) | добыча нефти (oil production) |
| Assessor 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| Assessor 2 | 1 | 0 | 0 | 0 | 0 | 1 |
| Assessor 3 | 1 | 1 | 1 | 1 | 1 | 1 |
| Assessor 4 | 1 | 1 | 1 | 1 | 1 | 1 |
| Assessor 5 | 1 | 0 | 1 | 1 | 1 | 1 |
| Assessor 6 | 1 | 0 | 0 | 1 | 1 | 1 |
| Assessor 7 | 1 | 0 | 0 | 1 | 1 | 1 |
| Assessor 8 | 1 | 0 | 0 | 1 | 0 | 1 |
| Assessor 9 | 1 | 0 | 0 | 0 | 0 | 1 |
| Assessor 10 | 1 | 0 | 0 | 1 | 0 | 1 |
| Total weight | 10 | 3 | 3 | 8 | 6 | 10 |
6 Discussion

We performed the analysis of raters’ label evaluation and tried to choose triplets of the best labels. To do this, we ranged total weights of labels as shown in Table 2.

Of 40 label triplets 20 contained 2 Yandex labels and 1 ESA label. In 13 triplets Yandex label was the best (e.g., Topic 16), in 6 triplets ESA labels got the highest scores (e.g., Topic 4), in 1 triplet Yandex and ESA labels got equal weights (Topic 27).

Topic 4: производство, предприятие, продукция, сеть, завод, технологии, оператор, производитель, продукт, оборудование (manufacture, enterprise, production, network, plant, technology, operator, manufacturer, product, equipment).

Label triplet: заводы (plants) (10, ESA), заводы и станки (plants and machines) (9, Yandex), завод пищевого оборудования (food equipment factory) (6, Yandex).

Topic 16: встреча, лидер, переговоры, саммит, КНДР, визит, состояться, обсудить двусторонний МИД (meeting, leader, talks, summit, North Korea, visit, take place, discuss, bilateral, Foreign Ministry).

Label triplet: встреча лидеров России и КНДР (meeting of Russia and North Korea leaders) (9, Yandex); Россия и КНДР (Russia and North Korea) (8, Yandex), Восточная Азия (East Asia) (4, ESA).

Topic 27: фильм, режиссер, картина, кино, кинотеатр, актер, театр, продюсер, фестиваль, зритель (film, director, picture, movie, cinema, actor, theatre, producer, festival, audience).

Label triplet: театр в кино (theatre in cinema) (5, Yandex), кино в театре (cinema in theatre) (5, Yandex), советское кино (soviet cinema) (5, ESA).

Further, 8 of 40 label triplets contained 1 Yandex label and 2 ESA labels. In 3 triplets Yandex label was evaluated higher (e.g., Topic 17), in 5 triplets ESA labels won (e.g. Topic 2).

Topic 2: следствие, адвокат, уголовный, дело, обвинять, следователь, СКР, арест, УК, фигурант, обвинение (investigation, lawyer, criminal, case, accuse, investigator, investigation committee, arrest, criminal code, person of interest, prosecution).

Label triplet: уголовное обвинение (criminal charge) (9, ESA) уголовный кодекс (criminal code) (8, ESA), дело в отношении адвокатов РФ (case against lawyers of the Russian Federation) (6, Yandex).

Topic 17: британский, Великобритания, Солсбери, Скрипаль, Лондон, отправление, дипломат, вещество, новичок, март (british, Great Britain, Salisbury, Skripal, London, poisoning, diplomat, novichok, substance, March).

Label triplet: Скрипали в Великобритании (the Skripals in Great Britain) (10, Yandex), новичок (novichok) (6, ESA), Лондон (London) (4, ESA).

The rest 12 triplets contained 3 Yandex labels (e.g., Topic 13), there were no triplets including ESA label only.

Topic 13: налог бюджет фондо доход налоговый Минфин трллион ставка НДС федеральный (tax, budget, fund, income, tax (attr.), Ministry of Finance, trillion, rate, VAT, federal).

Label triplet: налоги в бюджет (taxes to the budget) (10, Yandex), поступления в бюджет (budget receipts) (10, Yandex), рост доходов бюджета от повышения (growth of budget revenues from the increase) (7, Yandex).
In general, we got 28 triplets with both Yandex and ESA labels, where Yandex got the upper hand in 16 cases, ESA won in 11 cases, not to mention a single draw. That provides evidence in favour of Labels-Yandex and Labels-ESA ensemble.

Linguistic data analysis shows that Labels-Yandex algorithm gives quite concrete labels reflecting the agenda of the moment, while Labels-ESA works out labels with generalized content. Our experiments show that in certain cases Labels-Yandex algorithm fails to predict correct labels. Let’s consider Topic 10 and its labels:

**Topic 10:** губернатор отставка пост должность господин заместитель область администрация назначить совет (Governor, resignation, post, position, mister, deputy director, region, authority, appoint, board).

Label triplet: губернаторы (governors) (10, ESA); государственная администрация (state administration) (8, ESA), правительство в отставке (government in retirement) (6, Yandex).

In this case Labels-Yandex generated a label правительство в отставке (government in retirement) which reflected the news of the moment (January 15, 2020) [32], but could hardly correspond to a news cluster related to the topic from the hypercorpus. At the same time, Labels-ESA provided rather consistent labels like государственная администрация (state administration) or губернаторы (governors). So, news dynamics may be responsible for the shifts in the content of labels extracted from the search engine, while Wikipedia texts remain stable.

That’s why it is reasonable to use both Labels-Yandex and Labels-ESA as an ensemble of algorithms in automatic topic label assignment: a) internal ranking provided by both algorithms allows to choose 3 reliable labels from a mixed set of candidates, b) both algorithms provide multiword expressions alongside with single terms as reliable candidate for topic labels, c) the procedure is suitable for topic models of any type, irrespective of the topic modelling scheme and type of models (unigram or n-gram models), d) working together, two algorithms reduce losses in the quality of labels caused by the specificity of external sources and amend the total result.

7 Summary

The automatic topic label assignment is an important stage in the development of the semantic network theory of e-hypertext. This approach makes it possible to mark out multiple nodes of hypertext, rubricating them by topic, which will be the basis for studying the semantics of hypertext transitions and, as a result, it will be possible to determine the potential of a text to reveal its hypertextuality in a specific type of discourse.

Our experiments with the ensemble of topic labelling algorithms Labels-Yandex and Labels-ESA represent the first full-scale research performed on a large Russian e-hypertext corpus. We are the first to perform the procedure of topic label assignment working with the Russian e-media texts.

The choice of topic labelling algorithms Labels-Yandex and Labels-ESA which use external sources of labels (web-documents produced by Yandex search engine and the Russian Wikipedia dump) is justified by the fact that the Russian e-hypertext corpus is constantly extending and requires synchronization of its topics with the data from the web.
Results achieved in our experiments may serve as a starting point for further elaboration of automatic topic labelling techniques applied to Russian corpora as we provide both datasets, evaluation procedure and baseline scores. Thus, our research fills in the gap in contemporary Russian NLP research and may draw attention of computational linguists towards generalization and linguistic interpretation of topic models.

Acknowledgements. The reported study was funded by RFBR according to the research project № 18-312-00010. The authors express their deep gratitude to Aliia Erofeeva (CCG.ai, Cambridge, UK) and Kirill Sukharev (ETU «LETI», Saint-Petersburg, Russia) for their help in the development of topic labelling software.

References

1. Nelson, T.: Literary Machines. Mindful Press, Sausalito (1993)
2. Salmerón, L., Kintsch, W., Cañas, J.J.: Reading strategies and prior knowledge in learning from hypertext. Mem. Cognit. 34, 1157–1171 (2006)
3. Vandendorpe, C.: From Papyrus to Hypertext: Toward the Universal Digital Library (Topics in the Digital Humanities). University of Illinois Press (2009)
4. Shulginov, V.A., Shulginov, V.A., Mitrofanova, O.A.: Topic organization of c-hypertext media: corpus driven research. In: R. Piotrowski’s Readings in Language Engineering and Applied Linguistics (PREAL 2019), CEUR Workshop Proceedings, vol. 2552, pp. 299–312 (2019)
5. BeautifulSoup. https://pypi.org/project/beautifulsoup4/. Accessed 08 July 2020
6. NLTK. https://www.nltk.org/. Accessed 08 July 2020
7. re. https://github.com/python/cpython/blob/3.8/Lib/re.py. Accessed 08 July 2020
8. gensim. https://radimrehurek.com/gensim. Accessed 08 July 2020
9. t-Distributed Stochastic Neighbor Embedding. https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE. Accessed 08 July 2020
10. DBSCAN clustering. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html. Accessed 08 July 2020
11. Non-negative matrix factorization. https://radimrehurek.com/gensim/models/nmf.html. Accessed 08 July 2020
12. Lau, J.H., Newman, D., Karimi, S., Baldwin, T.: Best topic word selection for topic labelling. In: COLING’10 Proceedings of the 23rd International Conference on Computational Linguistics, Stroudsburg, PA, Association for Computational Linguistics, pp. 605–613 (2010)
13. Aletras, N., Stevenson, M., Court, R.: Labelling topics using unsupervised graph-based methods. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Baltimore, Maryland, ACL, pp. 631–636 (2014)
14. Mei, Q., Shen, X., Zhai, C.: Automatic labeling of multinomial topic models. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD 2007. pp. 490–499. ACM Press, New York (2007)
15. Cano Basave, A.E., He, Y., Xu, R.: Automatic labelling of topic models learned from twitter by summarisation. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 618–624 (2014)
16. Kou, W., Li, F., Baldwin, T.: Automatic labelling of topic models using word vectors and letter trigram vectors. In: Zuccon, G., Geva, S., Joho, H., Scholer, F., Sun, A., Zhang, P. (eds.) AIRS 2015. LNCS, vol. 9460, pp. 253–264. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-28940-3_20

17. Nolasco, D., Oliveira, J.: Detecting knowledge innovation through automatic topic labeling on scholar data. In: 49th Hawaii International Conference on System Sciences (HICSS), Koloa, HI, pp. 358–367. IEEE Computer Society (2016)

18. Magatti, D., Calegari, S., Ciucci, D., Stella, F.: Automatic labeling of topics. In: ISDA 2009 9th International Conference on Intelligent Systems Design and Applications, Pisa (2009)

19. Lau, J.H., Grieser, K., Newman, D., Baldwin, T.: Automatic labelling of topic models. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA, vol. 1, pp. 1536–1545. Association for Computational Linguistics (2011)

20. Hulpus, I., Hayes, C., Karnstedt, M., Greene, D.: Unsupervised graph-based topic labelling using DBpedia. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining WSDM 2013, pp. 465–474 (2013)

21. Bhatia, S., Lau, J.H., Baldwin, T.: Automatic labelling of topics with neural embeddings. In: 26th COLING International Conference on Computational Linguistics, 2016, pp. 953–963 (2016)

22. Allahyari, M., Pouriyeh, S., Kochut, K., Arabnia, H.R.: A knowledge-based topic modeling approach for automatic topic labeling. Int. J. Adv. Comput. Sci. Appl. 8(9), 335–349 (2017)

23. Mao, X., Hao, Y.-J., Zhou, Q., Yuan, W., Yang, L., Huang, H.: A novel fast framework for topic labeling based on similarity-preserved hashing. In: COLING 2016, pp. 3339–3348 (2016)

24. Kriukova, A., Erofeeva, A., Mitrofanova, O., Sukharev, K.: Explicit semantic analysis as a means for topic labelling. In: Ustalov, D., Filchenkov, A., Pivovarova, L., Žižka, J. (eds.) AINL 2018. CCIS, vol. 930, pp. 110–116. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01204-5_11

25. Mirzagitova, A., Mitrofanova, O.: Automatic assignment of labels in topic modelling for Russian corpora. In: Proceedings of 7th Tutorial and Research Workshop on Experimental Linguistics, ExLing 2016/A. Botinis, ed. Saint Petersburg: International Speech Communication Association, 2016, pp. 115–118 (2016)

26. Erofeeva, A., Mitrofanova, O.: Automatic Topic label assignment in topic models for russian text corpora. In: Structural and Applied Linguistics, Saint-Petersburg, vol. 12, pp. 122–147 (2019). (in Russian)

27. Kriukova, A., Mitrofanova, O., Sukharev, K.: Measuring semantic relatedness of Russian texts by means of explicit semantic analysis. In: Kalinichenko, L., Manolopoulos, Y., Stupnikov, S., Skvortsov, N., Sukhomlin, V. (eds.) Data Analytics and Management in Data Intensive Domains: XX International Conference DAMID/RCDL’2018 (October 9–12, 2018, Moscow, Russia): Conference Proceedings /, pp. 284–288. FRC CSC RAS, Moscow (2018)

28. Kriukova, A., Mitrofanova, O., Sukharev, K., Roschina, N.: Using explicit semantic analysis and Word2Vec in measuring semantic relatedness of Russian paraphrases. In: Alexandrov, D.A., Bokhanovsky, A.V., Chugunov, A.V., Kabanov, Y., Koltsova, O. (eds.) DTGS 2018. CCIS, vol. 859, pp. 350–360. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-02846-6_28

29. Scikit-learn. https://scikit-learn.org/stable/. Accessed 08 July 2020

30. Korobov, M.: Morphological analyzer and generator for Russian and Ukrainian languages. In: Khachay, M.Y., Konstantinova, N., Panchenko, A., Ignatov, D.I., Labunets, V.G. (eds.) AIST 2015. CCIS, vol. 542, pp. 320–332. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-26123-2_31
31. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1606–1611 (2007)

32. RIA News. https://ria.ru/20200115/1563456719.html. Accessed 08 July 2020