EVALUACION

ISSUES AND APPROACHES IN NLP EVALUATION

Marta Sáiz

Centre for Computational Linguistics, UMIST, P.O.Box 88, Manchester M60 1QD, England.
martas@ccl.umist.ac.uk

ABSTRACT

Reasons for the lack of evaluation methodology in NLP, and above all in MT, are offered. Important parameters to evaluation are discussed: task-dependent or non-task dependent applications, the object of MT is defined, glass box and black box evaluation. Also cost-effective evaluation and linguistic evaluation through the use of test suites and text corpora are tackled.

Keywords

NLP/MT evaluation
Adequacy evaluation
Task-dependent applications
MT output
Black box/glass box evaluation

1. INTRODUCTION

Research on Natural Language Processing has received increasing attention during the last thirty years and in its turn it has brought about an increasing awareness among the research community of how to evaluate NLP systems. NLP has been using different techniques in order to evaluate different tasks but it is becoming more and more clear that there is a lack of useful techniques for evaluation of NLP.

Evaluation of NLP is necessary to everyone in the field. System developers want to know if the system they have developed does what it is expected to do and how well it does it; sponsors will be interested in knowing whether the system they funded is worth the money they financed and purchasers will be interested in knowing whether the system they are thinking of purchasing will perform according to their needs.

NLP is a field of great complexity and heterogeneity: apart from there being different people interested in evaluation, there is a variety of approaches to the task: there is progress evaluation, diagnostic evaluation and adequacy evaluation (see below); there is also a variety of applications within NLP, all of which makes it extremely difficult to have a general methodology of evaluation appropriate for the entire field. One such application is Machine Translation. Despite forty years of research on MT there is still not a generally accepted, satisfactory and comprehensive evaluation methodology. One might say that the MT problem is also still not well understood -however, one should still be able to set up evaluation methodologies.

MT limitations are currently recognized and therefore it is not generally expected that an MT system produces a perfect translation without human intervention. In this sense there is also a growing acceptance of MT by users. However, on most occasions potential purchasers of MT systems are not allowed to access information about the advantages or disadvantages of buying a specific MT system. They will be interested in knowing whether an MT system will do what they require, how well it will do it and at what cost, therefore they will be concerned with adequacy evaluation. Adequacy evaluation is especially useful for the customer as it is oriented towards specific requirements and so a user will be able to choose the system which best meets their specific requirements.

Evaluating anything is difficult because it requires consideration of a variety of tasks, purposes and interests. Some insight that would be helpful for evaluation of MT could be gained from evaluation of other NLP applications but unfortunately as has already been pointed out, there is no well-developed evaluation methodology for any NLP application.

The purpose of this paper is to consider aspects relevant to the evaluation of MT in the context of NLP. In order to accomplish such an aim first some reasons for the lack of a consistent evaluation methodology for NLP are offered and also the usefulness that such a methodology would provide for the NLP field is considered. Some insight will be attempted to be gained from studying relevant past MT evaluations. Then some key debates that have been taking place lately among the research community will be discussed: whether there should be a general evaluation methodology for all NLP applications or whether differences between tasks and applications should be considered leading to different methodologies; which is the object of MT evaluation; and whether evaluation should be black box or glass box. As will be seen, focusing only on evaluation of system output is not enough because that would mean neglecting other parts of the components of a MT system which should be relevant for adequacy evaluation, that is, for a potential purchaser of a system trying to find out whether a system will perform according to his requirements.

2. STATE OF THE ART IN NLP AND MT EVALUATION AND PREVIOUS MT EVALUATIONS

2.1. Lack of methodology in NLP

Interest in evaluation extends to NLP systems as a whole because it is a matter of major importance to everyone in the field. This is so because evaluation of any kind of NLP system can show insights for the evaluation of other applications and therefore something could be learnt from evaluation in other applications. The problem is that there is no general well-developed methodology for the evaluation of most software which could be a source of inspiration. There are several reasons why this is so.

For some applications such as MT, which deal with general aspects of language, evaluation is more difficult than evaluating systems that deal with particular aspects of language. Compare, e.g. evaluating a grammar checker which could be done by developing test materials full of grammatical errors and then checking whether the system deals with such phenomena or not. However, constructing materials for MT, which deals with general aspects of language is not that easy. Also, the majority of evaluations are done under contract and usually under a confidentiality agreement so no criticism is contributed from the public domain which would be very constructive in helping to advance knowledge about evaluation in NLP. Another reason is that usually previous reports dealing with evaluations of previous methodologies conclude by saying that the methodology used was not appropriate for the current state of the art. However, there is no way of checking any evaluation methodology against a common standard simply because there are no common standards of methodologies for evaluation of NLP systems.

2.2. Usefulness of an evaluation methodology in NLP

A public discussion of evaluation methodologies would be of great use in the NLP field:
On the one hand it would help to be realistic about what the real capacities of systems are. In the special case of MT it would help to avoid statements made by developers saying that they have discovered the best translation system suitable for all purposes on the grounds that the system deals with a group of carefully chosen sentences and a limited vocabulary. On the other hand it would also help in the essential issue of making expectations about MT realistic. A proof of the damage the contrary might produce is provided by the ALPAC report (Pierce and Carroll, 1966), a product of the mistakes of the MT community who had believed and therefore encouraged beliefs in others about the possibility of attaining fully automatic high quality machine translation.

Evaluation of NLP systems should be also very useful for the user interested in discovering the functionality of the system. The designer of the system usually leaves the customer to figure out what the functionality is. As a result the user takes a large amount of time to discover the functionality (by feeding the system with vast amounts of data to examine the results) and if he does he finds himself at a loss because he has no basis for comparison and so he has no idea what the system is supposed to do or how it should behave.

2.3. Lack of evaluation in MT

The kind of process the user of an MT system has to follow is rarely found in other products of science and technology. It is hard to imagine a potential purchaser of a car measuring the capacity of the fuel tank to check how much fuel the car he is thinking of buying consumes per 100 miles. On the contrary, the vendor of the car will provide the customer with all the information available about the car performance, highlighting those aspects his product surpasses in comparison to those commercially available. However, in the case of MT systems there is no explicit, standardised methodology of evaluation or even an agreed list with the qualities an MT system should have.

This lack has so far been justified and to a certain extent excused with different kinds of arguments the most popular of which seems to be the one that alludes to the variety of purposes that MT/NLP can serve, the variety of approaches there are to the task and the variety of reasons that can exist for performing evaluation (Arnold et al, 1993). However, we can ask ourselves to what extent do the different purposes, approaches and reasons for evaluation represent much more difficulty for MT than for any other field of science and technology and whether this excuse has not so far been used as a way of avoiding tackling the problem. What must be pointed out is that every field of science and technology has different reasons, approaches and purposes for evaluating its products and therefore these factors should not be considered as an excuse for justifying the lack of or overlooking a systematic methodology of evaluation.

2.4. Previous MT evaluations

We now look at previous evaluations of MT systems. Different evaluation efforts have accompanied MT since its first appearance. Of these the Alpac report, Taum Aviation evaluation and the evaluation of Systran can be considered to be the most relevant.

The Alpac report was the first relevant attempt to perform evaluation of MT systems (Pierce and Carroll, 1966) which was concerned mainly with evaluating some aspects of system output. Although Alpac had disastrous effects as it stopped the funding and research in MT for the next decade in the USA it was also the start of an awareness, that since then has steadily increased, of the importance of practical methods for evaluation. Another positive contribution of the Alpac report was that it established the rating technique for evaluating intelligibility and fidelity and ultimately as a way to differentiate the quality of different types of translations.

One typical feature of earlier evaluations is the tendency to compare machine versus human translation. The Alpac report did this and the result was not a success: looking at the state of the art it might never be, at least in the near future. Also the Taum evaluation (Guida and Mauri, 1986) compared the quality of human and machine translation. The quality of MT output should not be assessed in terms of its identity with human products (Sager, 1994). One main reason for this is that two human translations do not produce two identical translations of a source text but two different versions of the same translation, therefore it would be very difficult to determine which of the two human products is of higher quality in order to compare it to the MT output.

Another typical feature of earlier evaluations is that they focus on the evaluation of MT output and the cost-effectiveness of the throughput. The evaluation of Systran carried out for the U.S. Airforce in 1970-80 (Wilks, 1991) can be considered to be a step forward in the sense that it assumed beforehand that SYSTRAN translated at a level suitable for some kind of customers and also considered extensions of the system to deal with a new type of text and how this could transfer to other texts of that type. To restrict evaluation to issues such as the quality and cost-effectiveness of MT output would mean to neglect other aspects of MT systems which require an examination of the inner components and design of the system and which cannot therefore be simply assessed by conducting a black box evaluation. These aspects are: reliability, usability, efficiency, maintainability, portability and extensibility. We will look at these aspects in due course.

3. SOME ISSUES ABOUT EVALUATION

3.1. Task dependent or non-task dependent applications?

In discussing evaluation of MT systems in the context of NLP systems there are several parameters which are worth discussing.

One is that there is an unresolved conflict between strategies for NLP evaluation relevant to different applications and strategies which consider that there are differences between tasks and applications. The issue here is whether it is possible to define a general evaluation methodology which is fair, reliable and applicable to any kind of system, such as MT, database query, MU etc, and which is informative for all purposes.

To start with, there is the question of whether a **commercial product and a research prototype** can be evaluated in the same way. A research prototype will be interested in the practicability of a specific approach while a commercial product will be mainly interested in answering some specific needs of a customer. Also both a commercial product and research prototype will be interested in time and efficiency. The lexicon is another aspect important to both because both will be concerned with covering a great variety of linguistic phenomena. The only difference relevant here between a commercial product and a research prototype seems to be that those systems designed to replace other commercially available ones must have certain characteristics which justify the fact that they are worth replacing the old one. They should be better, perform faster and be more economical than the existing one. However, these requirements although not immediately essential for a research prototype, should be considered if a prototype is expected to become an operational system in the near future.

More difficulties in considering a general methodology of evaluation come from the different groups who are interested in evaluation and who are therefore interested in different evaluation techniques.

System developers will be interested in detecting deficiencies of a system in order to correct them and to then check if those changes made to the system actually improve the system. System developers will have to answer the question: "Are our efforts making the system better?". They will attempt to look inside the system and find ways of measuring how well the system does something rather than just seeing whether it does it or not. This is why they will be mainly concerned with what is called glass-box evaluation.

System users would like to have answers for such questions such as "Which system is better, A or B?", "Will system A improve the quality of translation?", that is, they want to know if the system will perform effectively and economically. Because end users are not interested in the evaluation of individual components and are not interested in the inner workings of the system, it can be said that they are mainly concerned with black box evaluation.

Systems sponsors are those who fund a project and therefore will be concerned with knowing if their money has been well spent. **System researchers**, who carry out initial research, are concerned with issues such as the evaluation of the theoretical ideas of a system.

Moreover, in the context of EAGLES Evaluation Working Group (one of whose goals is to harmonise terminology in the field) we can differentiate several types of evaluation:

Progress evaluation is when the actual state of the system is assessed with respect to some desired state of the system or when successive versions of the same system are assessed in order to provide a way of measuring the system's progress over time.

Diagnostic evaluation is when the state of a system is assessed with the intention of discovering where it fails and why. This kind of evaluation requires an intimate knowledge of the system examined and this is why it is usually done by researchers. This evaluation is not done with the intention of comparing different kinds of systems but of comparing the effects of alternative versions of some system component.

Adequacy evaluation is when the adequacy of a system is assessed with respect to some intended use of that system by a customer wishing to know whether a system will do what he requires and how well it will do it and at what cost. This evaluation can take the form of comparative evaluation between two or more systems. This kind of evaluation is the most useful for the user, as it is oriented towards specific requirements although it will be hard to establish user needs. Adequacy evaluation is of great importance for NLP because in the future more and more end-users and buyers of NLP systems will have the problem of choosing which product best meets their specific requirements.

Some more evidence in favour of task dependent evaluation comes from the following:

1. If there are no general purpose NLP systems, why should there be a general evaluation appropriate for all purposes?

2. Evaluation becomes less complex as the nature of the comparison is restricted, so it can be said that general evaluation will be more informative because evaluation will be aimed at many applications but it will also be harder to design and there is always the risk of not covering all aspects.

3.The only common objective of evaluations for different applications, such as MT, MU, database query, etc, is that some input is processed in order to have an output. In MT an input in one language is converted into an output in another language.

4. The field of NLP is very complex and heterogeneous. Evaluations for different applications vary in purpose, scope and nature of the object being evaluated. There is the need of making a very strict analysis of what is involved and required for any individual evaluation. It is difficult to establish to what extent evaluation techniques can be transferred from one application to another. In this sense it is hard to see how relevant to MT evaluation is, for example, database query, in which the database can be queried and an answer obtained.

5. What could be common in the NLP field is the act of training an evaluator to provide him with appropriate evaluation methodologies.

6. If the application is well chosen then the evaluation is going to make the system look good and only in this way will it be possible to find.a satisfying evaluation paradigm.

We are faced with a lack of common objectives, the complexity and heterogeneity of NLP and the extensive amount of information needed to cover all the aspects relevant to all NLP applications. However, although no evaluation methodology seems to be appropriate for all purposes, what would be feasible is to develop and work towards a general framework for evaluation methodology, which then will have to be adjusted for the specific task some evaluation is intended for.

In this sense EAGLES Evaluation Group has established a general framework for the design of evaluation methodologies. This framework is seen as involving a series of steps:

1. Definition of relevant quality characteristics
2. Definition of attributes pertinent to each characteristic
3. Definition of measures able to provide values for each attribute
4. Definition of methods whereby the particular value of a particular system can be determined.

Quality characteristics are the basic characteristics of a system. The ISO proposal of the characteristics relevant in quality assessment of software (see appendix) has been followed but it has bene modified and extended when considered appropriate to deal with the special application. For example for the special case of MT extensibility including the notion of modifiability has been added to deal with the possibility of extending the systems.

Each of the quality characteristics can be more precisely refined in a set of relevant **attributes** with respect to which a system's performance will be judged.

The value of each attribute is expressed in terms of a **measure.** A measure should be valid and reliable, that is, it should measure what it is supposed to measure and it should do it consistently.

The evaluator should specify a **method** for each attribute to obtain the value defined by the measure.

3.2. What is the object of MT evaluation?

The evaluator, in order to perform an evaluation of an NLP system, must have clear what the aim or goal for his specific application is. This is necessary to establish the optimal level of performance for a given environment.

For some applications it has been possible to reach an agreement on the targets, e.g. in the case of MU the target has been text extraction from a database as in the USA MUC tasks. However, for other applications, such as summarising, it is difficult to establish and reach an agreement on the expected optimal level of performance. Another case is MT, where the difficulty in setting targets for evaluating translations shows how difficult it is to agree on targets.

The question is then what the correct output for a translation is, but this is a relative and unreliable issue, much more difficult to establish due to the fact that there is some disagreement concerning whether output must be divorced from context or not. To recapitulate the question, what is at issue here is whether the context should or should not be taken into account when trying to establish what the (correct output) for a translation is.

To this we could answer by relating this issue to the notion of sublanguage and document type. Therefore, if what we are translating is a self contained sentence, such as one part of an instruction manual, no context will need to be taken into account. METEO (Kittredge, 1987) is an example of a system which does not take context into account because the sublanguage of weather reports does not require it. TITRAN (Alexa, 1993), designed for the automatic translation of scientific and engineering titles from English into Japanese is another example.

However, if what we are translating is not a self contained sentence but is part of a bigger text then the context should be taken into account. This is the case of the translation of dialogues, in which although a translated sentence may be correct on its own, when considered it in relation to the context in which it occurs it may be nonsense. However, in the current state of the art it is not yet possible for MT systems to consider context in the process of translation. We have seen that sublanguage-based MT systems do not take context into account, therefore if MT systems were restricted to the translation of limited text domains or text types (Kittredge) the semantic domain of language would be restricted and therefore the context of dependency of translation would largely disappear.

3.3. Black box or glass-box evaluation?

Another very important issue is whether evaluation should be black box or glass box (Palmer and Finin, 1990).

Glass box evaluation is concerned with the inner workings of the system. It attempts to look inside the system and find ways of measuring how well the system does something rather than simply check whether it does it or not. It also measures the system's coverage of particular linguistic phenomena and the data structures used to represent them. The evaluator has access to all the inner workings of the system and can inspect intermediate results.

Black box evaluation is concerned with 'what a system as a whole does'. The evaluator has access to the output of the system and it is considered that if the output is accurate with respect to an input then that should mean that the system is performing correctly. Besides accuracy other criteria used for this kind of evaluation are intelligibility, fidelity and style. Because it is possible to perform a black box evaluation without knowing anything about the inner workings of the system, this is the kind of

evaluation that can be performed by system users and which therefore we are going to focus on. The question is whether black box evaluation provides enough information to a potential user of an MT system. In the specific case of MT, relying solely on black box evaluation would involve:
-ignoring the fact that in a system there is a variety of components that can give rise to an unsatisfactory output and this will not be solved by relying on black box evaluation and
-performing subjective judgements of the output.

3.4. Evaluation of MT output

In judging the quality of the output one of the most obvious ways of doing so is to consider characteristics like fidelity, intelligibility and style.

Usually the general method to measure these characteristics involves the questionnaire or investigation type of approach. Readers are asked to assess the raw translations for fidelity, intelligibility and style on a fixed scale, ranging from good to bad, the results being submitted to statistical analysis.

Some methods proposed for fidelity alone are performance evaluation, consisting in whether someone can carry out specific instructions using the translation as well as using the original. Another method is back translation, here the output is translated back into the original language and the result is compared with the original text.

For intelligibility, more objective texts than the ranking of output previously suggested employ readability scales such as flesh scales, based on average sentence lengths, use of complex nominalizations, etc, and cloze techniques, based on the masking of words in sentences and texts to create gaps which readers are asked to fill with appropriate words. Another objective test is the use of comprehension tests, used to test the reader's understanding of the content of a translated text as a whole.
Apart from these rather subjective measures of quality, some more objective, quantitative measures can be distinguished:

Error counting

More detailed evaluation of translations is obtained from error counting, which is the amount of work needed to correct MT output to make it acceptable as a translation. It consists of the reviser counting "each addition or deletion of a word, each substitution of one word by another, each instance of the transposition of words in phrases and calculates the percentage of corrected words in the whole text" (Hutchins & Somers, 1993). However, one big problem in relation to this method is that revisers differ in what they consider errors as there are different levels of acceptability depending on the particular circumstances in which the revision is taking place. Another problem is that it is up to the posteditor's attitude towards MT to decide what is an error and what is not. Overall, this method can be considered to be subjective and therefore not appropriate.

Classification of errors

A more objective method is classification of errors by type of linguistic phenomenon and by relative difficulty of correction (Ibid). In this sense some lexical errors are easily resolved by changing the dictionaries of the system, while others require altering the grammatical rules. The resolution of grammatical errors may require adjustment of some lexical entries while others may require adjustment of the basic design of the system. It is these kinds of mistakes related to linguistic structures which are difficult to correct because in trying to achieve the desired correct output, a ripple effect may be created when the basic design of the system is altered, which may cause a mistake somewhere else.

3. 5. Operational evaluation

A user oriented approach to evaluation which focuses only on quality assessment is not enough. Overall performance of an MT system has to be judged on aspects other than translation quality. This is why it has been suggested that it should be appropriate to conduct an operational evaluation, which would consist in establishing "cost-per-word figures for MT plus any necessary post editing, and to use this as a basis for comparison of two MT systems, or MT against human translation" (Arnold et al, 1993).

This kind of evaluation would be ideal to provide the user with information on whether MT would be profitable in financial terms and so whether it would fit in an organizational environment. This evaluation technique requires taking some issues into consideration: 1. The longer translators spend post-editing the MT output, the less profitable MT will be. As a result of extensive post-editing the differences between MT and HT will be minimal but the time required to achieve such a product will make MT unprofitable. A possible solution to this would be to conduct a rapid post-editing: even if the differences in quality between MT and HT are evident not to much importance should be attached to this as long as the output is suitable for a specific purpose.

2. Another issue concerns the updating of the dictionaries: the longer the time spent in updating the system's dictionaries the less profitable MT will be. Dictionary updating has several drawbacks. It is time consuming and requires experience; it is practically impossible to add all terms to the system dictionaries, therefore the evaluator will be forced to infer what the performance of the system will be by having a larger vocabulary. These inferences will be unreliable because only by doing the scaling up of the system will one reliably see how the system performs. Besides, there is the danger that the expansion of vocabulary will cause the system to degrade because of a 'ripple effect'.

3. Another issue worth considering is that the translator will have to be trained in the post-editing of the MT output, in any special text editing which the MT system incorporates and in the updating of the system's dictionaries in the handling of which some experience is required. This need for translator training, even before the actual task of evaluation starts, makes the task of post-editing and updating the system's dictionaries even more expensive and time consuming, factors which are discouraging for users attempting to conduct an operational evaluation of a particular system. This kind of evaluation would be more suitable for the designer or the owner of the system.

3.6. Test suite evaluation

Approaches to evaluation which focus either on translation quality or on economic factors have their weaknesses. An alternative method is to design an organised set of test inputs to test the syntactic coverage of the system. In this sense it is possible that the linguistic coverage of the system is incomplete or defective: it may be the case that a grammar rule is not appropriate for all circumstances and even though it works for a specific phenomenon, when different phenomena interact in the same sentence the rules intended to cover them fail to work together. Thus the use of test suites of specially constructed test sentences, in which each sentence in the suite contains one linguistic phenomenon or combination of phenomena, will provide with information regarding the syntactic coverage of the system.

Test suites are useful to: conduct a progress evaluation; test system coverage and performance as the system evolves through time: the system developer will check whether any changes to the system are actually improving the system or making it worse; conduct a diagnostic evaluation; correct or look for deficiencies in a system; check the suitability of a system for a specific task by the user, therefore it is also useful for adequacy evaluation. However, there are some drawbacks in using test suites for MT:

1. One of the problems with setting up test suites is when there is an interaction between different linguistic phenomenon. The solution used for this problem is to focus on the phenomena of direct interest and avoid where possible the linguistic complexity caused by other phenomena of no immediate concern. However, it should be worth considering whether this is an appropriate method, since real texts usually contain interaction of phenomena.

2. Construction of a TS is a long and difficult task even if only one input per phenomena is considered, therefore critical problems of size and difficulty arise if interaction of phenomena are covered thoroughly.
It has been proposed that general suites can be developed for different NLP products and they can be adapted for different applications. Nevertheless,

as different NLP products require different kinds of test suites because otherwise they would not reflect the particular characteristics of the application considered it seems more appropriate to construct specific test suites for each application. Even if we construct test suites for each application a problem arises if the intended application of the system is a sublanguage (as METEO or TITRAN). In order to design an adequate suite of test inputs to test a sublanguage MT system there would be a need to study the specific sublanguage the system is intended for in order to specify the nature and occurrence of its typical lexical and syntactic features and design the TSs in terms of them.

3. It is clear that a TS is design to test the syntactic coverage of the system, but unless the system shows an overall better performance in most of the linguistic phenomena tested, it will be hard to establish to what extent this information can be useful to a user considering purchasing a system because TSs may reflect that a system has good coverage of a certain linguistic phenomenon but bad coverage of another (for instance, if a TS shows that system A deals with passives correctly in 90% of the sentences submitted to the system but with relatives it does well in only 45% of the sentences and that system B's performance for the same linguistic phenomena is 70% and 60% respectively, does this mean that system A is better than system B? or is system B better than system A? A potential purchaser of an MT system may not find this information very illuminating.

4. TSs may be feasibly used to test source language coverage but the task turns out to be much more difficult to test translational behaviour. However, some translational difficulties arise when we are dealing with a bidirectional system. A test suite constructed to test such a system cannot ignore typical differences between the languages involved in translation. It is essential, then to pay attention to contrastively based test inputs and this is a most urgent issue since there are practically no existing test suites that deal with translational problems between two languages. Therefore, the evaluator is forced to adapt monolingual ones.

It should be worth mentioning the existence of the TSNLP project, a LRC project of the CEC, which aims to construct test suites for NLP. The project has began the elaboration of guidelines for test suite construction and has undertaken an analysis of existing, publicly available TSs. As a result the lack of a systematic annotation scheme has emerged. For this reason one of the main tasks of the project is the elaboration of a systematic annotation scheme. Once guidelines are produced, TSs will be constructed, which will be validated and revised by testing them on NLP applications.

3.7. Evaluation with text corpora

An alternative to the construction of TSs is to submit large quantities of real text to the system. The idea is that real text will contain any kind of linguistic phenomena. However, with test corpora there is the problem of representativeness. The reason for this is that only a particular set of texts is consider in order to test the system and when confronted with different sets of texts, then the behaviour of the system might be completely different.

TSs seems to be preferred in the NLP community as they are superior to text corpora in several aspects:

1. TSs allow the construction of test data focused on specific phenomena the evaluator is interested in testing. Alternatively they can also be constructed to test controlled combinations of phenomena. However, in corpora, as these represents real texts, arbitrary combinations of different phenomena are common, which makes it difficult to measure the performance of the system with respect to a particular phenomenon.

2. TSs allow for systematic variation for specific phenomena; however, in corpora, if such variation occurs at all, it is bound to have been produced accidentally.

3. TSs allow for the construction of test data in which ungrammatical sentences would be included and which the parser should recognise as ungrammatical. However, in corpora it is very unusual to find ungrammatical sentences.

4. TSs can be annotated with information in order to help in the judgement of the expected output (king and Falkedal, 1990).

3.8. More about evaluation

We noted above that restricting evaluation to issues such as the cost-effectiveness of the MT output would mean neglecting other aspects of MT systems which require an examination of the inner components and design of the system and which cannot therefore be simply assessed by conducting a black box evaluation. We said that these aspects were: reliability, maintainability, extensibility, usability, portability and efficiency. Consideration of these aspects should be essential for a user thinking in purchasing a system.

Reliability

There should be motivation for designers to provide reliability in MT systems. Unreliable software leads to the practice of follow-on contracts for software maintenance and as maintenance costs may become intolerable, apart from maintenance being frustrating and difficult to accomplish, users of MT systems will be looking for a system with the latest reliable tool development. It will be a sign of robustness if the system incorporates some debugging tools to help the system designer to go into the system when there is a fault in the software to see what is happening and therefore put it right. Another approach is to design the outline of a system involving rapid prototyping. The degree of reliability of a system will also depend on the degree of tolerance of ill-formed input by the system: if a system is tolerant of ill-formed input it will not be able to handle some kinds of well formed input since this will depend on the rejection of incorrect analysis by the system.

Maintainability

Unreliable software leads to the need of software maintenance. Although the ideal situation would be one in which there would be no need of any kind of maintenance, this characteristic also deals with the use of certain techniques to make maintenance more comfortable, easier and safer. In this sense the degree of maintainability could be expressed in terms of looking at the inner design of the system to check whether modularity and declaritivity are introduced into it.

The division of the system into components, that is, modularity of the system will make it easier to locate and correct or remove algorithmic errors because the modules can be developed and tested separately. Also an important aspect of modularity and one to look for in the system, is the design of interfaces between components, the modules being independent from each other. Another aspect, closely related to modularity is declarativity. This involves a clear separation between data and algorithms, the data having an interpretation independent from the algorithm. This helps in distinguishing two different kinds of errors typical of large computational systems: errors in the data and errors caused by an incorrect algorithm.

Extensibility

Extending the system coverage from one domain to a related domain would imply that even though some vocabulary and grammar will be common to both domains, the dictionary will have to be expanded to cover new technical terms and the grammar will have to be expanded if new structures are encountered. If the texts serve the same function perhaps not many syntactic structures will need to be introduced but if the texts serve a different function then the probability of having to deal with important changes in syntax is much higher. The problems encountered so far will be the same if only more so when extending the grammatical and lexical coverage of the system to an unrelated domain.

The degree to which a system can be extended is difficult to measure. However, some features, if present in the system under consideration, can indicate whether that system is suitable for extension:

-The system is based on some coherent theoretical foundation

-The system's syntax does not deviate from standard grammar, that is, it is not a sublanguage system.

-The system has a separation between the modules of analysis, transfer and generation, that is, it is a transfer or interlingua system.

-The system has a high modularity. It would be a sign of good engineering if interfaces between components were clean. For example a system may be easier to extend if the core words or rules do not have to be re-entered when the system is being extended to a new domain.

Usability

As translation organizations grow they are getting more and more dependent on MT systems. The ability of end-users to make use of these systems becomes critical and in some cases that ability may even lead to successful functioning of the whole organization. If end-users find that the system interferes with their work and causes them stress and frustration, then they may refuse to use it altogether. Usability plays an important role in determining users' decision in using an MT system and the importance for adequacy evaluation is clear, as it is concerned with the capabilities of the system to fulfil user's requirements.

System designers have become more and more aware of the need of creating systems which are easy to use by customers and they are providing more and more systems with an easy to use environment.
The user interface, which allows interaction between end-users and the computer, plays an essential part in the effective communication between the human and the computer. The user interface consists of facilities which allows information to be displayed to the user and facilities which allow the user to enter information into the computer, to manipulate the displayed information and to take control actions. User interfaces must have a good design and meet the requirements of those who are going to use the system. Evaluating the interface is relevant for MT, not only for interactive MT but also for pre-editing and post-editing.

Portability

A MT system should be integrated into the users' environment. If the system is mountable on the user's hardware and interfaces with the user's software then the system will be a cost-effective near-term solution. If the system can be easily ported to the user's computer environment then the system could be a possible medium-term solution (Arnold et al, 1993). However, if a system is difficult to port to the user's environment and causes a lot of disruption, it may not be cost-effective.

Efficiency

Also important for the user is the time related to any testing done to extend or improve the system. This can be measured in terms of how long it takes to encode, test and verify a new grammatical construction or dictionary entry.

4. CONCLUSION

NLP is a field of great heterogeneity and complexity: NLP system evaluation can be performed from the perspective of system developers, researchers or users; there are different types of evaluation: progress, diagnostic and adequacy evaluation. Moreover in NLP there are many different applications and evaluations for these different applications and evaluations for these different applications very in purpose, scope and nature of the subject being evaluated. Therefore, it seems that at the moment the creation of a general methodology for NLP is quite impossible as it is very difficult to combine the different perspectives involved in NLP into a general evaluation methodology appropriate for many applications. Instead, evaluations should take into account differences between tasks and applications.

This paper has focused on adequacy evaluation for the specific case of MT; other NLP applications have been considered in order to throw light in the task of evaluation. First we tried to determine the object of evaluation and considered whether the output should take the context into account or whether it should be considered independent from context. It was further observed that evaluation could be different in detail: black box evaluation focuses on what a system does and glass box evaluation goes into the system and examines how it works. Relying on black box evaluation would involve neglecting other components of the systems that can give rise to unsatisfactory output.

Operational evaluation, provides information about the rest of the whole translation process of one MT system comparing it to another system or to Human Translation.

As we pointed out syntactic coverage of the systems can be tested by submitting a suite of specially constructed sentences. An alternative method is to submit large quantities of real text to the system. TSs were shown to be preferred by the NLP community as they allow the construction of test data focusing on specific phenomena in a systematic way.

ACKNOWLEDGMENTS

I would like to thank Jock McNaught for his support and helpful comments in proof-reading this paper.

REFERENCES

Alexa, M. Corpus-based Sublanguage Analysis for a Multilingual Generation Systems, 1993.

ALPAC. Languages and Machines. Computers in translation and Linguistics. Washington: National Academy of Sciences National Research Council, 1966.

Arnold, D., Sadler, L., and Humprheys, R.L. "Evaluation: An Assessment". In Machine Translation, vol.8, Nos 1-2. Amsterdam: Kluwer Academic Publishers.

EAGLES Evaluation Working Group, Report of the Evaluation of NLP Systems Working Group, Pisa: Consorzio Pisa Richerche, Draft Interim Report, 1994.

Guida, G. and Mauri, G. "Evaluation of Natural Language Processing Systems: Issues and Approaches". Proceedings of the IEEE, vol.74, No.7, July 1986.

Hutchins, W.J. and Somers, H.L. An Introduction to Machine Translation. Cambridge: Academic Press Limited, 1992.

Kittredge, R.I. "The Significance of Sublanguage for Automatic Translation". In Nirenburgh, S. (ed) Machine Translation. Theoretical and Methodological Issues. Cambridge: Cambridge University Press, 1988.

King, M. and Falkedal, K. "Using Test Suites in the Evaluation of Machine Translation Systems". In Proceedings of Coling' 90, Helsinki, 1990, pp. 211-219.

Palmer, H. and Finin, T. "Workshop on the Evaluation of Natural Language Processing Systems". In Computational Linguistics, vol 16, No 3, Sep.1990.

Sager, J.C. Language Engineering and Translation. Consequences of Automation. The Netherlands: John Benjamins B.V., 1994.

Wilks, Y. "Systran: it obviously works, but how much can it be improved?". Las Cruces. New Mexico, 1991.

APPENDIX

Quality characteristics

Functionality: A set of attributes that bear on the existence of functions and their specified properties.
The functions are those that satisfy stated or implied need.

Reliability: A set of attributes that bear on the capacity of software to maintain its level of performance under stated conditions for a stated period of time.

Usability: A set of attributes that bear on the effort needed for use, and on individual assessments of such use, by a stated or implied set of users.

Efficiency: A set of attributes that bear on the relationship between the level of performance of the software and the amount of resources used, under stated conditions.

Maintainability: A set of attributes that bear on the effort needed to make specific modifications.

Portability: A set of attributes that bear on the ability of the software to be transferred form one environment to another.

PETICION DE COLABORACIONES

Próximo número de la Revista de la Asociación Española para el Procesamiento del Lenguaje Natural.

El objetivo de este número es proporcionar y estimular las distintas áreas implicadas en el Procesamiento del Lenguaje Natural. Los artículos pueden cubrir un amplio abanico de temas: tratamiento del habla, morfología, sintaxis, tratamiento de corpus, lexicografía computacional, semántica pragmática, interfaces en lenguaje natural, sistemas de diálogo, ayudas a la traducción, etc.

Este número aparecerá en Febrero de 1996 y contendrá, además de artículos sobre temas de investigación, secciones en las que se presentarán proyectos en curso y tesis doctorales leídas en los últimos meses; además, se publicará información sobre congresos, cursos, conferencias, mesas redondas, etc.

FORMATO DE LOS TRABAJOS

Artículos

El objetivo de los artículos es presentar los últimos resultados obtenidos en trabajos de investigación en curso en cualquiera de las áreas arriba mencionadas.

Los autores deberán enviar 2 copias de los artículos mecanografiados a doble espacio en A4 y estilo de letra "times 14" (ya que han de ser reducidos). Los escritos mantendrán 3 cm. en el margen izquierdo y **no han de sobrepasar las 15 páginas de texto.** Las subsecciones deben ir numeradas y tituladas.

Resumen de proyectos

En lo que se refiere a las presentaciones de proyectos de investigación en curso, se publicará un resumen de un **máximo de 3 páginas.**

Igual que en el caso de los artículos, los autores deberán enviar 2 copias de los resúmenes tecleados a doble espacio en A4 y estilo de letra "times 14" (ya que han de ser reducidos). Los escritos mantendrán 3 cm. en el margen izquierdo.

Tesis doctorales

En este caso, se presentará un resumen de un máximo de **2 páginas** sobre los contenidos de la tesis doctoral. El formato que deben seguir los trabajos se ajusta a lo especificado en los artículos.

Información de interés

Notas para facilitar la edición

•En los originales de los trabajos las páginas deben estar numeradas a lápiz.

•Todas las figuras y tablas se presentarán en hojas aparte, en el formato final para su publicación e indicando su colocación en el texto.

•Es preferible contar con el trabajo en diskette. Si se ha preparado en *Word* o en *WordPerfect*, se agradecería que se enviara el diskette además de las dos copias impresas.

Fecha límite para la presentación de trabajos: 11 de Octubre de 1995

Los originales pueden enviarse a:
Arantza Díaz de Ilarraza
Lengoaia eta Sistema Informatikoak Saila
Informatika Fakultatea. Euskal Herriko Unibertsitatea
649 P.K. 20080 Donostia (Gipuzkoa)
Tlfno: 943-218000. e-mail: aradiaz@si.ehu.es

SEPLN Impreso de inscripción

DATOS PERSONALES

Apellidos..

Nombre..

DNI..............................Fecha nacimiento..........................Teléfono..............................

Domicilio..

Código Postal .... Municipio ....

DATOS PROFESIONALES

Centro de trabajo..

Domicilio..

Código Postal .... Municipio ....

Teléfono..........................FAX..............................E-mail..

Areas de investigación o interés:..

....

La correspondencia de SEPLN debe ir dirigida a la dirección: Personal ___ Profesional ___.

DATOS BANCARIOS

Banco..Sucursal....................................

Dir. sucursal..

Cod. postal y municipio ....

Cod. Banco	Cod. Suc.	Dig. control	Número cueenta
, , ,	, , ,	,	, , , , , , , , ,

SEPLN Sociedad Española para el Procesamiento del Lenguaje Natural

Sr. diréctor de:
ENTIDAD .... Nº SUCURSAL ....
DOMICILIO ....
MUNICIPIO .... Cod. Postal ....PROVINCIA ....
TIPO DE CUENTA(Corriente/Ahorro)..................................
NUM. CUENTA...................... ....

Ruego a Vds. que a partir de la fecha y hasta nueva orden se sirvan abonar a la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) los recibos correspondientes a las cuotas vigentes de dicha asociación

Les saluda atentamente

Fdo: ....

.... de ....de 19.........

Cuotas de los socios:
Residente en España 3.000 pta, Residente fuera de España 4.000 pta , Institucional 50.000 pta

SEPLN Impreso de inscripción (socios institucionales)

DATOS ENTIDAD/EMPRESA

Nombre..

NIF...Teléfono...

FAX...E-mail...

Domicilio..

Código Postal .... Municipio ....

Provincia...

Areas de investigación o interés:...

....

DATOS DE ENVIO

Dirección...

Código Postal .... Municipio ....

Provincia...

Teléfono.........................FAX.............................E-mail......................................

DATOS BANCARIOS

Nombre de la entidad..

Domicilio...

Cod. postal y municipio ....

Provincia...

Cod. Banco	Cod. Suc.	Dig. control	Número cueenta
, , ,	, , ,	,	, , , , , , , , ,

SEPLN Sociedad Española para el Procesamiento del Lenguaje Natural

Sr. diréctor de:
ENTIDAD ....Nº SUCURSAL ....

DOMICILIO ....
MUNICIPIO ....Cod. Postal ....PROVINCIA ....
TIPO DE CUENTA(Corriente/Ahorro)......................................
NUM. CUENTA........................ ....

Ruego a Vds. que a partir de la fecha y hasta nueva orden se sirvan abonar a la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) los recibos correspondientes a las cuotas vigentes de dicha asociación

Les saluda atentamente

Fdo: ....

.... de ....de 19 ....

Cuotas de los socios institucionales: 50.000 pta