Machine Reading of Warranty Claims for Classification
Robert McCurdy, Mack Trucks, Inc.

ABSTRACT

A newly-developed SAS® program reads the free-form text of warranty claims for specific facts to better identify part failure causes, reducing the need for a person to read the claims. Parsing the text is challenging because there is no standard writing style. For example, a typical claim could read "inspected cam," "cam shaft inspected," "install new cam," or it could appear in other more complex forms. This program was developed (1) to determine whether the part was replaced; (2) to identify part failures due to an inherent weakness of the part; (3) to classify specific failure causes such as, "broken in half" or "lobe worn"; and (4) to flag ambiguous claims for human review. The program recognizes verb key words (replace, R&R, install, renewed); object key words (cam, camshaft, cam shaft); as well as noun key words (cam, camshaft, lobe); and adjective key words (bad, damage, worn), and matches them. Then an internal score card determines various classifications.

BACKGROUND

In use at Mack Trucks, Inc., this SAS program can be best explained in the context of Mack's warranty claims handling process.

After a Mack® truck is repaired at a dealer location, personnel enter data into Mack's computer system. The computer system's standard fields require information including the truck's serial number, engine serial number, and replacement part number. In addition to the system's standard fields, there are fields for free-flowing narrative where the description of the repair is entered.

INTRODUCTION

Currently the free-form narrative of the claim has to be read by a person in order to get an accurate picture of failures of a specific part. The reader must decipher the various shorthand codes used on the form to determine if there is an inherent part failure, or if there are other contributing factors. Further, the reader looks for descriptions that point exactly to where on the part the failure occurred. Once these determinations are made, the reader hand codes the warranty claim record. Human claim-by-claim reading is labor-intensive and error-prone. This SAS program reduces this labor-intensive effort.

SAMPLE PART

The part chosen for use as an example in this paper is the camshaft. The same approach works for any part in the warranty claims process. However, a verb-object vocabulary or a noun-adjective vocabulary would have to be developed for each part included in the automated claim classification program.

PART DESCRIPTION AND DESIRED INFORMATION

The camshaft under discussion is used on the in-line 6 cylinder engine which has 18 lobes--6 lobes for exhaust valves, 6 lobes for intake valves and 6 lobes for the fuel pumps. The fuel pumps go by the names unit pump and EUP. The facts the program searches for as it reads the warranty claims are as follows:

1) Was the cam replaced or not replaced?
2) Which type of lobe failed?
   A) Exhaust
   B) Intake
   C) Unit Pump (EUP)
   D) Not specified
3) Which lobe or lobes failed (1 through 6)?
4) Were there other causes for the failure?
   A) Coolant in the oil
   B) Fuel in the oil
   C) Dropped valve
   D) Dry cam (run without oil)

The program then codes the warranty claim as (1, 2, or 5) explained in the following table:

Cam Failure as the result of an intake lobe or an exhaust lobe failure or non specific statement of cam failure	1
Cam failure as the result of a unit pump lobe failure	2
Cam failure due to other causes such as, coolant in oil, or the claim was not a cam claim	5

A secondary benefit is obtaining the exact problematic lobe number or numbers (1, 2, 3, 4, 5, or 6).

PROGRAM ORGANIZATION

The program is written in base SAS version 6.09 and runs on a mainframe. It is divided into sections. Most sections do a search for a specific cause of part failure. One section searches for a general statement of failure, and another section searches for a statement of

actual replacement of the part. Using camshaft as an example, the sections are as follows:

    A) Exhaust lobes
    B) Intake lobes
    C) EUP lobes
    D) General statement of failure
    E) Determination of actual replacement
    F) Failure due to other causes
    G) Summary of findings

Each section parses the text for its specific task.

CHALLENGES TO PARSING TEXT

Parsing the text is challenging because there is no standard writing style. We will explore some phrases concerning a bad exhaust lobe. Treatment of the exhaust lobe is typical of treatment of the other lobes. Here are some examples as they might appear in the warranty narrative:

#5 EXHAUST BAD ON CAM
#4 EXHAUST LOBE WORN
NUMBER 6 EXHAUST LOBE SCORED
NO 2 EXHAUST LOBE ON CAM SHAFT DOWN

Then some may be reversed such as:

CAM BAD ON #2 EXHAUST
LOBE DOWN ON #3&5 EXHAUST

The word 'EXHAUST' can be abbreviated. Typical variant forms may be 'EX' and 'EXH'.

APPROACH TO PARSING TEXT

This program tentatively identifies a key word in the free-flowing narrative. Once the key word is found, the program then looks for other key words in close proximity.

In the camshaft example, the first key word search might be 'EXHAUST' or its common shorthand codes EX, or EXH. The search function of choice is the INDEX function because it returns the position of the key word within the string. An example is:

POSITION = INDEX(TESTSTR,'EXHAUST');

When the program identifies the first key word, it sets up three substrings. The first one is for the lobe number search. The other two strings are for the main key word searches such as 'cam' and 'bad.' One of these two strings is on the left of the key word 'EXHAUST,' and the other string is on its right.

When identifying a specific lobe number, first a search is done for the keyword 'NUMBER' and its variants. If found, the program will do a search for numbers between the key word 'NUMBER' and the key word

'EXHAUST.' Then the number or numbers are placed in variables for later analysis.

In finding the key words related to 'EXHAUST,' such as 'lobe' or 'bad,' the search must be restricted to the same sentence.

The search words are divided into noun and adjective groups: Group 1 and Group 2. Some examples of these words are listed in the following table:

Group 1	Group 2
CAM	BAD
CAMSHAFT	DAMAGE
CAM SHAFT	DOWN
LOBE	PITTED
	SCORED
	WORN

If one word out of each group is found, then the program has identified a bad cam--in this case based upon a bad exhaust lobe. The program then assigns 'True' to the first variable of its score card. An example is:

$$EXLOB = 1;$$

The same approach is used for intake lobes and EUP lobes, except the initial key word search for the intake lobe will be 'INTAKE' and its variants. The initial key word search for the EUP lobe will be 'EUP' and its variants. The score card variables will be INLOB and EUPLOB.

In the warranty claim, sometimes the statement 'bad cam' is made without specifying a particular lobe. In this case the initial key word search is for at least one of the words, 'LIFTER,' 'ROLLER,' 'LOBE,' and 'CAM,' and the commonly used forms of each. The score card variable will be BADCAM.

CAMSHAFT REPLACEMENT

In most camshaft replacement warranty narratives, two statements will usually be made: (1) the camshaft was bad, and (2) the camshaft was replaced. This program requires both. Otherwise the claim is flagged as ambiguous and the claim must be read manually.

A 'True' in any of the score card variables so far (EXLOB, INLOB, EUPLOB, and BADCAM) constitutes a statement that the camshaft is bad.

The program awards a second 'True' when it reads that the camshaft was replaced. Typical statements in the warranty claim could be:

REMOVED AND REPLACED CAM
R/R CAMSHAFT
INSTALLED NEW CAM KIT
CAM KIT WAS INSTALLED
RENEWED CAM

The program performs two searches, one for verb key words, and one for object key words in close proximity. The following table shows a partial list of each:

Verb	Object
INSTALL	CAM KIT
RENEWED	CAMSHAFT KIT
REPLACED	NEW STYLE CAM
R & R	NEW LIFTER AND CAM
RE / RE	CAMSHAFT ASSEMBLY
R/R	CAMSHAFT
REPLACED	

If the program finds both kinds of key words within the same sentence, the program assigns 'True' to the score card variable REPLACED.

OTHER CAUSES OF FAILURE

Since the program's objective is to track the number of failures due to inherent problems in the part, failures due to other causes must be ruled out. Some of these causes are:

1) Coolant in the oil
2) Fuel in the oil
3) Dropped valve

This program section also uses a key word search to locate key words in close proximity. However, if a counter reason is found to have caused part failure, then the program assigns 'True' to the score card variable CONTRA.

FINAL SCORE

After the program reads the warranty narrative, it assigns it a final score of 1, 2 or 5. In the case of ambiguity, it assigns the value 9. The following table illustrates how the scores from the various sections of the program are used to determine the final score:

EXLOB	1				1	
INLOB			1	1		
EUPLOB		1	1			
BADCAM						
REPLACED	1	1	1	1		1
CONTRA				1		
SCORE	1	2	1	5	9	9

CONCLUSION

The program will not deliver 100 percent accurate results. The old 80/20 rule probably applies here. Twenty percent of the programming effort gets 80 percent of the results. From there you get diminishing

returns for your effort. In undertaking a task like this, make adequate provision to flag the ambiguous cases so that they can be read manually.

REFERENCES

SAS Institute Inc., SAS *Language: Reference, Version 6, First Edition* Cary, NC: SAS Institute Inc., 1990. 1042 pp.

ACKNOWLEDGMENTS

CONTACT INFORMATION

Your comments and question are valued and encouraged.

Contact the author at:
Mack Trucks, Inc.
13302 Pennsylvania Avenue
Hagerstown, MD 21742 USA
Work Phone: 301-790-4788
Fax: 301-790-5702
Email: Robert.McCurdy@macktrucks.com
or RNMCCURDY@aol.com