Evaluating Symbiotic Systems: the challenge

Margaret King and Nancy Underwood

ISSCO/TIM/ETI,University of Geneva
40, bd. du Pont d'Arve,1211 Geneva 4, Switzerland
E-mail: Margaret.King@issco.unige.ch, Nancy.Underwood@issco.unige.ch

Abstract

This paper looks at a class of systems which pose severe problems in evaluation design for current conventional approaches to evaluation. After describing the two conventional evaluation paradigms: the "functionality paradigm" as typified by evaluation campaigns and the ISO inspired "user-centred" paradigm typified by the work of the EAGLES and ISLE projects, it goes on to outline the problems posed by the evaluation of systems which are designed to work in critical interaction with a human expert user and to work over vast amounts of data. These systems pose problems for both paradigms although for different reasons. The primary aim of this paper is to provoke discussion and the search for solutions. We have no proven solutions at present. However, we describe a programme of exploratory research on which we have already embarked, which involves ground clearing work which we expect to result in a deep understanding of the systems and users, a pre-requisite for developing a general framework for evaluation in this field.

1. Introduction

Over the last few years two paradigms have emerged for the evaluation of human language technology systems. The first of these, which we will call the functionality paradigm, focuses exclusively on what are seen as desirable functionalities of the systems being evaluated, and indeed, very often focuses on just one aspect of functionality, reducing evaluation to the application of a single metric, or of a handful of closely related metrics. The results obtained by applying that metric are then most frequently used to compare a number of systems. The functionality based paradigm is typified by a large number of evaluation campaigns. To cite but a few very old and quite new examples, campaigns evaluating speech recognition systems have used the word error rate metric as a standard for comparison (Pallet et al, 1993), document retrieval campaigns have used the twin metrics of precision and recall (Sparck Jones, 1995), machine translation campaigns have used BLEU (Papinieni et al, 2002) or its cousin the NIST metric (Doddington, 2002).

The second paradigm is heavily influenced by the ISO 9126 notion of quality in use (ISO/IEC, 2001), insisting on the idea that a system has to be evaluated in terms of its potential to help a user to achieve a task. Since users come in many shapes and sizes, have very different needs and work in very different contexts, this approach seems almost counter to the notion of setting up a general methodology for evaluating even the systems in one particular application area. To show that a general methodology is nonetheless possible, a great deal of energy has gone into finding ways of identifying classes of users and tasks, spelling out what it is that characterizes any given class and relating those characteristics to system characteristics and to metrics. We will call this approach the user-centred paradigm: it is typified by the various EAGLES and ISLE initiatives and in particular by the FEMTI framework for the design of evaluations of machine translation systems (Hovy et al 2002, Estrella et al 2005, Popescu-Belis et al 2006)).

In this paper we want to look at a class of systems which seem to pose severe problems of evaluation design for both of the conventional paradigms, although for different reasons. We have no proven solutions to present: our main aim here is to provoke discussion and the search for solutions.

2. Symbiotic Systems

We call the systems of interest to us here "symbiotic systems". They are characterized by being designed to work in critical interaction with a human expert user and to work over vast amounts of data, trying to discover from that data insights and information that a human mind would be unable to capture alone. Thus they are symbiotic in two senses; they cannot function alone but depend on the human user to achieve satisfactory results, and they are critically dependent on the data which provides the raw material for their search. Many kinds of knowledge discovery systems are symbiotic systems: it is trying to design evaluation methodologies for text mining systems in particular which has made us aware of the theoretical issues involved.

2.1 Accuracy Is Not Enough

The reason why symbiotic systems pose problems for the functionality based paradigm is fairly obvious. The system on its own attempts to sort data elements into clusters, to classify new elements based on training with previous exemplars, to discover associations and trends.

Often, the data which is mined has been previously extracted from a large quantity of free text, using natural language processing techniques which have been, where necessary, tailored to the particular application, through the use of domain specific ontologies, of specific terminology or the use of linguistic rules reflecting the nature of the users' interests and the nature of the text.

There are known metrics for evaluating the standard data mining components lying at the heart of such a system, looking, for example, whether the clusters are internally coherent and sufficiently distinct one from another or whether the associations found are valid given the nature of the data. But being able to produce a cluster or an association does not guarantee that it will be an **interesting** cluster or association. An example often quoted in the data mining literature is finding an association between being pregnant and being female. The validity of the association cannot be denied, but it hardly counts as an association leading to new insight. Another example frequently quoted is clustering people on the basis of the number of their bank account. Again, a common element which triggers creation of the cluster is certainly there, and the software can hardly be reproached for finding it. But the cluster is rather unlikely to be informative. For useful results, the intervention of a human expert user is required (expert both in the specific domain in which the application is being used and in the use of knowledge discovery tools). It is he who will look at a first set of results, identify what factors in the data have led to uninteresting information and direct the software to try again, ignoring the elements which lead in unpromising or false directions.

It is true of course, that the knowledge discovery system software must perform accurately its part of the task. The essential point here is that accuracy in the sense of producing results which conform to the software's specifications is, with symbiotic systems, largely divorced from suitability in terms of achieving the user's task in hand. The functionality based paradigm is promising in those cases where some core functionality can plausibly predict ultimate suitability; here that cannot in general be the case.

2.2 Dirty Data, Bad Results

Furthermore, the systems also live in symbiosis with the data over which they work. If the data is poor, the system will reach poor or even patently false conclusions, no matter how good its internal accuracy is, and cannot be blamed for doing so. To make a caricature example, if the data includes a number of articles suggesting that all red haired people have three feet, the software may well propose an association between having red hair and having three feet. This again, is a known problem with data mining systems, where much emphasis is put on the effort needed to collect and clean data before it can be reliably used. But there may be circumstances in which ensuring that the data is clean is impossible, either because of its size or because of its nature (think, for example, of a text mining system working over all the data available on the World Wide Web, when both the size and the nature of the data collection would be problematic).

Another way of looking at this problem is to think about metrics based on comparing results produced with a gold standard of (by definition) correct results. It is a little implausible to think of creating a gold standard from data perhaps riddled with dross, or from data so voluminous that it would take an inordinately long time to search out the elements of the gold standard.

In an evaluation context, it might be possible to create a data collection specifically for the evaluation and to ensure that it contained no patently false or nonsensical elements, a tactic already adopted in some evaluations faced with very large data collections. But this raises a new problem about the validity of any metrics used. How can it be guaranteed that any result obtained over the artificially constrained data will carry over to real life data?

In cases where the stated aim of an application is to help the user to discover previously unknown knowledge, relationships and tendencies (as with the type of system described by Hearst, 1999) the problem of building a gold standard is in fact insurmountable since by definition the new knowledge cannot be known before running the system and so a gold standard cannot be defined.

2.3 The Unacceptable Cost of Evaluation

The problem of cost is, of course, well known and difficult for both evaluation paradigms. Within the functionality paradigm, the creation of training and test material is known to be very expensive, and this is exactly why such resources are normally substantially re-used, both inside and outside the specific evaluation campaign for which they have been created. But there are other and perhaps more intractable expense problems with the user-centred paradigm.

We have already mentioned a number of times that the ultimate success of a symbiotic system once it is deployed depends critically on the nature of the data over which it will work. Typically, the data reflects the interest of the specific user: someone interested in finding indicators that a terrorist attack is being planned will not make use of the same data as the person interested in finding existing scientific research which is pertinent to his current interests. An assumption that a system which produces satisfactory results from a particular set of data will similarly produce satisfactory results from a quite different set of data is an extraordinarily risky assumption, the more so if either or both sets of data are unstable. This implies that the results of an evaluation executed using one set of data are unlikely to carry over to a different data set. Thus it would seem that design of an evaluation for symbiotic systems each time requires preparation and use of different data. The consequences on the cost of the evaluation are obvious.

2.4 How to define classes of users?

The tension between the needs of specific users and the desire to create a general evaluation framework such that resources and results can be shared across evaluations was already present in the EAGLES work. It was resolved there by trying to reason over classes of potential users and their needs. With symbiotic systems, the variety in types of users, the different competences of the expert users interacting with the system and in the different kinds of data of interest in the particular context of use make it prima facie at least extraordinarily difficult to distinguish clearly identifiable classes of contexts of use and consequent sets of needs, a problem made worse by the difficulty of disentangling the abilities of the expert user from the satisfactory functioning of the system and the need to take the nature of the data into account. In the worst case, we are back to having to set up individual evaluations based on specific data but now also with specific expert users – a solution which is neither viable economically nor satisfying intellectually.

3. The importance of finding solutions

It is an acknowledged fact that the major problem of our web-based knowledge cultures is not getting access to pertinent information but being able reliably to manage the huge mass of good and bad information available. Knowledge discovery systems offer a way out of this impasse. But the entry barrier to their use is currently far too high. The investment needed to determine whether a knowledge based system will actually be of practical use is prohibitive for too many potential users.

4. A glimmer of light?

Despite the gloomy tone of this paper we do not ourselves believe the problem to be unsolvable. We have started some ground clearing work in collaboration with NaCTem, the British National Centre for Text Mining in the hope of working towards an eventual solution. This ground clearing work takes the form of looking carefully at what applications are on offer, who uses them and how, in an attempt to understand the problem better as a preliminary to trying to solve it. The exploratory research we are carrying out within the ISO-inspired user-centred paradigm comprises the following 3 main pillars

4.1 User Modelling

Despite what has been said earlier about the difficulty of defining classes of users, it might be possible to find a level of generality at which some characterizations can be formulated. In order to discover whether this is feasible, the obvious starting point is to look closely at actual and present users of symbiotic systems. The goal of this activity then is to collect data on real (current or potential) users of text mining and convert that information into characteristics of classes of users which define their requirements for text mining systems. A promising way to achieve that characterization, we hope, is to separate the description of users into a number of different aspects. As we have discussed above, in evaluating the suitability of text mining systems for a particular user or class of user it is vital to consider the tasks that they need to perform including the data which is to be mined. Thus, we include a description of the task to be accomplished and of the data set from which knowledge is to be acquired as part of the description of the user and his needs.

As a first step then we propose to work with the following characterizations of users and of their context of use: tasks to be performed; target text data; the users' own available resources; His levels of expertise and experience; his current set-up and workflow. From these attributes we hope to be able to define classes of users. Note in passing that, in conformity with standard ISO use, user here is used in the widest possible sense, ranging from the individual end user interacting with the system to the whole of the organization of which he is a part.

4.2 Modelling applications, tools and resources

The sorts of applications we are investigating are typically very complex, comprising a variety of different components. Indeed within a single application it is normally possible to interchange particular tools and resources depending on a specific user's needs. Therefore as well as trying to model specific application types we need to consider the component tools which (potentially) make up an application.

We mentioned in the introduction that the types of systems of interest to us frequently rely on natural language processing techniques as well as on well-known data mining techniques. In turn, the natural language processing techniques may rely on the existence of external linguistic resources, just as the data mining techniques may rely on the existence of, for example, appropriate ontologies. Thus, characterizing a component means describing its requirements on the availability of other resources as well as its behaviour with respect to the standard individual quality characteristics, together with its requirements in terms of interactivity. Here we present a preliminary list of the types of components which could contribute to a full-scale application: classical NLP components (e.g. tagging, parsing); information and extraction; named entity recognition; term recognition; classification; association rule mining; clustering. This list is not exhaustive and we fully expect that as the field continues to develop new tools and techniques will continue to be developed and applied.

Given the highly data intensive nature of text mining and its reliance on large-scale resources it is vital, as we have said, to consider the available resources which are essential for producing satisfactory results: the question of how to model the nature of resources should be given as much attention as modeling tools and components. For the moment, we distinguish three types of resources: terminology: annotated corpora and ontologies. Again, as the field develops, it may be that other types of large-scale resources also become relevant. Our ultimate aim is to analyse and describe applications (including their component tools) and resources in such a way that the correspondence between them and the requirements imposed by classes of users can be exploited.

4.3 Mapping between users and systems

Continuing our adherence to the ISO standards for evaluation we take as a basic tenet that a specific evaluation requires the production of a quality model. Building such a quality model involves translating user requirements into quality characteristics of the software which in turn are decomposed into sub-characteristics and eventually evaluation metrics which can be applied to the software. Thus the third pillar in our exploratory programme involves investigating the mapping between user requirements and systems and resources. Conceptually the procedure for such a mapping can be sketched as follows: For a specific user characteristic (e.g. his expertise in the subject domain or the task to perform; the data to process, constraints on efficiency etc), identify which aspects of the tool or resource should respond to the user's requirements and then determine how those specific aspects can be evaluated wrt those requirements.

Essentially the process is one of mapping from user requirements to system characteristics which then need to be decomposed to make them tractable to evaluation. A very simple and informal (but realistic) example may help in understanding what is intended here. Imagine a user whose textual data comes in a variety of formats, including .pdf, .doc and .xls documents. For this user to exploit his data fully, the application must be able to deal with documents in all of the different formats of interest. First the evaluation designer identifies those of the application's components which are sensitive to text formats.Then for each such component, the metric becomes a series of simple yes/no questions: can the component deal adequately with each one of the formats present and important in the user's data?

Notice though that answering the yes/no question may involve carrying out tests, and that it may eventually transpire that a binary yes/no value for the metric is not adequate in terms of informativeness. Although the metric may be conceptually simple, defining it thoroughly together with a method for applying it may require considerable reflection. The expectation is that for more complex user requirements we could discover a need for much more complex evaluation metrics.

More complex user requirements abound. To mention just a few drawn from our own experience with users of text mining systems, a user might want to know which elements in the data caused a particular proposal to be made by the system, might want to know what has changed in the (textual) data being mined between one session and a previous session or might want some indication of the reliability of data on which a suggestion was based. Space constraints prevent us from going into any discussion of these more sophisticated requirements.

4.3.1 Evaluation Metrics

The mapping sketched above may sound deceptively simple to implement and use in order to carry out an evaluation. However it must be remembered that decomposing the quality characteristics imposed by the user's requirements will almost certainly be far more complicated than simply decomposing the system into its component parts or functionalities and measuring their performance. For example in evaluating an information extraction system one of the user requirements may be that he wants to be able to extend lexical coverage by adding terminology himself; this could then be translated into a metric which allows the evaluator to isolate the relevant component and test whether it is possible to do that. Contrast this with the situation where the user has requirements about how the system affects his productivity, which translates to some sort of efficiency characteristic. In this case, measuring a single aspect of the system's functionality cannot possibly be adequate. To try and evaluate potential productivity gains the evaluator needs to devise metrics ranging over the system as a whole, looking at interactions between quality characteristics as well as individual characteristics and perhaps even setting up experiments to model the context of use in order to compare productivity using the system with the user's current productivity.

A complementary approach to applying metrics the software and resources is to apply metrics to aspects of the class of user, in particular the suitability and tractability of the target textual data for processing with a particular system or application. A promising recent attempt has been made in the field of data mining to provide a semi-automatic service to users to determine whether their data is tractable to particular classification algorithms (METAL-KDD, http://www.metal-kdd.org. See also Vilalta et al, 2004). We intend to explore the development of similar metrics on a broader scale.

The shortcomings of some established functionality metrics traditionally applied to knowledge/information discovery systems in the context of user-oriented task-based evaluation have been briefly alluded to earlier in this paper. Developing valid and reliable metrics which can be used to predict the suitability of systems for classes of users is of course the core of the problem that we are addressing. It is our belief that to be successful in this endeavour requires the sort of ground clearing exploratory research that we have sketched above.

5. In the Longer Term

The modelling of both users and applications and the mappings between them as outlined above will result in structured descriptions which will form the skeleton of a general evaluation framework and which we intend to make available to the community at large to encourage input, and collaboration from all the stakeholders in the field: developers, users, acquirers, evaluators and all other interested parties. We believe that such a framework has implications not only for the evaluation of existing tools and resources but for guiding developers of new software and resources by clearly stating the needs of users and indicating which aspects of text mining solutions are important for particular classes of users

From our experiences in the ISLE project developing a framework for MT evaluation we have found that common structured descriptions support and encourage collaborative research. We hope in the future to implement the skeleton of the framework in a way similar to the web based tool created for MT evaluation. (Estrella et al 2005, Popescu-belis et al, this conference.)

References

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of HLT 2002*, San Diego, USA,. pp. 257-258.

Estrella, P., Popescu-Belis A.. & Underwood, N. (2005). Finding the system that suits you best: towards the normalization of MT evaluation. In *Proceedings of Translating and the Computer 27*. London:ASLIB, pp. 22-34.

Hearst, M. A. (1999) Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. Maryland: ACL, pp. 3-10.

Hovy, E. H., King M. & Popescu-Belis, A. (2002) Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, 17(1), pp. 1–33.

ISO/IEC (2001). *9126-1:Software engineering – Product quality – Part 1: Quality Model*. Geneva, International Organization for Standardization and International Electrotechnical Commission.

Pallet D. et al (1994) 1993 Benchmark Tests for the ARPA spoken Language Program. In *Proceedings of the ARPA Workshop onHuman Language Technology*, San Francisco: Morgan Kaufmann. pp. 51–73

Popescu-Belis, A Estrella, P. King M. & Underwood, N. (2006). A model for Context-Based Evaluation of Language Procssing systems and its Application to Machine Translation Evaluation. In Proceedings of the Fifth International Conference on Language Resources and Evaluation Genoa: ELRA (in press)

Papineni, K. Roukos, S. Ward T. & Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th ACL*, Philadelphia: ACL, pp. 311–318.

Sparck Jones, K. (1995) Reflections on TREC. *Information Processing and Management*, 31 (3), pp. 269 – 448.

Vilalta, R., Giraud-Carrier, C., Brazdil, P., & Soares, C. (2004). Using meta-learning to support data-mining. *International Journal of Computer Science Applications*, I, 3 pp. 1–45.