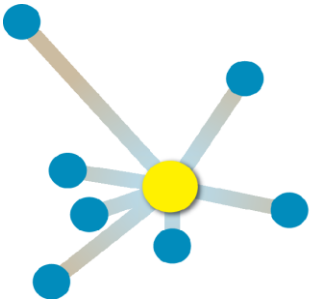
European Language Resource Coordination 3.0 (ELRC3.0) is a service contract operating under the EU's Connecting Europe Facility SMART 2019/1083 programme.

SMART 2019/1083
ACTION ON CEF AUTOMATED TRANSLATION
CORE SERVICE PLATFORM

D5.2.4 Report of Workshop 4

Author(s): **Tom Vanallemeersch (CrossLang)**

Version: **1**
Date: **2022-07-15**



European Language Resource Coordination
Connecting Europe Facility

Table of Contents

1. Introduction

The workshops organised in the SMART 2019/1083 programme support the continued development of the EC's eTranslation system and a wider deployment of DG CNECT's services in terms of language resources and tools. The fourth of these workshops, *Large language models: pre-training with a twist*, focused on models that incorporate information useful for understanding a language, such as its vocabulary and how it expresses meaning. Using deep learning, which is the state-of-the-art technology in AI, academic or commercial organisations with heavy computing infrastructure construct such language models (LMs) from large amounts of data (text or speech). Based on these LMs, other organisations can train additional models for their specific applications (e.g. automated translation, summarisation, dialogue interaction, speech recognition) or domains. The process of "specialising" a large model (the latter is called the pre-trained LM) requires much less data and computing power. Therefore, this type of specialisation is having a large impact on the field of natural language processing.

The workshop included 7 presentations, involving a total of 8 speakers. Besides the organisers and the speakers (1 from a public-private partnership, 2 from academia, 4 from industry, and 1 from a public administration), the 68 participants included staff of EU administrations and Member State administrations and representatives of academia and industry.

The following subtopics were dealt with in the workshop:

- how to make use of LMs available from repositories;
- how to specialise multilingual LMs, for instance for automated translation;
- how to leverage LMs in specific use cases, within public administrations and industry;
- how to take into account legal aspects of LMs.

The present document is structured as follows. Section 2 provides information on the setup of the workshop (preparation steps, agenda, participants, and practical organisation of the workshop day). Section 3 provides the abstracts of the presentations, questions of participants, and a description of the time slots with discussions between the speakers and other participants. Section 4 provides conclusions.

2. Setup

This section provides information on the setup of the workshop, i.e. the preparation, the agenda, the participants, and the practical organisation of the workshop day.

2.1 Preparation

CrossLang settled a date for the workshop in agreement with the Contracting Authority and set up an agenda including 8 speakers (1 from a public-private partnership, 2 from academia, 4 from industry, and 1 from a public administration). The agenda set up by CrossLang was submitted to and approved by the Contracting Authority. The speakers submitted their presentation abstracts in order to complete the agenda. One of the speakers (Georg Rehm) was asked whether he would be willing to provide the introductory talk, setting the ground for the other presentations, and he agreed to this role.

A page with the workshop summary and agenda was set up in the ELRC website (https://lr-coordination.eu/workshop4), as well as a page with the abstracts of the speakers and a registration page. CrossLang and DG CNECT created awareness of the workshop among potential participants inside and outside the EU administrations by providing them with a link to the workshop agenda. The workshop was promoted through social media activities (Facebook, LinkedIn, Twitter). While the audience of the workshop consisted of staff of EU and Member State administrations in the first place, the workshop was also open to a wider audience of people with an interest in techniques and research relating to the topic. Therefore, the speakers were invited to contact their colleagues. A total of 101 people registered: representatives from administrations, industry, and academia. A Zoom session was set up and CrossLang mailed the Zoom link to the people who registered.

Speakers sent their slides beforehand in order to anticipate on potential Internet connection problems during the workshop; for instance, in case of screen sharing problems, navigation through the slides can be performed by the organisers based on instructions by a speaker.

CrossLang set up two poll questions in order to increase interaction with the audience.

2.2 Agenda

10:00: Welcome by Tom Vanallemeersch (SMART 2019/0183 Project representative)

10:10: **Georg Rehm** (DFKI, Berlin): *Large Language Models and European Language Equality – Where do we stand and what do we need to do?*

10:50: **Nils Reimers** (Hugging Face, Frankfurt): *Cross-Lingual Semantic Search*

11:20: **Alexandra Chronopoulou** (Ludwig Maximilian University, Munich): *Improving Multilingual Machine Translation with Language-family Adapters*

11:50: **Denis Jouvet** (MULTISPEECH, Inria/LORIA, Nancy**)**: *Language Models for Speech Recognition*

12:20: Discussion

12:35: Lunch

13:50: **Love Börjeson** (KBLab, National Library of Sweden): *Large-scale, Open-access AI Models for Swedish*

14:20: **Jakub Zavrel** (Zeta Alpha, Amsterdam): *A New Generation of Neural Search and Knowledge Discovery Tools*

14:50: **Khalid Choukri**, **Mickael Rigault** (ELDA, Paris): *Legal Aspects of Language Models*

15:20: Discussion

15:50: Conclusion

16:00: End

2.3 Participants

101 people registered (including the speakers and organisers). The registered people have the following profiles (each number stands for the absolute figure as well as the percentage):

- staff of EU administrations: 31;
- staff of Member State administrations: 22;
- representatives of companies: 29;
- representatives of academia: 14;
- staff from public-private partnerships: 5.

Of those who registered, 68 people logged in to the workshop.

2.4 Practical organisation

From the part of CrossLang, two people participated in the workshop: one for chairing the workshop, and one for providing technical support to the chair and the attendees and for supporting the chair in following up the chat questions.

The speakers were given the *co-host* status in order to be able to share their screen. They were asked to log in 30 minutes before the start of the workshop in order to test whether screen sharing worked without problems.

When attendees logged in, the workshop organisers admitted them manually. Attendees used the chat function of Zoom for questions. After each presentation, the chair read questions from the chat window. The attendees could also use the *raise hand* button after a presentation in order to request the floor.

In order to give the workshop participants the possibility to interact in an informal way during the lunch break, CrossLang set up break-out rooms.

After the presentation of the second speaker, the organisers showed the attendees a poll question using Zoom functionality:

What kind of application in your environment could make use of a pre-trained model?

Possible answers (one or more per participant):

1. Machine translation
2. Search / question-answering
3. Speech processing (e.g. transcription)
4. Classification (e.g. sentiment analysis)
5. Other

33 people participated in the poll. The first answer was selected the most frequently (27 times), whereas the tree following ones were selected with a somewhat lower, similar frequency (answer 2 was chosen 15 times, 3 and 4 were chosen 18 times). The fifth option was selected 4 times.

After the lunch break, another poll question was presented to the audience:

If you would use a pre-trained model, what languages does it need to include?

Possible answers (only one per participant):

1. One language only
2. A few languages (e.g. local language plus English)
3. All official EU languages
4. As much of the world's languages as possible
5. Other

21 people participated in the second poll. Almost all of them selected an option in which more than one language is involved. The second and third answer were selected the most frequently (8 and 10 attendees). The fourth option was selected by 2 people, and the first one by 1 person. The fifth option was not chosen.

The presentations' slides were uploaded to the workshop page on the ELRC website. The link to the recordings of the presentations was also made available to the workshop participants.

3. Presentations

This section provides the abstracts of the presentations and the questions following presentations. The section also reports on the two discussion slots, before the lunch break and at the end of the workshop. These involved speakers as well as other attendees.

3.1 Welcome by project representative

The project representative, Tom Vanallemeersch (CrossLang) welcomes the audience, provides background information on the workshop (SMART 2019/1083 service contract, purpose of technical workshops), and provides the "housekeeping" rules for the workshop day.

3.2 Large Language Models and European Language Equality – Where do we stand and what do we need to do?

Prof. Dr. Georg Rehm, Principal Researcher in the Speech and Language Technology Lab at the German Research Center for Artificial Intelligence (DFKI) in Berlin, first summarised the EU projects European Language Grid (ELG; 2019-2022) and European Language Equality (2021-2022), which develop a technology platform for the European Language Technology community and a strategic research agenda for achieving digital language equality in Europe by 2030, respectively. The first part of his presentation concluded with a set of high-level recommendations, which also included a description of the current state of play of LM development in Europe. In the second part of his presentation, Georg presented the recently started German project OpenGPT-X, in which LMs for German will be developed. This project can potentially act as a blueprint for similar endeavours to be started in other European countries to create models for their respective languages.

The following questions were asked to the presenter:

You will deal mainly with German and English and a few other European languages. Do you already have an idea about the other languages you will deal with?

We have a shortlist, but it is not set in stone.

Will there be languages for which there is already a large amount of data, like French, or will you think about less-resourced languages?

Both. A couple of languages that are well-supported through data and probably also a few languages which are extremely badly supported. This way, we can compare how they stack up next to each other.

Do you already know on which data will you train the models?

We currently have a group that is collecting datasets. Datasets are also emerging. Different configurations, compilations of datasets have been published, that may include other dataset compilations. This is work in progress. We are primarily working on datasets that are well-known, like Oscar. They are not the biggest ones. We first have to set up the infrastructure, the tools, so we have the full assembly line set up, from

data to training to high-performance computing, so everything works together, and then we can scale it up.

Knowledge-driven approaches seem to be important in OpenGPT-X. Could you give a few pointers to what kind of knowledge will be used?

There are various approaches at how to make the best of both worlds. We have the subsymbolic level in the context of LMs. Europe, especially with the semantic web initiative fifteen years ago, invested a lot of money and resources in the development of what we called ontologies back then and knowledge graphs now. We still believe they have an importance because knowledge can be easily digested by humans, can be easily manipulated. There is quite a bit of research on combining, for instance on how to embed RDF (Resource Description Framework) triples in LMs. We think this is something that could be a good approach or niche for Europe to concentrate upon.

3.3 Cross-Lingual Semantic Search

Nils Reimers, who leads the research team on neural search in the company Hugging Face and is based in Frankfurt, explained that search across languages is challenging and keyword-based search methods will only return documents in the same language as the search query. Luckily, in the past few years, tremendous progress has been made in cross-lingual semantic search based on deep neural networks. Nils' talk gave an introduction on how state-of-the-art transformer networks can be created to allow search across many languages.

The following questions were asked to the presenter:

At some point you mentioned you are calculating similarity and then taking the entropy loss. Was that cosine similarity?

Yes, we used cosine similarity. The most common case is to use this type of similarity or the dot product between two vectors. This gives a score between minus infinity and infinity.

What is the impact of false positives and false negatives? For instance, English "die" and German "die" have a high surface similarity. How do you exclude false positives?

It depends on how your triplets look like. In general, the task of the model is to learn the difference, to learn that these two words, despite being spelt the same, have a difference in meaning. A bigger issue is when triplets are bad. For example, if we design triplets in which the model is forced to say that one of the answers is the only correct one to the question (p) and the second one is the wrong one (n), it hugely confuses the model, because the model needs to find out what the issue with the second option is. Why should it be distant in the vector space? It cannot be due to content because on that level they are the same. Then, the model will try to normalise the sentences by length, by grammar, by syntax. Having clean triplets is extremely important, triplets in which a human would say the first answer is the only positive one and the other answer is a negative one to my question. From the perspective of machine translation, you can have an English sentence (a), a translated Spanish sentence (p), and a Spanish sentence that is close but does not have the same semantics (n). The latter is a slightly different translation of the English sentence but would not be considered as a correct translation of that sentence.

Do you create the triplets manually?

It is a big challenge. Sometimes you can build the triplets from the data itself. For instance, if you look at the website stackoverflow.com (providing developers with the possibility to interact), people rate whether an answer is really good or really bad. Otherwise, an often-used strategy is to apply some kind of lexical search: you have a question and compare it to a large collection of possible sentences or text paragraphs, by measuring word overlap.

3.4 Improving Multilingual Machine Translation with Language-family Adapters

Alexandra Chronopoulou, who is PhD candidate at the Ludwig Maximilian University of Munich, explained that self-supervised learning using unlabeled data from multiple languages leads to powerful models that provide the basis for a wide range of cross-lingual natural language processing models. To create a machine translation system, it is necessary to further train the entire model using parallel data of one or multiple language pairs. This requires vast computational resources. Can a good translation model be created in a more efficient way? Alexandra discussed a novel method of creating translation models with adapters using pre-trained models. The target is to leverage the similarities between languages to permit positive cross-lingual transfer. To this end, adapters are added to the pre-trained model and trained on language families. Alexandra contrasted linguistic and automatic ways of forming these families. She discussed preliminary results on medium-resource and low-resource languages, as well as languages that are not part of the pre-training corpus of the self-supervised model.

The following questions were asked to the presenter:

How do the techniques you describe work for languages with completely different scripts? Is script an issue?

The vocabulary of the pre-trained model, MBart-50, covers a lot of languages, but we are limited by script. For instance, Bulgarian is not in the model, but its script is. Other scripts, like that of Amharic, are not, so you can not use the model to translate to this particular language. However, there are methods to extend the vocabulary of a LM.

How many languages does the model contain?

MBart-50 is trained on 50 languages, but we only used the model for 17 languages, some of which were not included in its training data. We extended the model to new language directions. For instance, a language like Serbian, which is not in the model's training data, is strongly related to many languages in the model. However, some degree of overlap with the languages already present is required.

Is the size of the training data important for clustering?

We only used one thousand sentences per language. It is surprising that even with such a small amount of data you can determine language groups almost perfectly. For instance, monolingual data can be taken from Wikipedia.

Did you perform experiments to find out at which level language families should be grouped (lower or higher)? For instance, in the Slavic family, you have morphologically rich as well as poor languages.

We requested the clustering algorithm to predict three clusters. When requesting it to predict more clusters than that, it clustered languages at a lower level. For instance, it created subgroups of the Balto-Slavic language family. The more you refine the clustering, the smaller and tighter the groups you get. There is definitely room for experimentation here.

3.5 Language Models for Speech Recognition

Denis Jouvet, who leads the MULTISPEECH team at Inria / LORIA (Lorraine Research Laboratory in Computer Science and its Applications) in Nancy, explained that conventional automatic speech recognition systems rely on acoustic and language models to extract the linguistic content of speech utterances. The best performance is achieved when these models are trained on in-domain data. The presentation by Denis focused on two aspects. The first one concerned the training of acoustic and language models from limited amounts of data and the use of uncertain speech recognition hypotheses. The second one concerned the adaptation of LMs that have been initially trained on privacy-transformed data. Privacy transformations modify named entities, typically referring to personal information, such as location names and person names. They therefore lead to a bias for models trained on such data.

The following questions were asked to the presenter:

Does the error detection module use both linguistic and acoustic information or the former only?

It makes use of both. Several years ago, people were mainly using posterior probabilities on the confusion network, which is a graph representing different word hypotheses from the recogniser before picking the best one. The neural network makes use of both acoustic information and LM information.

When dealing with in-domain data for speech, how do you define a domain? What are the most important factors: recording conditions, speaker, the actual content?

Everything that is characteristic to the application you are dealing with. The language, the speaker (native or non-native, this has an impact on the conversation), the acoustic conditions (people holding the telephone close to their mouth or using it hands-free, which introduces more noise). The in-domain aspect for the language part is mainly linked to the application but in terms of the acoustic model you have to consider the speakers and the acoustic conditions for recording.

Do you use standard labels for named entities or are there domains where you make use of specific ones (e.g. custom labels in the medical field)?

It could be any approach. It mainly depends on what the user would consider as personal information, what information needs to be hidden. In our case, we are simply relying on a named-entity recogniser to detect the words or sequences of words that correspond to specific entities, typically a person name, organisation name, location, time, and two or three other categories.

Is the replacement strategy also language-dependent to a certain extent? For instance, if you replace a name by another one, do you use a closed list of language-specific elements or do you use a more general resource?

You have to define ahead through statistics on some corpus the set of words that corresponds to a given entity in order to pick them for replacement. If we are anonymising dialogues, it is better to keep the same replacement to make the new dialogue somewhat consistent. If the word occurs again in another sentence of the same dialogue, we should use the same replacement.

3.6 Discussion based on morning session

The project representative and some speakers ask questions to the other attendees.

How do you see the evolution in the availability of training data? A lot of data is text-based now. There are also other modalities, but there is an unbalance.

For speech, you can use an unsupervised model called wav2vec. It helps in improving the performance of an application with a limited amount of training.

LMs are trained on a large amount of data. What happens when the input contains words not observed in the training data?

There is recent work that extends existing models, using an approach that gradually adds languages and extends the vocabulary. If you add more and more scripts, at some point it converges, i.e. the vocabulary remains more or less the same.

In morphologically rich languages, some inflections are not observed in the training data. How to deal with this?

If the text is in a script included in your training data, you can create zero-shot translations.

Can transliteration of data help to cover the gap between different languages?

There is some work on that, but, in general, it is not used in large models. If you do not have a script in your vocabulary, transliteration may help.

3.7 Large-scale, Open-access AI Models for Swedish

Love Börjeson, who is Head of KBLab at the National Library of Sweden, described the training of large-scale, open-access AI models based on the contents of the library, which holds the largest collection of Swedish data across modalities, i.e. text, sound, images, video, and games. Most notably, KBLab trains transformer-based LMs. They serve as a way to transfer the full potential of KB's collection to the Commons, thereby contributing to the digital transformation of society, and ultimately supporting high-quality research and democratic development. However, to be able to train such models, KBLab needs access to high performance computers (HPC). KBLab is the first public administration in Europe that successfully applied and received access to the EuroHPC JU systems, thus putting Sweden on the front line.

The following questions were asked to the presenter:

How is the general model specified, merely by hand annotation?

General models are trained with an algorithm that tries to mimic the cognitive process of attention. Given large text corpora, the training consists of taking away words, upon which the model should try to guess what word was taken away. This process is repeated many times. Another training objective of a standard BERT model is to guess which sentences go together in the same paragraph. This training process is unsupervised. You can tweak some parts, for instance the batch sizes. That pre-trained model does not have any useful capacity in itself. Therefore, it typically has to be further trained downstream, by providing it

with a dataset, for instance typical questions and typical answers. The model will recognise not only specific questions but also the general structure of a question. Since it also knows the general structure of an answer to a question, it can generate answers. This approach has been used in life sciences, they have taken our model and fine-tuned it on a hand-annotated Q&A dataset, in the context of a clinical process. The pre-trained model is slowly adjusted to new knowledge, i.e. fine-tuned.

Is your speech data hand-labelled? Could you tell more about your wav2vec model in general?

The general wav2vec model is trained in an unsupervised way, which is a huge leap forward, i.e. you do not need text running along with the speech. Similarly as for a pre-trained text model, you take away a bit of the sound and the model has to guess what sound was taken away; this is the main training objective. Again, the general model cannot really do anything. We have fine-tuned it for speech-to-text recognition, using sound coupled with transcribed text. The data for fine-tuning only comprises a small number of hours, whereas the ground model is trained on thousands of hours of Swedish speech. We used sound booths for the fine-tuning data.

National libraries in Europe do not own whatever is hosted, as it belongs to the people who deposit their data. How do you handle that?

Indeed, for most of the data, the copyright belongs to someone else. Different national libraries have a different take on this legally. There is a whole range, from libraries who think their legal situation allows them to train data in the cloud (not commercial cloud solutions) to libraries that think they cannot do any text mining. We are somewhere in the middle. A little bit of a risk management is involved.

How many hours of transcribed text have you used for your Swedish system?

The ratio between unlabeled and transcribed sound data is 1/1000, i.e. for 1000 hours of unlabeled data you need 1 hour of transcribed data. That is surprisingly little. There is great variation of Swedish in the ground model because we used local radio broadcasts, in which people from all over Sweden are calling in. However, when it comes to transcribed data, the variation is much smaller. Therefore, the model performs well on the biggest varieties of Swedish, but for certain dialects it does not work well.

3.8 A New Generation of Neural Search and Knowledge Discovery Tools

Jakub Zavrel, an AI researcher, technologist, and entrepreneur, who founded the company Zeta Alpha (Amsterdam) in 2019, explained that, for knowledge discovery in expert domains like AI R&D, a new generation of search technology based on large LMs (also known as neural search) presents a number of distinct advantages. First and foremost, it bridges the lexical gap between the searcher and the document collection, and also enables long complex discovery queries based on natural language. Jakub's presentation outlined how neural search is applied in the Zeta Alpha system and how it benefits query by document in the context of recommendations and in visualisation of large document collections. In addition, the presentation outlined recent research at Zeta Alpha on domain adaptation using GPT-3 and on multilingual search.

The following questions were asked to the presenter:

How much more does your solution cost with respect to the standard search indexation?

First, for small companies, a search solution would not be possible if we would not have an ecosystem in which pre-trained LMs would be generally available. It is unaffordable to train large LMs from scratch and reinvent the wheel. Second, we do a lot of fine-tuning using cloud infrastructure like Amazon or Azure, which is somewhat expensive but not prohibitively. You can fine-tune a BERT or T5 style model in a few days of GPU on Amazon for an order of magnitude of a few hundred euros. This is definitely affordable for smaller companies. Third, the quality/price ratio depends on the number of users, but it requires your whole infrastructure to be GPU-enabled. In a niche application (currently we are serving a few thousand users) it is not a massive undertaking, but it is definitely more expensive than just having traditional small CPU machines.

How do these models compare to semantic graph based search in terms of precision?

It depends on what you are searching for. There is no single right method for search. It might be that there are questions which are well served by knowledge graphs and include a lot of inference over the graph. If somebody is just looking for a title, keyword search or navigational queries may be better. We started out our work with the idea to combine neural, classical and knowledge graph search. We thought that we could keep up with the dynamics of modeling concepts in the domain of AI and data science, but the evolution is going so fast that we gave up. There is added value in having knowledge graphs, especially for aggregation and statistics, but for the type of knowledge discovery that we are currently seeing with our users, it only involves a small fraction of the queries. And precision depends on the exact question and the coverage and quality of your knowledge graph.

3.9 Legal Aspects of Language Models

Dr. Khalid Choukri, who has been the CEO of the European Language Resources Association (ELRA) and the Managing Director of its distribution agency (ELDA) since 1995, and Mickaël Rigault, who has been legal counsel responsible for matters on intellectual property and personal data related to Language Resources' collection and distribution since 2020 at ELRA/ELDA, explained that, beyond the technical questions that are well documented among the community, legal questions remain concerning LMs and doubts over what can be done may hamper current and future research. Their talk provided a few major points for discussion

around legal issues linked to LMs, whether they are copyright issues or issues related to the compliance with data protection regulations.

The following questions were asked to the presenters:

How hard is it to crack the data behind the models? Normally, when you have access to the LM, it does not necessarily entail that you access to the language data behind it.

In the legal community, the word "copy" in copyright is the most important. As soon as you copy data, and you derive models, even if they do not allow reconstruction of the original data, you have already infringed the law. You have taken someone's asset, you have used it for some purpose, you may not even compete with the initial owner, but you have taken something that belong to someone else. Under EU regulations, this is not allowed. In the US, you have the fair use doctrine, where you can take things for research purposes as long as you do not compete with the owners.

Can you provide any advice? What kind of license should be used when you want to make available a model?

You can ask advice to the legal helpdesk of ELRA.

There is the issue of copyright and the issue of privacy. Sometimes they are intertwined, but they are two different issues when building a model. In terms of copyright, we need reasonable control. In terms of privacy, people can retract their consent. We need practical rules: how to be as efficient and, if possible, as competitive as the US, or at least not be blocked from the legal point of view? In terms of privacy, what about repurposing of the data? An LM is a kind of data compression, so you cannot regenerate the data, for instance you lose track of privacy-related information, such as voice characteristics; you cannot recreate a voice from a model.

LMs and other statistical models consist of numerical figures. But to get the numbers, you need the private data, hence you need consent and a legal framework. Assuming you have all the rights, even if people withdraw their data, you have a skeleton of the model. You can then still translate information with personal data from one language to another without referring to the private information that you have on the people that were in the training data.

Assume you have set of data for which copyright and privacy aspect have been cleared: people have given you consent on paper to use their private information to build a model. You can use the data for purposes foreseen when people have given their prior consent. But if a user then does not want his/her data anymore for use in a model, the cost for removing this data is prohibitive.

12 years ago, ELRA conducted a study for DG Translation, together with two lawyers, one from academia (CNRS) and one practitioner at a court. They had two completely different opinions:

1. People knew their data is going to be incorporated in a big platform. If they want to be removed from the data, they need to cover some of the expenses.
2. People gave the data for a given purpose. If you go beyond that purpose or if they change their mind, you have to comply with this newer request.

We should check with LM experts. Data on one person should not have impact. But if you want to be fully sure, retraining model without this data is necessary.

It is easy to remove information on one person from the data, but not from the model. Retraining a large LM is very expensive, and it is has no significant statistical impact. Would a court allow for retraining when the cost is very high?

There are dispositions in the GDPR to ensure that a data producer does not need to take away information on a person from the database if it provides proper evidence that is either technically not feasible or too prohibitive in costs.

The possibility of automatically anonymising data may help to convince people that the model contains very little confidential information. If your training data contain named entities, such as person names, there is always a risk that they pop up in the output of the system. You should find a balance here.

Technically, you can find a balance, but legally, it is black and white. The definition of anonymisation is more pseudo-anonymisation in this context.

Note by an EC representative: a challenge for governmental institutions is to comply both with the GDPR and with the PSI (Public Sector Information) directive. All data that has been produced should be shareable, with the limitation of the privacy statement. The challenge is to find a statement that most of the people agree with.

3.10 Final discussion

While a 30-minute slot was provided at the end of the workshop for discussion between the presenters and other attendees on any of the topics presented during the day, the discussion following the presentation on legal aspects of LMs, the gist of which is provided in the previous section, was very extensive. Therefore, the organisers decided to consider this discussion as the final one, and to move on to the conclusions.

4. Conclusions

The workshop *Large language models: pre-training with a twist* focused on models that incorporate information useful for understanding a language and trained using deep learning algorithms, large amounts of data, and heavy computing infrastructure. These pre-trained LMs can be tuned towards specific applications with much less data and computing power. Subtopics covered in the workshop were LM repositories, multilingual LMs, use cases, and legal aspects. A substantial number of people registered for the workshop (around 100), of whom about 70 percent participated. The audience included staff of EU and Member State administrations and representatives of academia and industry.

The speakers had various backgrounds (academia, industry, public-private partnership, public administration). They discussed themes like LM development in Europe, LM projects and repositories, cross-lingual applications (search, translation), data sparseness, training efficiency, modalities, anonymisation, visualisation, copyright issues, and issues related to the compliance with data protection regulations.

The discussions between speakers and audience related to a variety of subjects, such as language selection, data selection, vocabulary of LMs, supported scripts, knowledge graphs, morphologically rich languages, and computational and financial cost involved in fine-tuning pre-trained LMs.

Take-home messages include the following:

- The introduction of neural networks has improved several AI fields to a large extent. For instance, in the field of search, it can expand the horizon of explorations compared to what has been done using keywords.
- Low-resource languages can benefit from the combination of languages into multilingual models.
- Training pre-trained LMs requires specific infrastructure, high-performance computing, and of course large amounts of data. It is important to know what data is available and what quality it has.
- It is interesting to consider a pan-European approach to LMs. What kind of computational resources could be provided and what kind of datasets could be made available and shared between organisations?
- Because of the GDPR, ways should be found to deal with personal data in the right way.