MINISTRY OF EDUCATION, YOUTH AND SPORTS

norway grants

The Research Council of Norway

D6.1: Project Evaluation Plan

Author:

Björn Gambäck
NTNU
gamback@idi.ntnu.no

Grant Agreement Number	7F14047
Project Acronym	HABiT
Project Title	Harvesting big text data for under-resourced languages
Deliverable Title	**D6.1: Project Evaluation Plan**
Responsible Partner	**Björn Gambäck, NTNU**
Dissemination Level	Public
Due Delivery Date	31 December 2016 (+30 days)
Actual Delivery Date	16 January 2017
Status	Final, v2.0

Principal Investigator	**Karel Pala**
Project Promoter	**Masaryk University**
Tel	+420 549 49 5616
E-mail	pala@fi.muni.cz
Project Website Address	www.habit-project.eu

Table of Contents

1 Introduction

The aim of this document is to provide a specification for the evaluation of the HABIT corpora and prototype system. The evaluations are required both to validate the overall project outcomes and in order to improve and revise the specifications of the HABIT prototype and its different modules. The entire project follows an iterative development approach, which entails the creation of intermediate system prototypes and modules, that will incorporate the results of the validation and evaluation activities. The pre-final system prototype will be used for usability studies and final system evaluation that will finish at the end of the project lifetime. Accordingly the evaluations need to cover a variety of situations and languages. These different requirements of the evaluations will require different measures to be used, some of which will be mostly quantitative and some qualitative.

Some of the work on evaluation will be carried out in connection to the development of the individual modules and published as separate deliverables (see in particular D4.1: "Methodology of Sketch Grammar evaluation"). The specific workpackage on "Requirements and evaluation" (WP6) aims to cover the overall system evaluation and the evaluation of the resources produced, as well as the evaluation of individual modules as influenced by their contribution to the project as a whole. The validation work is divided into three tasks:

Task T6.1: Project evaluation strategy

- Identification of quantifiable evaluation and success criteria
- Selection of appropriate evaluation methods of the different tasks

Task T6.2: Usability studies

- Quantitative user testing: A sample of test persons will provide early indication of whether the design meets end-user requirements
- Qualitative usability study: Longitudinal study of system usage

Task T6.3: System evaluation

- Integratability tests: Evaluate ease of system integration at external sites
- System evaluation: Performance evaluation of the complete integrated system prototype

These three tasks will in turn be reported in three separate and sequential deliverables, of which the present document is the first:

- **D6.1 Project evaluation plan:** Definition of evaluation criteria and methods
- **D6.2 Final Usability report:** Report on the overall usability of the system
- **D6.3 System evaluation:** Report on overall system performance

The second and third document will report be final evaluation results and be published at the end of the HABIT project. The main part of the evaluation will thus be carried out during

last half year of the project, that is, during the Spring of 2017. Hence the actual evaluation plan will be revised and further substantiated during that time period. In general, though, the evaluation process will be both *metric-centric* (primarily D6.3) and *user-centric* (primarily D6.2), where those concepts here are taken as:

Metric-centric The use of quantitative methods to determine values for metric data including system processing times and (sketch) grammar coverage rates, and corpora inter-annotator agreements, in conjunction with readily computable scores such as total corpora sizes, number of seed words for the collection and number of bigrams.

User-centric Qualitative methods used to acquire subjective impressions and opinions from the users of the HABIT prototype and corpora, including Likert-based surveys and interviews.

This report is basically split into to parts, where the first is concerned with system validation activities (Section 2) and the second part looks at overall evaluation activities (Section 3).

2 Validation

As mentioned in the introduction, the entire HABIT project follows an iterative development approach, where the refinement of intermediate system prototypes and modules will be achieved by incorporating validation and evaluation results. The overall project validation tests are expected to follow after testing of the individual modules, and their integration into a unified platform (where appropriate). Hence, the aim of validation is to ascertain the corrrect functionality of the software as a whole, which in turn is viewed as a black box by the persons performing the validation tasks. The test procedures must ensure that all functional requirements are satisfied, all performance requirements are achieved, documentation is correct and suitable for the user and any other requirements are met.

2.1 Internal validation

At the first level, a laboratory-level validation will be performed by the developers themselves, using a controlled environment but using the software will be in a natural run-time setting.

2.2 External validation

As a second level of validation, a validation approximating as closely as possible a real-world setting will be undertaken. In this case, the validation will be undertaken by persons with a profile closer to that of the typical HABIT system users, i.e., language professionals. These may be recruited from possible user groups or may be chosen from the personnel and students at the partner sites, but then be persons who have not been involved in any way in the software development or in the data collection and annotation.

Given that the persons at this validation level will not necessarily be computer analysts, the task will be much more demanding. Besides, if the users are external to the partner sites, no immediate assistance by members of the development teams will be available. Thus, it is necessary for the users involved in this type of validation to give clear and extensive feedback regarding the possible problems of the software.

2.3 Validation tasks

The validation tasks foreseen to be carried out involve the testing of all system functionalities available to the user, these functionalities including the following:

1. **Corpus collection** To use the corresponding tools that are available within the HABIT prototype in order to use a set of seed words to harvest a new language corpus from the web and clean it.
2. **Corpus modelling** In this case, the aim is to process a large monolingual corpus in order to extract information reflecting the language structure (such as word sketches, word senses and named entities).

3. **Corpus annotation** To test the available annotation tools, for example by creating word sketches for a corpus or for inducing word senses and named entities. In particular, the new Efficient Annotation Framework for under-resourced languages will need to be validated, in relation to the design and analysis of a new methodology for crowdsourcing annotation of previously uncovered language data as well as verification of the methodology by implementation of a new annotating framework.

The main part of the validation will be carried out in the form of a workshop in Brno during February 2017, and with annotators both on site in Brno and externally in Addis Ababa. For the validation, partner Masaryk willl prepare the tools with corpora in the four Ethiopian languages that the HaBiT project deals with (Amharic, Afaan Oromo, Tigrinya, Somali). The tools will allow for efficient preparation of new Part-of-Speech (PoS) taggers of these languages. The texts will need to be annotated (during the workshop) by native speakers, and the tools will then learn and improve automatically based on the manual annotations.

Masaryk will select texts for annotation from web crawled corpora. They will be from different domains, without boilerplate and tables. The texts will be tokenized and split into sentences. During annotation, each sentence will be annotated separately in random order, annotators will be able to see larger context (up to several sentences). Annotators will assign part of speech (PoS) tag for each token. The tagsets will be derived from existing corpora (such as the already annotated WIC for Amharic) or from Universal Dependencies. The tagsets could tentatively be modified later according to inter-annotation agreement.

For each language there will be from 5 to 10 annotators. Not all of them will have to annotate all tokens/texts. Some annotators could annotate only tokens where there is an disagreement between other annotators. A tagger will be trained using already annotated data and new texts will be pre-annotated using the tagger, which could further speed up the annotation. All annotation will be done using a web interface from any browser (including mobile phones or tablets) connected to a server at Masaryk University.

3 Evaluation

The main aim of HABIT project evaluation be to detect and accordingly modify potential system weaknesses, with the final results reported in two separate deliverables, where D6.2 will concentrate on the usability and user-centric aspects of the overall system, while D6.3 will highlight the performance and metric-centered aspects of the HABIT system and the created resources. (e.g., coverage and corpora sizes).

Normally, user-centered evaluation gives more relevant feedback, but tends to be subjective and can be both time-consuming and expensive. Metric-centered evaluation, on the other hand, can be fairly cheap and based on objectived figures, and can be easily employed for performing and testing changes during system development.

In general, there are two kinds of evaluation:

Quantitative / Intrinsic Evaluation:

- comparable evaluation metrics
- independent of the system and the task
- aim to correlate closely to human judgement

Qualitative / Extrinsic Evaluation:

- assess the (often indirect) effect of a component (through its performance as part of a functioning system)
- task- and context dependent
- e.g., measure user performance or satisfaction

3.1 Automatic evaluation

The automatic evaluation will be carried out internally, by members of the HABIT consortium. The data sets, measurements and the evaluation results will be posted on the HABIT website.

3.2 Human evaluation

The human evaluation will be carried out both in the form of interviews with individual users and through questionnaires. Interviews have the advantage of giving the option to go into detail of issues of specific concern, but are costly and time-consuming, and will hence only be carried out with a very limited number of users. Questionnaires can on the other hand be presented online (e.g., at `surveymonkey.com`) to larger user groups, and the answers can be given according Likert-scales (e.g., with five alternatives views on a specific statement along the lines of Strongly Agree, Agree, Neutral, Disagree, or Strongly Disagree). The evaluation results will be accessible via the HABIT website, while ensuring user privacy and anonymity.

References