Evaluating Interactive System Adaptation

Edouard Geoffrois
Agence Nationale de la Recherche (ANR)*
50 avenue Daumesnil, 75012 Paris, France
edouard.geoffrois@anr.fr

Abstract

Enabling users of intelligent systems to enhance the system performance by providing feedback on their errors is an important need. However, the ability of systems to learn from user feedback is difficult to evaluate in an objective and comparative way. Indeed, the involvement of real users in the adaptation process is an impediment to objective evaluation. This issue can be solved by using an oracle approach, where users are simulated by oracles having access to the reference test data. Another difficulty is to find a meaningful metric despite the fact that system improvements depend on the feedback provided and on the system itself. A solution is to measure the minimal amount of information needed to correct all system errors. It can be shown that for any well defined non interactive task, the interactively supervised version of the task can be evaluated by combining such an oracle-based approach and a minimum supervision rate metric. This new evaluation protocol for adaptive systems is not only expected to drive progress for such systems, but also to pave the way for a specialisation of actors along the value chain of their technological development.

Keywords: Evaluation methodologies; System adaptation; Interactive systems

1. Introduction

Interactive system adaptation, i.e., the ability of a system to learn from user feedback, is an important need for many applications. For example, in the domain of automatic speech recognition, a simple need is to enable users to provide new words in order to have the system recognize them. This need has been identified long ago (Asadi et al., 1991) and is still a topic of active research (Orosanu and Jouvet, 2015). Another example is computer-assisted translation, where the user edits an automatic translation (Isabelle and Church, 1997) and can expect to produce a correct translation more efficiently if the system learns from this feedback. Still another one is interactive information retrieval, where the user provides feedback on the relevance of retrieved documents to improve the overall search results, be it for textual (Salton and Buckley, 1990) or multimedia documents (Nguyen and Worring, 2008).

However, in general, evaluating the ability of systems to learn from user feedback in an objective and comparative way is difficult. Indeed, since human users are in the loop of the evaluation, such systems are often evaluated in a subjective way, which is expensive and not exactly reproducible. Another difficulty to compare approaches is that the improvements in performance obtained through the adaptation depend on the amount of information available for this adaptation, and in an iterative setting this amount is not under the control of the evaluator but of the user.

In practice, an objective evaluation protocol has been introduced only in some specific cases, such as interactive-predictive machine translation, which is a special case of computer-assisted translation where the user edits the automatic translation from left to right (Barrachina et al., 2009; Nepveu et al., 2004; Ortiz-Martínez et al., 2010) or interactive similarity-based retrieval (Nguyen and Worring, 2008). For most applications, for example computer-assisted translation in general and adaptive speech recognition, no evaluation protocol is available to date.

Nevertheless, interactive adaptation seems relatively intuitive to judge by humans and therefore not ill-defined, and experience in the domains of automatic speech recognition and machine translation shows that the objective evaluation protocols can be designed even after initially seeming out of reach (Pallett, 1985; Papineni et al., 2002). Furthermore, as also shown by experience in these domains, designing protocols to evaluate the ability of systems to learn from user feedback can be expected to be beneficial to the development of such systems.

The present article shows that this is possible by proposing such an evaluation methodology. After formalizing the problem, it describes solutions to overcome the above-mentioned issues and evaluate interactive system adaptation in an objective and reproducible way, before concluding and providing some perspectives.

2. Problem Formalization

The goal is to evaluate the ability of a system to learn from feedback provided by the user about its outputs. This interactive supervision setting in which the adaptation takes place is illustrated in Figure 1. In this setting, user feedback is used to update the system which then processes the input data again to produce an updated output, until the user stops providing feedback. The goal for the user is to get the best possible final output with the least effort.

Interactive system adaptation can be related to similar tasks. For example, interactive system adaptation is a special case of system adaptation where the feedback is about the system output, while in the general case of non interactive system adaptation the additional data is about the new environ-

*The author is currently on loan from the French national defense procurement agency (Direction générale de l'armement, DGA) to the French national research funding agency (ANR). The scheme described in this article was elaborated while the author was at DGA.

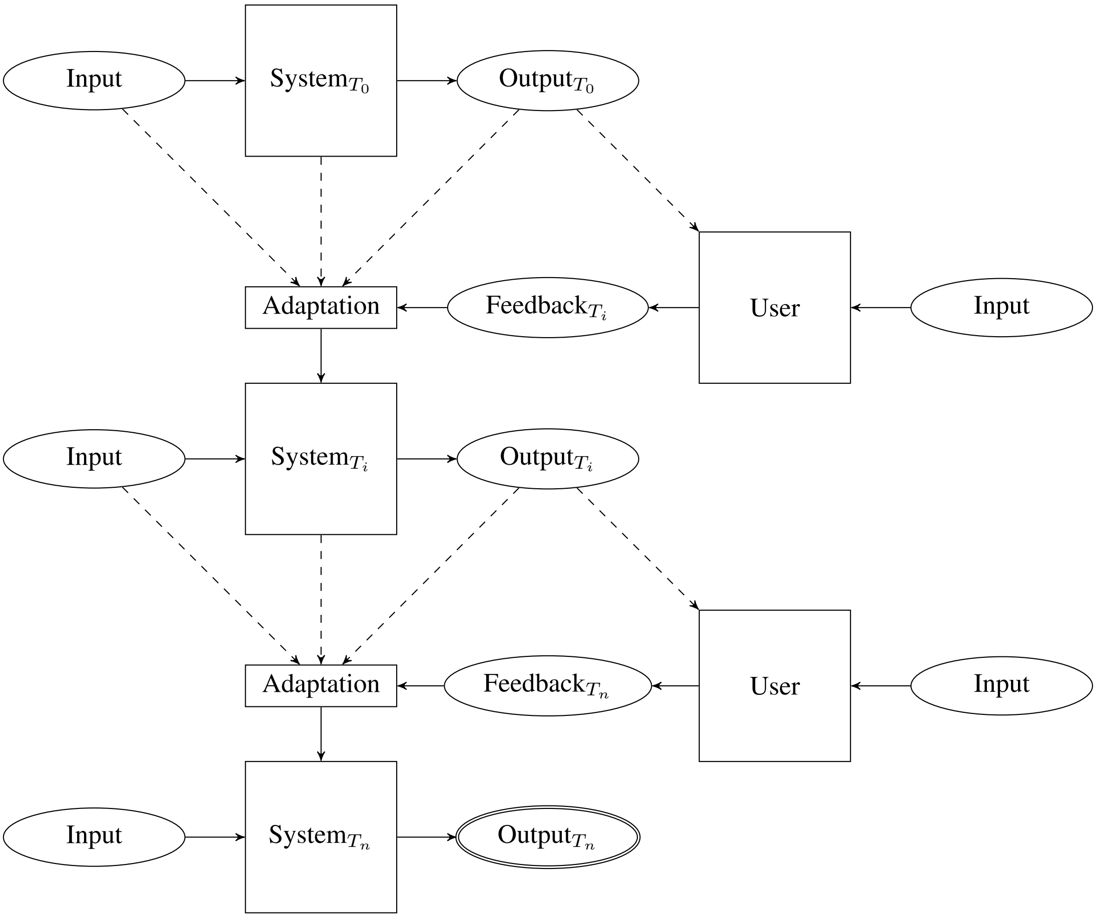

Figure 1: Illustration of the process of interactive system adaptation. Three systems states are represented, separated by dashed arrows: The initial state (at time T_0), an intermediate state (at time T_i), and the final state (time T_n). At each iteration, the user provides feedback to the system about its output, and this feedback is used by the adaptation module to update the system, which can process the input data again to produce an updated output. The process terminates when either the user or the system decides so, thus yielding a final output (shown by a double ellipse).

ment to which the system has to adapt.

Interactive system adaptation can also be compared to interactive systems in general. In a typical interactive system such as a dialogue system, the user feedback is a new input to the system, which is directly processed by the system while taking into account the context of previous interactions. In the case of interactive system adaptation, the user feedback is specifically about the errors of the system on the initial input, possibly constrained in its format, and it is used to improve the overall system performance. In other words, it is more the adaptation than the system which is interactive.

3. Problem Solution

As mentioned in the introduction, the problem is two-fold: The dependence of the improvement on the feedback received at each iteration and the involvement of a human in the loop. The main ideas to solve these issues are to measure the total amount of information needed to correct all system errors and to simulate the user by an oracle having access to the reference data. In the second case, this in turn raises futher issues, for which solutions can also be found. These ideas and solutions are developed below.

3.1. Relating Feedback and Error Corrections

As mentioned above, the goal of system adaptation is to maximize the system performance while minimizing the amount of feedback information needed to get the improved performance. However, since the feedback provided, and in particular the amount of information therein, is under the control of the user, if the task under study does not naturally lend itself to have this amount of information also controlled by the evaluator, no meaningful comparison between systems can be drawn.

A generic solution is to relate feedback to an error correction process and consider the residual errors after the last iteration as a last batch of correction to get an error-free output, i.e., to measure the overall amount of error correction needed to get a correct output, as illustrated in Figure 2. This measure integrates all error corrections into one figure of merit and yields a simple, scalar metric. The proposed name for this integrated metric is *Minimal Supervision Rate (MSR)*. Note that the details of the metric depend on the task under study.

If the metric for the basic, non interactive version of the task is in the form of an error rate, the MSR can be easily implemented by constraining the feedbacks to consist only of

error corrections. For most real applications, the user might want to also provide other types of feedbacks in addition to those of the basic task. Such other types of feedbacks can be allowed insofar as the costs associated to them are consistent with those of the error corrections of the basic task.
To give an example, in the case of automatic speech transcription, the standard metric, the Word Error Rate (WER), is composed of three types of errors: substitutions, insertions and deletions. A simple metric for measuring the performance of an adaptive transcription system is thus to measure the number of manual correction of such errors needed to get a correct output. But the user might also want to just give a word that is apparently not known to the system without giving any position in time. Such a feedback, which needs less effort from the user, should then be allowed but can be attributed a cost lower than 1.

Another approach, which can be especially useful if the basic task metric is not in the form of an error rate, consists in measuring more elementary user actions such as key strokes or mouse clicks, as is done in the specific case of interactive-predictive machine translation (Barrachina et al., 2009).

In all cases, the metric can be viewed as the minimal effort needed from the user to correct all system errors through successive interactions with the system. In this context, a good adaptive system is one which is able to use the user feedback to infer others corrections, i.e., to correct more errors that only those provided. If the basic task metric is in the form of an error rate, then a good adaptive system is one for which the MSR is lower than its initial error rate. On the graph shown in Figure 2, this means that the average slope of the evolution of error rates is steeper that -1.

3.2. Oracle-Based User Simulation

A second issue to tackle is the presence of a human user in the loop of the adaptation process, which prevents an easy reproducibility of experiments. The main idea to address this issue is to use an oracle approach, where the oracle is a computer system simulating the user by having access to the reference outputs. The oracle can thus automatically determine the errors of the system. This idea can be related to the more general one of simulating users, which is already widely used in the development of interactive systems (Azzopardi et al., 2011), but taking avantage of the specific form of the feedback to use the reference outputs in the user simulation.

While having access to the reference outputs enables the oracle to compute the error of the system in a deterministic way, in the general case, determining which feedback to provide in order to get an efficient adaptation still involves some choices. Different oracles can therefore have different behaviors, thus making the measurement dependent on the oracle. While this can in principle be acceptable if the oracle implements a representative user model, this effect should be minimized. Additionally, some oracles might have a limited ability to provide relevant feedback, which might make the results less representative of a real usage. Furthermore, if the oracle is designed using knowledge of the internal functioning of the system with which it interacts, the evaluation can be biased. However, these issues can be solved, as described in the remainder of this section.

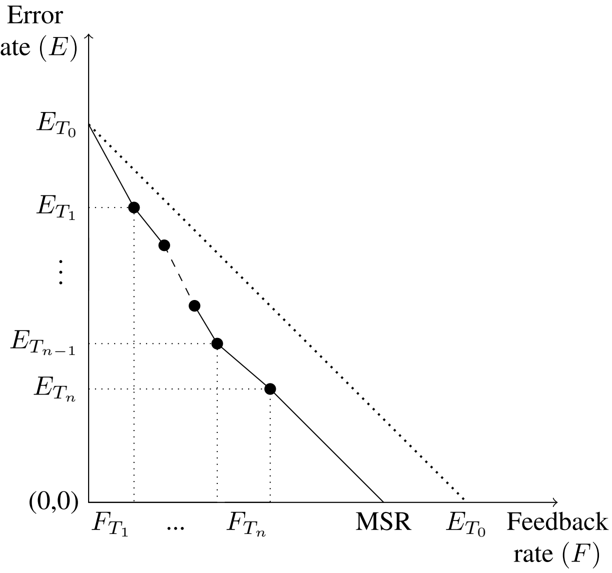

Figure 2: Computation of a Minimum Supervision Rate (MSR): Starting from the initial system error rate (E_{T_0}), some amount of information is provided as feedback (F_{T_1}), leading the system to update its models and yield a new error rate (E_{T_1}). Iterations take place until the adaptation process terminates, with an error rate E_{T_n} and after having provided a total amount of information of $\sum_1^n F_{T_i}$. The MSR is the sum of these two terms (MSR = $\sum F_{T_i} + E_{T_n}$).

3.2.1. Using Panels of Evaluated Oracles

In order to avoid too much dependence on a specific oracle, a solution is to use multiple oracles and average the results across them. In order to limit biases due to oracles with poor performance, each oracle should be itself evaluated and given more or less weight in the above-mentioned average depending on its performance. The performance of an oracle, i.e., its ability to efficiently enhance the systems with which it interacts, can be measured using the same metric as for the adaptive systems. In practice, the performance of an oracle can be defined as the best MSR achieved when interacting with any given system, i.e., $\min_{\text{sys}}$ MSR(ora,sys). The contribution of an oracle in the evaluation of a given system can then be weighted depending on this performance. The simplest weighting formula is 1 - $\min_{\text{sys}}$ MSR(ora,sys). Such a weighting scheme takes into account in the average representative oracles while strongly limiting the impact of others.

Note that this scheme encourages the development of oracles which are able to enhance the systems with which they interact, rather than directly modeling human behavior. This means that, as a side effect, the development of oracles leads to identify the best strategies to enhance the adaptive systems, which can then serve as guidelines for human users of these systems. For this reason, oracle development can be of interest as such, and not only for the sake of organising the evaluations.

3.2.2. Bootstrapping through Active Learning

To motivate the development of oracles, highly performing adaptive systems should be available, and vice versa. Bootstrapping the development and evaluation process is therefore an issue. A solution consists in requiring the systems to steer the oracles so that their functioning is deterministic, using an active learning approach (Settles, 2009). As a simple example, the systems can provide, together with their outputs, segments in which all errors would be provided as feedback by the oracle.

Using this possibility of relying on active learning capabilities in the systems, the protocol can be bootstrapped with a relatively simple, completely deterministic oracle. It can then progressively evolve towards a panel of more intelligent oracles, until such oracles become representative of real users.

3.3. Combined Solution

To summarize, the various issues identified can be solved, each by one main idea, as displayed in Table 1: In order to evaluate interactive system adaptation in an objective and comparative way while limiting biaises, one can develop oracles simulating users in their interactive supervision task, which can rely on the availability of reference outputs, evaluate these oracles to ensure that the feedback they provide enable the system to efficiently adapt, and form panels of such oracles. The overall process is detailed in Figure 3.

4. Conclusion and perspectives

A scheme for evaluating interactive system adaptation in an objective way has been proposed. It matches existing ones for some specific cases, but introduces innovative solutions to provide a generic framework for evaluating such systems. It is relatively complex compared to traditional evaluation schemes, by involving an interplay between the developement of systems to be evaluated and of oracles, and can be expected to need a few rounds of evaluation to reach maturity. However, this can also be seen as a richness and an opportunity to be representative of real usage.

The proposed scheme is expected to steer and support the development of systems able to learn from interactive user supervision. It opens the way to new evaluation campaigns for such systems in the many domains which remain to be covered[1].

This can in turn be expected to have a strong impact on the organisation of the research and product development in the domain. Indeed, the existence of highly performing and reliable adaptive systems not only enables user-driven adaptation, but also third-party system adaptation. The possibility to objectively measure the ability of a system to be adapted by a third party, outside of the research laboratory where the initial training took place, enables the creation of a chain of actors, each of them with a clear interface with the others. This thus paves the way for a specialisation of actors along the value chain, which would boost the development and spread of adaptive intelligent systems.

5. Acknowledgements

This work was partly supported by BPIFrance through the Quaero programme. The author wishes to thank Olivier Galibert and Juliette Kahn from LNE (Laboratoire national de métrologie et d'essais) for the exchanges on preliminary versions of the scheme described in this article.

6. Bibliographical References

Asadi, A., Schwartz, R., and Makhoul, J. (1991). Automatic modeling for adding new words to a large-vocabulary continuous speech recognition system. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 305–308.

Azzopardi, L., Järvelin, K., Kamps, J., and Smucker, M. D. (2011). Report on the SIGIR 2010 workshop on the simulation of interaction. *SIGIR Forum*, 44(2):35–47.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., and Vilar, J.-M. (2009). Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Isabelle, P. and Church, K. (1997). Special issue on new tools for human translators. *Machine Translation*, 12(1):1–2.

Nepveu, L., Langlais, P., Lapalme, G., and Foster, G. (2004). Adaptive language and translation models for interactive machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Nguyen, G. P. and Worring, M. (2008). Optimization of interactive visual-similarity-based search. *ACM Transactions on Multimedia Computing, Communications and Applications*, 4(1).

Orosanu, L. and Jouvet, D. (2015). Adding new words into a language model using parameters of known words with similar behavior. In *Proceedings of the International Conference on Natural Language and Speech Processing (ICNLSP)*, Alger, Algeria.

Ortiz-Martínez, D., García-Varea, I., and Casacuberta, F. (2010). Online learning for interactive statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 546–554. Association for Computational Linguistics.

Pallett, D. S. (1985). Performance assessment of automatic speech recognizers. *Journal of Research of the National Bureau of Standards*, 90:371–387.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Salton, G. and Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

[1] Such an evaluation campaign in the domain of automatic speech transcription is under preparation at DGA.

Table 1: Issues and solutions for evaluating interactive system adaptation.

Issues	Solutions
Feedback is not controlled by the evaluator	Count total amount of feedback needed to get correct output
Human is in the loop	Simulate user by an oracle having access to the reference output
Oracles have to make choices	Use several oracles, which are themselves evaluated
No oracle is available to start with	Bootstrap with active learning

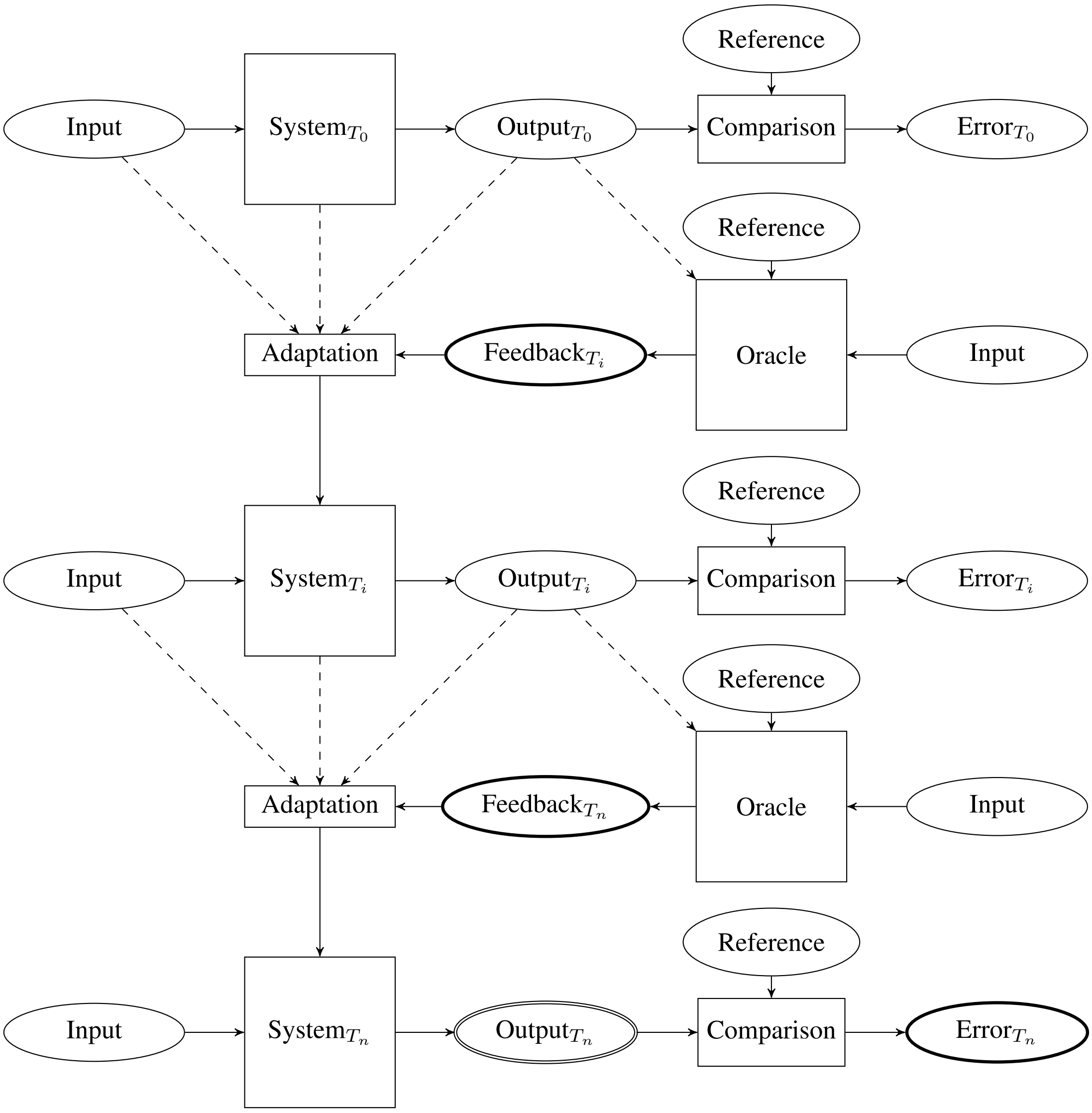

Figure 3: Illustration of the process for evaluating interactive system adaptation. In the first step (at time T_0), a classical evaluation process applies, where the system output is compared to a reference output to produce a figure of merit. In an intermediate step (at time T_i), the oracle provides feedback to the system, using the reference output in order to best represent the user. In the last step (at time T_n), the remaining errors (Error$_{T_n}$) correspond to the feedback which would be needed to get an error-free output, and can be combined to the cumulated feedbacks from previous iterations (shown in bold ellipses) to form a figure of merit reflecting the performance of the adaptation.