Co-funded by the European Community Programme eContent*plus*

europeana connect

Evaluation Criteria for Language Resources

- Europeana Connect Plenary Meeting : WP2 session
- Berlin, May 12th, 2010
- Anne Schiller (Xerox)

Coordinated by the Austrian National Library Österreichische Nationalbibliothek

Language Resources

- Memos:
  - Europeana Language Resources - Necessary Language Resources for a Multilingual Query Translation System in Europeana

    Alessio Bosca et al. (18/11/2009)
  - Europeana Language Resource Evaluation - Workflow for the integration of open source Language Resources
    - Juliane Stiller, Vivien Petras (19/01/2010)
  - **Europeana Language Resources - Evaluation Criteria**

    Anne Schiller (08/03/2010)
- Milestones:
  - M2.2.1: Europeana Language Resources Repository online

    Aaron Kaplan (28/04/2010)

Evaluation of Language Resources: Why?

- Translation modules (and indexing) rely on LR
  - Ensure "minimum" quality for LRs
- Compare open source vs proprietary resources

Evaluation of Language Resources: How?

- Intrinsic qualities
  - General information (size, functional specs, data description, ...)
  - Quantitative evaluation (speed, coverage, ...)
  - Qualitative evaluation (accuracy, interoperability, ...)
  - Maintainability
- Effect on application (query translation) quality

General Information about LR

- Information of the LR (independent of application)
  - Size (number of entries, ...)
  - Functional description (data format, API, ...)
  - Description of properties, e.g.
    - Lemmatizer/stemmer: what are base forms?
    - Dictionary: lexemes, categories
    - Named entities; classification, granularity, …
  - Interoperability
    - Output and input of interoperating LRs are compatible
      - Categories are identical or can easily be mapped

Easily Measurable Quantitative Criteria

- Requires a test corpus (application specific)
- Quick, automatic test
  - Does not require gold standard or manual checking
- Aspects:
  - Speed
  - Coverage
    - ***Note:***
      - Low coverage means low "usability" for the application
      - High coverage does **not** imply high quality
      - Example:
        - » A lemmatizer that provides no output for 50% of the input is "bad"
        - » A lemmatizer that returns a copy of the input for every word, has 100% coverage, but is not very useful either

Criteria Requiring More Effort to Evaluate

- Purpose:
  - Check the quality of LRs
    - Is the output correct?
    - Is the output complete?
    - What is the impact on Europeana modules? (--> task 2.5)
- Time consuming
  - requires a representative test corpus
  - requires manual intervention to
  - (a) evaluate results of LR modules *or*
  - (b) set up a gold standard for automatic evaluation

Maintainability

- Operations
  - Add new entries
  - Correct or adapt entries
  - Remove entries

- Modifications
  - … cannot be done locally (e.g. web service)
  - … consist in simple text editing (e.g. stop word list)
  - … require specific tools or compilers (e.g. finite-state tools)
  - … require specific skills or knowledge (e.g. POS taggers)

Test Corpus

- Extraction of ClickStreamLogs
  - Issue: Logs contain info about interface language, not query language
  - Question: number of queries?
- Test methods
  - Apply LR and check results manually

OR

  - Set up a "gold standard" and check results automatically

- *Note:* Re-use test corpus for Task 2.5 (Sandbox integration, testing and evaluation of translation modules) ?

TO DO

- Proposal *(to be discussed at the meeting)*
  - Set up wiki for further discussion (Anne)
  - Finalize memo (Anne)
  - Set up test corpus (Anne)
  - Choose LR for "test" evaluation (who ??)
  - ...
  - ...
  - ...