Setting up an Evaluation Infrastructure for Human Language Technologies in Europe

Kevin McTait & Khalid Choukri
ELDA
55 – 57 rue Brillat-Savarin
75013 Paris, France
{mctait,choukri}@elda.fr

Abstract

This paper describes ELRA/ELDA's vision of an evaluation infrastructure for Human Language Technologies in Europe. Drawing on its experience in national and Europe-wide evaluation projects and also its experience in the production, validation, packaging and distribution of language resources, such as electronic text corpora, lexica and speech databases, ELDA's evaluation department seeks to set up a European clearing house for evaluation related resources and software packages, in the same way that ELDA has become the European clearing house for language resources. ELDA's vision for a European evaluation infrastructure is inspired by both European and international evaluation initiatives, including the DARPA/NIST evaluation programme in the United States.

1 Introduction

In 1995, the European Language Resources Association (ELRA) was set up under the auspices of the European Commission as a non-profit making body with the aim of making language resources (LR) available to the language engineering community. Such resources are essential to both public research institutions and private companies wishing to construct, develop and test Human Language Technology (HLT) systems, such as speech recognisers, machine translation systems, terminology support tools etc. ELRA's operational body, the Evaluation and Language resources Distribution Agency (ELDA) was set up to act as the European clearing house for such LRs. ELDA is active in the specification, production, validation, packaging and distribution of LRs and also deals with the legal issues involved. Today, its catalogue contains several different types of LR, such as speech databases, electronic lexica and text corpora (monolingual, parallel multilingual, multimodal etc) in several different languages. ELDA's clients include not only large commercial organisations, but also public sector research laboratories and universities.

ELDA's evaluation department is active in evaluation projects on both the national (French) and European levels. For example, the *Technolangue* programme, funded by the French Ministries of Research, Industry and Culture, contains several projects dedicated to the advancement of HLT in France. One project under the *Technolangue* programme is entitled EVALDA, and is dedicated to setting up permanent and lasting evaluation protocols and packages for the major linguistic technologies, namely:

- Corpus Alignment
- Terminology
- Machine Translation
- Syntactic Parsers

- Q/A Systems
- Broadcast News Transcription Systems
- Speech Synthesis
- Dialogue Systems

In the context of the EVALDA project, players from academia, public sector research and the private sector are invited to take part in competitive and comparative evaluation campaigns, culminating in a workshop. A scientific committee was set up for each of the 8 linguistic technologies shown above, in order to discuss, define and agree on evaluation protocols including evaluation methodologies (whether automatic, assessed by human evaluators or both), metrics, evaluations tasks, resources and evaluation software.

In order that the evaluation campaigns are ethical and valid, an independent organisation with the necessary skills is required to oversee and manage the evaluation campaigns. In this case, ELDA is well placed to take on this role.

In addition to the EVALDA project, ELDA is involved in the TC-STAR_P project (Preparatory Action for the project Text and Corpora for Speech to Speech Translation). The purpose of this preparatory action is to write a proposal to the EU commission, under the 6th Framework, requesting funding for a 5 year project entitled TC-STAR.

In this project/proposal, ELDA undertakes all issues relating to LRs for Speech-to-Speech (SST) components i.e. speech recognition, speech centred translation and text-to-speech, and evaluation (of the SST system as a whole and the individual components). Therefore, ELDA undertakes the collection, specification, production and distribution not only of the LRs required by the research and development teams, but also undertakes to commission, produce and distribute resources for evaluation, including the software packages necessary.

ELDA is also involved in the CLEF project, a French national initiative whose purpose is to develop and maintain an infrastructure for the evaluation of cross-language information retrieval systems (CLIR). In this project, ELDA is responsible for data acquisition and negotiation of rights. With respect to information retrieval in French, ELDA was also involved in defining evaluation procedures in the French AMARYLLIS project.

Again, as an independent HLT organisation, ELDA is well placed to manage the evaluation campaigns. In fact, the networks for LRs and LR expertise set up by ELRA/ELDA prove to be an invaluable source of expertise in such projects.

2 Evaluation

2.1 Why Evaluate?

Evaluation forms a fundamental part of the development of language engineering products. It is essential for validating research hypotheses, for assessing progress and for choosing between research alternatives.

In more detail, it enables R&D teams to assess the impact of innovations on system performance. For example, does changing parameter *x* entail an increase in system performance validating the change?

Evaluation also identifies promising technology or research directions enabling industry to assess its market value. However, language engineering displays a paradoxical property in that in many areas, the state of the technology has reached a level barely sufficient to be usable in practice. Nevertheless, many commercial language-based applications do exist (e.g. machine translation, text summarisation, dictation, spoken dialogue systems). Comparative evaluation could help clear up the issues, where the advertised performance claims are difficult to assess and compare objectively.

Evaluation also allows funding agencies to determine whether their investment has led to significant progress. Many national, European or international projects require progress reports every *x* months. Therefore, the results of evaluation campaigns enable the progress of the project to be tracked. It also gives funding agencies the data necessary to quantitatively evaluate the progress made possible by their investment, and thus suggests priorities on where to plan research efforts and support for application development. Evaluation campaigns also provides useful input when deciding whether a technology is mature enough to be considered as a candidate for starting commercial application development.

A further side effect of evaluation campaigns is the production of high quality evaluation resources, in the form of training and test data

along with evaluation software packages, distributed or produced during evaluation campaigns. Also, the availability of evaluation packages enables *all* researchers in a particular field to evaluate, benchmark and compare the performance of their systems.

2.2 Evaluation in the US and Europe

In the USA, the DARPA government funding agency is active in the evaluation of the principal areas of HLT: speech dictation, spoken language understanding, broadcast news transcription, named entities extraction, topic detection and tracking, text retrieval, message understanding, machine translation, speaker verification, character recognition, etc. It organises competitive evaluation campaigns and publishes the results in a workshop. The tasks within the different language technologies have been made more and more difficult, in agreement with the improvement in the various technologies over time. In order to have the necessary logistics for such evaluations, two entities play a major role in this framework: NIST, the National Institute for Standards and Technology, and the LDC, the Linguistic Data Consortium, which was created for the purpose of distributing language resources.

It would appear that the US-based evaluation programmes follow a top-down strategy i.e. the US government strongly influences the campaigns, but provides abundant funding and a long-lasting infrastructure. In Europe, the strategy has been rather more bottom-up, starting from individual research groups and HLT systems.

The US campaigns have inspired efforts at creating a lasting and permanent evaluation infrastructure in Europe. However, the picture in Europe is more fragmented for several reasons. First, there have been much less resources devoted to evaluation and secondly, evaluation efforts have come from many different sources, the result of which is that there is no equivalent European evaluation infrastructure. However, there have been several initiatives, either at the EU level (CLEF, SQALE, TSNLP, the proposed EAGLES evaluation methodology, ETSI/Aurora, DiET, DISC, TEMAA, and SPARKLE etc.), or on a national level (Grace, Aupelf ARC, in France, Verbmobil and the Morpholympics in Germany and SENSEVAL/ROMANSEVAL co-sponsored by several EU-projects, ELSNET, ELRA and the British government). But all these initiatives were funded within limited duration projects, and there is no permanent entity designed to organise evaluation campaigns and capitalise on the resources and packages created during these independent initiatives. Therefore, the result is that European research teams are obliged to evaluate their technologies in US evaluation campaigns, using US evaluation packages which are subject to the geo-political incentives of the US research funding bodies.

However, inspired by the DARPA evaluation framework, the EU funded ELSE project was set up in order to draw up an executive summary for a general infrastructure for the evaluation of HLT systems in Europe. The idea was to focus on comparative as well as competitive evaluation techniques, taking into account the special situation in Europe i.e. multiple languages, a union of nations, industrial and commercial relevance, general EU programme policy etc.

From the analysis conducted within ELSE, comparative technology evaluation (in conjunction with DARPA style competitive evaluation) brings many interesting features. It forces researchers and technology developers to go deeper in their research field when they try to figure out how to measure the performance of a system for a given task. It gives technology developers objective information in order to make choices in system development. It gives industry the possibility of comparing their technology with others by participating in evaluation campaigns, or by acquiring the test data and comparing their systems performance with what has been achieved and reported so far. In particular, it provides SMEs with an efficient and easy market watch.

ELSE has provided recommendations for setting up such an evaluation infrastructure in Europe. It has identified the advantages of using the comparative evaluation paradigm and has listed several language technologies which could immediately make use of the evaluation infrastructure based on their relevance for research and industry.

The ELSE project report proposed two possible schemes for implementing this evaluation infrastructure. The first is a proactive approach,

responding to the needs of individual research groups or technologies and the second being reactive, responding to FP calls for proposals. The report recommends that both be followed in parallel, in a bootstrapping manner.

Finally, the ELSE project investigated the relationship between technology evaluation and usage evaluations, requiring best practice guidelines and handbooks. Also recommended is basic research in evaluation to be considered over a longer timescale in order to constantly improve knowledge about evaluation metrology.

3 A European Evaluation Infrastructure

ELDA has a proven track record in the efficient and cost-effective distribution of LRs on both a European and worldwide level. It has set up an *organisational model* for LR networks dedicated to the specification, commissioning, production, validation, packaging and distribution of LRs with the legal issues resolved.

Along with its experience in national and European evaluation projects, ELDA's evaluation department capitalises on this experience to create an *organisational model* for efficient and cost-effective evaluation management. This entails the creation of a European, even international, network or infrastructure of evaluation centres providing evaluation resources, software packages, technology, forums of scientific expertise and R&D centres for the independent, ethical evaluation of human language technologies.

A starting point for an evaluation infrastructure in Europe is the European ELSE project, whose aim as to draw up a blueprint for a comparative, as opposed to competitive, evaluation infrastructure for the major linguistic technologies. ELSE provides recommendations for the establishment of such an infrastructure. ELDA's vision for a European infrastructure is also inspired by the evaluation activities organised by the DARPA/NIST institutions in the US.

The European infrastructure, as the ELSE project, would be organised along two major principles, proactive and reactive evaluation schemes. ELDA's evaluation department is currently taking part in reactive evaluation in that it has answered calls for proposals for national and European projects, such as EVALDA, TC-STAR_P, CLEF and AMARYLLIS to be involved in the specification and production of evaluation resources, packages and protocols. An exit strategy is defined for each project where the evaluation resources, packages, software and knowledge (final project reports) produced in each evaluation campaign for each linguistic technology is made available to external players through ELDA's catalogue for a modest price. ELDA is well placed to carry out this mission due to its significant experience in the specification, production, packing and distribution of LRs – a related task.

In parallel, the evaluation infrastructure would be proactive. ELDA endeavours to make available evaluation resources and packages for all linguistic technologies in as many languages as possible. At the very least, a European evaluation infrastructure would have to make available evaluation resources and packages for the official EU languages.

A European initiative is required due to the international and multilingual nature of linguistic technologies. All major developers work on several languages even if they do not create truly multilingual systems. Furthermore, the major players operate on an international level. International cooperation has also been the key to the success of many projects or systems. Therefore, porting linguistic technologies across more and more language barriers leads to a greater need for a multilingual evaluation framework.

Finally, many European language markets are too small to sustain their own evaluation programmes. For example, a language with relatively few speakers i.e. Dutch or Danish, can only rely on European cooperation to organise the evaluation campaign that they need. With the arrival of the new member states in 2004, ELDA faces the challenge of providing evaluation resources and packages for these new languages and therefore seeks cooperation with the new national agencies, research centres and private concerns to make available, commission and produce language and evaluation resources in the new languages.

It would not have to stop there. ELDA's long term goal in this respect is to cover as many world languages and human language technologies as possible, therefore creating an international evaluation infrastructure, dealing not only

with European languages, but languages such as Chinese, Japanese etc.

In either case, the evaluation packages, in the form of training data, test data, test suites, evaluation protocols, software packages, toolkits, agreed methodologies, metrics and even *savoir-faire*, created through evaluation campaigns or by commissioning in a proactive manner, would be made available to the wider research community via ELDA's catalogue in the same way that ELDA makes LRs available. In this way, ELDA can take on the role of European clearing house or centre for evaluation technology, resources and expertise.

It is envisaged that a research or development team wishing to evaluation their system, whether for assessing development progress, focussing research efforts, providing feedback to a funding body or higher management etc., would contact ELDA's evaluation department and be supplied with the relevant evaluation packages. In the case of open source software or resources, the packages could simply be downloaded from ELDA's website.

Using its experience in the legal aspects of LR distribution, the legal issues pertaining to evaluation resources and packages would also be resolved by ELDA.

As the centre of a European evaluation infrastructure, ELDA would also become the forum or focus of knowledge on evaluation issues and evaluation metrology. In the course of evaluation campaigns and the commissioning of evaluation packages, ELDA will have acquired a good deal of expertise in evaluation over the entire range of linguistic technologies. In so doing, ELDA would become a centre of knowledge on evaluation in HLT and would be well placed to disseminate this knowledge.

In commissioning evaluation packages, agreement will have to be reached, in conjunction with other research groups, on evaluation protocols, methodologies and measures (as was the goal of the EAGLES project). Therefore, ELDA seeks to standardise evaluation protocols and make these standards available, along with the scientific justification behind it. Furthermore, using its expertise in evaluation, ELDA seeks to advance basic research in the subject of evaluation. In so doing, ELDA would be advancing the field of field of metrology in language engineering evaluation.

References

Adda, Gilles, Josette Lecomte, Joseph Mariani, P. Paroubek, M. Rajman. 1998. *The GRACE French Part-of-Speech Tagging Evaluation Task* in Proceedings of the First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998.

Harman, Donna. 1998. *The Text REtrieval Conference (TRECs) and the Cross- Language Track*, in Proceedings of the First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998

Kilgarriff, Adam. 1998. *SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs*, in Proceedings of the First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998

Mariani, Joseph. 1998. *The Aupelf-Uref Evaluation-Based Language Engineering Actions and Related Projects*, in Proceedings of the First International Conference on Language Resources and Evaluation (LREC), Granada, May 1998

Peters, Carol, Martin Braschler, Julio Gonzalo and Michael Kluck (Eds). 2001. *Evaluation of Cross-Language Information Retrieval Systems*. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 2001 (Revised Papers).

Walker, M., D. Litman, C. Kamm, A. Abella. 1997. *PARADISE: A Framework for Evaluating Spoken Dialogue Agents*, in Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics, ACL 97, 1997

Young, S.J., M. Adda-Decker, X. Aubert, C. Dugast, J.L. Gauvain, D.J. Kershaw, L. Lamel, D.A. Leeuwen, D. Pye, H.J.M. Steeneken A.J. Robinson, and P.C. Woodland. 1997. *Multilingual large vocabulary speech recognition: the european SQALE project.* Computer Speech and Language, 11(1):73-89.

http://www.limsi.fr/TLP/ELSE/

http://www.nist.gov/

http://www.ilc.pi.cnr.it/EAGLES/home.html

http://www.itri.brighton.ac.uk/events/senseval/