Automatic Recommendation of Prognosis Measures for Mechanical Components based on Massive Text Mining[1]

(Extended Version)

Jorge Martinez-Gil, Bernhard Freudenthaler, Thomas Natschläger

Software Competence Center Hagenberg GmbH

Softwarepark 21, 4232, Hagenberg, Austria

{jorge.martinez-gil, bernhard.freudenthaler, thomas.natschlaeger}@scch.at

Keywords: Text Mining; Pattern-based Information extraction; Fault Prognosis

[1]Extended version of our already published conference paper: Jorge Martinez-Gil, Bernhard Freudenthaler, Thomas Natschläger: Automatic recommendation of prognosis measures for mechanical components based on massive text mining. iiWAS 2017: 32-39

Structured Abstract

Purpose

Automatically providing suggestions for predicting the likely status of a mechanical component is a key challenge in a wide variety of industrial domains.

Design/methodology/approach

Existing solutions based on ontological models have proven to be appropriate for fault diagnosis, but they fail when suggesting activities leading to a successful prognosis of mechanical components. The major reason is that fault prognosis is an activity that, unlike fault diagnosis, involves a lot of uncertainty and it is not always possible to envision a model for predicting possible faults.

Findings

This work proposes a solution based on massive text mining for automatically suggesting prognosis activities concerning mechanical components.

Originality/value

The great advantage of text mining is that makes possible to automatically analyze vast amounts of unstructured information in order to find corrective strategies that have been successfully exploited, and formally or informally documented, in the past in any part of the world.

1 Introduction

According the ISO 13381-1 standard for condition monitoring and diagnostics of machines [9], the prognosis of future fault progression and the recommendation of proper corrective actions is a key challenge to prevent failures and avoid its consequences in a wide range of industrial domains. The problem here is that finding relevant, yet accurate, corrective actions among the huge pool of available resources (manual, books, technical reports, etc.) is a titanic task, even for domain experts. In this context, there is a great need of methods and tools to automatically process information from this domain. In fact, being able to help professionals by automating information search tasks that are often expensive, tedious, and error-prone could be of great relevance.

Mainly due to the vital importance of this task and its economic impact, there is a number of past works trying to address the problem of fault prognosis regarding mechanical components [5, 12, 13, 14]. However, most of existing approaches in this context are based on the identification and exploitation of sources offering structured information. The reason is that structured information presents a number of qualitative advantages, among which stands out that structured information is more likely to be machine interpretable [24]. This means that if the structure of the information is formalized in a way that a computer program can process, that computer program can accurately carry out tasks with this information with no human supervision.

However, structured information does not abound in today's world. There is a number of reasons for that, including but not limited to the great effort (in terms of time, money and resource consumption) required to generate and maintain such kind of information [18]. We think that these high costs are certainly a limiting factor to build solutions leading to the successful prognosis of future fault progression. Therefore, we have worked to find a solution that can find satisfactory results at much lower costs. This solution has come from the hand of the exploitation of unstructured data, which are certainly much more abundant mainly because these data are naturally generated within the daily activities of the human being.

Therefore, in this research work, our aim is to find useful ways to perform a prognosis of future fault progression to help practitioners in their daily work. The problem here is that it is not always possible to be accurate in the prognosis process, since that process requires a high complex mixture of assessment, development of degradation models, failure analysis, health management,

etc. which is far from being trivial [15]. However, there is a number of techniques from the text mining and pattern-based information extraction field that can support this process by correctly handling statistical or testimonial evidences [7]. Using these techniques as a basis, we have designed a novel method based on a massive text mining over different corpora that is able to automatically provide hints in order to guess the future fault progression and corrective actions.

In this context, this work is an extended version of our previous conference paper [17], where we described the design and development of this novel method, and we explained why such as method has several qualitative advantages over structured models. In this extended version, our contribution is intended to go a step further beyond as we summarize below:

- We have extended the Introduction and State-of-the-art sections to clearly reflect the last developments in this context.
- We provide deeper insights regarding the design and development of our method for automatic recommendation of prognosis activities based on a Q&A paradigm being able to exploit huge text corpora in order to help overcoming some of the existing limitations in the field of prognosis suggestion regarding mechanical components.
- We have extended the initial empirical evaluation for including different configurations over different text corpora, in addition to the use of an additional well-known data set. The rationale behind this extended evaluation is to provide additional hints on the feasibility of the proposed approach in real settings.

The rest of this paper is organized as follows, Section 2 describes the related works concerning fault prognosis activities. Section 3 formally presents the problem we need to address to successfully providing a solution for automatic prognosis suggestion. Section 4 explains the technical details of our contribution, and the implementation details concerning our method. Section 5 shows the experiments that we have performed in order to validate our approach. Section 6 initiates a discussion about the pros and cons of this approach. And finally, we present the concluding remarks, and the possible future lines of research in this context.

2 Related Works

As the amount of available information in the context of industrial and mechanical engineering grows, better methods and tools are needed to appropriately handle all this information, therefore we are facing here a challenge which could help professionals by automating information search tasks that are often expensive, tedious, and error-prone. In the particular case addressed here, there is a great demand of techniques to find specific prognosis activities regarding mechanical components. In response to this need, researchers and practitioners have developed a number of methods and tools to face this challenge. We review here the body of existing literature in this regard.

One of the most popular techniques is the adaption of knowledge-based approaches, i.e. approaches that use structured knowledge bases to automatically derive facts. The reason is that these approaches have been already successfully applied in a number of scenarios concerning detection of problems in machinery. In fact, knowledge based-models aims to undertake tasks on fault diagnosis, operation decision-making and maintenance of mechanical components, based on knowledge facts by comparing present and past measurement data. According the surveyed literature, these models seem to work very well on situations concerning fault diagnosis. Among existing works, there are solutions that have proven to be successful in a wide range of fields including power transformers [20], windmills [27] and railway vehicles [26].

Unfortunately, two major problems remain: first of all, the amount of structured information that may allow us to build knowledge based approaches is very limited. Secondly, the limited number of solutions in this context are appropriate for a successful fault diagnosis, but there are not suitable for recommending prognosis activities. In fact, knowledge based models works well in fault diagnosis situations for a number of reasons, including the fact by appropriately analyzing existing (although possibly incomplete) data is possible to derive many facts on the nature of a given failure. However, prognosis involves guessing what is going to happen in a near future with regards to a particular mechanical component. Such an activity involves a high degree of uncertainty. This means that just analyzing existing data could not be enough for our purposes. This makes this task very difficult, since it requires experience, but also creativity and intuition to interpret facts that are fuzzy, and therefore, it is not always easy to quantify them (e.g. disturbing

noise, black smoke, strange power loss, and so on).

In summary, knowledge-based models are able to understand and classify failures in mechanical components, but they currently fail in the process of suggesting measures for anticipating potential problems. Additionally, these knowledge-based methods have a number of technical drawbacks that do not facilitate the design, implementation and testing of fault prognosis strategies. These drawbacks are certainly a limiting factor that does not allow to build real solutions. Some of these major drawbacks are:

- Building a knowledge base is expensive in terms of resource consumption
- It is difficult to find experts with enough knowledge of each existing mechanical component for creating or curating the knowledge base
- Building a knowledge base is subject to errors
- A knowledge base is difficult and expensive to maintain and update
- A knowledge base for a particular mechanical component is hardly reusable

For all these reasons, it makes sense exploring alternative approaches. In fact, we propose to work with the automatic analysis of patterns from text fragments which are assumed to contain meaningful information [19]. We show how corpora of different nature can be exploited beneficially and how the nature of the patterns influences the selection of the most promising prognosis activities in this context.

Nevertheless, there are also a number of technical limitations and problems that make our approach difficult. For example, the large variability of language requires accounting for an infinite amount of possible expressions that imply the same information [2]. Or the ambiguity of terms and sentences can make interpretation difficult [3]. However, by overcoming these technical limitations and problems, the possibilities of this approach could be of greater caliber, i.e. delivery of accurate results at extremely cheap cost of terms of human and computational resources. In the next sections, we explain the way that we have envisioned to successfully address the problem.

3 Problem statement

The problem we are facing can be formally defined as follows: Given a specific binary relation R, find instances

$$(x_1, x_2) \rightarrow Domain\ (R) \times Range\ (R)$$

that stand in the relation *R*. Thereby, *Domain (R)* and *Range (R)* need to be known in advance. The approach, i.e. getting an extraction model means finding a relation-specific mapping

$$R:\ T \rightarrow 0, 1$$

that decides for each fragment of text

$$t \rightarrow T$$

whether or not a given relation is expressed and in addition, an extraction function

$$extract\ (R) : T \rightarrow Domain(R) \times Range(R)$$

that determines the relation instance that is present.

According to the literature, there are several features that can be exploited to build such an extraction function [3]:

- Token-based features are those features in which these features belongs to the set of all individual minimal textual units (tokens). The most clear example of token-based feature is the token string itself.
- Mention-based features encode information that holds for the entire mention (i.e. the text fragment which is under consideration) which is relevant for deciding whether or not the target relation is present at that position.
- Structural features usually need to be encoded as combinations of several token-based or mention-based features.

Since we are aiming to build an universal approach, i.e. an approach that can be used to suggest prognosis activities regarding every kind of mechanical component in every kind of situation, exploiting domain-dependent token-based measures is not appropriate in this context. For

the same reason, structural features that analyze the position of tokens in a given text fragment do not allow us to build a method that can be exploited in every possible scenario. However, mention based features are exactly the kind of feature that can help us to recommend prognosis activities in this context. The reason is that if the mechanical component and potential prognosis activities are frequently mentioned in the same text frame of a text corpus, we can assume that it is a solution that has been already exploited in the past. We will explain the rationale behind this solution in the next section.

4 Text mining approach

To overcome the current limitations of knowledge based approaches, we propose to work with the automatic discovery of patterns from text fragments belonging to different corpora of unstructured text. Therefore, our text mining approach being able to mine massive amount of data in order to search of patterns to infer potential prognosis activities concerning mechanical components. The reason to propose such an approach is that we have identified that this way of proceed has a number of advantages over the state-of-the-art. For example, there is no need to formalize knowledge, which is usually a very time consuming task, and it is often subject to many errors. Moreover, a text mining approach like ours is able to analyze vast amounts of raw unstructured data in order to suggest a number of prognosis activities for a given mechanical component leading to save a great amount of resources (time, money and effort), since such an approach can benefit from the past (documented) experience of many people around the world, in order to suggest measures that lead to the successful prognosis in the mechanical domain.

Our text mining approach works under an adaptation of a well-known assumption (a.k.a. the distributional assumption [1]) that has proven to perform well for a number of problems in the past: mechanical components and prognosis methods will physically co-occur in a small fraction of the existing literature represented by means of a given corpus. Our goal is to identify and analyze this co-occurrence, in order to present to the expert our suggestions based on the interpretation of this co-occurrence.

The problem here is how to design a co-occurrence mechanism that can help to identify promising prognosis activities. The solution we propose is somehow inspired in how Q&A sys-

tems work[10], i.e. we propose to divide the process into two parts: the processing of a question and the formulation of a number of potential answers for that question. In this way, the question represents the $Domain(R)$ and at the same time the potential answers represents the $Range(R)$ that we have already defined in the Problem Statement.

It is important to note that the question will be formulated by the person who wish to receive suggestions regarding prognosis activities, whereas the pool of answers can be either manually introduced by the user or automatically generated by a solution such a Word2Vec [6], which is a model used to produce word embeddings, and in our particular context, it can be used to automatic generate the words related to a given concept [16]. In this way, we will automatically analyze huge corpora of unstructured text in order to identify what of the potential choices that have been generated has more potential to be useful in the context of the formulated question.

In summary, we are addressing here a variant of one of the classic sub-problems from the Question Answering (Q&A)[2] field focused in a very particular scenario, i.e. the challenge of multiple choice question answering. This problem has very clearly defined boundaries: given a question, and a set of answers which is made up of a correct answer plus as set of distractors, the aim is to figure out what the right answer is and what the distractors are. The inherent difficulty of the problem is that distractors are incorrect options leading to confusion or revealing a poor understanding of the topic, so they are usually chosen so that they are easily confused with the correct answer. Otherwise, poorly chosen distractors in a multiple choice setting can make questions almost trivial to solve. This is the fact that makes this research problem very similar to the problem that we are addressing here, with two slight nuances: all the possible answers have to be automatically generated, and it may be tolerable not to always get the right answer, as long as the right prognosis measure has a significant impact on the result, as we just aim to provide suggestions.

Therefore, it is clear that we are looking for a technique for choosing the correct candidate answer from among a small set of possible answers. Although the concept seems to be easy to understand, there is huge technical limitation for its development; such an approach is subject to an important number of technical obstacles which should be overcome [4]. These limitations are inherent to the process of massively text mining and include:

[2]Please do not confuse Question Answering (Q&A) with Quality Assurance (QA)

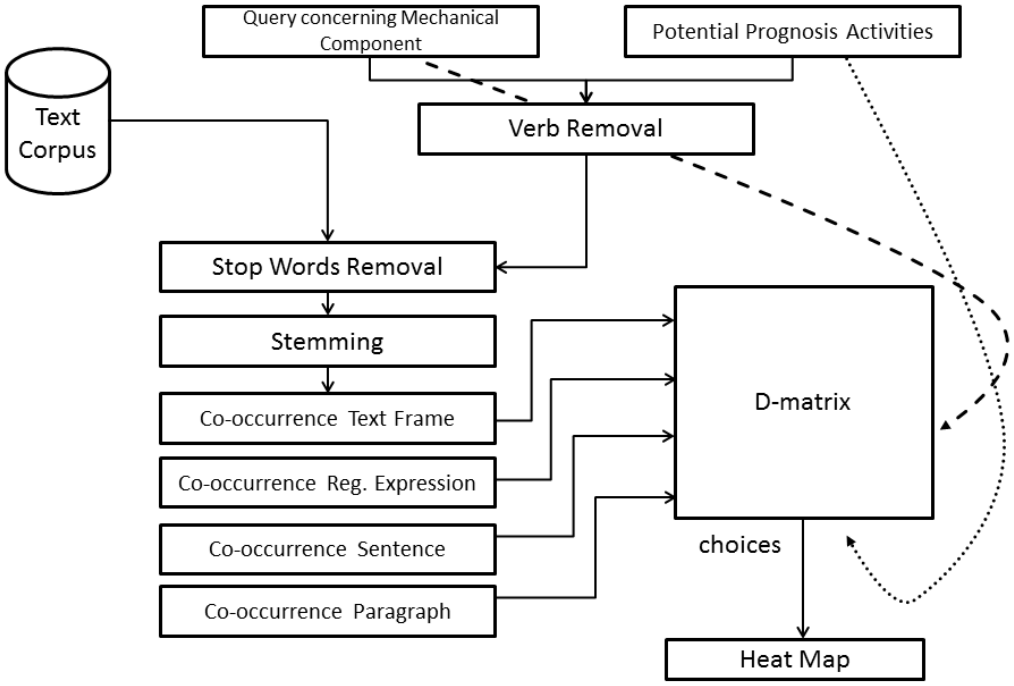

Figure 1: Overall view of our proposed solution. A question and some potential answers have to be initially formulated. Then, we analyze a corpus of unstructured text to detect the most promising co-occurrence patterns between the processed question and the potential answers. The result is achieved by selecting the most popular pair

- *Limitations concerning the corpus size*. It is clear that the size of the corpus have an impact on the time requested for dispatching a query. The reason is that extraction mechanisms operate under linear complexity in the best of cases. The reason is that all data has to be analyzed in order to determine if there is useful information to extract.

- *Variability inherent to the processing of natural language*. The reason is that our methods for information extraction try to detect patterns from the text to analyze. The problem here is that natural language is so rich and complex, that it is not always possible to detect all the possible variants that the same pattern can represent.

- *Issues concerning domain nomenclature*. One of the major problems for methods trying to exploit information extraction strategies is that they should be adapted to each different domain. The reason is that there is always jargon and other issues that just can be recognized

from experts in that field.

- *Degree of uncertainty on the accuracy of the contents*. There is an important number of issues to organize and work with different confident levels when managing information of textual nature. In fact, there are a number of features including but not limited to inconsistencies, errors, and even problems related to spam. All these factors make the information extraction processes even more complex since they need to operate with the concept of trust (or uncertainty).

- *Language in which the information is represented*. A first solution could be to restrict the information extraction processed to information sources using English since our intuition is that it is one of the most widely used languages in this field. However, a solution of this kind could sometimes face risks concerning the acquisition of very valuable information that is represented in other languages.

4.1 Contribution

For all these reasons, the design of proper methods in this context is far from being trivial. However, our strategy of rapid prototyping and testing using a number of representative experiments has shown us that it is possible to reach a reasonable level of success. According to our experience, the solution that works best is a method with four levels of confidence:

1. Mechanical component and prognosis hint co-occur in the same text frame (where the text frame is subject to parametrization)

2. Mechanical component and prognosis hint co-occur following a pre-defined regular expression (where regular expression can be chosen)

3. Mechanical component and prognosis hint co-occur in the same text sentence

4. Mechanical component and prognosis hint co-occur in the same text paragraph

Figure 1 shows us a overall view of our proposed solution. A question and potential answers are the basis for creating a decision matrix (D-Matrix). On the other hand, this D-Matrix is

populated by a pool of methods (each of them with a different level of trust) that analyze the co-occurrence of the question and answers in a text corpus. When the process is complete, it is possible to generate a heatmap from the D-Matrix in order to see what are the prognosis activities with a greater potential regarding the corpus of unstructured text.

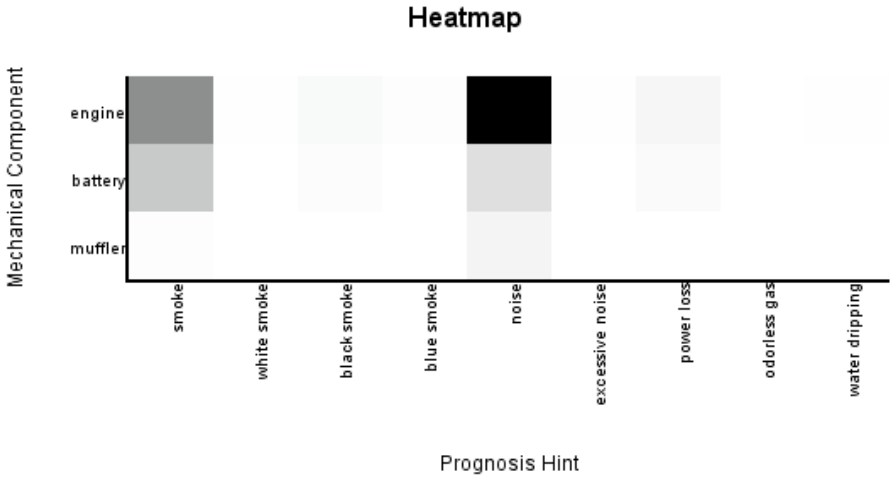

Figure 2: This automatically generated heatmap should be interpreted in the following way: if the practitioner experiences some smoke (specially black smoke), noise and a possible power loss in a given vehicle then a problem with the engine is expected. For batteries, just smoke and noise are expected. Whereas for mufflers, just smoke is expected

It is important to remark that we handle the concept of trust in terms of physical proximity [19]. For example, if a given mechanical method and a potential prognosis method appear in the same paragraph of a technical paper addressing a problem, we will have, at least, low evidence that could be a relation between them. But if this pair (mechanical component versus prognosis method) appears together frequently, in the same text frames or in the same regular expressions, then we can infer that the literature automatically analyzed suggests that the prognosis method is commonly used to monitor the given mechanical component. Please note that this is just a hypothesis that has to be validate by means of an exhaustive empirical evaluation.

Figure 2 shows us the resulting heatmap for a small use case where the most common symptoms of malfunctioning car components have been automatically identified. In this figure, it is

possible to see of interesting issues. For example, if the experience some smoke (specially black smoke), noise and a possible power loss then you have a problem with your engine. For batteries, just smoke and noise are expected. Whereas for mufflers, just smoke. The great advantage of using heatmaps for visualizing the solution is that allows the practitioner to see which are the most plausible solutions and their degree of reliability. Therefore, our approach's output does not give a single hint but a plot that gives an idea of the measures that could be most likely to be successful. Nevertheless, please note that this is one example extracted from a particular corpus, and results will present a great variation when other different corpora might be analyzed.

There are still, however, some technical difficulties that need further attention. For example, building a high quality corpus of material concerning fault prognosis of a wide variety of mechanical components is not an easy task. The reason is that published literature is usually very fragmented and it has been written in different languages and styles. Other serious problem is the word stemming, i.e. although two different pieces of literature can refer to the same mechanical component or prognosis method, these pieces can be written using plurals, temporal forms, slang words, and so on. For this reason, it is necessary to develop methods that can identify mechanical components and prognosis methods independently of how they appear written in literature. In general, we propose to avoid the processing of verbs (which can usually adopt a wide variety of forms) and focus on nouns that usually adopt a much more homogeneous representation.

4.2 Implementation details

We explain here the implementation decisions that we have taken in order to achieve a prototype for testing our hypothesis. The most important implementation details of our approach include:

- Limitations concerning the corpus size. With the emergence of new paradigms for parallelization and big data management, this kind of problems are losing importance.
- Variability inherent to the processing of natural language. It is widely assume that meaning is usually represented by nouns (and noun phrases) so that it is common to built retrieval methods based on noun representations extracted. For this reason, we have implemented some functionality to avoid processing verbs, and common stop words.

- Issues concerning domain nomenclature. The problem for methods trying to exploit information extraction strategies is that they should be adapted to each different domain. However, we are explicitly avoiding Token-based and Structural features, so that it is possible to partly alleviate this problem. However, there is still an issue concerning the use of very common word in either the question or the answers. The problem with common words is that they do not have a great meaning, and therefore we have a list of common words to be removed.

- Degree of uncertainty on the accuracy of the contents. In this work, we assume the fact that it is quite likely that the corpus to be analyzed might have some errors or inaccurate information concerning the information to be discovered. However, we foresee that the impact of these errors might be blurred by the overwhelming presence of correct information.

- Language in which the information is represented. To overcome this limitation, we have decided to use only English in this first version of our approach. Considering other existing languages remains as a potential future work.

5 Results

In order to validate our proposal, we have performed a set of experiments following the Question Answering style that our approach needs to appropriately suggesting prognosis activities in the mechanical domain. Since there is no a proper data set from this specific domain, we have managed to retrieve ten samples from the Stanford Question Answering Dataset (SQuAD) [22], and ten other samples belonging to the Open Trivia Dataset [3] (in particular we have focused on the Technology category).

With the help of one of our industrial partners, we have tried to select those questions that are more related to fault prognosis in the mechanical field, and that could be formulated at any given time by a practitioner wishing to have some hints. In this way, these questions simulate the situation whereby a practitioner could ask itself what is the way to proceed for assessing the likely status of a particular mechanical component in a given situation.

[3] https://github.com/uberspot/OpenTriviaQA

It is important to note that for the configuration of the system that we have used in these experiments, we have determined the following parameters:

- The text frame for determining the first kind of co-occurrence has been set up to 5 (what means that source and target expressions can be separated by up to five words)
- We use just the regular pattern is-a for determine the second kind of co-occurrence
- Every feature is weighted equally (no training has been performed in this work) what means that every kind of co-occurrence pattern detected when analyzing the corpus, increase the counter in just one unit
- Stop words and punctuation symbols are ignored
- The stemming library that we have chosen is Krovetz Stemmer [11]

Table 1 shows us the results of our approach. From the ten questions picked from the SQuAD dataset that we aimed to solve, our automatic approach has been able to guess the correct choice in 7 different cases. This means that we have achieved an accuracy of 70 percent.

Table 2 shows us the results of our approach from a different experiment. From the ten questions picked from the Open Trivia dataset, our automatic approach has been able to guess the correct choice in 7 different cases. This means that we have also achieved an accuracy of 70 percent.

However, in the next subsection, we will see that the results are even more positive in a real environments where a single response is not usually expected, but only prognosis recommendations.

Question	**Choices**	**Correct**	**Provided**
What device is used to recycle the boiler water in steam engines?	- Piston - Water pump -Cylinder -Valve	Water pump	Cylinder
What is often needed to make combustion happen?	- Condenser - Crankcase - Aluminum alloys - Ignition event	Ignition event	Ignition event
What motion does a steam engine produce?	- Rotary - Linear - Reciprocating - Oscillating	Rotary	Rotary
What are the stages in a compound engine called?	- Seasons - Chain changes - Expansions - Shortcuts	Expansions	Seasons
Where is the combustible material burned within the engine?	- Steam turbine - Firebox - Steel chamber - Muffler	Firebox	Firebox
What kind of device is a dry cooling tower similar to?	- Automobile radiator - Piston ring - Connecting rod - PCV valve	Aut. radiator	Piston ring
What is another term for rotors?	- Tractors - Rotating discs - Steering gears - Spokes	Rot. discs	Rot. discs
In an atmospheric engine, what does air pressure push against?	- Condenser - Seal - Plug Valve - Piston	Piston	Piston
What is a clear example of a pump component?	- Yoke - Gearbox - Injector - Bunker	Injector	Injector
What is a term that means constant temperature?	- Isothermal - Heat capacity - Combustion - Steam	Isothermal	Isothermal

Table 1: Results automatically achieved by our text mining approach when solving a subset of ten questions from the Stanford Question Answering Dataset (SQuAD) concerning mechanical engineering.

Question	**Choices**	**Correct**	**Provided**
What of these chemical elements is tasteless, colorless and odorless?	- Chromium - Magnesium - Oxygen - Hydrogen	Hydrogen	Hydrogen
What substance is naturally present in water?	- Nitrite - Nitrate - Chlorine - Fluoride	Fluoride	Chlorine
What force slows down and stops a moving object?	- Friction - Gravitational - Magnetic - Molecular	Friction	Friction
What unit is the base unit of electrical current?	- Ohm - Frequency - Ampere - Capacitor	Ampere	Frequency
This gas is produced mainly by automobiles	- Acid sulfates - Carbon oxide - Carbon monoxide - Carbon dioxide	Carbon monoxide	Carbon monoxide
What does a hydrometer measure?	- Air temperature - Liquids temperature - Liquids weight - Liquids density	Liquids density	Liquids density
What electronic components replaced vacuum tubes in computers?	- Diodes - Chips - Transistors - Processors	Transistors	Transistors
How are floppy disks read?	- Laser - Beam of Nuclear Radiation - Magnetically - Sensor	Magnetically	Magnetically
What electric power is produced by running water?	- Nuclear - Aquatic - Hydrothermal - Hydroelectric	Hydroelectric	Nuclear
Which of these is an output device that produces audible sound?	- Printer - Microphone - Speaker - Monitor	Speaker	Speaker

Table 2: Results automatically achieved by our text mining approach when solving a subset of ten questions from the Open Trivia Dataset.

These good results have been achieved by using the Wikicorpus [23], a large general purpose data set created from Wikipedia in order to test different approaches from the text mining field. This corpus has a size of near 4 GB of raw text (approx. 140 million words). However, it is not always possible to get so good results. In fact, we have performed more experiments using smaller corpora. However, these results were not complete satisfactory. Bad results in this context are given because these corpora are very small or so specific that do not contain the nomenclature necessary to reply our questions. Please note that when our approach is not able to find any solution, it is always possible to choose one answer in a random manner, this means a accuracy rate of approximately 25 percent for the case of dealing four possible choices. However, for facilitating the reproducibility of our work, we prefer to avoid this method when reporting our results.

5.1 Less strict evaluation of the results

Although in the previous subsection, we have strictly evaluated our approach in terms of correctly guessing the answer or not, in real environments our approach is able to fulfill its purpose even more often. Let us look at the specific case of the question *What electric power is produced by running water?* which, in fact, seems reasonably easy for a person, but that our approach has not been able to guess. Figure 3 shows us the heatmap that has been automatically produced for this particular sample.

In theory, our approach fails since it indicates that *nuclear* is the correct choice: 97 co-occurrence points versus 92 co-occurrence points for the second ranked answer. But, by manually inspecting the heatmap, a domain expert would be able to see that the reason for this answer is that the *nuclear* option is very strongly linked to the *electric power* segment of the question and weakly linked to the *running water* segment. In addition, the domain expert can easily see that

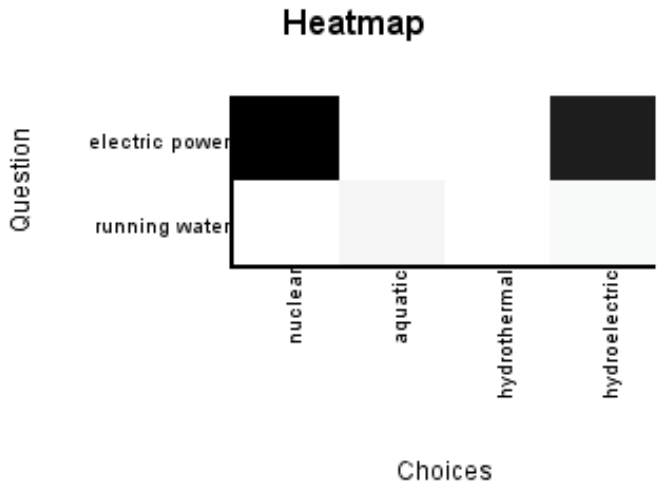

Figure 3: Automatically generate heatmap for the question *What electric power is produced by running water?*

the correct option has a very strong impact on the heatmap as well. In fact, it presents the same order of magnitude as the correct answer in the segment *electric power*, and it is also linked to the *running water* segment. Therefore, an expert who would know how to properly inspect and interpret the heatmap, might not so easily discard the option that is the correct one.

6 Discussion

This section is devoted to analyze the pros and cons of our text mining proposal in relation to a knowledge base approach. In particular, our approach presents a number of qualitative advantages. However, it is not less certain that there is still a number of technical limitations that should be faced in the future.

6.1 Advantages

Concerning our approach, we think that it is possible to envision five major advantages:

1. *Building a knowledge base is expensive in terms of resource consumption*. However, our approach for massive text mining does not involve the development of formal models from scratch, including entities, relations, instances, axioms, and so on. We just need to adapt or improve well-known text mining methods for getting the first meaningful results.

2. *It is difficult to find experts with enough knowledge of each existing mechanical component for creating or curating the knowledge base*. However, with our approach there is no need of creating or curating the (already) existing corpus of technical literature implicitly contain the knowledge necessary to perform our tasks.

3. *Building a knowledge base is subject to errors*. However, in our approach although it is possible to find errors in the vast amount of technical literature that we analyze, its impact is blurred by the overwhelming presence of correct information.

4. *A knowledge base is difficult and expensive to maintain and update*. However, our text mining approach does not need any kind of maintenance and the updates can be done programmatically.

5. *A knowledge base is hardly reusable*. However, our text mining approach can be used for any mechanical component that exists with no extra cost.

6.2 Limitations

It is also possible to identify a number of disadvantages; evaluating all text fragments one by one making the amount of processing time grow linear with the amount of text. This means that

for scenarios working with huge text corpora, the response time could be not reasonable enough. Fortunately, the emergence of new paradigms for parallel computation in big data environments might help to greatly mitigate this problem.

Concerning the use of verbs and its variations, our approach is not able to properly work with the different personal and temporal forms that are inherent to the nature of these verbs. Maybe, recent advances in natural language processing for the automatic recognition of word roots could face this kind of limitation.

Moreover, it is important to remark that the choice of the different alternatives for answering the questions is a critical point. Therefore, it is necessary to evaluate the fairness of the choices to be evaluated. In the future, we want to use the knowledge bases YAGO [25] and YAGO2 [8]. These knowledge bases should allow us to automatically extract the different parts of a mechanical device. It is supposed, that in that case, the fairness of the multiple choices to be evaluated is high, since every part of the mechanical device is likely to present future problems.

Finally, it is also worth mentioning that some kind of sentiment analysis [21] should be performed. The rationale behind this idea that if two text expressions co-occur in the same physical space but with a negative polarity, then we should discard that the original author referred to a potential prognosis activity.

7 Conclusions and future work

In this work we have described our novel approach for massive text mining that tries to face the challenge of assisting experts on the prognosis of future fault progression regarding mechanical components. To do that, we have designed an approach which is based on the analysis of vast amount of written information to discover textual patterns, i.e. explicit descriptions of text frag-

ments, that may allow us to automatically provide suggestions leading to a successful prognosis of mechanical components.

Our research has concluded that an approach based on mining vast amounts of technical literature presents a larger number of advantages, including: less resource consumption, no need of expert support, (almost) error-free data, no need of manual maintenance, and high level of re-usability. As a disadvantage, evaluating all text fragments one by one making the amount of processing time grow linear with the amount of text being analyzed.

The results that we have achieved from our experiments seem to be promising. In this context, our approach has been able to successfully address of a subset concerning mechanical components from the Stanford Question Answering Dataset with a 70% of accuracy and another subset from the Open Trivia Dataset[4] with a 70% of accuracy. Although the results may vary depending on the configuration and the corpora being chosen.

As future lines of research, we need to work towards improve the technical limitations that we were not able to overcome in this work. This includes the work with textual corpora from different languages at the same time, the proper consideration of verbs when formulation questions and preparing potential answers, the sentiment analysis of the text expressions, and the proper weighting of the different features by means of a training phase. We think that by successfully addressing these research challenges, it is possible to build solutions that can help to the industry to overcome one of the most serious problems that they have to face in their daily activities. It is also possible to envision that the community needs a standard benchmark dataset to properly assess the accuracy of new developments.

[4]https://github.com/uberspot/OpenTriviaQA

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments and suggestions. The research reported in this work has been carried out in the frame if the project PROSAM funded by the Austrian Research Promotion Agency (Project Number 845578) and by the Austrian Ministry for Transport, Innovation and Technology, the Federal Ministry of Science, Research and Economy, and the Province of Upper Austria in the frame of the COMET center Software Competence Center Hagenberg (SCCH).

References

[1] R. Basili, M. Pennacchiotti: Distributional lexical semantics: Toward uniform representation paradigms for advanced acquisition and processing tasks. Natural Language Engineering 16(4): 347-358 (2010).

[2] S. Blohm, P. Cimiano, E. Stemle: Harvesting Relations from the Web - Quantifiying the Impact of Filtering Functions. AAAI 2007: 1316-1321.

[3] S. Blohm: Large-scale pattern-based information extraction from the world wide web. Karlsruhe Institute of Technology 2010, ISBN 978-3-86644-479-9, pp. 1-236.

[4] S. Blohm, P. Cimiano: Using the Web to Reduce Data Sparseness in Pattern-Based Information Extraction. PKDD 2007: 18-29.

[5] C. Chen, D. Brown, C. Sconyers, B. Zhang, G. J. Vachtsevanos, M. E. Orchard: An integrated architecture for fault diagnosis and failure prognosis of complex engineering systems. Expert Syst. Appl. 39(10): 9031-9040 (2012).

[6] K. W. Church: Word2Vec. Natural Language Engineering 23(1): 155-162 (2017).

[7] D. Freitag: Machine Learning for Information Extraction in Informal Domains. Ph.D. dissertation, Carnegie Mellon University (1998).

[8] J. Hoffart, F. M. Suchanek, K. Berberich, G. Weikum: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artif. Intell. 194: 28-61 (2013).

[9] ISO. Condition monitoring and diagnostics of machines-prognostics - Part 1: General guidelines. Int. Standard ISO13381-1, 2015.

[10] O. Kolomiyets, M.-F. Moens: A survey on question answering technology from an information retrieval perspective. Inf. Sci. 181(24): 5412-5434 (2011).

[11] R. Krovetz: Viewing morphology as an inference process. Artif. Intell. 118(1-2): 277-294 (2000).

[12] Y. Lei, Z. He, Y. Zi: Application of an intelligent classification method to mechanical fault diagnosis. Expert Syst. Appl. 36(6): 9941-9948 (2009).

[13] P. P. Lin, X. Li: Fault Diagnosis, Prognosis and Self-Reconfiguration for Nonlinear Dynamic Systems Using Soft Computing Techniques. SMC 2006: 2234-2239.

[14] S. Huang, K. K. Tan, T. H. Lee: Automated Fault Detection and Diagnosis in Mechanical Systems. IEEE Trans. Systems, Man, and Cybernetics, Part C 37(6): 1360-1364 (2007).

[15] C. Ly, K. Tom, C. S. Byington, R. Patrick, G. J. Vachtsevanos: Fault diagnosis and failure prognosis for engineering systems: A global perspective. CASE 2009: 108-115.

[16] L. Ma, Y. Zhang: Using Word2Vec to process big text data. Big Data 2015: 2895-2897.

[17] Jorge Martinez-Gil, Bernhard Freudenthaler, Thomas Natschlger: Automatic recommendation of prognosis measures for mechanical components based on massive text mining. iiWAS 2017: 32-39.

[18] J. Martinez-Gil: Automated knowledge base management: A survey. Computer Science Review 18: 1-9 (2015).

[19] J. Martinez-Gil: An overview of textual semantic similarity measures based on web intelligence. Artif. Intell. Rev. 42(4): 935-943 (2014).

[20] M.-A. Mortada, S. Yacout, A. Lakis: Fault diagnosis in power transformers using multi-class logical analysis of data. J. Intelligent Manufacturing 25(6): 1429-1439 (2014).

[21] A. B. Pawar, M. A. Jawale, D. N. Kyatanavar: Fundamentals of Sentiment Analysis: Concepts and Methodology. Sentiment Analysis and Ontology Engineering 2016: 25-48.

[22] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang: SQuAD: 100, 000+ Questions for Machine Comprehension of Text. EMNLP 2016: 2383-2392.

[23] S. Reese, G. Boleda, M. Cuadros, L. Padr, G. Rigau: Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus. LREC 2010

[24] F. M. Suchanek, G. Ifrim, G.Weikum: Combining linguistic and statistical analysis to extract relations from web documents. KDD 2006: 712-717.

[25] F. M. Suchanek, G. Kasneci, G. Weikum: YAGO: A Large Ontology from Wikipedia and WordNet. J. Web Sem. 6(3): 203-217 (2008).

[26] Y. Zhao, T. Xu, H. Wang: Text mining based fault diagnosis of vehicle on-board equipment for high speed railway. ITSC 2014: 900-905.

[27] A. Zhou, D. Yu, W. Zhang: A research on intelligent fault diagnosis of wind turbines based on ontology and FMECA. Advanced Engineering Informatics 29(1): 115-125 (2015).