The use of artificial intelligence techniques to support control activities

Luís André Dutra e Silva is an employee of the Federal Court of Accounts – Brazil. He has a B.A. in Computer Science from the Brasilia University Center (UniCeub). He has certifications in Software Engineering from IEEE and Project Management from Stanford University.

ABSTRACT

Cognitive services are an alternative based on artificial intelligence, with the purpose of obtaining solutions that are capable of detecting any kind of patterns in texts, images or any other source of data. This article describes the experiments on the use of artificial intelligence with unstructured data carried out by the TCU. The project stages are detailed and linked to the future applications and possibilities. The results of these experiments were very promising and we intend to adopt the techniques used during this period in new technological initiatives for the TCU and for the Public Administration in general.

Keywords: Cognitive services. NLP. Artificial Intelligence. Machine Learning. Text mining.

1. INTRODUCTION

A big part of the information consulted and produced by the TCU (Federal Court of Accounts) is received from bodies under its jurisdiction, registered in reports, votes, rulings, orders and other documents. These records are textual and complex, and demand sophisticated interpretation resources to obtain linguistically represented knowledge, especially because the data is unstructured. This characteristic demands countless analyses and combinations to explore and

add value to the information and to the decision-making process.

This fact entails considerable effort from the Federal Court's employees in order to structure these data and to systematize knowledge for use and decision-making. An example of this is the effort required for the performance of certain management and control activities, such as monitoring of TCU's deliberations and the classification and sorting initiatives of the Special Rendering of Accounts (*Tomada de Contas Especial* – TCE) cases.

In this context, the use of tools and algorithms supported by machine learning models for automation of document interpretation turns out to be essential and strategic for classifying and automatically extracting information contained within unstructured data sources.

The artificial intelligence techniques explored and systematized in this context allow the machine to learn more complex characteristics of concepts present in different documents. In view of this, the intention is to structure information that was initially scattered around different documents and formats and make it available and useful.

For information purposes, the development of initial experiments using techniques of Deep Learning has shown great promise. The prototype was executed during the period from July 1st to December 31st, 2015, and 257 thousand rulings from 1993 to 2013 were used as a training basis in the performed concept test, which classified deliberations contained in the text of 5,300 rulings delivered between 2014 and 2015. The result obtained revealed an average accuracy of more than 96%.

2. DEVELOPMENT OF WORK

Given the exploratory and pioneering nature of utilization of artificial intelligence techniques in the TCU context, the development of the work was divided into stages of experimentation with initial focus on cases of Special Rendering of Accounts. Given their experimental nature, many of the analyses performed were dismissed so that the best solution could turn up. With this in mind, the idea of this work has always been to replicate it in other control objects and business challenges.

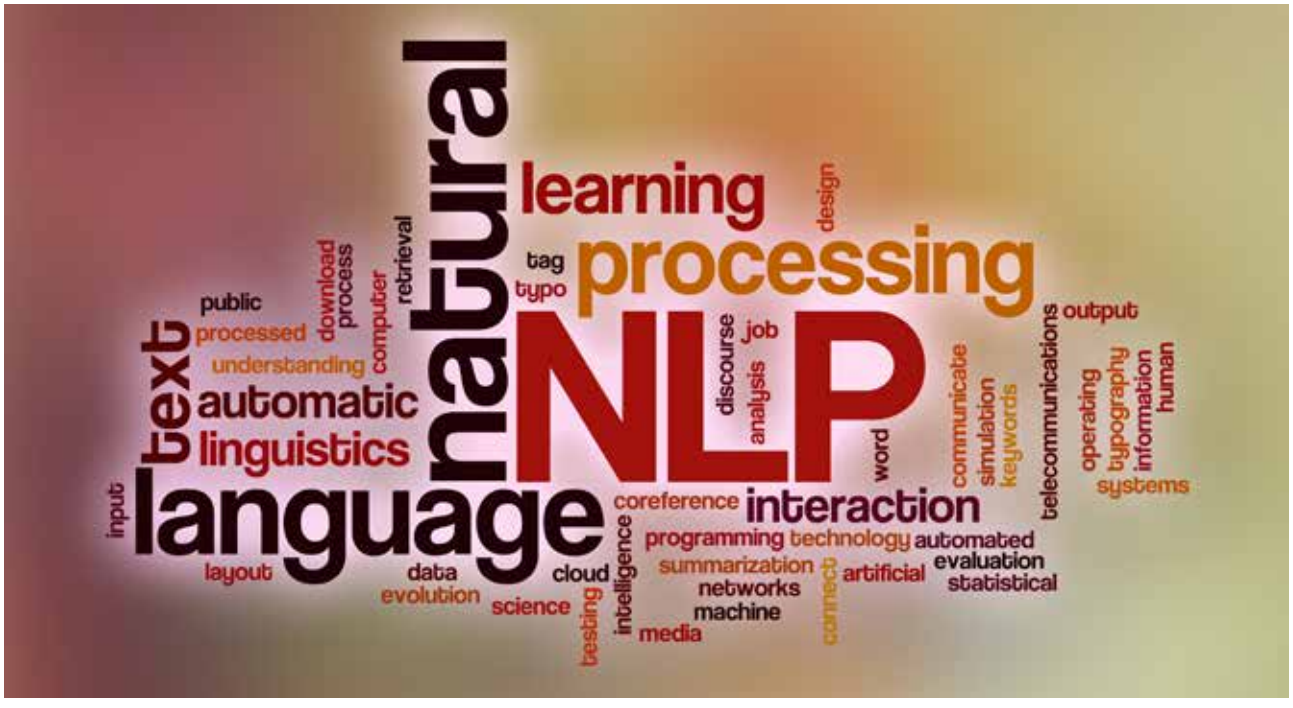
The first stage consisted in obtaining the documents of Special Rendering of Accounts cases directly from the EDM (Electronic Document Management) repository. The purpose of this was to establish a local base for future mining of the TCE case documents. The number of case documents used in this stage was around 680 thousand. To extract them, an algorithm in Phyton language was used to access the development database and the EDM in production, with the purpose of downloading all items into folders related to each TCE case. This stage was carried out between February 20 to 24, 2016, and the main challenge faced was the frequent interruption in the download of the case documents, which occurred due to connection issues. In addition, at first it was not possible to download all

documents. In the end, however, all documents of all TCE cases were downloaded.

In the second stage, the TCE's Registration database in Apex and the CGU's (Comptroller General of the Union) TCE Registration database were obtained, as well as the CGU's report documents in Microsoft Word file format. This was done with the purpose of composing the database associated with the TCE cases to perform text mining on the documents obtained. The number of documents used in this stage was around two thousand items, and an SQL (Structured Query Language) script was used for the databases dumps and for access to the CGU's shared data through local network sharing. The actions of this stage were performed between March 2 and 3, 2016, and the main challenge faced was that not all of the CGU's documents in Word format had matching metadata in their registration system. The system's registration data in Apex matched a small fraction of the liability present in the TCE cases. Due to these actions, the downloaded data became unavailable for further analysis.

In the third stage, the documents in TCE cases were grouped according to the government function and the types of irregularities found in the registration system in Apex. The purpose here was to attempt to classify automatically the non-annotated case documents, which comprise most of the cases, aiming to speed up sorting and distribution of the cases. The amount of case documents used was around 680 thousand items, analyzed through neural networks of the Self-Organizing Maps or Kohonen Maps type, using the free software Somoclu in a Linux environment. These actions were performed between March 4 and 10, 2016, and the main challenge faced was the fact that a large portion of the analyzed documents demanded several consecutive days of processing. Furthermore, the proportion between the number of documents that had previous classification and the number of documents without classification would be too divergent to obtain a useful result. Notwithstanding the difficulties encountered, it was possible to classify a great part of the documents by government function based on the registration obtained from the Apex system, which was used, until then, for the TCEs.

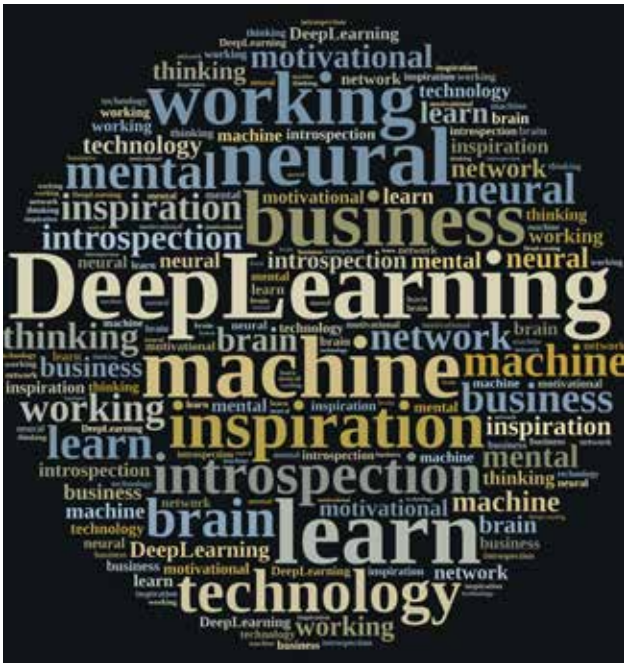
The fourth stage consisted in analyzing the irregularities in the cases of the Healthcare government function, aiming to classify the cases, sent in Excel spreadsheet format, that were pending distribution. Around 3,146 case documents were analyzed through the clustering technique using Self-Organizing Maps. This stage was carried out between March 11 and 21, 2016. The main challenge faced was the fact that, due to the small number of cases previously classified in comparison to the existing total and to the diffuse configuration of the irregularity classifications, the result would probably not be very accurate. Thus, of these 3,146 cases identi-



fied in the Healthcare government function, all of them would have irregularities types II, III, and IV, and most of them would have type V. None of these 3,146 items would have the type I irregularity, which, according to the specialists, should be detected in at least 10% of the cases.

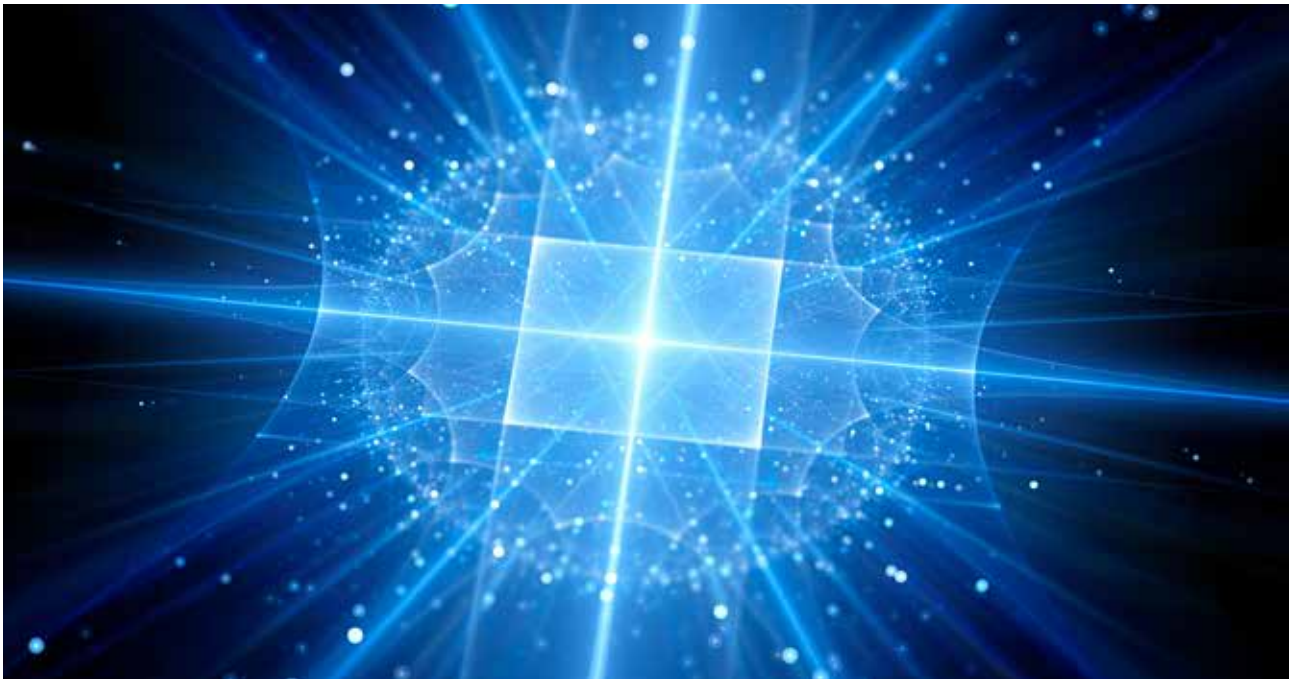
The fifth stage consisted in creating the deliberations extraction service. The objective was to make available a web service to extract deliberations from the texts of court rulings, aiming at performing the semi-automatic insertion of the Determination to Agency/Body deliberation types into the Radar system, using the Apex interface,. Around 200 rulings are delivered per week and, to analyze them, Deep Learning in Python was used along with the Theano/Keras framework, POS tagging using the Aelius free library, regular expressions and the Flask framework in Python with the Werkzeug infrastructure. This stage was carried out from March 22 to 28, 2016. During its execution, the challenge was the existence of too many linguistic ambiguities to be dealt with through specific rules when extracting attributes of types of ruling deliberations. A few of these ambiguities need some kind of text comprehension that is at a level close to that of human intelligence. Nonetheless, the largest part of the Determination to Agency/Body deliberation items is identified with over 96% of accuracy and, according to the specialists, their feature are also extracted with high precision. Therefore, the type of deliberation chosen to demonstrate the usefulness of this service is considered the most adequate to provide a high level of success in the initiative to automatize the extraction process.

The sixth stage was dedicated to establishing a correlation between the texts of opinions given by the TCU contained in the TCEs and the 37 standard paragraphs, corresponding to each theme, available as aid on the website. The number of documents used was around 5,997 TCE opinions, and the techniques used in this approach were the ones of lexical and semantic similarity, using the Levenshtein edit distance algorithm and the cosine between the Bag-of-Words vectors. These actions occurred between April 1st and 7, 2016, and the main challenge faced was the high level of similarity between each theme and the opinions regardless of the chosen theme. Thus, it is too difficult to distinguish precisely the presence of standard paragraphs in each analyzed opinion. Therefore, the result was the fact that the percentage of similarity found for each opinion/paragraph, as indicated by the TCE specialists, was unsatisfactory because the similarities were too close to



each other. Therefore, the experiment was considered concluded and this approach was hence dismissed.

The seventh stage was an experiment whose purpose was to analyze the structure of the CGU reports, and determine which information could be extracted and structured and apply text correction algorithms with a high level of noise derived from OCR (Optical Character Recognition). Other purposes include elaborating an information extraction algorithm similar to the one used for the Deliberations Extractor; providing a validation interface for the TCE specialists; defining classification and routing criteria based on the extracted information; classifying and suggesting routing through the validation interface; and validating the classification and routing by the specialists. The number of documents involved was around 3,800 reports, and the technique used was information extraction using regular expressions and crossing of structured data based on the entities extracted from the documents. This stage was carried out from April 8 to 26, 2016, and the main challenge faced was the fact that the quality of the scanned documents was not sufficient to ensure extractions that matched the same level of reliability achieved by the reading performed by the auditors that analyze TCE cases. In the end, this stage rendered several results. With the help of the TCE specialists, we determined what information would be structured and extracted from the CGU reports in Word format and extracted information from these reports making it possible to fill in with precision an opinion draft for the irregularity "Omission in



the duty of rendering an account". In addition, the crossing of extracted information with structured systems was performed, such as the Sisobi (Death Control System), the Sincov (Administrative Agreement and Transfer Contract Management System), and the Siafi (Federal Government's Financial Administration System). In addition, the same extractions performed in Word format documents started being used for the PDF scanned reports, which have less quality due to the OCR process; and a validation interface of the extractor was provided to the TCE specialists. The extractions performed by this experiment were considered useful for the TCE inventory management. According to the specialists, the task that is performed manually to this end can be partially automated, thus reducing the time needed to distribute/sort the TCE cases. However, the specialists considered automatic generation of drafts troublesome because it presented data that do not completely fulfill the needs of the auditors that analyze TCEs.

In the eighth stage, carried out between April 27 and May 25, 2016, the first deliveries had the purpose of providing the first versions of the services stipulated in the project's schedule. According to the schedule of the cognitive services creation project, the following products, in their first version, are available for validation and access: REST service for extracting deliberations of rulings; Deliberation Extractor validation interface; REST service for extracting TCE reports of entities; TCE Content Extractor validation interface.

The ninth stage, carried out between July 8 and 15, 2016, had the purpose of improving the scanned documents by using outsourced services. A sample of ten reports with major scanning issues was used, and the Google Vision API was employed for this attempt. The main challenge was to improve the quality of the text extracted from the TCE scanned documents without resorting to a new scanning process. However, in terms of the amount of errors, the achieved result was no better than the OCR software used in the TCU context (Adobe).

The tenth stage, performed between July 16 and 25, 2016, had the purpose of developing the first version of the recognition service of mentioned entities, NER (Named Entity Recognition), which is capable of extractingfrom any text the names of natural persons, legal entities, CPF (National Identification Number), CNPJ (National Registry of Legal Entities), and normative references. The number of documents used for the training of the entity recognition neural models was around 58 thousand rulings and the Amazon base of the University of Lisbon, which contained the manual annotations of each type of entity mentioned. To that end, the machine learning framework Apache OpenNLP, in Phyton and Java, was used with manual and automatic annotations in texts of rulings. With the use of machine learning algorithms instead of preset rules, the accuracy and the other associated metrics may not be ideal if there is not a large amount of manually annotated texts. The first

version of the generic entity extractor was exposed as web service in the production infrastructure. Seeing that an untrained validation base of approximately 10 thousand sentences was set apart to assess the quality of the NER web service, a FI score of around 81% was obtained for natural persons, which reflects the state of the art in terms of extraction of mentioned entities. The F1 score is the most adequate to demonstrate the balance between accuracy and recall, which should be the pursued goal.

In the eleventh stage, carried out between August 1 and September 14, 2016, there was the development of the second version of the generic entity extractor in pure Java, locally using the OpenNLP library and the previously trained models for the first NER version. The main challenge was the fact that portability from code Phyton to Java is not always possible in many customizations. Therefore, this initiative may not have been very successful if one or more frameworks used in Phyton did not similarly exist in Java. Nevertheless, the NER service, according to the standard TCU service architecture, Reference Architecture 8, was developed and is currently in production in the JBoss 6 EAP environment.

The twelfth stage, which occurred from September 16 to October 11, 2016, aimed to construct a service capable of detecting possible material errors in rulings delivered by the TCU before they were formalized. Around 600 documents were used to test the tool chosen from among one thousand rulings - which an inspection by the TCU's Internal Affairs Office pointed out as having material errors in their elaboration. The result achieved by the cognitive service was a capture of 40% of the material errors present in the rulings, after verifying the correct spelling of the names of the responsible parties, non-existent CPF and CNPJ or ones belonging to deceased natural persons or to inactive legal entities.

The thirteenth stage, which is ongoing since October 18, 2016, aims to provide a service of automatic extraction of ontologies in OWL format, of text documents, using machine learning algorithms. Around 280 thousand TCU precedent documents are being used. The Latent Semantic Indexing technique is being used by decomposing the matrix W of documents x concepts, which is obtained by the Bag of Words technique and the TF-IDF normalization, using Singular Value Decomposition. Latent Semantic Indexing is a statistical method that connects terms in a useful semantic structure, with no syntactic or semantic analysis and no manual intervention. By using this method, each document is represented not by terms, but by concepts that are truly and statistically independent in a way that the terms are not.

3. CONCLUSION

The cognitive services that were developed can be used in a successful way by other systems, with the purpose of improving the TCU and the Public Administration work processes. The need to structure texts that are produced continuously and depend on the classification and extraction of the information contained in these unstructured bases is inherent to the work of the TCU and of other bodies.

Therefore, making these services available may give more accuracy to the work of analyzing cases and make available to providing External Control professionals contexts that are relevant to every work developed in textual form.

For example, while producing a certain report all precedents on the topic could be automatically related to the text in creation, thus simplifying the process of searching for information relevant to the content being elaborated. Another example of use would be the automatic elaboration of summaries of texts received from bodies under TCU jurisdiction, which would speed up the processes of analysis of these external documents.

Moreover, by means of cognitive services, customized clippings can be elaborated about each subject addressed in a certain work. Thus, the auditors would always have the most updated information to support the elaboration of audit reports and all the other necessary documents.

REFERENCES

NOVELLI, Andreia, OLIVEIRA, José. Simple Method for Ontology Automatic Extraction from Documents. International Journal of Advanced Computer Science and Applications. Vol 5. No. 12. 2012

MADDI, Govind, VELVADAPU, Chakravarthi. Ontology Extraction from text documents by Singular Value Decomposition. ADMI. 2001

BERRY, Michael, DUMAIS, Susan, O'BRIEN, Gavin. Using Linear Algebra for Intelligent Information Retrieval. SIAM Review. Vol 37. No.4, pp-573-595, December 1995.