A variety of factions may be interested in evaluating natural language processing (NLP) systems, ranging from funding authorities who must choose between competing research projects and justify their choices through the results subsequently obtained, to the end user who needs to choose between competing products. If the product is expensive, or if it implies a major reorganization of workflow, the user may also need to provide post hoc justification.

Surprisingly, the literature on evaluation is relatively sparse for several reasons. First, evaluations are often carried out under consultancy arrangements for a particular customer. Not only is the evaluation then tailor-made to suit that customer and therefore not considered to be of general interest, but the customer may be reluctant to have the results of the evaluation made public, either because of an agreement with the manufacturer whose product has been evaluated, or an unwillingness to reveal the results to competitors.

Secondly, evaluation of research proposals and of projects in the academic area has traditionally been by peer review. This has changed somewhat in particular areas, under the influence of the DARPA/ARPA series of evaluations, but is still the most common pattern. The only result of a peer review evaluation is a report which is often confidential. Thirdly, evaluation acquired a bad name as a result of the Automatic Language Processing Advisory Committee’s (ALPAC)
evaluation of machine translation in the mid-1960s. Subsequently, those involved in evaluation were understandably reluctant to discuss their techniques unless they were very sure of their ground. Finally, serious evaluation is expensive. Those who pay for the evaluations have sometimes been reluctant to share what they have purchased with third parties.

Partly under the influence of the ARPA efforts and partly because of a growing interest in being able to share resources such as test materials and techniques which are expensive to produce, the European Commission included a working group on evaluation among the groups created under the Expert Advisory Groups for Language Engineering Standards (EAGLES) initiative launched in 1992. This article has been heavily influenced by the work of that group; an outline of the EAGLES’ proposals presented here serves as a conceptual framework for this piece.

The EAGLES proposal [4] builds on the ISO 9126 standard [12] for quality characteristics to be used in the evaluation of software. The evaluation process model given as a guideline in ISO 9126 starts with the definition of quality requirements. Relevant technical and standards documentation serve as input to this definition, as do the stated or implied needs of those who will use the results of the evaluation. Quality requirements are expressed in terms of quality characteristics, each of which is a set of attributes which may be further refined as necessary. Each attribute leads to the choice of one or more metrics by which a value for that attribute may be obtained.

Let us make this more concrete. Imagine that a railway timetable system is to be evaluated to determine whether it should be installed in every main railway station for use by the general public. The user will type in a question, such as “When is the next train to Lugano?” and will get an answer like “At 15:30, leaving from platform 3. There is a change in Lausanne. The train arrives in Lugano at 22:00.” Two relevant quality characteristics here are the functionality of the system (whether it does what it is supposed to do) and its usability (how easy it is to use). Attributes of functionality could include: does it give the correct information? Attributes of usability could include: is it easy to understand the information given?

Defining relevant attributes is of no use if there is no way to measure the system’s value with respect to those attributes. Measures must be both valid and reliable. They must measure what they are really supposed to measure and they must measure it consistently. To continue with the railway timetable example, a measure of whether the system gives the correct information might be to formulate a series of questions which the system should be capable of answering, and find out whether it can in fact answer them correctly. But that measure will not be valid if the questions only concern travel to major cities, neglecting less important destinations, or if they concentrate on trying to catch the system out by asking about villages which cannot be reached by train. Analogously, a measure of whether it is easy to understand the information given might be to give a list of questions which the system is known to be capable of treating correctly to a statistically significant set of representative potential users and measure how long it takes them on average to obtain answers to all the questions on the list by using the system. This measure will not be reliable if, through carelessness, some users are asked to read a screen in bright sunlight and others in a darkened room.

Some measures, like those mentioned, give a reassuring air of objectivity by being expressible in quantitative terms. Others call explicitly for subjective judgement and are qualitative by nature. Some attributes will have their value determined by purely factual information, such as the language a spellchecker uses.

The ISO 9126 evaluation model sees evaluation in terms of judging a system’s adequacy: whether it can meet a set of stated or implied needs. We might also distinguish progress evaluation, typically carried out by a system developer in order to determine whether progress has been made towards some desired goal state of the system, and diagnostic evaluation, designed not only to discover whether the system fails but why it does so.

Some Past Evaluations
With these considerations in mind, let us now look briefly at some past evaluations. In an article of this length, it would be impossible to give an exhaustive review; for that, the reader is referred to [5] and [9]. The first gives a critical review of a number of machine translation evaluations, the second is concerned with natural language processing in general, using many insights from the field of information retrieval which has a mature and well-established evaluation methodology. Both have valuable extensive bibliographies.

The intention behind the choice of example evaluations here is to illustrate the variety of evaluation scenarios and to introduce some common evaluation tools and techniques; no judgement should be inferred as to the quality of the evaluation itself.
ALPAC’s evaluation of machine translation was one of the first [1]. It also created enormous controversy, and is still often reproached with having been both politically motivated and deliberately biased. Nonetheless, it does offer us a very clear example of an attempt at rigorous evaluation, and some of the techniques used are still quite common. Furthermore, its impact on the research community highlights the importance of developing sound and well-accepted evaluation methodologies. For a fuller account, see [5, 11].

The conceptual framework of the evaluation was a comparison of machine translation and human translation on the three dimensions of speed, cost, and quality. Here we shall focus only on Carroll’s quality assessment experiment, which introduced techniques and notions which are still quite commonly used.

The measure used was to take a set of translations, some produced by machine, some by human translators, and ask a group of test persons to rate the translations on two scales, one for intelligibility and one for fidelity. The test material was 144 sentences randomly selected from 4 different passages of a Russian book. Six different translations were produced for the 144 sentences, three by human translators and three by different machine translation systems. The translations were then merged randomly into six sets, with the constraint that each sentence appeared only in one translation in each set. Each set was then given to three monolingual and three bilingual test persons, all of whom had had one hour’s training using a set of 30 sentences drawn from the same material as the test set. There were 96 test persons in total.

A definition in English was given for each of the points on each of the rating scales. Thus, intelligibility was rated on a nine-point scale from “perfectly clear and intelligible” to “hopelessly unintelligible.” Fidelity was defined over a 10-point scale in terms of informativeness, the basic idea being that if the original conveyed additional information if it was read after the translation had already been read, the original was low in informativeness relative to the translation, and the translation was therefore good. That this definition is somewhat counter-intuitive can be deduced from how frequently it is inaccurately reported in the literature on evaluation.

The committee reached extremely negative conclusions. The validity of the evaluation and of the conclusions has frequently been challenged. In our terms, it is at least clear the measures used were entirely qualitative, and that the perverse definition of the fidelity scale must cast some doubts on their validity.

There have been many more recent machine translation evaluations, most frequently adequacy evaluations carried out on behalf of a potential customer. In adequacy evaluation, a great deal of effort is typically required to determine what the potential customer’s needs really are. Among many others, a customer normally will intend to use the system to translate only certain kinds of texts. This need is often reflected in the use of one or more test corpora—collections of naturally occurring machine readable text which can be submitted for translation.

Before leaving the topic of adequacy evaluation, it is worth making a point which is valid for commercial NLP systems in general. This is true even for modest products like spellcheckers, which, although they may aim at covering the general vocabulary of a given language extensively, cannot realistically be expected, for example, to cover all of a particular client’s terminology needs. Thus evaluation is aimed at finding out not only what the system currently does but also how easily it can be modified. This leads to another common evaluation technique, whereby errors are analyzed into classes intended to indicate which errors will be easy to correct, which not. A recent example of error classification used for this purpose can be found in [7].

Between 1992 and 1994, a series of evaluations of ARPA sponsored machine translation projects was carried out [20]. The intention was to stimulate development of the core technology by comparative evaluation. A basic problem faced by the designers of the evaluations was the diversity of the systems being evaluated. The three research systems were based on radically different linguistic technologies. Their intended mode of use was also fundamentally different: one was designed as an interactive system, with human intervention to resolve problem cases, one was intended as a fully automatic translation system, one was designed primarily as an aid to a human translator. This problem was compounded by inviting operational systems, including commercial systems, to participate in the evaluation.

Given the focus on core technology, no specific community of users was taken into account. The only quality characteristic to be evaluated was functionality, and its attributes were taken to be definable in absolute terms. In the 1992 evaluation, the attributes singled out were comprehensibility and quality of translation.

The systems to be evaluated also translated from different languages into English. The evaluation design tried to compensate for this by using as test material a corpus of newspaper articles, originally written in English, but which had been professionally translated into the appropriate source language for each system. These translations were then input to the machine translation systems for translation back into English.

Comprehensibility was measured by presenting the results to a set of monolingual speakers of English, in the form of a comprehension test, in which they had to answer multiple choice questions about the content of the articles.

This measure proved not to be valid. The human
translations into the systems’ source languages inevitably modified the content of the original, and it was impossible to separate the effect of the initial human translation from that of the subsequent machine translation. Later evaluations have retained the comprehension test, but have not used human translations as test material. Even so, the measure remains vulnerable to criticism. Comprehension tests are typically used to measure human intelligence. One might ask to what extent the intelligence of the person answering the multiple choice questions affects the results and thereby still obscures measurement of the system’s performance.

The translation quality attribute was measured by asking a panel of judges, composed of professional translators, to grade the machine translation outputs on a scale based on one normally used for grading human translations. This measure proved to be neither valid nor reliable. The scale had to be modified to take into account the nature and proliferation of the errors made by the systems, and the panel of judges found it extremely difficult to agree on the grade to be awarded. The measure was subsequently dropped.

In later evaluations, the translation quality attribute has been refined into two subattributes: adequacy and fluency. The measure of adequacy is to ask literate monolingual English speakers to judge, fragment by fragment, the degree to which the information in a professional translation can be found in a machine translation of the same text. The measure of fluency is to ask the same evaluators to determine on a sentence-by-sentence basis whether the translation reads like good English. They have no means of checking the accuracy of the translation: their task is simply to determine whether each sentence is well-formed and fluent in context.

As reported in [20], the three evaluations completed by 1994 raised a number of issues about the validity and the reliability of these measures. By far the most important, however, was the general question of human-assisted measurement. Any human involvement raises the issue of the extent to which the human is being evaluated as much as the system’s performance. For example, a human’s abilities as a translator and his familiarity with the system’s interfaces will both affect the results obtained from an interactive system, as well as the time required to obtain the results. Such considerations finally led the evaluation designers to drop all human-assisted measurements, and to evaluate each system only on fully automated outputs, whether or not the system was initially intended to produce output without human intervention.

Whatever reservations one may have about the proposed solution to the dilemma, much credit must go to the ARPA evaluators for insisting on the critical importance of the issue, and for their willingness to discuss their techniques openly.

Space constraints prevent any fuller discussion of the wide variety of evaluation scenarios and techniques reported in the literature on machine translation. A recent collection of papers can be found in [6], and [14] reports on a recent Japanese effort to take users and their requirements into consideration during evaluations.

The ARPA machine translation evaluations are the latest in a tradition created and maintained by DARPA and by ARPA of an attempt to define careful and rigorous evaluation techniques, to apply them to comparative evaluation of a number of systems which may all work according to quite different principles, and to discuss both the evaluation techniques and the results openly. Other evaluations in the same tradition include the ATIS [2] and TREC [10] evaluations, the first in the domain of database query, emphasizing a spoken language component, the second in the domain of text retrieval. A third set of evaluations—the MUC evaluations of fact extraction systems—is reported by Cowie and Lehnert in this issue and is reported in detail in [18]. The test material for these other evaluations in the ARPA tradition is, however, critically different. All of them make use of a test collection, a set of inputs with a predefined set of correct outputs to which a particular system’s actual outputs can be compared, semi-automatically or even automatically in some cases. It is interesting to note that it would be impossible to follow the same plan for evaluation of machine translation. It is in the nature of translation that for any given text, potentially many translations would all be equally acceptable: it is not possible to define a “correct” translation as being the only acceptable output from the system.

Database query is another application of natural language processing with a long history of evaluation. Informal field testing of the Lunar system through monitoring the treatment of 110 queries during demonstration of the system is described in [21], and [3] reports more extensive field testing of the Transformational Question Answering (TQA) system—a transformational grammar-based front end linked to a pre-existing database of town planning data—over a period of two years from late 1977 through 1979. Both of these were clearly adequacy evaluations, with the interesting characteristic of being executed in close collaboration with the end-user community. The emphasis on field testing of database query systems is reflected also in recent work [13, 22].

In this article we shall concentrate on a proposal made in the context of progress and diagnostic evaluation by a group at Hewlett Packard [8]. They argue that although no evaluation tool could be developed for use with NLP systems in general, it should be possible and useful to develop a methodology for a single application domain (database...
Even the very small number of incidences described here are enough to support the conclusion that evaluations vary enormously in their purpose, in their scope, and in the nature of the object being evaluated.

Resources for Evaluation

Even the very small number of incidences described here are enough to support the conclusion that evaluations vary enormously in their purpose, in their scope, and in the nature of the object being evaluated. Consequently, it is hardly surprising that evaluation techniques in their turn differ widely, as do the resources they require. It is this observation which led Flickinger et al. [8] to the conclusion that it is in principle impossible to envisage the design and construction of some general evaluation tool, into which any NLP system could be plugged in order to obtain data relevant to a set of informative measures. However, this conclusion is not universally accepted: Neal et al. [19] report on an attempt to create a more general evaluation tool.

In any case, we might ask whether it is not possible to share some of the resources used in evaluation. It is clear that test materials, like test collections or test suites, are expensive to produce and maintain. There would be obvious interest in producing materials which could be shared by the community as a whole and reused in different evaluations. Furthermore, use of the same test materials might be expected to produce evaluation results which could be compared more easily, thus leading to another kind of shared resource. In this final section, we consider some of the issues surrounding attempts to produce shared resources.

Test corpora, in the sense used here, are collections of naturally occurring text in electronic form. Some of the best known corpora available are the Brown corpus of English, the Trésor de la Langue Française corpus of French, and the bilingual (English-French) parallel corpus drawn from the Canadian Hansard. With the power of increased computing capacity to process and store large amounts of text, a strong interest has been developing recently in collecting corpora and in defining tools able to make use of them [17]. In line with this, the Association for Computational Linguistics has launched a Data Collection Initiative (DCI) and the Linguistic Data Consortium is specifically concerned with collecting corpora. In Europe, the European Commission supports the collection of multilingual and plurilingual corpora.
These efforts are all concerned with collecting what might be called general corpora; that is, they do not aim to reflect some particular pattern of needs or some specific set of uses to which the corpus may be put, but rely on text that can be found in machine-readable form in large quantities. Although this does not detract from the value of such collections of linguistic data, it does raise questions about their representativeness. For example, the Hansard corpus is clearly representative of the English and French used in the Canadian Parliament, but is highly unlikely to be representative of the French used in technical documentation or the English used in school textbooks. Why this is a problem is most clear in the case of evaluations designed to test a system’s ability to deal with some particular type or types of text; it will be pure serendipity if a general corpus contains texts of the relevant types.

On the other hand, there is a strong sense in which every corpus is representative of something, and this in its turn can lead to misleading results if the corpus is mistakenly taken to be representative of, say, English or French in general.

By a test suite, we mean sets of inputs, artificially constructed and designed to probe the system’s behavior with respect to some particular phenomenon. The main problem associated with test suites is the complexity of their construction. Even at the level of syntactic phenomena, there are problems in defining inputs which will test precisely what one wants to test, and once semantic, pragmatic, or translation phenomena are taken into consideration, test suite construction becomes a very delicate matter.

Furthermore, test suites can quickly become unmanageably large. A principle usually adopted is to design one input per linguistic phenomenon to be tested, in order to isolate the system’s behavior with respect to that phenomenon. However, real text rarely contains one interesting linguistic phenomenon per sentence, and much of the real interest in a system’s behavior is in looking precisely at what happens when the input contains interacting phenomena. A test suite based on constructing one input per phenomenon is already large if any serious attempt is made to cover a language exhaustively. Once interactions are to be accounted for, the problem of size becomes critical. Notice too that size is a problem not only in constructing the test suite but in administering it and in analyzing the results. These and some of the problems specifically associated with constructing test suites for evaluation of machine translation systems are discussed in [15].

All these considerations constitute a good argument for thinking of general test suites of the sort envisaged by the Hewlett Packard group as good candidates for collaborative development. A note of caution is in order, however, if collaborative development is taken to mean simply pooling such test suites as exist. Test suites are often designed in the context of testing a specific system. There is a danger in that case that, deliberately or inadvertently, they are attuned to that particular system, thus limiting their applicability when other systems are to be evaluated.

By a test collection we mean a set of inputs associated with a corresponding set of expected outputs. In information or document retrieval, for example, a test collection consists of a set of documents, a set of queries or topics and a set of “relevance judgements,” which identify the individual documents relevant to the individual topics or queries. Typically, these elements of the test collection are divided into “training” and “test” sets.

The most costly element in creating such a test collection is the creation of the relevance judgements. When a large set of documents is concerned, the effort involved is so great that means are employed to allow the evaluation to be conducted with an incomplete set. Substantial effort is also required to develop the queries and topics. For the TREC evaluations mentioned earlier, they were designed by information analysts to present varying degrees of difficulty and to cover a wide range of subject matter.

In the MUC-3 evaluation [18], the filling of the templates for the terrorism test collection required not only making relevance judgements but also determining how many relevant terrorist incidents were reported in a given document (and therefore how many templates to generate) and which passages contained explicit or implied information pertinent to each of these incidents. This task was carried out in a cooperative venture involving the evaluators and the conference participants. Virtually every text presented difficulties of interpretation due to vagueness, ambiguity or downright self-contradiction, or due to inadequacies of the template representation or in task documentation. Therefore, as the template filling task proceeded, the documentation containing the task and output specifications had to be refined. (For a lively discussion of this effort, see [16]).

The value of such a test collection in evaluating systems comparatively to discover their adequacy with respect to the task defined by the test collection cannot be denied. The effort and cost of constructing the test collection leads one to wonder how the test collection can be reused elsewhere. First, of course, such a collection provides an obvious and valuable source of test materials to research groups working in the domain, independently of participation in the comparative evaluations themselves. However, a new set of issues arises if the collection is to be used for progress evaluation. In order to obtain direct comparability of results, the collection and the evaluation metrics should remain absolutely stable. Yet freezing
the collection prevents it from being as good as it might be. For example, for information retrieval, additional relevance judgements will improve it, and for information extraction, higher quality filled templates will improve it.

Although a test collection can be reused for a series of ARPA-like formal evaluations (usually changing the test set), a change must eventually be made. If no change is made, the collection may at some point start to hinder progress in the field, as it tends to encourage researchers to focus on certain key problems and to ignore others, and, once a system reaches a certain level of maturity, it may encourage researchers to spend more time tweaking the system than tackling the remaining major issues. However, it is certainly true that the TREC and MUC collections are sufficiently large, challenging and well-defined to support research and development in information retrieval and information extraction for a long time, even after they are no longer used for formal evaluation.

However, the high cost of designing and constructing test collections makes it hard to imagine their being constructed outside the evaluation guided research paradigm, where the investment implied by a number of different groups working at essentially the same task over a considerable period can be used to justify the expense involved.

With all the test materials we have discussed, there is a tension between constructing test material which is in some sense general and can be shared across different evaluations and the common-sense feeling that most evaluations are specific, at least in the degree that the system has been constructed to carry out some specific task and should be evaluated on its ability to do so. This is as true for evaluation methodologies as it is for test materials. Designing and carrying out an evaluation is costly both in time and in money; it would be helpful to everyone concerned if ways could be found to share methodologies. But at the same time each evaluation is very specific to a particular system, and, perhaps even more importantly, to a specific environment in which the system should work. Increased experience and widespread discussion of evaluation techniques as topics worthy of consideration in their own right should lead to a better consciousness of what can be shared.

References


About the Author:

MARGARET KING is Director of ISSCO and associate professor at the University of Geneva School of Translation and Interpretation.

Author’s Present Address: ISSCO, 54, route des Acacias, Ch-1227 Carouge, Geneva, Switzerland.

Permission to make digital/hard copy of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication and its date appear, and notice is given that copying is by permission of ACM. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or a fee.

© ACM 0002-0782/96/0100 $3.50