Automated Operations Classification using Text Mining Techniques

Bilal Esmael
University of Leoben
Leoben 8700, Austria
Bilal@stud.unileoben.ac.at

Mohammad Arghad Arnaout
TDE GmbH
Leoben 8700, Austria
Arghad.Arnaout@tde.at

Rudolf K. Fruhwirth
TDE GmbH
Leoben 8700, Austria
Rudolf.Fruhwirth@tde.at

Gerhard Thonhauser
University of Leoben
Leoben 8700, Austria
Gerhard.Thonhauser@unileoben.ac.at

***Abstract*- Classifying textually-described drilling operations presents big challenges because drilling personnel use a specialised way of describing the drilling process as part of their daily reporting routine. They typically use short summaries to describe the activities. In most times they write a short technical text that includes the most important measurements related to current situation to describe the activity in hand.**

In this paper we propose an approach to extract useful information from drilling activities described in daily morning reports, then making high level classification using a Support Vector Machine.

The proposed approach extends the "bag of word" representation used in text mining methods. We represent each activity as features vector which contains three groups of features namely: measurements features, keywords features and phrases features. Extracting all these types of features will be done by exploiting the internal semantics from the original text and external knowledge bases such as WordNet[1] and drilling ontologies.

Measurements features will be extracted using a patterns-based parser; whereas keywords and phrases features will be extracted by analysing the text syntactically.

Experimental evaluation with real daily morning reports shows that our approach has ability to construct an efficient features space and build an accurate classifier for drilling operations classification task.

Index Terms – Text Mining, Short Text Classification.

I. INTRODUCTION

Text classification methods base on the feature-based representations of original text and not the text itself. Many potential features can be employed to represent texts, but the features most commonly used are: bigrams, trigrams, words, terms and phrases [1].

The simplest feature-based representation is the "bag of words" representation. That is, each text is represented by a vector of words appearing in the text. This vector contains the words and their weights. The weight of a word is the frequency of that word in the original text.

The "bag of word" (BOW) representation is often insufficient and has several limitations when directly applied to short text [2, 4, and 5]. Short texts consist of a few phrases or 2–3 sentences, thus they do not provide sufficient term co-occurrence information [3].

Many research investigations show that external knowledge bases, such as ontologies and dictionaries, can be used to enrich the bag of words representation and fill the semantic gap in text representation [5, 6, and 7].

Although most of machine learning classification methods such as Neural Networks, Decision Trees and Bayesian Classifiers can be used for text classification, the Support Vector Machine is the preferred one in most cases, because it can deal with high dimensional data [8], like found in features-based representation.

The rest of the paper is organized as follows: Section II presents the general framework of the proposed approach. Sections III, IV and V show the details of Measurements Features Extraction, Keywords Features Extraction and phrases Features Extraction. Section VI shows the details of Activities Classification task, and the last section VII shows the experimental results of the proposed approach.

II. THE GENERAL FRAMEWORK

In this section we will introduce the proposed approach that aims to classify drilling operations by extracting three different groups of features and utilizing the internal and external semantics.

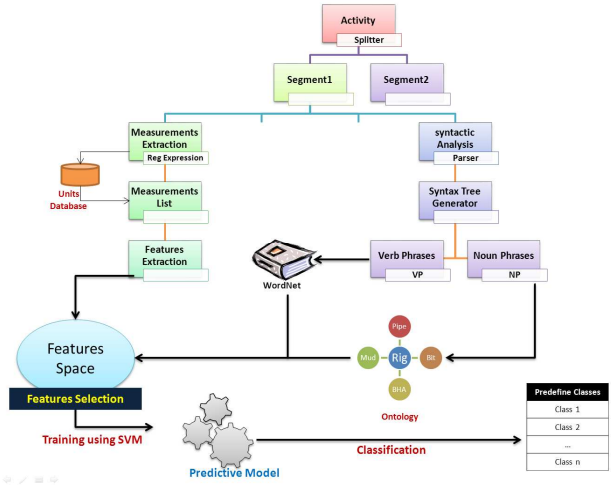

Fig. 1: The general phases of our approach

[1] http://wordnet.princeton.edu/

The workflow consists of three sequential phases, including Activities Segmentation, Features Extraction and Activities Classification. Figure1 describes these phases.

In the first phase, the original text is split into segments with the help of punctuations. Each segment contributes to provide a subtopic of the original text. For example the following text:
"Slugged the string, 570 lpm / 17-23 bar. POOH with the 8 1/2" BHA on 5 1/2" DP from 1871 m to 798 m, 692 m/hr. Unable to function the DDM."
will be divided into three segments, each one represents a separate sentence:

- Slugged the string, 570 lpm / 17-23 bar.
- POOH with the 8 1/2" BHA on 5 1/2" DP from 1871 m to 798 m, 692 m/hr.
- Unable to function the DDM.

In the second phase the features space will be constructed by extracting informative features as far as possible. To detect what the informative features are; we analyzed a huge amount of activities, we observed that:

- Most activities have measurements such as Mud Weight, Hole Depth, Pump Pressure, etc.
- Most activities start with a verb, or contain a verb which represents a meaningful feature to classify the drilling operations.

Measurements features and keywords features such as verbs will play an important role in activities classification task, so these features should be extracted correctly and accurately.
In addition to measurements and keywords features we found that using key phrases as features will help us in building an accurate classifier, because phrase-level features are informative and more focused on one concept. To extract phrases features efficiently, stemming and stop-words removal processes should be executed.

In the third phase, after constructing the features space, Support Vector Machine technique or any others techniques can be used to train a classifier to classify any new activities.
Next Sections will present the details of these three phases.

III. Measurements Features Extraction

Measurements should be extracted from the text because they have special meaning and cannot be interpreted using text mining techniques. These measurements need special dealing, to give us very useful information that can be used later to classify the drilling operations.

For measurements features extraction, a simple patterns-based parser has been developed. This parser uses regular expressions to search for measurements and extracting them. Two groups of patterns have been supported. The first group is used for single value measurements extraction, and the second group is used is for range measurements extraction. For example, using the developed parser with the following text: *"Drilled 12 1/4" hole from 2111 m to 2265 m with 35 - 50 m/hr using 3250 lpm, 268 bar."* will give us one range and five single values namely: *"12 ¼", "from 2111 m to 2265 m", "35-50 m/hr", "3250 lpm"* and *"268 bar"*.

In most cases the extracted measurements are ambiguous because they contain unknown units such as "m", "m/hr", "lpm", etc. To clarify the extracted measurements, external knowledge bases should be used. We used a special drilling database that contains all possible units and the corresponding measurement name. For example, form this database we can know that "bar" is the unit of "pressure".

The final step in this phase is measurements features extraction. For range measurements we check the first value and last value, and then construct the appropriate feature. For example from this extracted range "from 2111 m to 2265 m" we generate "Depth increased" feature.

The developed parser can be extended by adding new patterns to find new unconsidered cases.

IV. Keywords Features Extraction

Although short technical texts do not provide sufficient term co-occurrence information, they contain keywords which can be employed as features in text mining. For keywords features extraction the traditional syntactic technologies is used. Full parsing to analyze the original text is done, then a stemmer to return each word to its basic form (root) is used. Finally the WordNet dictionary was used to select the type of each word.
Three categories of part of speech were created: verbs, nouns and adjectives, and stored all extracted words under the appropriate category to be used later. All words with their frequencies were stored to be able to select the keywords based on the frequency. After analysing a huge amount of activites, and extracting all verbs, nouns and adjectives, we got 300 adjectives, 700 verbs and 1200 nouns.
In this phase the most frequent verbs as keywords features will be used. Figure 2 shows the most frequent verbs such as "drill", "test", "pump", "run", etc. Nouns and adjectives will be ignored in this phase. They will be used as phrases features in the next phase.

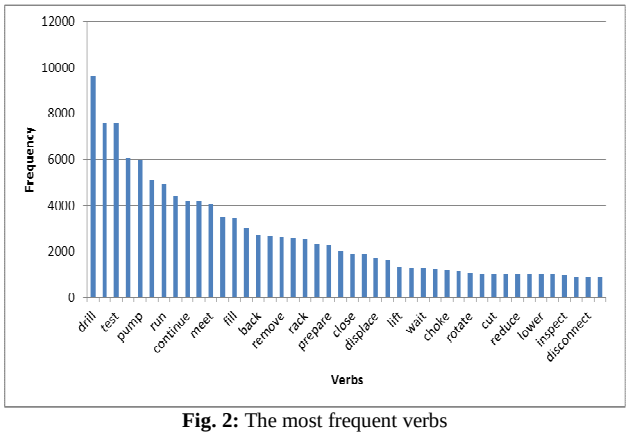

Fig. 2: The most frequent verbs

The disadvantage of keywords features is that they are too general to generate meaningful features, but on the other hand, combining these features with measurements and phrases features will increase the classification accuracy.

V. Phrases Features Extraction

In this phase the Link Grammar Parser[2] to analyze the text syntactically and extract the "key phrases" was used.

Syntactic analysis of the text will produce the syntax tree that divides sentences into a series of words that together compose a grammatical unit, such as noun phrase, verb phrase, etc. For example the syntactic analysis of the following activity: *"Drilled 3 m new formation from 992 m to 995 m with 4000 lpm/ 195 bar, 87 rpm/ 6-9 kNm and 3-5 ton WOB. Circulated cuttings behind BHA."* will result the syntax tree shown in Figure 3.

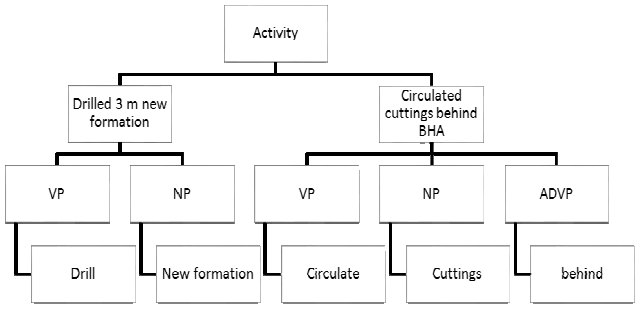

Fig. 3: A Syntax tree

Before the phrases features extraction phase, all measurements should be deleted from the sentence, otherwise we will get a confused output because numbers will be considered as nouns and units mainly will be unknown word.

In this phase all noun phrases will be added to the features space and verb phrases will be ignored because they are already in the features space.

VI. Activity Classification

After constructing the features space by extracting the three different types of features, we are ready to start classification process. Classification task has been done using Support Vector Machine classifier which is a state-of-the-art classification method. We used SVM due to its high accuracy and ability to deal with high-dimensional data like we have.

SVMs belong to the general category of kernel methods. The most important point should be taken into account when using SVM is selecting an appropriate kernel, and determining the best parameters.

Most people randomly try a few kernels and parameters, and in most cases they cannot build an accurate classifier. The authors of paper [8] propose the following procedure when using SVM tools:

- Transform data to the format of an SVM package
- Conduct simple scaling on the data
- Consider the RBF kernel $K(x,y)=\mathrm{Exp}(-\gamma|x-y|^2)$
- Use cross-validation to find the best parameter C and γ.
- Apply the best C and γ to train the whole training set
- Test the classifier

We applied this procedure and used different values for C and γ. We got different accuracy each time. In addition a linear kernel instead of RBF kernel was used and found that in case of features the count is big. Like found for the text classification task, using linear kernel for training is better, than using RBF kernel

VII. Experimental Results

To evaluate our approach, it was tested with a data set of real daily morning reports. We used a dataset consisting of 87470 instances and 19 classes. Each class represents one drilling operation.

At the beginning we analysed the dataset and generated the histogram of the classes. Then we took the first seven classes and eliminated the rest as shown in Figure 4.

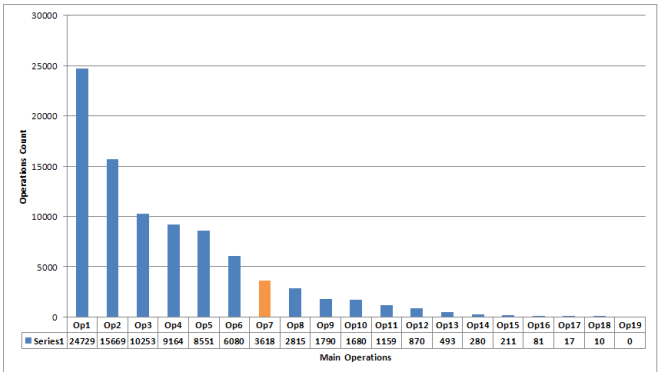

Fig. 4: The histogram of drilling operations

We divided the whole dataset into three datasets. The first one is for training and the others for testing.

We applied our approach to construct the features space, then we used LIBSVM[3] library to train a SVM classifier.

We followed the procedure mentioned in section VI and used two types of kernels: linear kernel and RBF kernel and compared the results. We found that with linear kernel we can build a more accurate classifier than RBF kernel. Finally, we tested the trained classifier with tested datasets. The following Table I shows the result:

Dataset	Instances	Accuracy (RBF kernel)	Accuracy (linear kernel)
#1	15000	52.7%	82%
#2	20000	50.9%	80%

Table I. The results

[2] http://www.link.cs.cmu.edu/link/

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

VIII. Further Work

Our aim in future work is to improve this system and extend it to be able to detect the drilling problems described in daily morning reports, and extract the steps that have been taken to solve these problems.

Acknowledgment

We thank TDE Thonhauser Data Engineering GmbH for the permission to publish this paper.

References

[1] R. Feldman, J. Sanger, "The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data", Cambridge University, 2007.

[2] X. H. Phan, M. L. Nguyen, S. Horiguchi: "Learning to classify short and sparse text & web with hidden topics from large-scale data collections." WWW 2008: 91-100

[3] X. Hu, N. Sun, C. Zhang, C. Tat-Seng. "Exploiting Internal and External Semantics for the Clustering of Short Texts Using World Knowledge". In Proceedings of the 18th ACM Conference on Information and Knowledge Management (*CIKM 2009*). November 2-6, 2009, HongKong, China. pp. 919-928.

[4] S. Aubin , A. Nazarenko, C. Nedellec "Adapting a General Parser to a Sublanguage". Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05) (2005) 89-93

[5] A. Hotho, S. Staab, G. Stumme, "Ontologies Improve Text Document Clustring", Third IEEE International Conference on Data Mining (ICDM'03), 2003

[6] S. Banerjee, K. Ramanathan, A. Gupta "Clustring Short Texts using Wikipeda" 30th Annual international ACM SIGIR conference, 2007

[7] L. Wenyin et al., "A short text modeling method combining semantic and statistical information", Information Sciences Journal, Volume 180 Issue 20, October, 2010

[8] C. Hsu, C. Chang, C. Lin, "A Practical Guide to Support Vector Classification", National Taiwan University, Taipei 106, Taiwan, http://www.csie.ntu.edu.tw/~cjlin