

# Quality Gates: a New Device for Development and Evaluation in Cross-Language Information Retrieval ?

René Schneider

University of Hildesheim  
Information Science  
Marienburger Platz 22- D-31141 Hildesheim, Germany  
rschneid @ uni-hildesheim.de

## Abstract

This paper outlines and discusses the perspectives quality gates offer in cross-lingual information retrieval to ensure that the development process benefits from evaluation. Adequate evaluation in this context is possible through a combination of modular quality gates at the inch-pebble level, their linear connection in networks and re-organisation during different development cycles. As a consequence, the strict separation between development and evaluation disappears.

## 1 Introduction

Cross-lingual information retrieval (CLIR) has attracted increasing interest during the last decade not only through national and international evaluation initiatives such as TREC<sup>1</sup>, CLEF<sup>2</sup> and NTCIR<sup>3</sup> but also through the rising amount of multilinguality in the world wide web and a stronger user interest on information from non-english and non-indoeuropean languages. Simultaneously, evaluation techniques have become “a must” for any research in natural language processing and development of applications in human language technology.

However, the situation is by no means satisfying. Evaluation in CLIR has only partly gone beyond the scope of precision, recall and f-measure or remains too modular for such complex systems where a multitude of resources, processes and methods merge in applications that differ considerably from each other. These differences emerge not only from the fundamental concepts resulting in different system architectures but also from the changes and adaptations of a single system during the different development cycles, keeping in mind that the “evaluation method is intimately connected to the software life cycle of the emerging technologies.” (Hirschman & Mani, 2003, p. 415)

This paper discusses the perspectives that quality gates offer for the area of cross-lingual information retrieval. After a brief overview of the different approaches used in evaluation in general and their application to CLIR, we will give an introduction to the concept of quality gates before outlining their concrete installation within the context of CLIR.

## 2 Evaluation in Cross-Lingual Information Retrieval

Evaluation in cross-lingual information retrieval offers a wide field of application due to the complexity and the large number of factors that have an implication for the retrieval results. For the sake of illustration we will name a few following the distinction of (Hirschman & Mani,

2003) and show their potential as well as their limitations in CLIR.

- *Gold standard based measures*: This method of defining, training and evaluating a gold standard has proved to be of great benefit in different fields, esp. for the task of named entity recognition. As recent work shows, high precision in named entity recognition has an equivalent impact on the overall quality of retrieval results. Nevertheless, the distribution, relevance and meaning of named entities varies considerably from genre to genre so that the definition of a generic gold standard is far from being found so far.
- *Feature based metrics* consist of checklists that record “critical features for different functional properties of components to be evaluated” (Hirschman & Mani, 2003, p. 418). Although they are usually set in opposition to corpus based methods, the gathering, annotation and evaluation of language corpora or any other data collection can easily be supported by this means to ensure representativeness and quality in test sets. As will be seen later (see section 3), feature based metrics are a basic element and starting point for the settlement of quality gates.
- *Embedded component evaluation* plays a vital role in the area of information retrieval. With the help of this evaluation device it is possible to track different scenarios of a system and to compare their corresponding results. Furthermore this strategy is strongly connected with the concept of user relevance feedback to either adapt a system according to the specific needs of a user or to measure general acceptance and usability respectively. Thus embedding component evaluation may collect experiences from the life cycle partly as a reaction to first results from evaluation metrics (such as the difference between the systems output and a gold standard) or from the interaction progress between the user and a system.

---

<sup>1</sup> trec.nist.gov

<sup>2</sup> www.clef-campaign.org

<sup>3</sup> www.ntcir.org

- Due to the fact that machine translation (MT) is a “*conditio sine qua non*” in CLIR the varying strategies that have been developed for *output evaluation* are of great importance for any evaluation of a cross-lingual retrieval engine. Translation occurs at different points of the system with either the queries and/or the retrieval results (or parts of them) being translated. Different translation strategies may be appropriate at different points of time and for different input.
- As already mentioned earlier, *user relevance feedback* is a very powerful means to discover the interactivity and usability a retrieval system has – and in case of being tracked appropriately – it will affect any further development of a retrieval system considerably. As a result it offers the combination with machine learning strategies to enable fast adaptation to the specific needs of a user. Unfortunately their implementation remains difficult so far.

As the enumeration listed above shows, CLIR offers a large number of interfaces towards evaluation, whose number and complexity changes with any further language that is integrated into an existing system. In some cases a method used so far may remain useful for a language pair (such as n-gram based methods that have proved successful for cross lingual retrieval of indo-european languages), but sometimes this method will lose its power and new components will have to be applied (e.g. in hamito-semitic languages with root-inflection n-gram based feature extraction becomes less valid). This will necessarily have an impact on the evaluation method used so far.

As a consequence the need for a framework arises that captures the dynamic complexity of CLIR in a synergetic system without being too complex but rather basic and feasible itself. The following section will propose a solution to this specific problem.

### 3 The Concept of Quality Gates

Quality gates had their origin in car manufacturing before being used metaphorically in quality assurance and project management. Generally, a Quality Gate (QG) is a checkpoint consisting of a set of predefined quality criteria that a project must meet in order to proceed from one stage of its life cycle to the next. Quality gates thus serve as amendments to milestones and deliverables which meet predefined quality benchmarks to

- support planning,
- improve status visibility,
- measure the current project status and
- control necessary changes or improvements.

Each quality gate is characterized by its own entry and exit criteria. A typical entry criteria is the completion and baseline of deliverables while an exit criteria can be the removal of the identified defects. By including metrics at

every stage of the development process projects are monitored against their stated goals. Another important feature of quality gates is that they can be installed at any point during the life cycle of a project. The appropriate linking and enlargement of their simple structure allows project planning, control and measurement, whereas three different levels of complexity may be differentiated:

- *Binarity*: A very simple but extremely powerful realisation consists in “binary quality gates at the inch-pebble level” (Suzuki, 2003), where “binary” refers to meeting a requirement with no partial credit being given in order to avoid any variance between the planned and the actual performance and the “inch-pebble level” refers to a detailed tasks of short duration to prevent long term periods without control. As mentioned earlier, binary quality gates can be compared with feature based metrics.
- *Interconnectedness*: Growing complexity of a system leads to a connection of sequential or parallel quality gates resulting in a network with semaphores at the intersections to direct further activities and to highlighten the status of a system depending on the fulfilment or missing of a task. This idea corresponds strongly to that of “embedded components” as described earlier, whereas interconnections enable activation of a component or vice versa. Different results will be compared and transposed into appropriate conditionals for further use in equal settings.
- *Recursion*: Finally, since every project is far away from being terminated with a first yield, quality gates show a big part of their potential in keeping record of the whole project life cycle to prevent the repetition of failures. Monitoring of “lessons learned” from previous development cycles enables adaptive and re-active control mechanisms for succeeding activities, e.g. follow-ups, re-implementation or up-dates of a system. These gates – located at the end of a test suite or a life cycle – are used for output evaluation and become input for any refinement to occur.

As can be seen from this short introduction, quality gates have a local aspect (i.e. their distribution and interconnectedness over a system) as well as a temporal aspect (modification in form and content over time). Nevertheless, they are characterized through formal simplicity consisting in binary features in combination with semaphore logic. Thus they can easily be visualized to serve developers, project managers and users for the creation of different plug-and-play settings, the design of different test suites or the creation of user-specific search engines.

## 4 Integration and Transparency in CLIR

### 4.1 Lessons Learned from Evaluation Campaigns

Similar to complex manufacturing processes or product development the release version of a retrieval system consists of many separate components, which may be developed at different times and are based on concurrent, sequential or recursive applications of several development patterns. This is esp. true in cross-lingual information retrieval, where the number of critical parameters and test suites is multiplied by the number of different languages a system is designed for and the implications that these languages have for retrieval strategies.

Thus the successful implementation of a retrieval system and the corresponding participation in an evaluation initiative (such as CLEF, TREC or NTCIR) depends considerably on a large number of quality criteria. Quality gates ensure that the project deliverables meet the criteria necessary to carry out subsequent project activities. In this context it should be noted explicitly, that quality gates are not considered for evaluation and comparison of several participants during an evaluation campaign, but have to be installed before and after the participation in an evaluation campaign.

The results that a system generates during participation will lead to many requests for changes: by developers as they realize something can be done in a different way, by comparison to strategies that other participants applied, etc. The collection and validation of these experiences will be discussed and transferred into appropriate alternations of the system. Sometimes these changes are small and a decision can easily be made whether to implement the change: but the changes to specification should be noted. Some requests may be kept open depending on the projects progress against timetable and/or some will be deferred as taking too long to implement. All of these specifications should be kept appropriately and probably be converted into quality criteria for further development circles.

The remaining question is then: How can we transfer our different experiences to objective quality criteria, that improve the development, testing and deployment of retrieval systems and avoid making the same mistakes again?

### 4.2 First steps

Our vision is that of using quality gates as a concrete method not only for project management but also for development and implementation, i.e. that – after a first period of intellectual and manual specification - of using their potential for effective planning, control and measurement in CLIR. After the definition of desired quality criteria, different components of a retrieval system will be connected via a network of coupled quality gates to control system parameters, to steer information flow and to document learning effects. To illustrate this vision, we dedicate the following paragraph to a first outline of quality gates in CLIR according to the differentiation in

section 3. Due to space limitations we will restrict ourselves to three examples already mentioned, namely data collections, fusion of strategies and user relevance feedback.

- CLIR is a heavily data-oriented approach. Consequently, results in CLIR depend to a big part on the data collections used for development and testing. Therefore, the quality of the corpora used for test suites and system development have to be described in terms of binary criteria concerning quantity, heterogeneity, data format, conversion, compression etc. A growing number of fulfilled criteria reflects growing validity of retrieval results and maturity of the system. Concrete realisation of this task might be achieved through simple templates that report on the adequacy of the data collection.
- Secondly, the overall system has to have knowledge concerning the components being used and coupled. To attain this, system components will be linked via gates that have information about the use and purpose of the specific components within a given context (e.g. the languages were n-gram based feature extraction has proved of great benefit) and allow steering and retaining of the information flow. Information flow and fusion of strategies might be controlled via conditionals and their correspondence to semaphores.
- Concurrent, sequential or recursive application of different system components will necessarily lead to different results. Combined with user relevance feedback (where users e.g. show their satisfaction by clicking on respective buttons to label documents as relevant or irrelevant) this information will serve as input to the whole system and has to be stored adequately and will lead to a reorganisation of the system and a change of the criteria and connections of lower-level quality gates.

## 5 Conclusion

The paper presents some reflections on the integration of quality gates into the process of developing and evaluation in cross-lingual information retrieval. This methodology is certainly not limited to this area, but promises to be helpful: on the one hand due to the high complexity of CLIR, on the other hand due to the fact that those systems – in the context of evaluation - initiatives have to be redesigned at least in a yearly interval.

While many of the techniques described can be found in the literature concerning evaluation, the ambition of the concept here is to bundle experiences and methodologies within a single framework (based on the metaphor of quality gates) to ensure adaptive and re-active project management.

## References

- Braschler, M., Harman, D., Hess, M., Kluck, M., Peters, C. & Schäuble, P.(2000). The Evaluation of Systems for Cross-Language Information Retrieval. In Proceedings of the Second Conference on Language Ressources and Evaluation (LREC-2000).
- Carbonell, J., Yang, Y., Frederking, R., Brown, R. D., Geng, Y. & Lee, D. (1997). Translingual information retrieval: A comparative evaluation. In Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence.
- Charvat, J. How to use quality gates to guide IT projects. ZDNet Australia. Builder. <http://web.zdnet.com.au/builder/manage/project/story/0,2000035082,20271712,00>.
- EAGLES. 1996. *EAGLES: evaluation of natural language processing systems*. Final Report EAGLES Document EAG-EWG-PR.2. <http://issco.www.unige.ch/projects/ewg96/ewg96.html>
- Hirschman, L. et al. (1997). Evaluation. In Ron Cole et al. (Eds.), *Survey of the State of the Art in Human Language Technology*, Cambridge University Press and Giardini.
- Hirschman, L. & Mani, I. (2003). Evaluation. In Mitkov, R. (Ed.), *The Oxford Handbook of Computational Linguistics*. Oxford University Press, pp. 414-429.
- Kando, N. (2000). NTCIR-Workshop: an Evaluation of Cross-Lingual Information Retrieval. In Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, Aug. 30-Sept.1, 1999, Tokyo.
- Quality Gates (2003). [http://www.compulink.co.uk/~querrid/STANDARD/quality\\_gates.htm](http://www.compulink.co.uk/~querrid/STANDARD/quality_gates.htm)
- Saracevic, T. (1995). Evaluation of Evaluation in Information Retrieval. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA, July 9-13, 1995, pp. 138-146.
- Suzuki, J. (2003). Best practices. Software Consulting. <http://members.cox.net/johnsuzuki/best.htm>.