Performance Evaluation of MAX: the Maintenance Administrator Expert System

Erica J. Wolin

NYNEX Science and Technology, Inc.
Expert Systems Laboratory

500 Westchester Avenue, White Plains, NY 10604

July 25, 1993

Abstract

MAX (Maintenance Administrator Expert) is an Expert System developed by NYNEX Science &Technology that screens and processes trouble reports in Telephone Maintenance Centers. MAX has been very well received and is currently deployed in more than sixty telephone Maintenance Centers throughout New York and New England(NE). But when the developers tried to quantify MAX's performance, unexpected roadblocks were encountered. This paper contains a discussion of different methodologies that were developed in an effort to evaluate MAX. Its purpose is to familiarize the reader with some of the complexities and issues associated with evaluation of a deployed Expert System.

Keywords:

Diagnosis
Evaluation
Expert System
Loop Maintenance
Telephone Maintenance Center

1.0 Introduction

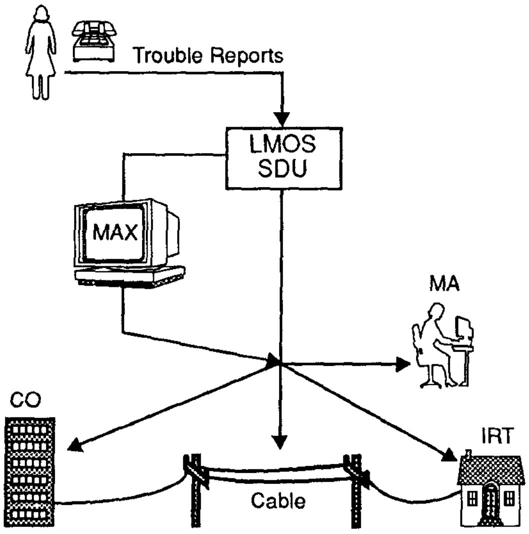
Traditionally computers have been used to solve problems that were straightforward. Programs were verified by comparing the answers that were produced against a set of correct answers. With the introduction of Expert System technology, a whole new problem area has emerged: the task of verifying a computer program when there does not necessarily exist a set of correct answers.

MAX (Maintenance Administrator Expert) is an Expert System developed by NYNEX Science and Technology that screens and processes trouble reports in Telephone Company Maintenance Centers (MCs). MAX was first deployed in 1989, and is now operational in more than sixty MCs throughout NY and NE Telephone.[1] MAX's success was obvious through its wide acceptance; but the task still before the developers was to produce tangible evidence that MAX was having a positive effect on MC operations.

This paper contains a discussion of various methodologies that were developed in an effort to evaluate the performance of MAX. Its purpose is to familiarize the reader with some of the complexities and issues associated with the evaluation process, and to present and assess those procedures that are currently available.

2.0 The Problem Domain

Telephone troubles containing line record information and test results are entered into Loop Maintenance Operation System (LMOS). The Screening Decision Unit (SDU) of LMOS performs the function of directing each trouble to dispatch or to a holding area for closer review by a Maintenance Administrator (MA). There are three dispatch options: to the Central Office (CO), to an Installation & Repair Technician (IRT), or to cable maintenance, plus the option of holding for review by an MA. With the introduction of MAX, the SDU now has a fifth option: to send the trouble to MAX and have the decision made there. An unnecessary dispatch or a dispatch to the wrong area of the loop inconveniences the customer and costs the company money. Of particular interest is avoiding a dispatch in a case where the trouble is caused by Customer Provided Equipment (CPE). The goal is to maintain high auto-flowthrough while keeping false dispatches and double dispatches to a minimum.



3.0 Evaluation Approaches

LMOS provides various mechanisms for tracking troubles, including DATHs, Microfiche, and TREAT. The DATH (Display Abbreviated Trouble History) is a history of closeouts by telephone number, indicating each final disposition code (FDIS), which is a four digit number that describes what was done to clear the trouble. The microfiche contains a more detailed record of the trouble, having a line of information for each time the trouble is handled. TREAT (Trouble Record Evaluation Analysis Tool) is a database query program that provides the user with the ability to pull reports from a database of closed out troubles.

MAX can be evaluated from three different perspectives:

- In isolation on a trouble by trouble basis; that is, for each trouble that it processed, did it or did it not make a correct decision; and how did it compare with the SDUs decisions.
- In comparison with Maintenance Administrators.
- From an overall perspective, that is, does the MC as a whole perform better with MAX than it would without MAX.

Each of the different perspectives had its own unique problems.

[1] In 1991 the paper "NYNEX MAX: A Telephone Trouble Screening Expert" was presented at the Innovative Applications of Artificial Intelligence conference, and the paper "The Lessons of MAX" was presented in 1993.

CIKM '93 - 11/93/D.C., USA
© 1993 ACM 0-89791-626-3/93/0011 ....$1.50

3.1 Evaluating MAX in isolation, trouble by trouble

The first problem that was encountered in doing a trouble by trouble evaluation was that many of the final disposition codes were inaccurate. Some miscodings were simply the result of careless lookup or careless typing. Others were of a more deliberate nature. For example, a technician might be dispatched to the house of a very elderly person and decide to fix a CPE trouble rather than mark it as CPE so the elderly person would not be billed. Telephone company experts estimate that the FDISs are only about 60 to 80% accurate.

The second problem that was encountered was in the case of subsequents and repeats[1], which comprise a considerable portion (perhaps 1/3) of the troubles. Suppose that MAX called for dispatch to the house, and the IRT found squirrel damage, repaired the wire and closed out the trouble. If it was an intermittent type trouble, the customer could be happy for several days, but then find that the trouble came back and had to be reported again; this time water leakage was found in a cable. Which of the two closeouts should be used to evaluate MAX? The repeat could be the result of the trouble never having been fixed, or it could be the result of an unrelated problem. There is no simple solution to this problem.

The third problem that was encountered was the situation of overlapping responsibilities. Some troubles are of a nature that they could be handled by either a Cable technician or by an IRT; but since Cable technicians are generally in shorter supply, a dispatch to IRT would have been 'better'. This problem was circumvented by allowing either dispatch to be counted as correct.

3.2 Comparing MAX with Maintenance Administrators

Some MCs have devised reports from the TREAT database to look at false dispatches to evaluate their MAs. One might conclude that the most obvious way to evaluate MAX is to run these same TREAT reports on MAX and compare the results with the theoretical "average" MA. This conclusion however turns out to be faulty. Consider for example a TREAT report that calculates:

%dispatch errors = #false dispatches / total #troubles screened.

By always taking a conservative line of action, that is whenever in doubt, send the trouble to an MA, MAX could keep its error rate very low; indeed in the absurd case it could not dispatch any troubles at all, and maintain a perfect record. Because the human MA does not have the luxury of being able to pass the trouble on to someone else, but is required to either dispatch it or close it, the percentages of the two entities cannot be compared. Suppose on the other hand we were to calculate:

% dispatch errors = #false dispatches / #troubles dispatched

Again we have an erroneous study, for we have not credited MAX for instances where it correctly held a trouble back from dispatch. The ability to detect a possible CPE trouble, which is a very significant part of MAX's expertise, would not be measured here.

We could perhaps propose some sort of composite score, similar to a typing test, where both speed and accuracy are accounted for in a formula. The formula could include both number screened and number dispatched in calculating an "error rate". However, using any such formula we fail to take into account the fact that MAX and the MA handle different types of troubles. MAX actually just "skims the cream" by handling the more routine troubles, leaving the more difficult ones for the MA to tackle. One would expect that the "average MA" score would go down once MAX is installed. But suppose that the "average MA" score were to improve after MAX is installed, and in fact compared well against MAX's score. What would this say about MAX? Suppose that the presence of MAX in the MC helps the MAs perform their jobs better?

This concept turns out to be not as intangible as one might think. Functional diagrams of MAX in the MC show MAX working in parallel with other MAs, screening and statusing troubles. But since MAX works so quickly, statusing many troubles within just seconds of receipt, in practice what actually happens is that MAX works in serial with the MAs, pre-statusing most troubles, as shown in the diagram on the first page. The effect is that any trouble that has gone through MAX and was not sent to dispatch is automatically suspect of being non-routine, and will be viewed more carefully by the MA. In addition, the narrative that MAX places on the newly statused record may be of some assistance to the MA or to the field technician in diagnosing the trouble; and since troubles are screened and dispatched more quickly and efficiently by MAX, there is less time for a trouble to just disappear.

Simply stated, because of MAX's complex interaction with MAs, it is not valid to compare the two entities, and the decision was made to not pursue this avenue.

3.3 Evaluating the Maintenance Center as a Whole

In viewing the MC as a whole, we gain the advantage of being able to see second order effects, such as a reduction in volume as MC personnel are freed up to do preventive maintenance type work. But we are faced with the question of how valid, for our purposes, are the criteria that are traditionally used to evaluate an MC.

In White Plains MC, a sample of double dispatches was taken, and review of the microfiche revealed several causes, including:

- The job spanned two days
- An IRT was dispatched first for verification of a cable trouble and then a cable splicer was dispatched.
- The IRT sent the trouble to cable, but s/he could have or should have done a pair transfer[2]
- The IRT failed to find the trouble the first time
- There were multiple problems: in the cable and in the CO.
- There was no access to the customer premises.

It's certainly valid to include these items when evaluating a Maintenance Center, since they are all contributing to the inefficiency of operations. But to include these items when one is evaluating the effectiveness of MAX is certainly questionable.

Suppose that we accept that there will be items of this nature in any set of data, and that with a large enough sample size, they will not sway the results in either direction. Then we are still faced with the problem of choosing a valid baseline, or basis of comparison.

One possibility is to run TREAT reports for a pre-MAX period, and compare them with a with-MAX period. Of major concern here is that weather and other environmental factors could influence the data, and we would never know if the same increase or decrease in errors would have occurred regardless. This would certainly be the case if the pre-MAX baseline were taken in February, when the ground is frozen, and the with-MAX data were taken in April when cables are wet and wire-gnawing squirrels are rampant. To minimize the weather factor we could take the with-MAX data exactly one year after the baseline was taken, but then we run the risk of other factors coming into play, such as new LMOS generics, new methods and procedures, or a change in management. Maintenance Centers are continually improving their operations,

[1] A subsequent is a report on the same telephone number that comes in before the original one has been closed. A repeat is a recurrence of a trouble on the same telephone number within 30 days of the closeout.

[2] A pair transfer is the swapping of a good line that is spare for a bad one, leaving the bad one unused. This is considered proper procedure in the case where the process of opening up an entire cable seems unwarranted.

and it would be difficult, if not impossible to determine just what portion of the improvement was directly attributable to MAX.

We could compare the MAX-equipped MC with a "similar" MC that does not have MAX. But since MCs vary so much in the distribution of troubles they handle, the way they handle those troubles, (for example the level of SDU Auto-flowthrough[1]), the makeup of the customer base, and overall volume, identifying such an MC may not possible; and the difficulty becomes more acute as MCs are cut over to MAX, and non-MAX MCs become more and more scarce. With 20/20 hindsight it is obvious that we should have pulled baselines while MAX was still under development.

4.0 Five Evaluation Methodologies

Five different methods were developed to evaluate MAX's success in diagnosing troubles. Though none of these methods were perfect, they each had some merit of their own, and it was hoped that the composite result would be revealing.

The first four methods use DATHs and microfiche to evaluate MAX in isolation, trouble by trouble. They all use a weighting factor to score instances of dispatch errors and "unnecessary" human intervention. They differ in whether they are performed on-line or off-line, what data is used to obtain the "correct answers", whether the procedure can be automated or must be done by hand[2], and what entity is being compared. The fifth method uses TREAT database information on dispatch errors to evaluate MAX's impact on the MC as a whole. As explained earlier, no studies were developed which attempted to compare MAX with the MA.

1. The off-line DATH/microfiche study provides a means for comparing what MAX would have done, with what the MC actually did, without interfering with current operations. Troubles are collected and fed through an off-line MAX, that is, a computer that has all the MAX software but is not connected to LMOS; then DATHs and Microfiche are viewed by hand to compare MAX's decisions with those made in live operations. One such trial was conducted in East Brooklyn in 1989, and the results indicated that installing MAX would result in an improvement in operations. It must be noted however that in an off-line test such as this, it is not necessarily clear just what constitutes an acceptable result.
2. The on-line DATH/microfiche study, is a means of evaluating, by hand, MAX's decisions as compared with the decisions that would have been made on those same troubles by the SDU. This type of study is most significant in an MC that operates with a high rate of auto-flowthrough. A study involving four sites and about a thousand troubles was performed. Results varied widely from one site to the next, but all four showed improvement with MAX.
3. The on-line automated DATH study is an automated version of the on-line DATH/microfiche study. The accuracy that is lost by using only the DATHs is made up for with the increased volume, and extrapolation is more appropriate.[3] One such study, involving 5000 trouble records was conducted in East Brooklyn, with positive results. Two more sites from different areas of the company are currently being evaluated.
4. The self-evaluation is an automated on-line DATH study, where MAX is evaluated on its own merits, and is not compared with any other entity. This type of study is generally used to evaluate individual rules in the knowledge base, and parameter settings. One such study was performed in White Plains on 4000 trouble records, and as a result the knowledge base was modified. The follow-up study showed marked improvement.
5. The TREAT evaluation is a means of evaluating the MC as a whole with MAX, against a baseline. The main advantage of this type of study is that it can be performed easily on a very large sample size. The disadvantage, as explained earlier is that establishing a meaningful baseline is non-trivial. In using TREAT there were difficulties in aligning the MAX logs with the TREAT data, since troubles enter the TREAT database only after they are closed. But the main problem was that since TREAT only maintains data for a short time,[4] the critical baseline information was quickly disappearing.

   Four separate TREAT analysis studies were done. The first two studies straddled a work-stoppage and were deemed unreliable. The third study involved ten New England sites, using the same baseline data as the first study, but the with-MAX data was pulled exactly one year later. The overall result was a reduction in dispatch errors, with a negligible increase in repeats.

5.0 Conclusions

Five methods for evaluating MAX have been presented here. Each has its advantages and its limitations. The following table represents an attempt to rate each method in five categories:

Method	Reliable	Conclusive*	Offers Comparison	Large Sample Size	Evaluate Rule Base
1	Y	N	Y	N	N
2	Y	Y	**	N	N
3	***	Y	**	Y	N
4	***	N	N	Y	Y
5	Y	Y	Y	Y	N

* When we see the results we know whether they are good or bad
** With the SDU rules only
*** Depends on the accuracy of the final disposition codes

In summary, the following were among the problems encountered:

- Inaccurate data
- Conflicting answers
- More than one action being correct
- Lack of a valid evaluation formula
- Lack of an entity with which to compare
- Side effects from the Experts System's presence
- Second order effects and intangibles
- Inappropriate evaluation criteria traditionally used
- Lack of a baseline
- Changing conditions
- Procedure too tedious to get a meaningful sample size
- Difficulty in aligning two separate databases
- Dealing with databases that are continually changing
- Dealing with databases containing transient data

The problem of evaluating MAX seems to resist simple solution, and we have had to use a combination of approaches, understanding and accepting the merits and shortcomings of each. Some of the difficulties are unique to telephone company operations; others may be common to many environments. It is hoped that the ideas presented here will be valuable to knowledge engineers in their efforts to evaluate their Expert Systems.

[1] Some Maintenance Centers had a policy of depending on the SDU for only 5% of the troubles (having an MA review the remaining 95%), while others allowed the SDU to handle 80% of the troubles.

[2] Since microfiche can only be reviewed by hand, studies that involve microfiche are severely limited in volume.

[3] A statistician has been involved in the proper extrapolation of the results.

[4] Data is held for 99 days in New England, and for only 40 days in NY.