Introduction

Extracting all necessary metadata from unstructured reports has plenty of advantages for O&G organizations. Unfortunately, the building of RDBMS architectures has some limitations. The attempt to index extensively the documents on arrival or before their archiving is never completely achieved due to a shortage of time and resources. The quality of the indexation is difficult to evaluate when the information is not sourced. The deployment of innovative acquisition technologies and field layouts means that the database design has to be revisited frequently to accommodate unforeseen changes to data hierarchies.

One of our customers faced this type of issue with a relational database designed to store seismic design parameters. On a representative sample of it, we observed that only 78% of the more important targeted metadata was collected despite their presence in the documents, and in addition 7% of them were not correct (i.e. 7% of false positive values in data scientist jargon) despite having been extracted by skilled technicians.

Therefore, we proposed to test our solution to check if:

- A ML system can extract more metadata than a human based document indexing process
- The extraction quality could be better (less false positive values)

While being able to:

- Qualify the confidence of the extracted metadata (qualified information)
- Establish active links between the extracted values and their location in the source documents (sourced information)

Principles of metadata extraction by a machine learning system

Our ML system manages two different processes: a training process to create the learning models (one per targeted metadata item) and the indexing process to extract qualified and sourced metadata under the guidance of the LM. Most of the tasks are similar for the two processes. Some, display as a square in fig 1 are computed in a parallelized IT environment easy to scale according to the volume of documents to index. In addition, a Graphic User Interface allows to prepare or update LM.

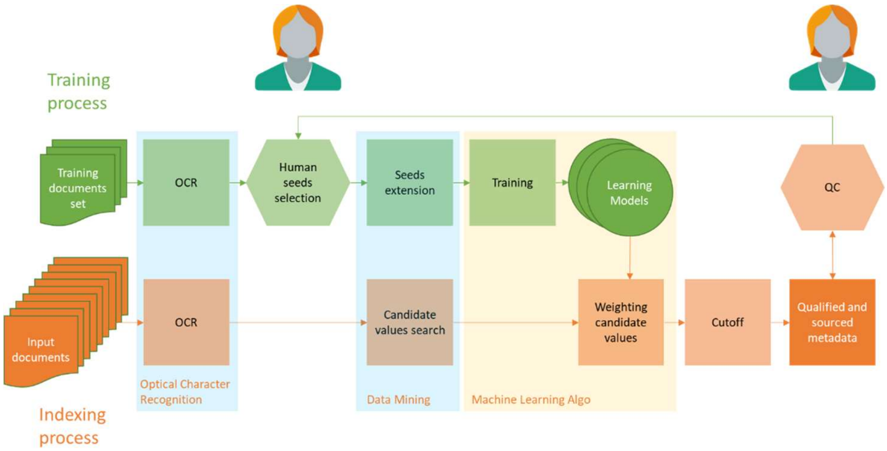

Figure 1: iQC machine learning workflow.

- **OCR (Optical Character Recognition).** Most of the tasks of the training and indexing processes are similar. Both processes start by an OCR we run using a parallel processing environment. The OCR creates an 'image' of the original document that will be processed.
- **Seed selection and seed extension.** Once the documents have gone through the OCR step, the "trainer" can start tagging positive and negative examples of metadata value, the seeds.
  To limit the training effort to 100 to 300 tagged values per metadata, an algorithm extends automatically the number of seeds using the metadata properties. As an example, if the user has tagged the value 12 as the positive example of the number of streamers for a particular survey, all the other integers of the document in the range [1;30] but different to 12 will be considered as negative seeds. Conversely, the other occurrences of the value 12 in the survey report will be considered as potential positive seeds and weighted according to their context.
- **Creating the learning models.** A Learning Model is created for each targeted metadata item. Each LM is a collection of quantitative features characterizing that particular item. In our particular case, the features are the average distances in the document between the metadata item's values and key words frequently occurring in the surrounding space.
- **Selecting the metadata candidates.** Once the LMs are created, the system is ready for indexing a first batch of documents. The next task after the OCR is a search of metadata value 'candidates'. This search is guided by the metadata properties. Each metadata can be defined as:
  - Integer range. Example: number of streamers, number of sources …
  - Floating point range. Example: shot interval, source volume …
  - Dictionary. Example: the client name, the seismic contractor name, the streamer vessel name …
  - ReGex. We use this category for text metadata values which cannot be summarized by a dictionary. In such a case, the search for candidate values can be more or less constrained by a "regular expression".
  - Category. This type of metadata represents the documents' class according to a given taxonomy.

- **Weighting the candidates, selecting the best ones.** All identified candidates and their surrounding context features are submitted to the learning model in order to 'weigh' their confidence factor. Only the best candidates for each document or survey are selected. If necessary, a defined cut-off for the confidence factor may be used to ensure that most of the false positive values are rejected.
- **QC and Learning Models updating.** The QC is made easy since a confidence factor is associated to all metadata values. Using this factor, the trainer can focus on the ambiguous extractions. He has the possibility to validate, edit or reject the results. Doing so, he doesn't only update the indexing results, he also adjusts the seeds, giving the possibility to recreate a new version of the training model.

The training curve

The possibility of editing, validating or refuting the results allows the ML system to learn with experience. Similar to the training of humans, it is possible to measure the training progress by asking the machine to go through some benchmarks.
In the case of our pilot performed for the indexing of seismic acquisition documents, the measurement of the learning curve was done using the holdout method. In this type of simple cross validation method, the set of documents provided by the customer to build the model is divided into three subsets:

- **The training set.** It is the largest subset (70%) used to tag the positive and negative seeds on which the learning models are built.

- **The benchmarking set** is another subset of the documents provided by the client to assess regularly the performance of the model built on the training set.
- **The blind test set.** To avoid the risk of bias due to the knowledge by the trainer of the benchmarking set, the last evaluation is done using a unknown set of documents.

During the pilot, the benchmarking has been done 12 times on each individual learning model and the number of extractions and the percentage of true positive values per metadata item were averaged to produce the learning curve of the indexing of seismic acquisition reports (fig 2).

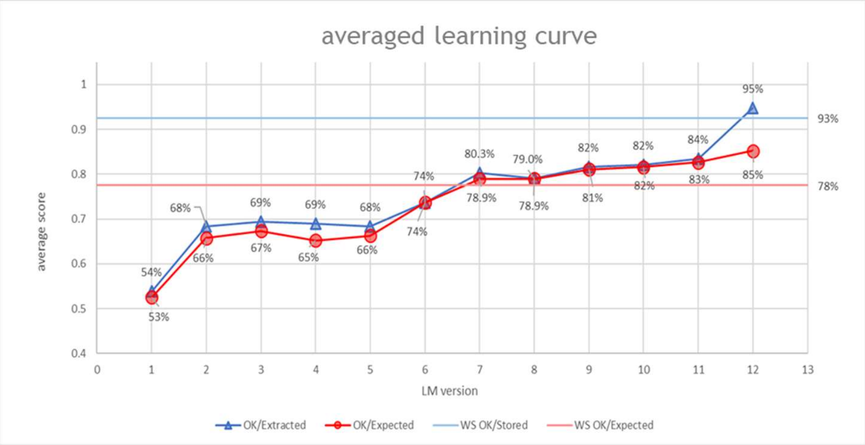

Figure 2: Automatic indexing training curve for seismic acquisition reports

The averaged result illustrates a progress of the indexing quality after just about every loop. Finally, the machine has been able to extract more metadata than was achieved by the human operator on the same benchmark dataset (85% instead of 78%) and to extract less wrong values (false positive) than the human operator (95% of true positive values instead of 93%).
But the average quality factor doesn't illustrate the variability of the training progress for each of the targeted metadata, as it could be observed for some of them in figure 3.

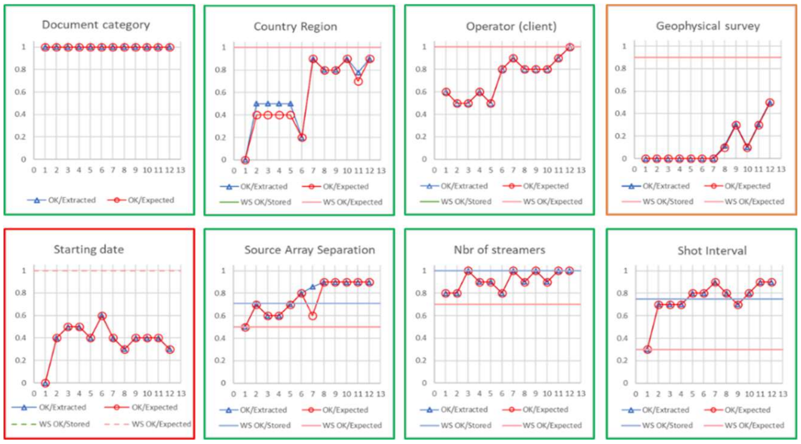

Figure 3: Learning curves of some targeted metadata

The ease of training of the machine was not the same for all types of metadata. The easiest one was the detection of the document category (Acquisition report, QAQC report, HSE report, Navigation report ...). Since the first version of the learning model, the detection was done without any error and remained stable all along the training.
The improvement of the training for the metadata associated to dictionaries goes with the dynamic dictionary expansion.

We have been surprised by the good result of the detection of the numerical values. The machine avoided rapidly some errors made by the technicians when detecting the distance between the sources or the shot interval, and therefore produced a very low amount of false positive values.
We got some issues training the machine on some metadata frequently obvious for the human like the survey name (geophysical survey). This comes from the fact that within the same document, the survey name may be written in various ways. This point makes the seeds extension a little bit hazardous. Nevertheless, a training effort can solve this issue as the shape of the training curve shown in fig. 4.
The only metadata for which the training curve didn't show some progress was the survey starting date survey. After analysis, it appeared this was due to the training strategy: The trainer considered any date as the mobilization date, the date of first shot point, the date of kick-off were some valid seeds for the starting date. This approximation multiplied the number of possible contexts where the metadata could be located and made the LM inconsistent. Just like humans, machines can make errors after poor training!

Conclusion

A generic ML system for text detection has be applied with success to automatically index documents related to seismic acquisition. This automatic indexing, faster than a manual indexing and much easier to scale up thanks to the parallel processing architecture, has proven to be equal or better in terms of quality compared to the indexing done manually: it extracts more information and produces less false positives values. In addition, all extracted information is evaluated in terms of confidence and it is also accurately sourced (located) within the original documents. The indexing is easier to QC and more trustworthy for the end user.
This result opens the possibility to use the large volumes of technical documents archived by O&G organizations for data analysis faster and at a substantially lower cost. In addition, the information extraction doesn't need to be done at the time of archiving but can be done at the time of data consumption, when there is a purpose to the data extraction exercise. The machine learning systems are agile enough to rapidly produce efficient learning models for specific application.

References

Blinston, K., H. Blondelle, 2017, Machine learning systems open up access to large volumes of valuable information lying dormant in unstructured documents: The Leading Edge. March 2017, p257-261

Juneja, A., J. Micaelli and J. Johnston, 2017, Method and system for extracting, verifying and cataloging technical information from unstructured documents: US patent 20170169103 A1, www.google.com/patents/US20170169103

Su, F., et al., 2015, Attribute Extracting from Wikipedia Pages in Domain Automatically *in* V. E. Balas, L. C. Jain, X. Zhao, eds., Information Technology and Intelligent Transportation Systems: Springer International Publishing, 433-440.

Vapnik, V.N., 1999, An overview of statistical learning theory: IEEE Transactions On Neural Networks, **10**, no. 5, 988-999, http://web.mit.edu/6.962/www/www_spring_2001/emin/slt.pdf.

Zhong, B., J. Liu, Y. Du, Y. Liaozheng, and J. Pu, 2016, Extracting attributes of named entity from unstructured text with deep belief network: International Journal of Database Theory and Application **9.5**, no. 5, 187-196, http://dx.doi.org/10.14257/ijdta.2016.9.5.19.