An Economic View on Human Language Technology Evaluation

Edouard Geoffrois

DGA / Direction de l'Expertise Technique / Centre d'Expertise Parisien
16 bis av Prieur de la Côte d'Or
94114 Arcueil cedex, France
Edouard.Geoffrois@etca.fr

Abstract

This paper analyses some general issues about human language technology evaluation, focusing on economic aspects. It first provides a scientific rationale for the need to organize evaluation in the form of campaigns, by relating this need to some basic characteristics of human language technologies, namely that they involve learning to process information in a way which reproduces human capabilities. It then reviews the benefits and constraints of these evaluation campaigns. Borrowing concepts from the field of economics, it also provides an analysis of the economic incentives to organize evaluation campaigns. It entails from this analysis that fitting evaluation campaigns to the needs of scientific research requires a strong implication in term of research policy and public funding.

1. Introduction

Evaluation campaigns were introduced in the field of speech and natural language processing more than twenty years ago (Moore, 1986; Pallett, 2003). In this framework, several research teams agree on common evaluation protocols, simultaneously submit the outputs of their systems for scoring according to these protocols, and gather for a debriefing workshop. This methodology is more and more widely accepted for evaluating various types of human language technologies, much experience has been gained over the years (Martin et al., 2004), and evaluation campaigns are organized by an increasing number of organizations in various countries. It also gradually extends to neighboring domains such as image processing.

However, in comparison to many other fields of science and technology, this organization of evaluation in the form of campaigns appears to be quite specific. One can therefore wonder why it is needed, and which fields are concerned. In other words, one can raise the following questions: Why is it that there is a need to organize campaigns for evaluating human language technology research and development, and why letting researchers publish results and share them in conferences would not be enough? What characteristics of the domain lead to this specific organization, and which other domains share the same characteristics? In practice, given the growing importance of evaluation campaigns in the field, such a need is generally taken as granted. However, an explicit and widely shared line of reasoning supporting it is still lacking.

Furthermore, despite their success, evaluation campaigns and more generally the evaluation methodology is sometimes a matter of debate (Förstner, 1996; special issue, 1996). These debates are all the more important since one issue at stake is how research directions are selected and funded. It is therefore useful to summarize why evaluation campaigns are beneficial, while remaining aware of their limitations or constraints.

Insofar as evaluation campaigns are needed and beneficial, a related question is who should organize and support them. At first sight, if their usefulness is acknowledged, one might hope that the "offer" in evaluation campaigns will naturally adjust to the "demand". However, in several parts of the world and in particular in Europe, despite the increasing number of initiatives to organize evaluation campaigns, the need for a strong and durable evaluation infrastructure such as the one established by the National Institute for Standards (NIST) and supported by the Defense Advanced Research Projects Agency (DARPA) in the US is still often expressed. It thus seems useful to analyse the incentives and impediments for organizing and supporting evaluation campaigns.

This article reviews the above issues, i.e., the scientific rationale for evaluation campaigns, their benefits and constraints, and the economic driving force behind them, after situating these campaigns among the various types of evaluations. It builds on the lessons learned from setting up and steering evaluation-oriented programs for both human language and image processing technologies (Technolangue[1] and Techno-Vision[1], Quaero[2]) and campaigns withing these programs (Technolangue/ESTER (Galliano et al., 2005), Techno-Vision/RIMES (Grosicki et al., 2006)), as well as on many discussions with actors involved in well established international campaigns such as those organized by NIST or CLEF (Cross-Language Evaluation Forum). It also relies on some basic concepts borrowed from the field of economics.

2. Types of evaluation for human language technologies

The term "evaluation" covers various types of activities. In the domain of human language technologies, it is customary to distinguish "technology evaluation" and "usability evaluation". However, this distinction is not specific to the field. Evaluation through experiments yielding quantitative and reproducible results, reported together with the experimental protocols in the publications, applies to any experimental science. Since the aim is to develop technology, this type of evaluation can also be called technology evaluation. Besides, evaluation of usability through usage stud-

[1] http://www.enseignementsup-recherche.gouv.fr/
[2] http://www.quaero.org

ies pertains to any applied science. Technology evaluation and usability evaluation are complementary. The former is measuring to what extent a given technology is appropriate for some application, whereas the latter measures how users react to a given technology. Neither one can replace the other, and both are needed for ensuring the success of a new technology.
More specific to human language technology is the organization of technology evaluation. Measurements can be performed in two ways: in the framework of evaluation campaigns, but also on an everyday basis in the research laboratories using development data sets or test sets from previous campaigns. Both are important, but the former requires a more complex organization. The present paper focuses on this aspect.

3. The need for evaluation campaigns

The technology being about data processing, experimental measurements are done by confronting ideas implemented in systems against data representative of the task under study. However, getting reproducible and unbiased measures raises several issues:

1. Knowledge is infinite: Since in practice test data can only be a sample of the potentially infinite and to some extent unpredictible set of data corresponding to the task, the evaluation results depends on the data set, and experiments cannot be precisely reproduced without the data set. For the sake of reproducibility, any published measurement should therefore be done on a publicly accessible data set.

2. The judge is human: Since the technology is about reproducing and automatizing human capabilities, at least partially in order to help the users, testing it requires manual work to judge automatic systems and in particular to define references against which system outputs can be compared. The evaluation tool thus implies human intervention and cannot be fully automatized.

3. Systems involve learning: Since learning is involved in system development, be it automatic learning or information gained by the developer, acquiring more information about the test data than one would have in a real setting before ending the development would introduce a bias. To put it simply, the measurement disturbs the next version of the system. Therefore, in order to garantee that no bias can be introduced, a protocol where the test data is not known in advance should be used.

These characteristics are key differentiators compared to most other scientific domains, for which

1. experimental measurements rely on measurement instruments which are precalibrated, and the results are supposed to not depend on the very instrument which is used,

2. benchmarks and units are defined by physical phenomena,

3. knowing the exact measurement protocol and measuring several times using the same tools does not introduce a bias.

The combination of two of these constraints, i.e., that the test data relies on human intervention but should not be known in advance, implies that an independent third party is necessary to garantee that there is no bias. This is best achieved when the organizer of the evaluation is from a different institute than the developers.
The combinations of two other of the constraints, i.e., that the test data should not be known in advance but should be publicly accessible after the measurement, implies that all measurements should be done simultaneously. This explains the need for a specific organization in synchronized evaluation campaigns.
These characteristics are shared by domains other than speech and langage processing. Similar domains include image processing, which has started to adopt the methodology of evaluation campaigns. Other related domains such as knowledge processing and artificial intelligence are also concerned. In all these domains, when the above characteristics are met, the same line of reasoning should lead to the conclusion that there is a need to organise evaluation campaigns.
Note that one could compare the organization of campaigns for evaluating automatic knowledge processing systems to the widespread organization used for evaluating “human knowledge processing systems”, i.e., the examinations for evaluating students. Indeed, students acquire knowledge in order to perfom certains information processing tasks for which the reference is set by teachers, and they are tested simultaneously in groups on specific problems, defined by the teachers and kept secret until the start of each examimation.
To summarize, the need to organise evaluation campaigns for making measurements comparable and unbiases stems from the fact that the technology to evaluate involves learning to process data in a way which reproduce an intelligent behaviour beyond analytic modeling.

4. Benefits and constraints of evaluation campaigns

Evaluation campaigns provide the infrastructure for objective evaluation, and as such they bring several benefits. Indeed, quantitative evaluation in general is commonly acknowledged to drive progress: It makes issues explicit, allows to validate new ideas, identify missing science, reproduce experiments and compare approaches. It avoids duplication of effort and allows to judge funding efficiency. It also provides the basis for determining the maturity of the developments for a given application, and ease technology transfer.
In addition, evaluation campaigns are also a specific organization where several research teams gather their efforts toward a common challenge. This is useful to increase visibility, organize a community, foster exchanges, create emulation and encourage innovation.
Evaluation campaigns can also have some limitations (Hirschman and Thompson, 1996) or imply some constraints which can be seen as negative aspects. For example,

the research teams have to synchronize their work, both in terms of calendar and in terms of tasks under study, which can be perceived as a lack of freedom. The very fact of defining objectives can sometimes also be felt as stifling creativity. However, this could be argued in any domain, since focusing on some research issues is always done at the expense of others. Nevertheless, in the case of human language technology the constraints are stronger because the choice of priorities is done by a community, not by each research team independently. More importantly, evaluation can become useless or even drive in a wrong direction if the measurement tools are biased, which can happen especially if not enough means are devoted to it.

5. Evaluation campaigns as public goods

Insofar as evaluation campaigns are needed and have a globally positive impact, one might expect that they would naturally adjust to the needs of research. In practice, however, this does not prove to be the case. Several factors can explain this situation.

A first impediment to the setting up of an evaluation infrastructure is that the research teams and their funding organizations might not be accustomed to evaluation campaigns and are reluctant to get involved or support them because of the implied constraints and apparent complexity. From an economic point of view, this is a classical situation of imperfect and asymmetric information. However, when a team gets involved once, it most often considers that the advantages clearly outweight the inconvenients. Lack of exposure is therefore is less and less of an impediment as the paradigm becomes well known.

A more structural issue is about why would an organization set up or fund an evaluation infrastructure. Indeed, as mentioned above, the evaluation tool cannot be fully automatize and therefore has a non negligible and recurrent cost. In addition, the evaluation tools, like the technologies they have to evaluate, are based on software. Once they are designed, the level of human intervention is limited, even though there are exception such as for the edit distance used in the GALE project (Przybocki et al., 2006). The measurement tools are thus costly to develop but not to duplicate. This implies that there is not only a scientific need but also a strong economic benefit in sharing them. Furthermore, when the goal of an evaluation infrastructure is to promote the advancement of science, every developer should be granted access to the evaluation infrastructure. Indeed, for a given task, in order to ensure that results are comparable, there can be only one reference evaluation campaign and not several smaller ones.

In economic terms, the above properties are called non-rivalry and non-excludability. A good is non-rival if consumption by one consumer does not prevent simultaneous consumption by other consumers. It is non-excludable if no one can be prevented from accessing and using the good. Non-rivalry and non-excludability are the two important characteristics of public goods. When both are fulfilled, the good is a pure public good. When only non-rivalry is fulfilled, it is called a club good. The more these properties are fulfilled, the less the law of offer and demand applies and naturally adjusts the level of activity to the needs (Jones, 2001). These different types of goods are represented in Table 1.

Table 1: Private and public goods with typical examples (after Wikipedia entry on public goods)

	Excludable	Non-excludable
Rivalrous	Private goods *food, clothing, cars*	Common goods *water, fish*
Non-rivalrous	Club goods *cable TV*	Public goods *free-to-air TV*

Using these economic terms, an evaluation infrastructure aimed at the promotion of the advancement of science is a pure public good and cannot exist without a strong support from public funding. An evaluation infrastructure aiming at supporting the technological progress of selected partners is a club good, and partial funding might be enough an incentive to have several developers gather and share an evaluation infrastructure, but not to have them share it more widely. On that respect, evaluation activities are closer to project management than usual research activities.

The above analysis thus creates a link between the adequacy of an evaluation infrastructure and financial instruments. Indeed, if the evaluation infratructure is not supported strongly enough, the result is a scattered effort mainly based on isolated initiatives and good will rather than a coordinated strategy. In practice, individual researchers sometimes wish to get involved in organizing evaluation campaigns, motivated by the fact that this a way to boost their domain. However, a limitation of this scheme is that once the evaluation protocol stabilizes, there is no new idea to publish and the activity stops, even though the community needs recurrent evaluation campaigns to achieve long-term progress.

To sum up, evaluation campaigns organized for supporting scientific research constitute a research infrastructure which has all characteristics of a public good, and thus requires a strong support from public funding and dedicated structures working for the benefit of others.

6. Conclusions

An analysis of some specific aspects of human langage technology evaluation was conducted with an emphasis on economic aspects. Two main conclusions can be drawn from this analysis.

First, the basic reasons for organizing evaluation campaigns are that the technology to evaluate is about information processing and involves learning to reproduce human capabilities. These properties are shared by several scientific fields. It is thus not surprizing that the methodology introduced in the domain of speech processing has gradually spread to natural language processing and image processing, and should continue to grow to encompass the whole domain of artificial intelligence.

Second, the evaluation infrastructure is an expensive investment that single developers are not enclined to pay for even partially, or, if they do, would not be enclined to share. This means that setting up an efficient evaluation infrastructure

suited to the needs of scientific research is a matter of public policy.

To put it in a nutshell, evaluation campaigns are needed and globally beneficial, but require a strong public implication in order to fit the needs of scientific research in the domain.

Acknowledgements

The author wishes to thank the many persons who contributed to the elaboration of this paper through numerous and fruitful exchanges, as well as the reviewers for their useful comments.

7. References

Wolfgang Förstner. 1996. 10 pros and cons against performance characterization of vision algorithms. In *Workshop on Performance Characteristics of Vision Algorithms*.

Sylvain Galliano, Edouard Geoffrois, Djamel Mostefa, Khalid Choukri, Jean-François Bonastre, and Guillaume Gravier. 2005. The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In *European Conference on Speech Communication and Technology (Interspeech-Eurospeech)*.

Emmanuèle Grosicki, Edouard Geoffrois, Matthieu Carré, Emmanuel Augustin, and Françoise Prêteux. 2006. La campagne d'évaluation RIMES pour la reconnaissance de courriers manuscrits. In *Colloque International Francophone sur l'Ecrit et le Document (CIFED)*.

Lynette Hirschman and Henry S. Thompson. 1996. Overview of evaluation in speech and natural language processing. In R. Cole et al., editor, *Survey of the State of the Art in Human Language Technology*. Cambridge University Press.

Charles I. Jones. 2001. *Introduction to Economic Growth*. W. W. Norton, 2nd edition.

Alvin F. Martin, John S. Garofolo, Jonathan C. Fiscus, Audrey N. Le, David S. Pallett, Mark A. Przybocki, and Gregory A. Sanders. 2004. NIST language technology evaluation cookbook. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*.

Roger K. Moore. 1986. The NATO research study group on speech processing: RSG10. In *Proc. Speech Tech 86 (Media Dimensions)*, pages 201–203, April.

David S. Pallett. 2003. A look at NIST's benchmark ASR tests: past, present, and future. In *IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 483–488.

Mark Przybocki, Gregory Sanders, and Audrey Le. 2006. Edit distance: A metric for machine translation evaluation. In *International Conference on Language Resources and Evaluation (LREC)*, pages 2038–2043.

Speech Communication special issue. 1996. Comments on "Towards increasing speech recognition error rates". *Speech Communication*, 18(3):232–255.