The Extent of Legal Control over Language Data: the Case of Language Technologies

Aleksei Kelli
University of Tartu
Estonia
aleksei.kelli@ut.ee

Arvi Tavast
Institute of
Estonian Language
Estonia
arvi@tavast.ee

Krister Lindén
University of Helsinki
Finland
krister.linden@helsinki.fi

Kadri Vider
University of Tartu
Estonia
kadri.vider@ut.ee

Ramūnas Birštonas
Vilnius University
Lithuania
ramunas.birstonas@tf.vu.lt

Penny Labropoulou
ILSP/ARC, Greece
penny@ilsp.gr

Irene Kull
University of Tartu
Estonia
irene.kull@ut.ee

Gaabriel Tavits
University of Tartu
Estonia
gaabriel.tavits@ut.ee

Age Värv
University of Tartu
Estonia
age.varv@ut.ee

Abstract

The article aims to increase legal clarity concerning the impact of data containing copyrighted content and personal data on the development of language technologies. The question is whether legal rights covering data affect language models.

1 Introduction

The development of language technologies (LTs) relies on the use of language data (LD). Language data is often covered with several tiers of rights (copyright, related rights, personal data rights). The use of this kind of data can be based on consent or exemption model (for further discussion, see Kelli et al. 2015; Kelli et al. 2018).

The relevant issue here concerns the impact of data's legal regime on LTs. The question is whether legal restrictions applicable to data apply to the language technologies that are developed using them as well. The article aims to reduce the legal uncertainty regarding how far, in the pipeline of developing language technologies, the original copyright and personal data protection[1] regulations apply. If we take a recorded phone call, for instance, it is obvious that copyright and data protection apply to a copy of that recording. At the other extreme, it is equally obvious that they do not apply to the Voice UI (User Interface) of a new fridge, even though the latter was trained on a corpus containing the former. The line

[1] The GDPR defines personal data as "any information relating to an identified or identifiable natural person ('data subject')" (Art. 4 (1)).

where the original rights cease to apply has to be somewhere between these points, and it is vital for researchers and developers to know where.

2 From language data to language technologies

The development of data-driven/data-based language technologies contains:

1. Collection of raw data (written texts, speech recordings, photos, videos, etc.). These often contain copyrighted material and personal data. Their development usually does not involve any other activities than the actual recording, initial cleaning and sanity-checking of the data.

Dangers for both copyright and personal data can be very real: re-publication of copyrighted works, surveillance by governments or insurance companies, etc.

Almost impossible to anonymise or pseudonymise completely, so that it would become mathematically impossible to identify any persons or reproduce any significant portions of copyrighted works.

2. Compiling of datasets, or collections of data (raw text corpora like Google News, Common Crawl or OpenSubtitles, speech corpora like the Prague DaTabase of Spoken Czech, etc.). The above, but collected and organised with a specific criterion in mind (e.g. speech recordings on a specific topic by residents of a certain region in order to capture the accent of the region); these datasets usually come in such quantities that any individual piece of data constitutes a negligible part of the whole, and could in principle be removed without affecting the usability of the dataset.

For copyright and personal data purposes, not different from raw data[2]. The main practical difference is that the sheer volume of data may make it technically difficult for an individual to become aware that their data has been included in the dataset.

Creation of a dataset often involves a nontrivial contribution in gathering, organising, indexing, presenting, hosting etc. of the data.

3. Creation of annotated datasets (POS-tagged corpus of written texts like the ENC17, syntactically parsed corpora like the Universal Dependencies treebanks, etc.). The above, augmented with some kind of analysis.

Again, not different from raw data in terms of copyright and personal data, although the copyright holders of the raw data and the annotations may be different. The annotation layers may be stored separately and may even have some use on their own, but normal practice is to process copies of the original data together with the annotation layers so that the resulting dataset contains all of the original data.

Creation of an annotated dataset includes analysis of the data, either manual, semi-automatic or automatic.

4. Models. Data products developed from some sort of processing on the above, but not necessarily containing the above, which try to *model*, i.e. represent or describe, language usage. Examples: dictionaries, wordlists, frequency distributions, n-gram lists like Google ngrams, pre-trained word embeddings like in Grave et al. 2018, pre-trained language models like in Devlin et al. 2018.

Creation of a model involves significant amounts of work, expertise and (computational) resources. Steps include at least creation and/or selection of the algorithm, implementation of the algorithm in software, hardware setup (may even include custom hardware development), hyperparameter optimisation, model validation.

In rare cases, some model types may be consumer products of their own (e.g. dictionaries). Mainly, however, models are used in downstream tasks to create other products.

[2] In fact, it can be argued that data-sets qualify for database protection (for further discussion, cf. Eckart de Castilho et al. 2018).

5. Semi-finished products (text-to-speech engine or a visual object detector) and **finished products** (talking fridge). Out of scope for the current analysis, because their status as original works should be beyond doubt.

3 The legal status of models

The focus of the article is on models. It is crucial to determine whether the use of data containing personal data and copyright content influences the subsequent utilisation of the model. Therefore, copyright and personal data regulations are analysed.

3.1 Copyright perspective

From the copyright perspective, there are three relevant issues. Firstly, whether copyright material is used. Secondly, if it is used then whether there is a legitimate ground for the use. Thirdly, how to define models themselves within the copyright framework.

To answer the question about the copyright law impact on models, the requirements for copyright subject matter should be briefly outlined. The main and long-established requirement is that of **originality**. Work is protected if, and only if, it is original. Therefore, the originality requirement defines the copyright status of the input data. Oddly enough, this general requirement was never defined in international treaties or European *acquis*[3]. The task to define the legal meaning of originality for copyright purposes was mainly taken by the Court of Justice of the European Union (CJEU). As was explained in the seminal decision in the *Infopaq* case (C-5/08), originality means the author's intellectual creation. Another important explanation in the *Infopaq* case was that an extract consisting of eleven words could constitute an original work. The Court has also explained that a single word cannot be regarded as original and protectable work.

In the context of the current research, the originality requirement is important from two different perspectives. First, if originality is missing, the pre-existing text contained in a dataset is not protected and can be used without authorisation. Therefore, even if parts of this text are reproduced in the model, they are not protected as well. Second, even if a text as a whole is original and, therefore, protected, the question remains, whether the fragments used in the model are original on their own. If they are not, then again, they can be used without authorisation. Thus, originality must be established not only concerning the original work but also as regards the parts used.

In addition, in its latest case law CJEU has underlined, that, besides originality, a work also must meet the second requirement in order to be copyright-protected, i.e. it "*must be expressed in a manner which makes it identifiable with sufficient precision and objectivity, even though that expression is not necessarily in the permanent form*" (C-310/17). Arguably, this requirement in practice will be present in the majority of cases, because the texts used for models normally are expressed in a fixed form.

We consider that the development of the model is done through a data mining activity, according to the definition of the Copyright Directive, which defines text and data mining as "any automated analytical technique aimed at analysing text and data in digital form to generate information which includes but is not limited to patterns, trends and correlations" (Art. 2 (2)).

To answer the question of whether models are copyright protected, by the previous section, we must establish whether they meet the requirement of originality also on their own (irrespective of the input dataset).

One of the criteria that can be used for assessing originality has to do with the degree of human intellect invested in the process: how far is the model a unique product, the result of the intellectual creation of the author (developer) and not the result of a process that any other qualified engineer could also create? Building a model (as presented in the previous section) includes a number of choices and actions on the part of the developer: choice/creation of the dataset, choice/creation of the programme to be used for the training and development of the model and various cycles of testing and validation by tuning the parameters of the training programme. So, the question is: if the same program is used by another qualified user (engineer) on the same dataset, would they arrive at the same results, i.e .produce the same model? The main differences lie in the tuning of the parameters of the algorithm/program

[3] Although it was defined in several EU directives with regard to specific categories of works, such as computer programs or photographic works.

which is linked to the cycles of testing and validation; so, if this tuning is "original/creative" enough it can be considered a copyrighted program. If the choice of parameters is limited, as it happens in specific cases, the program and the model could not be considered original (cf. Eckart de Castilho et al. 2018). In other cases, this may indeed involve a substantial intellectual effort on the part of the developer; if so, it can be argued that the resulting program is "original" and thus copyright-protected. But what about the output of the processing, the model? Is this also "original"? The application of the same algorithm with the same parameters on the same input will result in the same model. In fact, this can be seen as similar to using a Part-of-Speech tagger to automatically annotate a corpus without any human intervention: the tagger itself may be copyrighted, the input dataset may also be copyrighted, but the annotations themselves cannot be considered as "original"; what is copyright protected is the part of the input dataset that remains in the annotated dataset. We can thus argue that the model itself is not original in this sense.

Also, we need to establish if any substantial original parts of the input dataset remain in the model and thus qualify for copyright protection of these parts. If none or very small parts remain in the model (and thus does not contain any original parts), then we can conclude that as far as this point is concerned, the model is not copyrightable.

In case considerable parts of copyrighted works remain in models, they can be considered derivative works. There is no clear definition of derivative work in international or European legal acts, and different jurisdictions have a quite different understanding of this concept (for further discussion, see Birštonas and Usonienė, 2013; Echart de Castilho et al. 2018). It is not clear, how much of the original work should remain to categorise a model as a derivative work. However, this issue is not very practical since the copyright protection of the primary work (copyrighted content in the dataset used for the development of the model) is not dependent on the fact whether the model is a derivative work or not. The only important question is whether the original part of the primary work has been used in the latter work (the model).

It is generally not possible to recreate copyrighted works or personal data contained in a dataset from the model that has used it. Some small excerpts of the original data may remain in the model, and it is important to see if these violate the regulations of copyright and personal data. Very idiosyncratic language use would be hard to filter out in a guaranteed way, but cases of this are very rare and lose significance even more with the increase of data volumes.

To give a definite answer, we should have a closer look into all the model types and the processes and resource types and modalities they have been built upon, which is not possible in the limits of this article. It can be argued though that models by definition try to capture *generalities* of language use and *abstract* from the original texts as far as possible, producing mainly lists of words or phrases and patterns with statistical measures.

3.2 Personal data perspective

Regarding personal data, it is theoretically possible that small but identifiable bits of information make it to the model. A wordlist might contain a name or e-mail address, for instance. This is easy to avoid using anonymisation or pseudonymisation.

However, it should be kept in mind that for personal data, there is no minimum segment in the audio synthesis. Even if the voice is synthesized using neural networks without any remnants of the person’s original voice recording, having trained the network for research purposes using a publicly available radio transmission as training data, one is still using the personal data of that person when the person can be identified based on the synthesized output despite the fact that there is no single bit in the network which could be attributed to the person’s voice.

The main issue here is how to substantiate the processing[4] of personal data contained in a model. Generally speaking, the compilation of datasets containing personal data used to create models can be based on the consent, public interest research and legitimate interest (see, GDPR Art. 6 (1) a), e), f)). In

[4] The GDPR defines processing as “any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction” (Art. 4 (2)).

case there is consent to process data for research purposes, or processing relies on public interest and the resulting model is used for research purposes as well (it is not made available to the public or used for commercial purposes), then there is no problem. There is also no problem if consent covers commercial use and public dissemination.

However, the situation becomes complicated when a dataset containing personal data is processed based on consent asked for research or on the public interest research exception, but the resulting model (where the personal data may remain) is planned to be used for commercial purposes or be made publicly available. If the personal data is in the form of speech, then anonymisation is rather difficult. In the described case there are the following scenarios:

1) Argue that voice without any identifying information is not personal data (it is anonymous data). The key here is how to interpret the concept of identifiable natural person (Art. 4 (1)) [5];
2) Ask for consent for commercial use;
3) Argue that the use of voice in the model is based on the legitimate interest. Especially bearing in mind that the identification is impossible or almost impossible and the voice does not contain any data which would affect the data subject negatively.

4 Conclusion

It is clear that raw data, datasets and annotated datasets are affected by copyright and personal data regulations. To some extent, models rely on datasets. They do not usually contain copyright-protected content. However, models containing speech need to address personal data issues.

When creating a language model, there are several activities involving complex human intellectual activity such as choosing and annotating datasets as well as choosing the software and tweaking its parameters. The outcome of the preparatory software activities is applied to a prepared dataset to compile a language model.

The outcome of the preparatory software activities is usually encoded in a piece of software. This software becomes the model trainer that embodies the copyright of the preparatory software activities. The model trainer is applied to the prepared dataset but does not inject the copyright in the model trainer into the language model as the model trainer is a piece of software, which is mechanically applied to the dataset.

If the compiled language model contains sufficiently long pieces of the original data, there may be some copyright left from the dataset.

[5] See e.g., Article 29 Working Party Opinion 4/2007 on the concept of personal data.

References

Birštonas, R., Usonienė, J. 2013. Derivative Works: Some Comparative Remarks from the European Copyright Law. *UWM Law Review*, Volume 5.

Case C-310/17. Levola Hengelo BV vs Smilde Foods BV (13 November 2018). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555243886259&uri=CELEX:62017CJ0310 (14.4.2019).

Case C-5/08. Infopaq International A/S vs Danske Dagblades Forening (16 July 2009). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555243488182&uri=CELEX:62008CJ0005 (14.4.2019).

Directive (EU) 2019/… of the European Parliament and of the Council on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Text with EEA relevance). Available at http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2019-0231+0+DOC+XML+V0//EN&language=EN#top (14.4.2019).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [Cs].

Eckart de Castilho, R., Dore, G., Margoni, T., Labropoulou, P. & Gurevych, I. 2018. A legal perspective on training models for Natural Language Processing. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, ELRA. Available at: http://www.lrec-conf.org/proceedings/lrec2018/pdf/1006.pdf (17.4.2019).

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. 2018. Learning word vectors for 157 languages. ArXiv Preprint ArXiv:1802.06893.

GDPR. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, p. 1-88. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555312258399&uri=CELEX:32016R0679 (15.4.2019).

InfoSoc Directive. Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. *Official Journal L 167*, 22/06/2001 P. 0010 – 0019. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555254956114&uri=CELEX:32001L0029 (14.4.2019).

Aleksei Kelli, Krister Lindén, Kadri Vider, Pawel Kamocki, Ramūnas Birštonas, Silvia Calamai, Chiara Kolletzek, Penny Labropoulou, Maria Gavriilidou. 2018. Processing personal data without the consent of the data subject for the development and use of language resources. *CLARIN Annual Conference 2018 Proceedings: CLARIN Annual Conference 2018, 8-10 October 2018 Pisa, Italy*. Ed. Inguna Skadin, Maria Eskevich. CLARIN, 43−48. Available at https://office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf (14.3.2019).

Aleksei Kelli, Kadri Vider, Krister Lindén. 2015. The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure. 123: *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland*. Ed. Koenraad De Smedt. Linköping University Electronic Press, Linköpings universitet, 13−24. Available at http://www.ep.liu.se/ecp/article.asp?issue=123&article=002 (28.3.2018).