

Language Technology Evaluation: which Funding Strategy?

Edouard Geoffrois, DGA

Position paper
FLaReNet launching event
Vienna, February 12th, 2009

Introduction

Quantitative evaluation or measurement is at the heart of experimental sciences and is necessary to drive scientific and technological progress. In the case of Language Technologies (LT), objective evaluation requires a specific organization often referred to as evaluation campaigns. This organization has been in use for more than 20 years in the DARPA/NIST programs. It has proven to be attractive to European research teams participating in these programs and naturally started to be used in Europe. However, it has not developed to the same extent there: in practice the evaluation efforts are more limited and scattered, relying more on local if not individual initiative. As a side effect, evaluation also remains a subject of more debate. One can therefore step back and wonder what are the basic reasons to organize evaluation campaigns, what are the impediments to such an organization, and what can be done to overcome them.

The rationale for evaluation campaigns

When analysing the reasons for organizing evaluation campaign, one can distinguish two types of arguments corresponding to two questions: why evaluation is beneficial, and why should it be organized in the form of campaigns?

Arguments about the benefits of evaluation are generic ones: It allows researchers to objectively compare approaches and to reproduce experiments, and more generally to make issues explicit, to validate new ideas and to identify missing science; It is also an important tool to judge funding efficiency and to determine the maturity of the developments for a given application.

The reasons for organizing evaluation in the form of campaigns in the case of LT are more specific to the domain, and are manifold. First, the results must be measured using common test data and protocols in order to be comparable. Furthermore, since developing the technologies implies some learning, the test data should not be known in advance. But for the sake of reproducibility and scientific progress, this data should also be published after the test and discussed. As a further consequence of having some data unknown before the test but discussed after, the testing period should be common to all systems under measurement. All this implies a specific organization where all activities must be synchronized. Such an organization might be perceived as complex at first sight, but results from intrinsic properties of the domain and is necessary for a sound evaluation.

The lack of LT evaluation infrastructures

LT evaluation can be analyzed, in economic terms, as a market failure. This was further analyzed in a separate article [1]. The analysis can be summarized here as

follows. Since the technology is about tasks which are not yet automatized, providing the testing infrastructure necessarily implies an important cost in human expertise. This is almost entirely a fixed cost, while the marginal cost per system under measurement is negligible. In addition, the testing infrastructure should be easily accessible to new research teams without long-term advance planning. As a result, neither the participants to the evaluation nor its organizer have a direct interest in investing in an infrastructure which will be useful to others and which does not clearly generate a future business.

In that context, the partial funding grant model, which relies on the intrinsic interest of the beneficiaries to incite them to cover the rest of the costs, does not provide enough incentive effect. In fact, the economic actors which are the most interested in funding the evaluation infrastructure are the ones who seek to foster progress in the domain as a whole. In some specific cases where there is a niche market, one actor might be dominant enough to take the lead, but for more general purpose technologies with a wide range of applications and many actors, fostering the evaluation infrastructure is the role of the research funding agencies. However, to actually do this, the adequate instruments must be available.

How to adapt the infrastructure to the needs?

The main factor to get evaluation infrastructures adapted to the research needs is to grant it 100% public funding. There are at least two different, complementary ways to do this. One is to adapt the existing funding instruments to include evaluation activities as a special case, in a similar way to program management activities. Another one is to rely on dedicated public bodies which have LT evaluation as a part of their public missions and fund their additional costs through each specific program. In any case, it is a matter of funding strategy rather than a purely technical issue.

A secondary factor to get suitable evaluation infrastructures is to tightly connect them to the research they serve, and in particular to design all types of activities together when preparing a new research program. Additionally, since all the research activities on the same topic should share the same evaluation data and protocols, gathering similar research in single large programs can be expected to be more efficient than scattering it in different smaller ones.

Conclusion and perspectives

To summarize, objective LT evaluation is beneficial and should be organized in the form of evaluation campaigns, if possible embedded into large integrated programs. However, the traditional partial grants combined with a lack of dedicated public structures results in a shortage of evaluation infrastructures. Different funding strategies are required to ensure that these infrastructures are suited to the needs of research and development. It is therefore of critical importance to set up new funding strategies for LT evaluation in Europe to get the full benefits of the large investments in the domain.

References

- [1] Edouard Geoffrois. 2008. An Economic View on Human Language Technology Evaluation, in *Proc. International Language Resources and Evaluation Conference (LREC)*.