A NLP-Oriented Methodology to Enhance Event Log Quality

Belén Ramos-Gutiérrez[1(✉)], Ángel Jesús Varela-Vaca[1], F. Javier Ortega[1], María Teresa Gómez-López[1], and Moe Thandar Wynn[2]

[1] IDEA & ITALICA Research Groups, Universidad de Sevilla, Seville, Spain
{brgutierrez,ajvarela,javierortega,maytegomez}@us.es
[2] Queensland University of Technology (QUT), Brisbane, Australia
m.wynn@qut.edu.au
http://www.idea.us.es

Abstract. The quality of event logs is a crucial cornerstone for the feasibility of the application of later process mining techniques. The wide variety of data that can be included in an event log refer to information about the activity, such as what, who or where. In this paper, we focus on event logs that include textual information written in a natural language that contains exhaustive descriptions of activity executions. In this context, a pre-processing step is necessary since textual information is unstructured and it can contain inaccuracies that will provoke the impracticability of process mining techniques. For this reason, we propose a methodology that applies Natural Language Processing (NLP) to raw event log by relabelling activities. The approach let the customised description of the measurement and assessment of the event log quality depending on expert requirements. Additionally, it guides the selection of the most suitable NLP techniques for use depending on the event log. The methodology has been evaluated using a real-life event log that includes detailed textual descriptions to capture the management of incidents in the aircraft assembly process in aerospace manufacturing.

Keywords: Natural Language Processing · Event log quality · Process mining

1 Introduction

Event logs include the footprints generated by an organisation's information systems, being possible to store a wide variety of information [4,6] related to the tracked events, e.g., textual descriptions, timestamps or used resources. In general, event logs need to be adapted for a later (process mining) analysis, for instance, to discover processes. Thereby, the assessment of the quality of an event log [7] is the very first and crucial step for any subsequent analysis. The application of any process mining technique over incorrect or inaccurate event logs, e.g. process discovery, will produce incorrect or inaccurate process models [32].

Several authors have defined criteria to assess the data quality in general [9,21] and event log quality in particular [7,32], such as completeness, correctness, security and

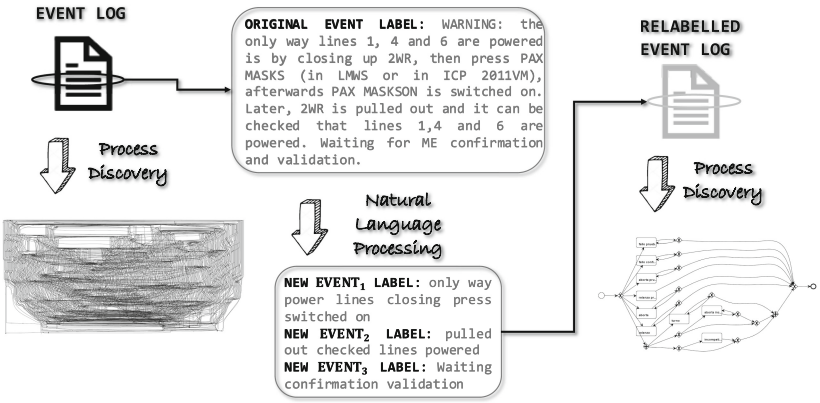

Fig. 1. Relabelling application example.

trustworthiness. The Process Mining manifesto [7] introduces the quality of an event log as a quality maturity level.

The imperfections than can produce a low event log quality might be improved analysing the activity labelling, timestamps, case identification, etc. In this paper, we focus on event logs that include some textual descriptions which detail what happened in various moments of process execution. We propose to first identify these activities and their inter-relations from textual constraint descriptions using NLP techniques and then relabelling activities in the event log to extract these details in an easy to handle way.

Figure 1 illustrates how the use of Natural Language Processing (NLP) techniques and the relabelling of activities in an event log, can improve the results of automated process discovery. For this reason, we not only adapt a general methodology to measure and assess the data quality of event logs [27], but also propose a decision-support system to assist in the selection of the most suitable NLP techniques according to the current quality level and the expected assessment. The research question is: *What are the most suitable NLP techniques to use for relabelling an event log in order to improve its quality?* This question does not have a simple answer, since it depends on the current quality of the event log and the dimension or dimensions that must be improved and how.

To answer this research question, it is necessary to define metrics to measure the event log quality and a mechanism to assess how good is this quality level in each context. This is well-known as the fitness-per-purpose [19], where the level of quality must be customised according to the needs, as for example: (1) determining the average length of the label of the activities; (2) the level of noise allowed; and (3) the usual number of activities per trace.

With this goal in mind, we propose a methodology, called LOADING-NLP, which assists in the decision-making for the application of NLP techniques over raw event logs for relabelling activities in accordance with the decision rules about data quality described by experts. A set of fitness-per-purpose metrics and dimensions are

proposed to measure and assess the quality of an event log. Both, the metrics and dimensions can be customised, or extended for other examples. In this regard, the measurement and assessment can be adjusted and alternative NLP techniques can be applied. Our methodology also assists the user to select the most suitable NLP techniques. To validate the proposed methodology, we applied it to a real case study based on the management of the incidents produced during the aircraft assembly processes in aerospace manufacturing.

The rest of the paper is organised as follows: Sect. 2 includes the related work in the area. Section 3 introduces the methodology; the measurement and the assessment of the event log quality are discussed in Sect. 4. Section 5 reviews the NLP techniques that can be applied in this context and outlines to what extent they can affect the quality assessment. Section 6 presents the evaluation results and Sect. 7 concludes the paper.

2 Related Work

In order to understand the advantages that this proposal offers, it is necessary to know the level of maturity in the following areas.

Event Log Quality. The necessity to have data with suitable quality is crucial for any process and necessary for later analysis, such as process mining [10]. How to measure and assess the possibility of leveraging their quality is an important topic which has been a focus of study during the last decades. However, event logs appear in new contexts [30] and include features that make it necessary to define new metrics to measure, extend and adapt the dimensions (e.g., completeness, accuracy, simplicity) to the business process context [8,31].

Event Log Improvement. Once the data quality level can be assessed, various are the techniques that can be applied to improve the event log quality [26]. Some solutions are based on timestamp [12,13,15], case identification, and activity relabelling. Regarding the activity relabelling, the solution presented in [25] proposes to detect synonymous and polluted labels in event logs, but no techniques were proposed to improve the quality in accordance to the previous detection, and [24] uses a gamification approach to repair the labels. The types of improvements over event logs depend on the case and the later use of them [18].

Use of NLP in Business Process Management. Previous works have studied the extraction of declarative [2] and imperative business process models [3] from texts. In those works, the NLP techniques have been used in order to facilitate the automation of tasks that require a significant effort detecting patterns of relational order between the activities involved [1]. In addition, the detection of activities and their associated labels is crucial for further analysis and refactoring of the terms to enable an automatic analysis [17]. The text analysis in the context of the business process has also been focused on the detection of inconsistencies between the textual descriptions and the graphical representation [3], as a mechanism of misalignment detection.

Use of NLP to Improve Process Discovery Results. Some works have studied the preprocessing of the event data to improve the discovery task when using real-life logs that are written in natural language. In [23], is done by automatically detecting and classifying eight different semantic roles in event data. In [14], semantic-based techniques are applied to aggregate and normalise event log text information. Other types of analysis have been made to improve the labels of the activities in a process model by detecting erroneous ways of labelling activity that lead to ambiguity and inconsistency [22]. Contrary to our proposal, in all cases, these proposals start from event data in natural language with a very process-oriented construction and a simple and correct syntax.

To sum up, to the best of our knowledge, this is the first work where NLP techniques have been used to improve the event log quality according to a set of proposed metrics, guided by a decision-support system to ascertain the best techniques to apply depending on the event log and its quality level.

3 LOADING-NLP: Methodology for Assessing and Improving Event Log Quality with NLP

When the labels of the events within the log include natural language texts, it is necessary to analyse and treat them to become the log useful. The NLP techniques used to this aim will cause a direct impact on the different number of labels in the log, the number of events per trace, or the similarity between the labels.

Therefore, the best NLP techniques to use depends on the meaning of event log quality in each context, the event log quality before the application of the techniques, and how the experts want it to evolve after the application of the techniques. To support these three aspects, we propose the methodology presented in Fig. 2, described through a BPMN model.

The first step is related to the definition of when the event log is usable (cf., Determine the usability of the event log quality), according to the measurement and the assessment described by a set of decision rules about data quality. If the event log has sufficient quality, it can be used for a process mining analysis. However, if the event log is deemed unsuitable, the expert must determine the dimension or dimensions that must be adjusted and the required assessment (cf., Introduce dimensions and assessment to achieve). Using this information, we propose a decision-support system (cf., Infer NLP techniques to apply) that provides possible NLP techniques to use for improving the event log quality according to the described decision rules about data quality and the requirements of the experts. This process can be repeated until the resulting event log achieves the required level of quality.

4 Determine the Usability of the Event Log Quality

The question of when an event log has sufficient quality does not have a single answer. It will depend on the meaning of event log quality in each context. There are solutions as [27] that provide mechanisms to describe the decision rules related to the measurements and assessments adapted to each context and requirements. At first, we need to define

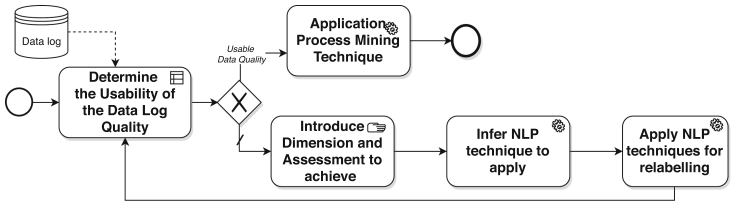

Fig. 2. Methodology for the application of NLP techniques.

the decision rules according to data quality using a set of metrics extracted from the event log. Thereby, it is necessary to define a set of metrics to evaluate the dimensions, as detailed in Subsects. 4.1 and 4.2. These measures enable us to perform an assessment according to the meaning of quality defined by the expert, as described in Subsect. 4.3.

4.1 Metrics to Measure the Quality of Event Logs

Based on [11], an ***event log*** is a set of traces that represent different instances of the same process. A ***trace*** is an ordered sequence of events that represents a process instance. Every trace is associated with a unique case identifier. The execution of an activity in a business process is represented as an ***event*** in an event log. Similarly, an event is the representation of the execution of an activity in a business process. Each event is associated with a case identifier, one timestamp and can also have many other contextual attributes. Usually, each event has associated at most one timestamp, which represents the start or the end of the execution of an activity.

To measure the dimensions, some metrics must be extracted from the event log [5,11], as the mentioned in [8]. In our case, the used metrics are:

- **Number of traces.** Total number of traces in the log, and the trace j is represented by τ_j.
- **Number of events** (ε). Total number of events in the log, and the event i is represented by ε_i. It helps to know the size of the log.
- **Number of different labels.** Number of different labels that occur in every trace.
- **Number of unique labels.** Number of single (unique) labels that appear in the log.

4.2 Quality Dimensions for Event Logs

In general, the data quality dimensions describe the relevant aspects for a data set and typically consist of accuracy, completeness, consistency and uniqueness. However, we cannot guarantee that an event log with a high level of quality in those dimensions will produce valuable business processes. For this reason, other dimensions are included to assess the event log quality in process mining, as was defined in [31]. Those dimensions can be affected by the application of NLP techniques. Based on them, we propose the following dimensions albeit others can also be used together with our methodology:

$$m_{Uniqueness} = \left(\frac{Number\ of\ Unique\ Labels}{Number\ of\ Events}\right) \quad m_{Complexity} = \left(\frac{Avg.\ Number\ of\ Events}{Number\ of\ Traces}\right)$$

$$m_{Relevancy} = \left(\frac{Number\ of\ Different\ Labels}{Number\ of\ Events}\right) \quad m_{Consistency} = \sum_{i=1}^{\varepsilon}\left(\frac{|l(\varepsilon_i) - \overline{l(\varepsilon)}|}{Number\ of\ Events}\right)$$

Uniqueness. If every label in an event log is the same, the discovered process will only include one activity. However, if each label is unique in the traces, the discovered process will have one branch per trace, thus not very useful. The uniqueness dimension, that we propose in a range between $[0..1]$, measures the percentage of single labels regarding the total number of labels. When the values are in the extremes (i.e., 0 or 1), it implies that the process may be too simple (low uniqueness) or too complex (high uniqueness).

Consistency. When the labels of activities are dissimilar, especially for textual formats, they can imply that the descriptions have different granularity. For this reason, the measurement of consistency that we propose is based on the average length[1] of strings in these textual descriptions, and the mean of the distance to the average. Therefore, this dimension is bounded by the length of the longest string.

Relevancy. The relevancy of each label depends on the number of times that it occurs. It is important to analyse the number of different labels according to the total number. It is related to the uniqueness, but it is not exactly the same. The dimension is bound between $[0..1]$.

Complexity. There are several metrics that can represent the complexity of an event log [26], such as the average of events per trace. We propose to measure the complexity by the mean of the number of events per trace. A higher mean implies a higher concentration of events per trace, therefore representing a more complex process.

4.3 Customising the Measurement and Assessment of Event Log Quality

As commented previously, the data quality is an aspect highly related to the later use of the data, hence it must be customised according to the necessities. Following the DMN4DQ proposed in [27,28], DMN (Decision Model and Notation) [20] can be used for facilitating the description of data quality divided into measurement and assessment rules. DMN is a declarative language proposed by OMG to describe decision rules applied to a tuple of input data to obtain a tuple of outputs according to the evaluation of a set of conditions described in FEEL. A DMN table is composed of rows that describe a decision rule as an if-then condition so that, if it is satisfied, the output is returned. Also, DMN permits a hierarchical structure where the output of a DMN table can be the input of another. Using the methodology DMN4DQ, we propose to split up the measurement and assessment in two different levels for each involved dimension. Additionally, the final assessment is obtained by aggregating the assessment of every dimension, as described in Fig. 3.

[1] We use $l()$ function to define the length of a label description.

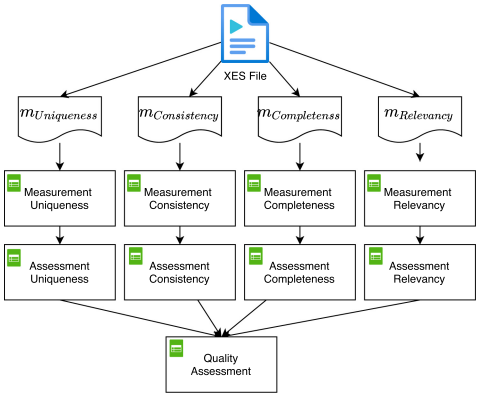

Fig. 3. DMN for describing the Quality Assessment.

DMN tables have various types of columns (`orders`, `inputs`, and `outputs`). The first column establishes the `order` by assigning an index to each row, and includes the hit policy to determine how to act when more than one is satisfied, (cf., *F* to describe that the evaluation of the condition is in order). Each `input` column represents the input variables that are evaluated the condition of the row. An example of DMN tables for the measurement of each dimension is detailed in Tables 1a, 1b, 1c and 1d, that illustrate the four dimensions proposed in this paper. Each dimension is described by a set of domains of the metric values, where the measurement of the metrics is the output value of the table. For example, the consistency dimension, in this case, describes that if the average of the numbers of characters is greater than 30, the measurement of the consistency will be *Very low*. For the measurement of each dimension, only one metric is required as input. In each row of the input column, the conditions are described in FEEL, for instance, the first row in Table 1b establishes that the valid range for the metric $m_{Uniqueness}$ is between 0 and 1. We use the metrics defined in the previous subsection as the inputs. The `outputs` represent the obtained values depending on the condition satisfied, for instance, if the input for $m_{Uniqueness}$ is 0.5 the outputs is *High*. The values and conditions established in Table 1 have been adjusted according to the know-how of experts for the use case at hand.

We should bear in mind that the measurement of a metric does not represent whether the metric is good or not, this is why a later assessment is necessary. Table 2 includes a proposal for the assessment of each dimension, for example, both a *Low* and *Very Low* number of events will imply an *Excellent* assessment in the event log according to the Complexity metric. As previously commented, this assessment has been defined based on the experts' knowledge of the event logs but other assessments can be accommodated. The assessment of each dimension needs to be aggregated to determine a global one. A possible set of decision rules for the aggregation of the assessment of the four dimensions is described in Table 3, albeit another combination can be applied according to the necessity of the organisation.

Table 1. Decision tables for measuring each dimension.

(a) Measurement of Uniqueness Dimension

F	Input $m_{Uniqueness}$	$Output_U$
1	[0, 0.1]	Very Low
2	(0.1, 0.2]	Low
3	(0.2, 0.4]	Medium
4	(0.4, 0.6]	High
5	(0.6, 1]	Very High

(b) Measurement of Consistency Dimension

F	Input $m_{Consistency}$	$Output_{Cs}$
1	[0, 5]	Very High
2	(6, 14]	High
3	(14, 20]	Medium
4	(20, 30]	Low
5	(30, ∞)	Very Low

(c) Measurement of Relevancy Dimension

F	Input $m_{Relevancy}$	$Output_R$
1	[0, 0.1]	Very High
2	(0.1, 0.2]	High
3	(0.2, 0.4]	Medium
4	(0.4, 0.6]	Low
5	(0.6, 1]	Very Low

(d) Measurement of Complexity Dimension

F	Input $m_{Complexity}$	$Output_{Cx}$
1	[0, 4]	Very Low
2	(4, 6]	Low
3	(7, 10]	Medium
4	(11, 15]	High
5	(16, ∞)	Very High

Table 2. Decision tables for the assessment of each dimension.

(a) Assessment of Uniqueness Dimension

F	Input $Output_U$	$Assess_U$
1	Very Low	Fair
2	Low $\vee$ Medium	Excellent
3	High	Poor
4	Very High	Very Poor

(b) Assessment of Consistency Dimension

F	Input $Output_{Cs}$	$Assess_{Cs}$
1	Very Low	Very Poor
2	Low	Poor
3	Medium	Fair
4	High	Good
5	Very High	Excellent

(c) Assessment of Relevancy Dimension

F	Input $Output_R$	$Assess_R$
1	Very Low $\vee$ Very High	Very Poor
2	High $\vee$ Medium	Fair
3	Low	Poor

(d) Assessment of Complexity Dimension

F	Input $Output_{Cx}$	$Assess_{Cx}$
1	Very Low $\vee$ Low	Excellent
2	Medium	Good
3	High	Poor
4	Very High	Very Poor

Table 3 is designed in such a way that, when at least 3 assessment values for the dimensions is qualified as Excellent, and the remaining one is qualified as Good or Fair, the quality outcome of the log is Excellent. Similarly, when we have three dimensions qualified as Excellent or Good and one as Fair or Poor, the quality outcome of the log will be Good. On the other hand, when we find out two dimensions as Excellent or Good and other two as Fair or Poor, the quality outcome of the log will be Fair, while,

Table 3. Aggregation of the Dimensions for the Quality Assessment

	Inputs				Output
F	$Assess_U$	$Assess_{Cs}$	$Assess_R$	$Asses_{Cx}$	$Quality_{assessment}$
1	Excellent	Excellent	Excellent	Excellent ∨ Good ∨ Fair	Excellent
2	Excellent	Excellent	Good ∨ Fair	Excellent	Excellent
3	Excellent	Good ∨ Fair	Excellent	Excellent	Excellent
4	Good ∨ Fair	Excellent	Excellent	Excellent	Excellent
5	Good ∨ Excellent	Good ∨ Excellent	Good ∨ Excellent	Good ∨ Excellent	Good
6	Poor ∨ Fair	Good ∨ Excellent	Good ∨ Excellent	Good ∨ Excellent	Good
7	Good ∨ Excellent	Poor ∨ Fair	Good ∨ Excellent	Good ∨ Excellent	Good
8	Good ∨ Excellent	Good ∨ Excellent	Poor ∨ Fair	Good ∨ Excellent	Good
9	Good ∨ Excellent	Good ∨ Excellent	Good ∨ Excellent	Poor ∨ Fair	Good
10	Poor ∨ Fair	Poor ∨ Fair	Good ∨ Excellent	Good ∨ Excellent	Fair
11	Good ∨ Excellent	Poor ∨ Fair	Poor ∨ Fair	Good ∨ Excellent	Fair
12	Good ∨ Excellent	Good ∨ Excellent	Poor ∨ Fair	Poor ∨ Fair	Fair
13	Poor ∨ Fair	Good ∨ Excellent	Good ∨ Excellent	Poor ∨ Fair	Fair
14	Very Poor	Very Poor	Very Poor	Very Poor	Very Poor
15	Very Poor	Very Poor	Very Poor ∨ Poor ∨ Fair	Very Poor ∨ Poor ∨ Fair	Very Poor
16	Very Poor ∨ Poor	Very Poor	Very Poor	Very Poor ∨ Poor ∨ Fair	Very Poor
17	Very Poor ∨ Poor ∨ Fair	Very Poor ∨ Poor ∨ Fair	Very Poor	Very Poor	Very Poor
18	-	-	-	-	Poor

when, at least, two dimensions are qualified as Very Poor, the quality outcome of the log will be Very Poor. In any other case, the quality outcome of the log will be Poor.

5 Improving Event Log Quality: NLP Techniques for Relabelling Activities

Our proposal aims to guide selection the NLP techniques for the relabelling of activities to extract the most meaningful and representative words for each process activity, but first, we need to introduce some NLP techniques.

For the sake of clarity, we take the following description of an incident as an example to show the effects of each NLP technique proposed in the paper: *"WARNING: the only way lines 1, 4 and 6 are powered is by closing up 2WR, then press PAX MASKS (in LMWS or in ICP 2011VM), afterwards PAX MASKS ON is switched on. Later, 2WR is pulled out and it can be checked that lines 1, 4 and 6 are powered. Waiting for ME confirmation and validation."*

The NLP techniques that are being proposed to be applied are described below:

Sentence Detection. This technique splits the text into its main components (i.e., sentences) to make it easier for the next steps to extract rich information from them. For our example, the application of this technique provides the next output:

- *Sentence 1.* WARNING: the only way lines 1, 4 and 6 are powered is by closing up 2WR, then press PAX MASKS (in LMWS or in ICP 2011VM), afterwards PAX MASKS ON is switched on
- *Sentence 2.* Later, 2WR is pulled out and it can be checked that lines 1, 4 and 6 are powered.
- *Sentence 3.* Waiting for ME confirmation and validation.

Each sentence is shorter and contains fewer verbs (actions) than the original, so they should be easier to analyse afterwards.

Part-Of-Speech (POS) Tagging. It consists of determining the grammatical function of each word in a text (i.e., noun, verb, adjective, preposition, pronoun, etc.), choosing for each word its corresponding class from a set of predefined tags[2]. The result of a POS-tagging on our example, selecting those words that are tagged as "NOUN", "VERB" or "ADJ" (adjective), therefore excluding the rest:

- *Sentence 1.* Only way power lines closing press switched on
- *Sentence 2.* Pulled out checked lines powered
- *Sentence 3.* Waiting confirmation validation

With this technique, we can keep those words that we consider relevant according to their grammatical category.

Lemmatisation. This technique normalises or substitutes the inflected forms by its lemma. In this way, it is easier to compare texts or even to group together different inflexions of the same lexeme. Lemmatisation can provide a more normalised text which can be better suited for relabelling. The results of the lemmatisation of our example are:

- *Sentence 1.* The only way that power the line 1 , 4 and 6 to be close 2WR, then press PAX MASKS (in LMWS or in ICP 2011VM), afterwards this switch on in PAX MASKS
- *Sentence 2.* Later pull out 2WR and check that the line 1 , 4 and 6 now to be power
- *Sentence 3.* Wait ME confirm valid.

Dependency Parsing. It determines the syntactic relationships between the words in a sentence by obtaining a *dependency tree* which provides information about the root verb of the sentence, its subject, the different objects and complements that it could contain. These are the roots of each sentence in our example detected by a dependency parser: *Sentence 1:* press; *Sentence 2:* pull; *Sentence 3:* Waiting. In this case, the dependency parser is used to extract the main verb of each sentence, so we could identify the action that characterises the corresponding activity.

Acronyms Detection. We have implemented a simple rule-based acronym detection that retrieves those words that are written in upper-cases and their lower-case form do not exist in the target language. Next, we show the acronyms detected by our approach

[2] All the tag sets used in this work come from the community open project called *Universal Dependencies* (https://universaldependencies.org/).

Table 4. Expected impact of NLP techniques on the log quality dimensions.

Dimensions	Sentence detection	POS tagging	Lemmat.	Dependency parsing	Acronym detection
$m_{Uniqueness}$	↓	↓	↓	↓	↑
$m_{Consistency}$	↓	↓	↓	↓	-
$m_{Relevancy}$	↓	↓	↓	↓	↑
$m_{Complexity}$	↑	-	-	-	-

in the example: *2WR, LMWS, 2011VM*. Depending on the context, these acronyms could be useful for detecting relevant entities in the domain.

It is important to bear in mind that the application of some of these techniques to certain texts may lead to the generation of void labels. When this happens, those events with an empty label are not included in the new log. On the other hand, when the NLP technique splits a label into several new ones (e.g., the acronyms 2WR, LMWS, 2011VM), one new event is generated for each new label, but maintaining the other attributes from the original event (e.g., the timestamp attribute).

5.1 Decision-Making for the Application of NLP Techniques

Our proposal for relabelling an event log consists of the application of the aforementioned NLP techniques to the incident descriptions, filtering out or editing them, to produce new simplified texts that are used to replace the original activity labels.

As previously commented in Sect. 5, bear in mind that the application of NLP techniques may generate or delete events according to the new labels generated. We can observe in Table 4 how the detection of sentences produces shorter texts, reducing their diversity but increasing the number of events (we will have one event per sentence, in the same order they appear within the description). Therefore, it can decrease all the proposed dimensions, except the $m_{Complexity}$ which will be increased. Concerning detection of acronyms, it can increase the $m_{Uniqueness}$ and the $m_{Relevancy}$, while the $m_{Complexity}$ and the $m_{Consistency}$ may not be noticeably affected due to the usually short length of the acronyms. The rest of NLP techniques (POS Tagging, lemmatisation, and dependency parsing) are applied to filter out irrelevant elements of the texts, keeping the important ones, and also to unify different inflexions of words into a common meaningful form. Therefore, their effects on the dimensions would be fairly similar. The $m_{Uniqueness}$ and the $m_{Relevancy}$ can be decreased since the normalisation of texts achieved by these techniques should decrease the number of unique events as well as the number of different events. $m_{Consistency}$ can be decreased as well because the length of the texts will be reduced and so will be the difference in length between them. $m_{Complexity}$ is not significantly affected because the number of traces and the number of events stays almost the same.

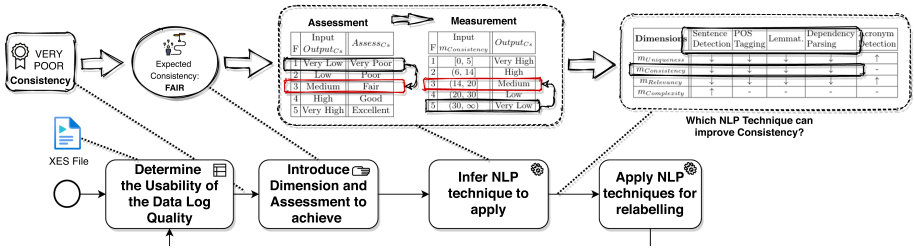

Fig. 4. Inferring the NLP technique to apply.

5.2 Inferring NLP Techniques to Improve Quality Dimensions

The relabelling of event logs through the application of NLP techniques and the definition of the dimensions and metrics discussed in previous sections provide a useful tool that guides us in the selection of the most suitable set of techniques to be applied achieving a certain assessment of quality.

Let's go back to the example proposed in Sect. 4.3. Let assume that there is an event log with a Consistency assessment ($Assess_{Cs}$) equal to $Very\ Poor$ and we want to improve it to $Fair$ as shown in Fig. 4. According to Table 2(d), we will need to take our event log from $Very\ Low$ to, at least, $Medium$ in terms of the measurement of the Consistency ($m_{Consistency}$). Then, looking at Table 1(d), it is necessary to reduce the value of the $m_{Consistency}$. Finally, according to Table 4, we can find out the NLP techniques that decrease the $m_{Consistency}$, and hence can be applied to our event log to achieve our objective. In this case they are *Sentence Detection*, *POS tagging*, *Lemmatisation*, and *Dependency parsing*.

In summary, given an event log with an assessment of quality and the dimensions to be improved, our proposal can help us to choose and apply the proper set of NLP techniques to achieve our objective.

6 Evaluation of the Proposal

For the evaluation, we use an event log (hereinafter Log_{desc}[3]) that represents the description of the evolution of the incidents in the aircraft assembly process which was presented in [29]. For instance, the following text represents a real incident description: *"When reading, the F1 error appears, wiring is verified according to FAQ and there is no continuity in any pinning in sections from 1509VC (pin 16) to 1599VC (pin 12). It is also appreciated that the colour coding concerning the plane (P1) does not correspond. Between FLKC1 and 250VC the wiring is correct"*. It can be easily observed that in this description of an incident, 3 sub-incidents are recorded: *(i) the F1 error appears*, *(ii) there is no continuity in any pinning from 16 to 12*, and; *(iii) the colour coding is incorrect*. For these reasons, textual descriptions can be useful to improve event logs

[3] Characteristics of the event log: 11.342 cases, 114.473 events, number of different labels 78.012, and 10.811 variants.

Table 5. Metrics of the event log used in the example, tagged as Log_{desc}.

Description	Total
Total number of textual descriptions	4,022
Total number of words	72,435
Out-Of-Vocabulary (OOV) words	10,832
Number of descriptions with OOVs	3,468
Grammatical and syntactic errors	1,642
Number of descriptions with errors	1,233

and discovered processes quality, but they must also be carefully processed to obtain relevant results.

The NLP is carried out with the aid of *spaCy* [16], a state-of-the-art Python library with pre-trained language models. Specifically, we have used for this work the largest Spanish pre-trained spaCy model, "es_core_news_lg"[4].

In order to illustrate the complexity of the problem, we show some metrics about Log_{desc} used for our evaluation in Table 5. We can see that 14.95% of the terms in the log are Out-Of-Vocabulary (OOV), which means that those words do not belong to the language at hand (they do not appear in the language model). In this point, domain-dependent enhancements could have been applied to the log, but the evaluation of our technique could have been biased by the nature of the domain, so we decided to keep the log as is. This complicates proper processing of the texts since 86.22% of the descriptions contain such terms. An additional difficulty from the point of view of NLP is the length of the descriptions, since too short texts may be insufficiently informative, and too long texts may be noisy for the task at hand. In this sense, Log_{desc} contains 110 descriptions with 3 or fewer words and 212 descriptions with more than 50 words. Finally, the event log has also been analysed using a grammar and spell checking tool[5], detecting a total of 1,642 errors (apart from the errors provoked by OOV words), which affects the $30,65\%$ of descriptions.

In order to evaluate the proposed steps, they have been applied to Log_{desc}, performing different sets of NLP techniques. At first, we have applied five techniques, thus, five new event logs have been created: Log_{acro}, the acronym detection is used to keep only these keywords; Log_{dep}, the dependency analysis is applied to keep only the root word of each description; Log_{lemma} contains the lemmatisation of the words; Log_{pos} applying POS tagging and keeping only those words tagged as "NOUN", "VERB" or "ADJ" (nouns, verbs and adjectives); and Log_{sent}, the sentence detection is used to split up each description into its constituent sentences. Second, we propose several pipelines of NLP techniques for the improvement of the event log quality. The Log_{sent} is used as the first step for all the proposed pipelines: Log_{sent_dep}, we simplify sentences only maintaining the root word; $Log_{sent\_dep\_lemma}$, we apply a lemmatisation to the root word previously obtained; $Log_{sent\_dep\_lemma\_acro}$, in addition to the lemmatised

[4] https://github.com/explosion/spacy-models/releases//tag/es_core_news_lg-3.0.0.

[5] Language-Tool: https://github.com/languagetool-org/languagetool.

form of the root words of each description, we also keep the acronyms within them; for Log_{sent_pos}, the POS tagging is applied to keep the nouns, verbs and adjectives of each sentence; $Log_{sent\_pos\_acros}$ the acronyms detected are added to the previous log; with $Log_{sent\_pos\_lemma}$ we keep the lemmatised forms of the words within Log_{sent_pos}; finally, $Log_{sent\_pos\_lemma\_acros}$ adds the acronyms to the previous log.

We have obtained the results of the quality assessment for each log previously described as shown in Table 6. The results show the value for each metric and the final quality reached. The results presented support how the application of the NLP techniques affect the measurements as estimated in Table 4. However, there exist dimensions, such as Relevancy, whose relation among the metric and the assessment is not lineal, e.g., when an NLP is applied to increase the relevancy metric, the assessment can become *Very Poor* instead of *Fair*.

Finally, the implementation of our framework used in this evaluation is available on a website [6].

Table 6. Dimensions values for the event logs applying NLP techniques.

Event log	Complexity	Uniqueness	Relevancy	Consistency	Quality assessment
Log_{desc}	6.734 (Excellent)	0.621 (Very Poor)	0.734 (Very Poor)	58.916 (Very Poor)	Poor
Log_{acro}	2.708 (Excellent)	0.229 (Excellent)	0.399 (Fair)	3.818 (Excellent)	Good
Log_{dep}	6.384 (Excellent)	0.177 (Excellent)	0.291 (Fair)	4.373 (Excellent)	Good
Log_{lemma}	6.707 (Excellent)	0.474 (Poor)	0.620 (Very Poor)	36.091 (Very Poor)	Poor
Log_{pos}	6.697 (Excellent)	0.439 (Poor)	0.586 (Poor)	26.671 (Poor)	Poor
Log_{sent}	8.886 (Good)	0.500 (Poor)	0.643 (Very Poor)	30.104 (Very Poor)	Very Poor
Log_{sent_dep}	8.227 (Good)	0.090 (Fair)	0.183 (Fair)	1.831 (Excellent)	Fair
$Log_{sent\_dep\_lemma}$	8.227 (Good)	0.057 (Fair)	0.126 (Fair)	1.733 (Excellent)	Fair
$Log_{sent\_dep\_lemma\_acro}$	8.325 (Good)	0.104 (Excellent)	0.190 (Fair)	2.560 (Excellent)	Good
Log_{sent_pos}	8.533 (Good)	0.413 (Poor)	0.566 (Poor)	18.924 (Fair)	Poor
$Log_{sent\_pos\_acro}$	8.564 (Good)	0.422 (Poor)	0.576 (Poor)	19.415 (Fair)	Poor
$Log_{sent\_pos\_lemma}$	8.533 (Good)	0.399 (Excellent)	0.552 (Poor)	18.563 (Fair)	Fair
$Log_{sent\_pos\_lemma\_acro}$	8.564 (Good)	0.410 (Poor)	0.562 (Poor)	19.051 (Fair)	Poor

7 Conclusions and Future Work

The preparation of an event log by carefully paying attention to its quality is crucial for the later (process mining) analysis. One of the difficulties is to ascertain when the quality is sufficient for a specific purpose, and which techniques to use to improve the quality. In this paper, we focus on the improvement of the event log quality by using NLP techniques that affect both the measurement and assessment. We propose: (1) a set of metrics to measure the quality of an event log; (2) a mechanism to describe both, data and process rules, for assessing the event log quality; and, (3) a guide for selecting the more proper NLP techniques to apply. The viability of the proposal has been demonstrated by an implementation applied to a real event log from an industrial context. As an extension of the paper, we plan to analyse how the quality of the event

[6] http://www.idea.us.es/loading-nlp.

log can be aligned with the quality of the process discovered. In addition, we will extend the number of metrics and mechanisms to improve the quality level of the event log, not only contextualised to the data textual analysis.

Acknowledgement. Projects (RTI2018-094283-B-C33, RTI2018-098062-A-I00), funded by: FEDER/Ministry of Science and Innovation - State Research, and the Junta de Andalucía via the COPERNICA (P20_01224) project.

References

1. van der Aa, H., Carmona, J., Leopold, H., Mendling, J., Padró, L.: Challenges and opportunities of applying natural language processing in business process management. In: Proceedings of the 27th COLING 2018, Santa Fe, New Mexico, USA, 20-26 August 2018, pp. 2791–2801 (2018)
2. van der Aa, H., Di Ciccio, C., Leopold, H., Reijers, H.A.: Extracting declarative process models from natural language. In: Giorgini, P., Weber, B. (eds.) CAiSE 2019. LNCS, vol. 11483, pp. 365–382. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-21290-2_23
3. van der Aa, H., Leopold, H., Reijers, H.A.: Comparing textual descriptions to process models - the automatic detection of inconsistencies. Inf. Syst. **64**, 447–460 (2017)
4. Van der Aalst, W.: Process Mining Discovery Conformance and Enhancement of Business Processes. Springer-Verlag, Heidelberg (2011). https://doi.org/10.1007/978-3-642-19345-3
5. van der Aalst, W.: Extracting event data from databases to unleash process mining. In: BPM - Driving Innovation in a Digital World, pp. 105–128 (2015)
6. van der Aalst, W.: Process Mining - Data Science in Action, 2nd edn. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-662-49851-4_1
7. van der Aalst, W., et al.: Process mining manifesto. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) BPM 2011. LNBIP, vol. 99, pp. 169–194. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28108-2_19
8. Andrews, R., van Dun, C.G.J., Wynn, M.T., Kratsch, W., Röglinger, M., ter Hofstede, A.H.M.: Quality-informed semi-automated event log generation for process mining. Decis. Support Syst. **132**, 113265 (2020)
9. Batini, C.: Data quality assessment. In: Liu, L., Özsu, M.T. (eds.) Encyclopedia of Database Systems, 2nd edn. Springer, New York (2018). https://doi.org/10.1007/978-1-4614-8265-9
10. Bose, R.J.C., Mans, R.S., van der Aalst, W.M.: Wanna improve process mining results? In: 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp. 127–134. IEEE (2013)
11. Chapela-Campa, D., Mucientes, M., Lama, M.: Discovering infrequent behavioral patterns in process models. In: Carmona, J., Engels, G., Kumar, A. (eds.) BPM 2017. LNCS, vol. 10445, pp. 324–340. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65000-5_19
12. Conforti, R., La Rosa, M., ter Hofstede, A.: Timestamp repair for business process event logs (2018). http://hdl.handle.net/11343/209011
13. Denisov, V., Fahland, D., van der Aalst, W.M.P.: Repairing event logs with missing events to support performance analysis of systems with shared resources. In: Janicki, R., Sidorova, N., Chatain, T. (eds.) PETRI NETS 2020. LNCS, vol. 12152, pp. 239–259. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-51831-8_12
14. Deokar, A.V., Tao, J.: Semantics-based event log aggregation for process mining and analytics. Inf. Syst. Front. **17**(6), 1209–1226 (2015). https://doi.org/10.1007/s10796-015-9563-4

15. Fischer, D.A., Goel, K., Andrews, R., van Dun, C.G.J., Wynn, M.T., Röglinger, M.: Enhancing event log quality: detecting and quantifying timestamp imperfections. In: Fahland, D., Ghidini, C., Becker, J., Dumas, M. (eds.) BPM 2020. LNCS, vol. 12168, pp. 309–326. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58666-9_18
16. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python (2020). https://doi.org/10.5281/zenodo.1212303
17. Leopold, H., Pittke, F., Mendling, J.: Ensuring the canonicity of process models. Data Knowl. Eng. **111**, 22–38 (2017)
18. Martin, N., Martinez-Millana, A., Valdivieso, B., Fernández-Llatas, C.: Interactive data cleaning for process mining: a case study of an outpatient clinic's appointment system. In: Di Francescomarino, C., Dijkman, R., Zdun, U. (eds.) BPM 2019. LNBIP, vol. 362, pp. 532–544. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-37453-2_43
19. Mocnik, F.B., Fan, H., Zipf, A.: Data quality and fitness for purpose (2017). https://doi.org/10.13140/RG.2.2.13387.18726
20. OMG: Decision Model and Notation (DMN), Version 1.2 (2019). https://www.omg.org/spec/DMN
21. Otto, B., Lee, Y.W., Caballero, I.: Information and data quality in networked business. Electron. Mark. **21**(2), 79–81 (2011). https://doi.org/10.1007/s12525-011-0062-2
22. Pittke, F., Leopold, H., Mendling, J.: When language meets language: anti patterns resulting from mixing natural and modeling language. In: Fournier, F., Mendling, J. (eds.) BPM 2014. LNBIP, vol. 202, pp. 118–129. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-15895-2_11
23. Rebmann, A., van der Aalst, H.: Extracting semantic process information from the natural language in event logs. CoRR abs/2103.11761 (2021)
24. Sadeghianasl, S., ter Hofstede, A.H.M., Suriadi, S., Turkay, S.: Collaborative and interactive detection and repair of activity labels in process event logs. In: 2nd ICPM, pp. 41–48 (2020)
25. Sadeghianasl, S., ter Hofstede, A.H.M., Wynn, M.T., Suriadi, S.: A contextual approach to detecting synonymous and polluted activity labels in process event logs. In: Panetto, H., Debruyne, C., Hepp, M., Lewis, D., Ardagna, C.A., Meersman, R. (eds.) OTM 2019. LNCS, vol. 11877, pp. 76–94. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33246-4_5
26. Suriadi, S., Andrews, R., ter Hofstede, A., Wynn, M.: Event log imperfection patterns for process mining: towards a systematic approach to cleaning event logs. Inf. Syst. **64**, 132–150 (2017)
27. Valencia-Parra, A., Parody, L., Varela-Vaca, A.J., Caballero, I., Gómez-López, M.T.: DMN4DQ: when data quality meets DMN. Decis. Support Syst. **141**, 113450 (2020)
28. Valencia-Parra, Á., Parody, L., Varela-Vaca, Á.J., Caballero, I., Gómez-López, M.T.: DMN for data quality measurement and assessment. In: Di Francescomarino, C., Dijkman, R., Zdun, U. (eds.) BPM 2019. LNBIP, vol. 362, pp. 362–374. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-37453-2_30
29. Valencia-Parra, Á., Ramos-Gutiérrez, B., Varela-Vaca, A.J., Gómez-López, M.T., Bernal, A.G.: Enabling process mining in aircraft manufactures: extracting event logs and discovering processes from complex data. In: Proceedings of the Industry Forum at BPM, Vienna, pp. 166–177 (2019)
30. Vanbrabant, L., Martin, N., Ramaekers, K., Braekers, K.: Quality of input data in emergency department simulations: framework and assessment techniques. Simul. Model. Pract. Theory **91**, 83–101 (2019)

31. Verhulst, R.: Evaluating quality of event data within event logs: an extensible framework. Master's thesis, Rijksuniversiteit Groningen, echnische Universiteit Eindhoven (2016)
32. Wynn, M.T., Sadiq, S.: Responsible process mining - a data quality perspective. In: Hildebrandt, T., van Dongen, B.F., Röglinger, M., Mendling, J. (eds.) BPM 2019. LNCS, vol. 11675, pp. 10–15. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-26619-6_2