УДК 81'33

Technical Writer in the Framework of Modern Natural Language Processing Tasks

Larissa Beliaeva[a] and Valeria Chernyavskaya[b]*
*[a]Herzen State Pedagogical University
48 Moika river Emb., St. Petersburg, 182100, Russia
[b]Peter the Great St. Petersburg Polytechnic University
29 Politekhnicheskaya Str., St. Petersburg, 195251, Russia*

Received 05.12.2018, received in revised form 28.12.2018, accepted 10.01.2019

This study focuses on technical writer competences and necessary specialized language resources, supporting any language worker in the framework of modern natural language processing domain (terminologist, translator, lexicographer, technical writer, grammarian, teaching language specialist etc.). These specialists are new generation professionals with a basic linguistic/philological education, prepared to meet the demands of modern technology and science which are defined by the potential of industrial processes automatization (Industry 4.0) and appropriate presentation of information (Information 4.0). New concept of Information 4.0 reveals that information to be researched and/or created can be presented as a set of information molecules that are examined under the conditions of form, production and curation. Information 4.0 injects new life into the profession of technical communicator (technical writer) and language workers as a whole: those, who are prepared to solve the text processing tasks in this new technology space. The underlying assumption of the paper is that the text is the result of information transfer and the starting point of information mining and extraction.

Keywords: Information 4.0, technical communication, knowledge mining, text structure.

Research area: philology.

Citation: Beliaeva, L., Chernyavskaya, V. (2019). Technical writer in the framework of modern natural language processing tasks. J. Sib. Fed. Univ. Humanit. soc. sci., 12(1), 20-31. DOI: 10.17516/1997-1370-0377.

Introduction

The current situation in science and technology, namely new investigation avenues and new knowledge domains result in sharp backlog both in scientific and

* Corresponding author E-mail address: lauranbel@gmail.com; tcherniavskaia@rambler.ru

technical writing competences and in specialized language resources, supporting any language worker studies and work. Modern state of technology and science is defined by the potential of industrial processes automatization (Industry 4.0) and appropriate presentation of information on the project under development and implementation. There have been considerable changes in the circulation of scientific and technical information. Previously a major focus was brought on rhetoric of promoting research outcomes, text types/genre forms related to abstracting, summarizing and peer review of scientific knowledge, for more detail see (Bell, Housel, 2001; Chernyavskaya, 2016; Chernyavskaya, 2017; Rude, 2009; Velasco, 2005). A present-day researcher has to be focused mainly on the technology of information presentation rather than on the text in its linguistic, pragmatic and hermeneutic sense. This technology should be considered as secondary with regard to the previously adopted understanding of text as a structure. It is presumed that technology is primarily and essentially information (Rogers, 2002: 326), and that every technological activity that is introduced, negotiated, and applied within specific sociocultural contexts contains a core communication system. Technology transfers must be communication-driven, requiring mediators who will define approaches for a particular innovation to be introduced. Scientific and technological advancements must be properly relayed to their intended users.

In the framework of Industry 4.0 and Information 4.0, see for example (Gollner, 2016) the major task to be solved is to maximize the flexibility of information exchange. In the process of scientific and technical communication, information is generated in the form of documents at all stages of the project development. The quality of the documents produced in the source language and those translated into all the languages where the product to be distributed determines the opportunity to apply high-level automation in the process of their interpretation and publication.

Problem statement

Information 4.0 is seen as a cloud of information molecules rather than a set of structurally complete documents (“no documents, just information molecules”); it is dynamic, i.e. regularly updated; offered rather than delivered; ubiquitous, interactive, easily accessible and findable; uncensored, i.e. produced by contexts, profiled automatically, cf.: (Gallon, 2016). Thus, specialists should proceed from formal analysis of text semantics to generating and structuring the text in accordance with new contexts of the text usage and its requirements. A major concern for the specialists within the concept of Information 4.0 is to produce a text according to the rigorously set structural models and content rather than to analyze ready-to-use text structures. Therefore, information provided in a natural language (usually in a controlled language; Muegge, 2009) in the form of scientific and/or technical documentation must be prepared to be used in various situations and be quickly adapted to different scenarios of production, maintenance and security, both financial and information. Thus information must be presented in the way that it can be exchanged at any stages of project implementation. This gives a different view on the text structure and the role of separate components in structuring of the text information. Extracting information from texts in the age of its automated processing and transfer demands specialists of a new kind who would be concerned with text processing. Information 4.0 injects new life into the profession of technical communicators and language workers – those, who are prepared to solve text processing tasks in this new technology space.

Theoretical framework

This paper concerns the whole potential of applied linguistics in order to answer the following broad question: What new approaches to content structuring are seen in advancing from the procedures of formalizing text semantics to the procedures providing machine-readable automatic information structuring and text production? The study intends to answer this question by providing answers to the following secondary question: How do technical writers serve as mediators in the transfer of the same information into different text types.

Thus the paper addresses one of the general problems of linguistic education caused by the changes in modern situation with text generation and use in the framework of Information 4.0. In the context of growing information volume, an evident conflict of "academic" approach to educating a language specialist and the need for competent language workers, the aim of the present paper is also to justify the need for training new specialists that results from emerging high-powered technologies of extracting information from texts in scientific and technical communication.

The studies performed made it possible to define several special competences for a modern language worker.

Discussion and Results

Technical writers core competences. Specialists who deal with new forms of information presentation are referred to as *language workers*. This notion is used as a common nomination for terminologists, translators and all those who produce technical documents – *technical authors, technical writers*, and transfer technical information – *technical communicators.* In the communication of technology, technical writers serve as the gate-keepers to mediate information between innovation centers and society. In line with the product-centred orientation of technology transfer (Rogers, 2002) the output of technical writers were a planned effort to answer particular information needs of technology receptors in assimilating innovations. Information that accompanies the software is the realm of technical communication. Technical publications allow the product knowledge base to provide a holistic understanding of the technology itself from installation and administration to general daily usage. Technical communication should be understood as an important component of technology transfers. Technical writers perform a balancing act between managing internal information within development centers and assisting software users with understanding complex technology. Technical writers are involved in reporting research, managing knowledge based content and producing end-user documentation (Evia, 2008; Hughes, 2002; Jablonski, 2005; Jones, 2005). Without their intervention adopters of an innovation would have no starting point for the information and instructions necessary to use a particular technology.

Linguists working with Information 4.0 are required to have new competencies that can be described as follows:

- ability to collect, analyze and select relevant information to develop an information product,
- ability to choose the most appropriate product manufacturing strategies in order to design corresponding information products for various purposes and consumers,
- ability to ensure that information is extractable and available, that it is a coherent model, and comport with products and contexts,
- ability to select appropriate hardware and software to be used in scientific and technical communication,
- ability to design and assess e-learning courses,
- knowledge of the process of information products publishing and stages,
- sufficient level of understanding of subject domains relevant for specialists in technical information distribution (information science, mechanical engineering, physics, etc.) to be able to cooperate with experts in these subject areas,
- knowledge of basic principles and methods of the science of terminology,
- ability to build up linguistic resources, lexicographical databases and text corpora to solve professional tasks.

The two latter competencies concern working with terminology since in a new information environment exactly a technical writer, a product manager and a terminologist reveal new terms. They appear as a result of developing, certifying and documenting new products, thus all types of documentation are to be considered: catalogues, user manuals and reports, user interfaces, error reports and system messages, etc.

Competences in automatic text profiling and content structuring. Since the pragmatic approach to the text structure in terms of text grammar and semantics claims, the text should contain structural and meaning-oriented emphasis of certain text components which bear the most significant, in the author's viewpoint, information. This approach determines the standard pattern of production and perception of a scientific and/or technical text. It is a generally accepted IMRAD formula (introduction, method, results and discussion). Potential of automated processing sets a task of promoting a research outcome and its algorithmic marking differently.

Within the Information 4.0 framework a major concept is *structured content authoring*. It means structuring of the content into sections referred to as *topics* which are further automatically assembled in *maps*. These components allow produce a final draft of content to be used in a certain function and in a specific type of document. Topics should completely correspond with the text subjects. Then 'information molecules' can be marked algorithmically and form the ground for texts of a different kind. The latter can be used automatically.

Research of the text structures in order to extract information has become an important objective since the information flow expanded to the extent that its prompt and high-quality processing appeared to be really sophisticated for those who need this information. Among them are specialists and analysts who deal with data and knowledge extracting and processing. Texts began to be analyzed based on their content in the 70s of the 20th century when information retrieval (IR) became automatic. Information retrieval was used to select texts from previously produced and constantly supplemented text collections according to some topic. The topic has to be chosen by the user's query or extracted from a set number of topics. Texts could also be selected based on specific factual information, etc. IR was supposed to select automatically the texts to correspond with a certain query/topic.

A new method is based on the *productivist approach*. It implies that "the granularity of the topics is determined by production issues" (Lacroix, 2016: 103), by the objectives of scientific and technical documents. It is also potentially decorrelated from the content itself, i.e. from those topics that are actually discussed in the text.

Depending on working conditions with the information flow or with the existent text corpus, preliminary text indexing can be conducted in two ways. One way is to preset a list of topics relevant for the user – analyst, researcher, librarian, etc. (manually or automatically). In this case indexing is done according to this list. In the second case the indexing itself is to be carried out automatically. This means that the main content of the text is described as its search profile that forms the ground for its further processing.

According to the approaches mentioned above, text indexing can be based:

- on semantic analysis of texts using special dictionaries; this method allows to establish a relationship pattern (frame) between components of a sentence and/or a text, and thus determine the topic of the whole text, i.e. its search profile;
- on lexical, statistical and/or dictionary techniques; these allow to refer a text to a certain topic relying on whether it is relevant to the preset dictionary models and texts.

Thus, when indexing, firstly, it is possible to implement a statistical approach. In this case indexing is done by selecting the key terms from a text and referring them both to diagnostic features and frequency of their occurrence in the recognized text and in the reference collection (Krippendorff, 2013). Here a special dictionary of stop-words plays a significant role. This dictionary contains a list of words which do not bear information on a certain topic regardless of their frequency in the texts of specific subject domains (SD). Secondly, a dictionary approach is to be considered. It enables to produce a hierarchical system of models, thesauri and frames. This system represents a set of diagnostic mechanisms.

As a rule, indexing is done in terms of a certain set of topics and corresponding dictionaries. It should be noted that solving the problem of attributions of any type requires a preliminary description of the subject domain structure where indexing takes place. This description can be done in the form of the system of local dictionaries that contain such lexical units which have been selected based on preliminary research of a reference corpus and are more relevant when referring some document to a certain topic.

Therefore, to generate such a system, the first thing to be done is to investigate a set of topics and texts corresponding to these topics. A set of reference texts in each topic is selected based on consultations with experts, and the size of this set is determined by a standard approach accepted in linguistic statistics. Based on sets of reference texts frequency dictionaries are to be received. To reduce the amount of texts to be analyzed, all structural words (stop-words) and the words that do not bear information on the topics in question should be rejected from these dictionaries.

When this task is solved, specialists have quite a large amount of statistical data in order to observe the general topics and the direction of information flows that reflect the spread of interest in a particular subject domain. In addition, this system allows select the most important documents for further detailed semantic analysis. Indexing subsystem produced in this way can operate efficiently if quite a full and rigidly structured description of a certain subject domain has been made. These methods make it possible to determine which key terms are characteristic of each text and what topics they reflect. Words and documents in several topics are clustered as both the key word and the text document can correspond with several topics with different probabilities.

One of the major procedures is preliminary indexing of both queries and texts. This means that topics are formulated in advance. Topics correspond with the system of arranging the topic molecules of the text when it is produced automatically. It makes the queries and retrieval corresponding the sets of key terms and relations between them that are the result of preliminary collaboration between analysts and linguists.

Topics are selected based on an assumption that specific (key) units, in modern terminology – key terms which meanings are fully described, can be pointed out in the text structure. Today choosing the key words is regarded as an important objective of formulating the result obtained, cf.: (Krippendorff, 2013). Whether a term or a term combination belongs to a set of key units chosen to characterize the text content, is a critical factor in further use of the text by specialists and in spreading the results. All modern metrics used to point out key units are based on this idea. Only certain approaches and procedures vary as they employ more specialized methods of statistical analysis and big data.

Solving standard tasks both in content analysis and information retrieval traditionally relies on small sets of particular key terms (key words) used to mine relevant information on the text collection. The information can be pre-arranged as a text bank or be a permanent text flow. Automatic analysis and further information retrieval on the uses' queries is based on the procedure of attributing formally arranged information to the text about its content (indexing). This information is organized in the way that it can be used in automatic solution of different tasks in information retrieval and mining. This means that such procedure has to be multi-level and oriented to the text markup with respect to its syntactic (acoustic and graphic), semantic (combinatory and lexemic) and pragmatic (contextual) information. This procedure is based on the method of content analysis.

Both in terms of automated document generation and text topic determination, indexing is done with regard to a certain set of topics and corresponding dictionaries. It should be noted that solving the problem of attributions of any kinds requires preliminary description of the structure of the subject domain where indexing takes place. This description could be built in the form of a system of location dictionaries. Such dictionaries contain those lexical units that had been selected based on preliminary research of a reference corpus and are more relevant when referring some document to a certain topic.

Therefore, to generate a similar system, the first thing to be done is to investigate a set of topics and texts corresponding to these topics. A set of reference texts in each topic is selected based on consultations with experts, and the size of this set is determined by a standard approach accepted in linguistic statistics. Based on sets of reference texts alphabetical frequency dictionaries are arranged. To reduce the amount of texts to be analyzed, all structural words (stop words) and the words not bearing information on certain topics must be eliminated from the dictionaries. When producing a text, preliminary analysis of a similar set of reference texts should be done. This allows to define a set of topics – documentation ‘molecules’. Thus, text indexing and semantic analysis of a text are not only required for solving modern tasks in the concept of Information 4.0, but are also compulsory procedures.

When solving this task, specialists have quite a large amount of statistical data in order to observe the general condition of topics and the direction of information flows to reflect the spread of interest in a particular subject domain. Moreover, this system allows select the most important documents for further detailed semantic analysis. Based on probability topic modeling it is possible to determine which lexical units, namely key terms, are characteristic of each text and which topics they reflect. Topic models enable to cluster words and documents in several topics-clusters as both the key word and the text document can correspond with several topics with various probabilities.

Methods applied in analyzing texts and generating documentation are implemented by preliminary manual selecting of a bank of documents. This text bank can be regarded as a reference collection of a certain system. Search in such a reference collection is conducted in three stages:

1. Extracting a set of key terms from a reference collection. Terms in this collection are estimated as relevant/irrelevant to the texts. It allows to reveal whether they are

relevant or not in analysis and to assess whether they are necessary to be used in synthesis.

2. Mining of data on the key terms co-occurrence in the reference collection as well as in the national text corpus representative for the given language.

3. Quantitative estimation of relevance of each document based on the dictionary and co-occurrence data used to generate an estimated list of documents, for more detail see: (Beliaeva, 2003; 2009; Wiedemann, Niekler, 2014).

Conclusion

According to the predictions made by specialists, transfer to the fourth technical revolution will inevitably cause employment structure changes and new specialists demand. Permanent increase in demand on technical communication specialists, that is a pertinent and actively developing knowledge domain, gave rise to occurrence of new occupations of technical communicator (technical writer). Experience in development of professional educational programs for this profession acquisition has extended mainly in the domain of additional professional education, realized, mainly, by private companies. Academic training and accreditation of University programs in this domain falls substantially short.

At the same time the necessity to revise and comprehend the professional technical writer skills arose, as, first, the range of his problems increases permanently, and second, he goes to be not the unique professional, involved in the technical communication process. New competence structure gives the base to indicate that new specialist skills should be much broader.

Professional language worker training can be realized on the basis of philological education. Conception of new professional educational program “Language worker in the educational space and scientific and technical domain” (master’s level) is based on the new requirements to information preparation and on the research of necessary professional competences.

References

Beliaeva, L. (2009). Scientific Text Corpora as a Lexicographic Source. In *SLOVKO. NLP, Corpus Linguistics, Corpus Based Grammar Research, Proc. from the Intern. Conference, Smolenice, Slovakia. 19-25.*

Bell, A., Housel, T. (2001). *Measuring and Managing Knowledge.* New York: McGraw-Hill.

Beliaeva, L. (2003). Machine Translation Versus Dictionary and Text Structure. In *Journal of Quantitative Linguistics*, 10, 2, 193–211.

Chernyavskaya, V. (2017). Towards methodological application of Discourse Analysis in Corpus-driven Linguistics. In *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology,* 50, 135–148. DOI: 10.17223/19986645/50/9.

Chernyavskaya, V.E. (2018). Discourse analysis and corpus approaches: a missing evidence-based link? Towards qualitative and quantitative approaches. In *Voprosy Kognitivnoy Lingvistiki*, 2, 31-37. DOI: 10.20916/1812-3228-2018-2-31-37.

Evia, C (2008). The changing face of technical communication in the global outsourcing economy. In *Outsourcing Technical Communication, Evia C., Thatcher B.(eds.).* Amityville, NY: Baywood.

Gallon, R. (2016). Information 4.0, the Next Steps. In *Towards a European Competence Framework. tekom-Jahrestagungund tcworld conference in Stuttgart. Zusammenfassungen der Referate – Stuttgart:tcworld GmbH Verantwortlich*, pp. 95-97.

Gollner, J. (2016). Information 4.0 for Industry 4.0. In *Towards a European Competence Framework. tekom-Jahrestagungund tcworld conference in Stuttgart. Zusammenfassungen der Referate. Stuttgart:tcworld GmbH Verantwortlich,* pp. 93-94.

Hughes, M. (2002). A new value proposition for technical communicators. In *Technical Communication,* 49(3), 275-285.

Jablonski, J (2005). Seeing technical communication from a career perspective the implications of career theory for technical communication theory, practice, and curriculum design. In *Journal of Business and Technical Communication,* 19(1), 5-41.

Jones, S (2005). From writers to information coordinators: Technology and the changing face of collaboration. In *Journal of Business and Technical Communication,* 449-467.

Krippendorff, K. (2013). *Content Analysis: An Introduction to its Methodology.* 3rded. LosAngeles; London: Sage. 441 p.

Lacroix, F. (2016). Writing for the 21st Century. In *Towards a European Competence Framework. tekom-Jahrestagungund tcworld conference in Stuttgart. Zusammenfassungen der Referate – Stuttgart:tcworld GmbH Verantwortlich,* 102–106.

Muegge, U. (2009). *Controlled language – does my company need it?* Available at: www.tekom.de/artikel/artikel_2756 html. 2009.

Rogers, E (2002). The nature of technology transfer. In *Science Communication*, 22, 323-341.

Rude, C.D. (2009). Mapping the Research Questions in Technical communication. In *Journal of Business and Technical Communication*, 23, 2, 174–215. DOI http: 10.1177/1050651908329562.

Velasco, T. (2005). *Communication of Scientific and Technical Information*. Quezon City: Open University, University of the Philippines.

Wiedemann, G., Niekler, A. (2014). Document Retrieval for Large Scale Content Analysis using Contextualized Dictionaries. In *Terminology and Knowledge Engineering*, 29-32.

Технический автор в решении современных задач обработки текстов на естественном языке

Л. Беляева[а], В. Чернявская[б]

[а]Российский государственный педагогический университет им. А.И. Герцена
Россия, 182100, Санкт-Петербург, наб. р. Мойки, 48
[б]Санкт-Петербургский политехнический университет Петра Великого
Россия, 195251, Санкт-Петербург, ул. Политехническая, 29

Фокусом этого исследования выступают компетенции технического автора и необходимые специализированные лингвистические ресурсы, поддерживающие работу любого специалиста в области современной обработки текстов на естественном языке: терминолога, переводчика, лексикографа, технического автора, филолога, методиста в области преподавания языка и т.д. Такие специалисты являются профессионалами с базовым лингвистическим (филологическим) образованием. Они готовы соответствовать требованиям современной технологии и науки, которые определяются потенциалом автоматизации производственных процессов (Промышленность 4.0), и современным требованиям к представлению информации (Информация 4.0). Информация 4.0 характеризуется как «молекулярная», то есть формируемая не из структурно завершенных документов, но из информационных молекул, как непрерывно обновляемая, интерактивная, доступная и удобная для поиска; спонтанная, т.е. вызываемая контекстами и профилируемая автоматически.

Авторы показывают, что новое понятие «Информация 4.0» требует, чтобы создаваемая и/или исследуемая информация была представлена как набор информационных молекул, рассматриваемых по условиям формы, продуцирования и поддержки. Исходное предположение авторов статьи состоит в том, что текстовая структура является результатом передачи информации и начальной точкой поиска и извлечения информации. Новые технически обусловленные возможности извлечения информации *из текстов в научной и технической коммуникации в контексте Информации 4.0 порождают новые требования к профессии технического коммуникатора (технического писателя) и к специалистам в области современной обработки текстов, готовых решать задачи автоматической обработки текста в этом новом технологическом пространстве. Для работы специалиста в контексте Информации 4.0 существенное значение имеет не анализ уже готовых текстовых структур, но построение текста по заданным структурным моделям и контенту. Следовательно, информация, представленная на естественном языке в виде научной и/или технической документации, должна динамично приспосабливаться к различным сценариям производства.*

Ключевые слова: Информация 4,0, научная и техническая коммуникация, структура текста, авторская разработка структурированного контента.

Научная специальность: 10.00.00 – филологические науки.