Automated Context and Text Analytics by Applying Cognitive Language Processing Tools

Ambruzs Csaba, Herczeg Dominik, Dobos Zoltán, Dr. Hajnal Éva

Óbuda University, Alba Regia Technical Faculty
ambruzs.95@gmail.com
herczdom96@gmail.com
hajnal.eva@amk.uni-obuda.hu

Abstract—**IBM Watson** [1] **is a Cognitive system, what can analyze and interpret data - similarly to a human being - including unstructured text, images, audio and video, can learn and reason. This system has several capabilities (accessible through interfaces), which allow to execute context based natural language processing and interpretation. The purpose of the project is to provide support for human by significantly minimizing the effort, which is currently needed to analyze and understand large volume of unstructured audit text. IBM Watson gives an opportunity to identify IT risk factors, and compliance problems automatically, finding trends, and provide solution to improve problem detection while it helps to decrease the faults from manual human processing and improves efficiency.**

In this article, the authors propose a method, which utilizes advanced natural language processing by using cognitive systems. Need to emphasize that the activity had to be carried out in an IT service management specific language area. The possible structures of dictionaries were investigated to adapt best the natural language processing capabilities and the required categorization, also the necessary pre-processing actions were reviewed. Within the dictionaries hierarchical mapping of the categorization levels (related to the IT risk and compliance area) is built up. Furthermore, the optimal combination related to the usage of nouns and verbs is determined to achieve higher hit ratio.

I. INTRODUCTION

Our future is the automated world[2], so we are trying to automate processes in many areas of the life. Automated machines, robots are used to reduce failure of the human working, and moreover to make the work easier and faster. There are two different approaches, the first one is the software agent and the second one is the hardware agent.

There are certain areas, which are highly automatable as they have many repetitive tasks, like IT services, which can be modelled into small executional parts. Similarly to other industries IT services or service management is also very cost sensitive, which is also a good driver for running automation and serves like a good business need. Automation not only makes several repetitive tasks to be easier for us, but it means that employees can focus on tasks with more added values. This increases the success of a company.

The project what we are working on is a process automation. With the help of different text analytics tool, we have made a previously manual process faster, progressive, and more continuous. In this case our main goal is an automated text evaluation on a specific lingual area.

Solution selected for this task is a Watson based solution, what can analyze, interpret unstructured texts. These functions make it possible to apply for our task.

Similar solutions are being used in everyday life, but we are not thinking about how they work. The plainest examples: library and web searchers, these examples operating on the same principles, which is a keyword, apart from the algorithm, of course. Methods are searching for patterns, what comes from the searching criteria. This an efficient method while we are not looking for the context of the information along the interpretation

In this paper, the following section introduces data and the proposed method, Section 3 shows you how to create the model, Section 4 is about the results, Section 5 discusses and conclusions, Section 6 introduce the limitations of the solution, also introduce the planned future works, and finally section 7 say thank you for supporting our project.

II. DATA AND METHODS

A. Data to Analyze

The problems revealed during the audits, the related additional information, conditions and the descriptions of systems covers the data which are involved in the processing. This specific data is restricted to IT areas and within our project; it is limited to English text processing. The data are not structured, as they have free text nature. During processing what makes the challenge to be more complicated, that the concerns identified through the audit process are described in compound sentences or a single paragraph might be referring to several problem areas. Because of these, it is required to define a splitting method based on rules, and the analysis must run on these data for the efficient processing.

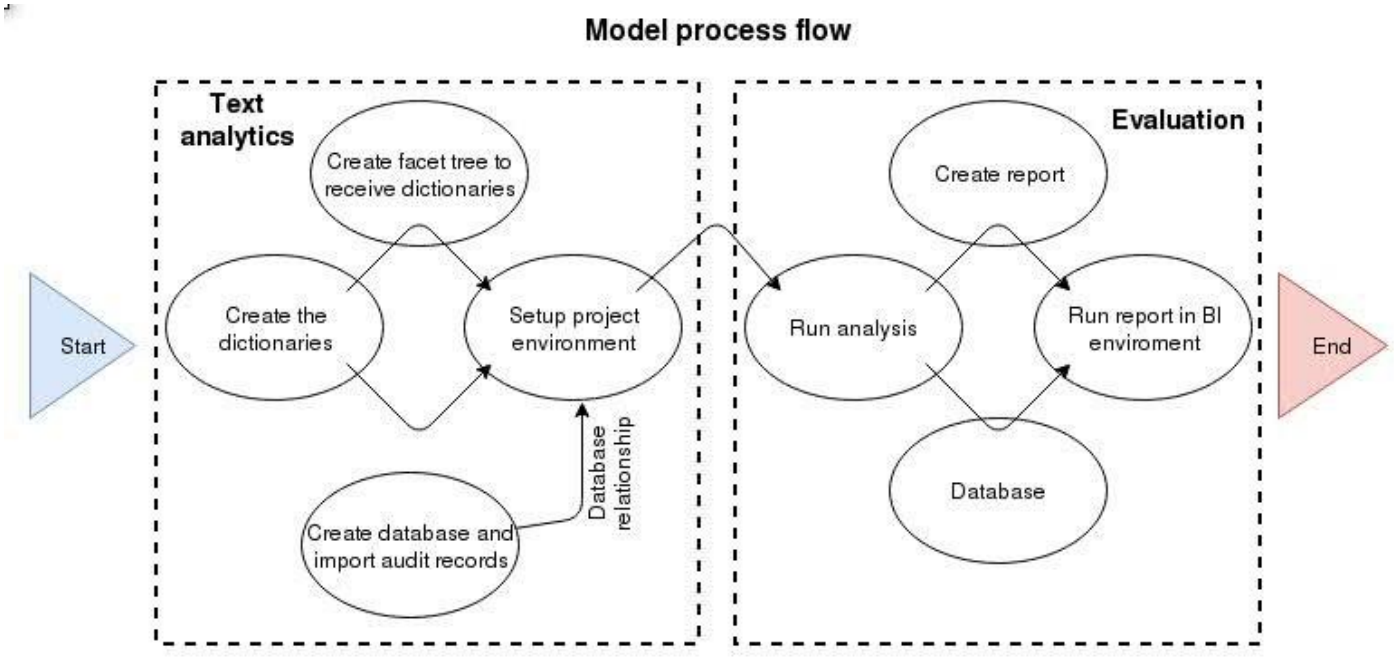

Figure 1. Overview. Shows the steps: Create Dictionaries, Export Pipeline, Run analysis, Run report based on our statistical model

The purpose of the analysis is to interpret these audit data, categorize texts and identify the risk issues in an automated way. As mentioned earlier this was done manually.

B. An overview of the model

The task is to create a process or model that can interpret texts in the described specific language area. The specialty given by two things, first text cannot be interpreted based on normal daily speech, second this IT service related language area is continuously improving and changing. We need to think of a solution that can support these specialties.

During the planning, we looked at some alternatives, and selected two of which we wanted to deal with:

- Traditional language analytics based solutions (key words, dictionary-based)
- Contextual based analytics (machine learning, context analysis)

It has been decided that both solutions will be analyzed and implemented, but as the first step the keyword driven analysis has been chosen to start with, considering that it will provide lots of experiences about the specific linguistic forms, which also required for the next step to apply machine learning and contextual analysis. The topic of the article is to introduce the approach applying traditional language analytics.

The figure 1 shows the overview of the process model: It is quite important to be familiar with this specific language area to be capable to prepare our processes for linguistic interpretation. The essence of the process is to identify important, meaningful words from the audit texts, which can be used for interpretation in this given linguistic area for making decision. The purpose of the decision is to define what problematic area is described by the analyzed sentence in terms of IT risk and compliance. The decision is made on the bases of majority principle, which means to determine those keywords in the sentence, which are fitting the best for a known IT risk and compliance area.

The approach assumes that the sentence - describing a given problem - mainly contains specific terms for that area. There are exceptions like complex sentences, which are focusing on more than one problem or other sentences, which contains enumerations. Such sentences require data cleansing, data structure interpretation. The key in our procedure is to define this majority principle. Therefore, it is required to define first the environment, where these sentences, problem descriptions are analyzed, so this environment can be considered as a linguistic space for interpretation. This interpretation space can be mapped by a hierarchy, because the IT risk and compliance areas themselves are also mapped into specific categories, which can be divided further into subcategories along a hierarchy tree.

Our interpretation space contains 14 main categories and most of the categories contain 4 levels of subcategories. The success criterion of this process is to be able to map and associate the analyzed sentences into one specific element of the given category of this interpretation space. In the output, any sentence that corresponds to an element of the hierarchy is meaningful to us, in all other cases we are confronted a sentence what is not fitting in the given language area. Of course, in order to have a perfect decision, we need to cover this interpretation space perfectly, thus it is mandatory to

know the terms within this specific area. Since this is constantly evolving, the goodness of the model depends on how deep we can cover the space. Thus it can be stated that our model as more accurate and efficient as more text interpretations is executed.

Due to the fact that a particular word can fit into more than one specific point of the hierarchy, since the meaning of a given word is impacted by other words, thus we need to think of a solution that can handle not only a single word, but also word sequences or meaningful phrases. Regarding categorization, it can be observed that the main categories are mostly nouns, while subcategories are rather verbs.
Illustrated on an example (Mgmt = Management)

"Approximately 39,667 privileged userids were not revalidated during 1Q 2017 continuous business need and privilege access revalidation activities."

Main category:
"access" = noun -->Identify and Access Mgmt

Category 1:
"privileged userids" = noun --> Privileged ID Mgmt

Category 2:
"not revalidated" = verb -->No revalidation

The exact finding:

Identify and Access Mgmt
Privileged ID Mgmt
No revalidation

However, in the description of the procedure, stay on a frequency-based decision [3], which is performed as follows:

In the first step, we have to count the number of main categories for the sentence, and then summarize the same(1):

```
A=count(distinct(main) for sentences)
  main = sum((A) for main)                (1)
```

Then calculate the total number of main categories in one sentence(2):

```
B = count(distinct(main) for sentences)
  all main = sum((B) for sentences)       (2)
```

After these steps count a percentage of the main categories for sentences (3):

```
main% = main/all main *100                (3)
```

With repeating the same steps doing this on category level 1-2-3, making sure that the parent changes each time because of the hierarchy of multiple levels of the tree (4):

```
A = count(distinct(cat1) for main)
    cat1 = sum((A) for cat1)
B = count(distinct(cat1) for main)
    all cat1 = sum((B ) for main)
    cat1% = cat1/all cat1 * 100

A = count(distinct(cat2) for cat1)
    cat2 = sum((A) for cat2)
B = count(distinct(cat2) for cat1)
    all cat2 = sum((B )  for cat1)
    cat2% = cat2/all cat2 * 100

A = count(distinct(cat3) for cat2)
    cat3 = sum((A) for cat3)
B = count(distinct(cat3) for cat2)
    all cat3 = sum((B) for cat2)
    cat3% = cat3/all cat3 * 100           (4)
```

If these results are available, then main%(own choice) is cut above a threshold to decrease the multitude of findings(5):

```
Where(main%>50)                           (5)
```

Thus a list is created with a reduced probability, but in many case it contains more than one hit per sentence, because of the multi-level structure (several subcategories belong to a major category). To overcome this problem, weighted percentages are calculated for each level. Based on our tests, the higher levels reach a more accurate hit, so they get the highest weight, down the levels these weights decrease (6):

```
If(main% is not null) then(main%*1000)
                  else(0)
If(cat1% is not null) then(cat1%*100)
                  else(0)
If(cat2% is not null) then(cat2%*10)
                  else(0)
If(cat3% is not null) then(cat3%)
                  else(0)                 (6)
```

Then sum these scores for each row (7):

```
Sum = sum(main%,cat1%,cat2%,cat3%)        (7)
```

In the last step the highest score is picked up for each sentence, so we reach our goal what is to identify only one finding for one sentence (8):

```
sort by (sentences, Sum)
running-count (sum for sentences)
count <2                                  (8)
```

III. CREATING THE MODEL

First step is to define data what we have to process. It is mentioned previously that data are audit records. Creating a database with specific parameters is needed for data storage. After that data have to be imported for processing from a centralized system.

IBM DB2 – database 2- is used to store data. After the installation we needed to create a non-default buffer pool[4] - cache - for reason, storing big volume of data. This solution is used to make our process faster.

In case of DB2, the objects of the relational database are organized into sets called schemas. A schema is a collection of named objects that provides a logical classification of objects in the database. So next step is creating a new schema based on the special buffer pool, to provide access to a bigger cache.

To store data, we had to create an intermediate table. Temporary data storage is not requirement in a complex database architecture in our case, so we created only one table for storage. Subsequently a star schema is used for making report, similar like this:

Figure 2. Star database schema

The table contains an ID with automate generate function yet, it is used for identifying and counting records. Table also contains additional information: Account name, Organization, Location, Date, Assessment, Tittle, Relative Size. These columns are stored in text format, and they contain our most information. Concern, Conditions are also stored in text format.

After creating the table, we have to upload it with data. It was solved with pull method, that imports data into our table from a centralized database. It makes capable to use this table to communicate with the natural language processing tool.

Next step is creating a dictionary, it is the most critical movement in our process. Analysis is based on keywords, so it needs to fill extensively. Watson Analytics Studio [1] capable to make unique dictionaries. To cover IT risk specific problem space, needed to design an efficient dictionary hierarchy. Our system needs to run along this IT risk hierarchy, to achieve fully qualified finding. While creating the dictionary we have to pay attention, that the specified dictionary contains only category specified words, otherwise it would result fake findings. At first, by analyzing the existing audit records, we searched for words fitting into categories, and after reviewing those words, they were added into dictionaries. After loading, the measurement gives us a feedback about our work. By repeating these steps, we derived all of the dictionaries, what we can.

System allows us to create a “custom annotator”. But what does it mean? It means we could insert our IT risk dictionaries into an annotator - pipeline -, in addition to the basic English dictionary. The difference between the two kinds of dictionaries is that basic one aims to grammatically analyze a sentence and identify the parts of the text, while our dictionaries aim to identify the specific category. This feature allows us to categorize the texts. To continue the process, another tools (Watson Explorer Content Analytics, Cognos BI Report Studio) have to be involved into the solution work flow. This is needed because the Studio cannot give opportunity to export the results into relational database for finalizing the report.

Watson Explorer Content Analytics[1] - WEX - collects and analyzes structured and unstructured content in documents, email, databases, websites, and other enterprise repositories. By providing a platform for crawling and importing content, parsing and analyzing content, and creating a searchable index, Watson Explorer Content Analytics helps you perform text analytics across all data in your enterprise and makes that data available for analysis and search. Working with the two applications there is a way to connect them together. WEX enables to import our “custom annotator”, for use it for analyzing. In the WEX firstly we have to create an empty project. In the WEX similarly to Watson Analytics Studio we have to create a hierarchy to map the dictionaries, it is called “Facet tree”. It is important to create a good hierarchy, because the software can recognize the parent-children relations. It is important for us to “draw” the fully qualified branch, in the lower levels.

Last step is exporting the result into a relational database, what a report system, in this Cognos BI can use. As the results of the custom annotation gives us more than one category for a given sentence, we need to select the best candidates as finding. That we achieve by using the earlier mentioned statistics model - weighted percentage -.

IV. RESULTS

In this section, we are presenting the result achieved by our solution. Test cases were based on existing audit records, with human validation. Our database table is used to provide data for text analytics. The figure 5, 6 visualize our report quality, based on the previous points.

Before presenting the final results, we introduce how the system works, and the intermediate steps what format the final results. Let's see an example (Table I):

„Root cause analysis was not always performed upon a service level failure. Further, root cause analysis when performed did not always identify the actual trigger of a failure."

Findings are the following words: identify, identify, was not always performed, root cause, root cause analysis, Root cause analysis

TABLE .I
RESULT OF AN EXAMPLE. FIRST COLUMN IS THE FOUND KEYWORD, SECOND IS THE TARGETED MAIN CATEGORY, AND THIRD IS THE FOUND CATEGORY. THE FALSE TRUE ANSWER IS BOLD.

Keyword	CategoryMain	Founded category
identify	IT_Risk_Management_ Services	**Exception_not_ identified**
identify	IT_Risk_Management_ Services	RCA_inaccurate_ or_incomplete
was not always performed	IT_Risk_Management_ Services	RCA
root cause	IT_Risk_Management_ Services	RCA
root cause analysis	IT_Risk_Management_ Services	RCA
Root cause analysis	IT_Risk_Management_ Services	RCA
Overall	6	1:5(false:true)

Figure 3 visualizing the Table I.

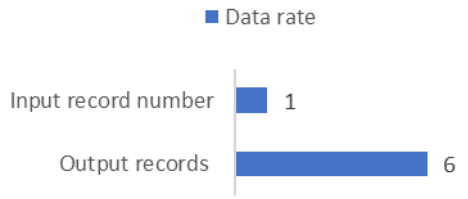

Figure 3. Findings per record

Word repetition is discoverable, because a word can belong to more than one category. To solve this problem, we need to apply a report system as we mentioned. Let see the "identify" word. This belong to two categories, however in the first case it reached only 16%, in the second case 84%. This difference is due to the other words which strengthens this branch for example: "root cause analysis", so the final vote is based on the 2nd "identify":

IT_Risk_Management_Services
IT_Risk_Management
RCA_inaccurate_or_incomplete

This solution ensures that finally only one result is displayed on the output. The test running on 97 records, then with summarizing the data we get the following result (Figure 4).

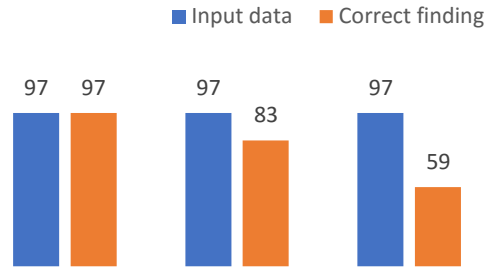

Figure 4. Correct Conversation Rate

In a percentage format, it looks like this (Figure 5), the correct finding percentage is significantly decreasing with the category level:

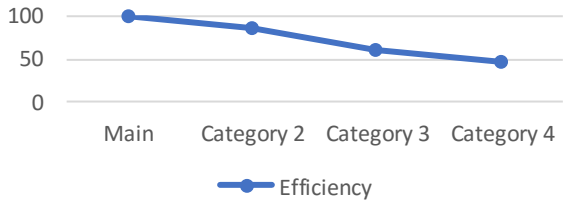

Figure 5. Percent hit rate

Summing up the results with hundreds of audit records, we get the Table II.

The upper blue part of the table shows a good ratio while the lower parts are getting weaker - the weakness can be attributed to the non-sufficient number of audit records what causes that the dictionary not filled up satisfactorily - but overall, despite the weakness of the lower part of the table, the average accuracy of 83.3 was achieved.

TABLE II : SUMMARIZED RESULT FOR 574 RECORD

Account	Accuracy	Manual	WEX
Account_1	100,0%	15	15
Account_2	96,2%	26	25
Account_3	94,7%	38	36
Account_4	94,4%	18	17
Account_5	93,8%	16	15
Account_6	92,7%	55	51
Account_7	91,7%	12	11
Account_8	90,9%	11	10
Account_9	90,5%	42	38
Account_10	89,3%	28	25
Account_11	87,5%	16	14
Account_12	87,5%	24	21
Account_13	87,5%	8	7
Account_14	85,7%	7	6
Account_15	85,0%	20	17
Account_16	83,9%	31	26
Account_17	81,5%	27	22
Account_18	77,8%	18	14
Account_19	76,0%	25	19
Account_20	68,2%	66	45
Account_21	68,0%	25	17
Account_22	66,7%	18	12
Account_23	66,7%	9	6
Account_24	47,4%	19	9
Grand Total	83,3%	574	478

V. CONCLUSION AND DISCUSSION

In this paper a complex text analyses was presented with its result. In summary the results are satisfactory.

Based on our model and approach, we have reached a relatively high hit ratio and accuracy by applying this language analytics model. Accuracy means in our case that this solution is able to identify the correct main category above 90% and in the second category the accuracy varies between 65-85%. This level of accuracy is good base for automation and decrease the level of manual effort. We still can identify space for improvements, which we can approach from two aspects. First aspect is the goodness of the dictionary itself, here we need external support having good level of knowledge about this specific area.

However it can be seen that at higher hierarchy levels - the top levels- we can achieve a flawless or almost completely flawed hit rate (fig 5). Towards the lower levels, this ratio was decreasing, which can be attributed to the degradation of processes in many branches, and the diversity of the content of the verb report. The weakness of the lower levels can be attributed also to the dictionary's weaknesses, which means that in those dictionaries there are not enough supporting words.

All test results were discussed concerning the success of the analysis, and it was concluded that in the further text analysis the improvement of the hit rate can be approached from two sides. First, let's take the easiest one to refine the dictionary's content. In this case, our knowledge about this specific area is not appropriate, so it is required to involve an external help - a person who had made manual analysis earlier - to refine the dictionary. We have also tried to consider the structure of the hierarchy when uploading the dictionary. The upper two levels of the hierarchy point to the definition of the subject, so in this case the given levels are filled with nouns defining the fundamental problem. The lower levels under each major categories refers to some parts of the processes, so it is better to moderate the number of adjectives, nouns in the dictionaries at these levels, but use rather verbs to create accurate results. This plays an important role because, if the dictionary is populated by this way, the approach will find more meaningful words for the higher level and for the lower levels, so the statistical model will give higher weight than any possible false results.Second, clearing the text can move to the desired direction. Input data contains a lot of specific characters - "; ‘ etc." that are removed to make the device work optimally.

In addition, the text contains many non-noticeable control characters - line breaks, etc. - which also have an affecting feature for the analysis; these are also removed from the text in some cases. As the last step of the content purification, it is necessary to mention that an input data consists of several sentences in most cases. These sentences are usually not describing only one fault, so it is necessary to analyze these sentences separately and to display them as separate texts between the output data, keeping in mind, of course, the related original text.

VI. LIMITATIONS AND FUTURE WORKS

Knowing that the tool is key word-based, so it looks for matching pattern in the text, it propounds various problems, the most significant of these being the spelling error. In the knowledge of the problem, we can put up a similar problem - a foreign word embedded in the word connection -, which means that we will not be able to recognize the specimen in this case, so after detecting these problems, we started looking for another similar text analytical method, and this is Watson's Cognitive Tool what is able to demonstrate the ability to use machine learning by training through number of samples.

Keeping in mind the success of our project in the following, we try to recognize this cognitive solution and increase the categorization goodness in this special language area. The cognitive path gives two paths, first one is also a keyword-based system that, with the help of the applied dictionaries, is able to further refine system by machine learning. Another one is a new one, which does not require any previously created dictionaries, it is based on machine learning only, of course, a sufficient number of samples is required, and then the approach will give us automatically the highest probability data.

VII. ACKNOWLEDGMENT

SUPPORTED BY THE ÚNKP-16-1 NEW NATIONAL EXCELLENCE PROGRAM OF THE MINISTRY OF HUMAN CAPACITIES

We acknowledge the financial support of this work by the Ministry of Human Capacities of Hungary

REFERENCES

[1] W.-D. J. Zhu and International Business Machines Corporation. International Technical Support Organization., “IBM Watson Content Analytics discovering actionable insight from your content.” 2014.

[2] R. Berger, “Of robots and men - in logistics.” 2016.

[3] Dudás László, “Alkalmazott Mesterséges Intelligencia,” 2011.

[4] J. Z. Teng and R. A. Gumaer, “Managing IBM database 2 buffers to maximize performance,” *IBM Syst. J.*, 1984.