Cleaning and Enriching Textual Databases from the Natural History Domain

Caroline Sporleder, Marieke van Erp, Tijn Porcelijn, Antal van den Bosch
{C.Sporleder,M.G.J.vanErp,M.Porcelijn,Antal.vdnBosch}@uvt.nl
Pim Arntzen
Arntzen@naturalis.nhm.nl

Information about cultural heritage collections is often stored in textual or semi-textual databases. Such databases are a potentially very valuable resource for researchers in the field, especially if the contents can be systematically searched and queried. Unfortunately, information retrieval from a database can be adversely affected by a variety of factors, such as errors and inconsistencies in the data or a suboptimal structure of the database itself. However, methods from data mining and natural language processing can help to detect and correct errors and inconsistencies and to add meta-data (such as named-entity tags) in order to aid information retrieval.

The Reptiles and Amphibians Database

- Database containing 16,870 records covering part of the collection of reptiles and amphibians
- Derived from: taxonomy, fieldbooks, logs, register book numbers, labels, publications and expert knowledge
- 37 fields (taxonomic, location, person name, free text, numeric and data fields)

Correcting Content Errors

- Aim: Detecting and correcting erroneous values in the database
- Method:
  - exploit interdependencies between columns (location: “Tafel Mountain”, country: “South Africa”)
  - train a classifier to predict the value of a cell based on values of other cells in the record
  - signal potential error in the database if the predicted value deviates from the original value
- Results:
  - recall up to 97%
  - precision up to 50%

Bootstrapping a Named Entity Tagger

- Aim: bootstrap a domain and database-specific named named entity tagger
- Method:
  - extract gazetteer lists from entity-specific columns (e.g. country → location)
  - use gazetteers to find training examples for database-specific entity tagger
- Results (F-Scores):
  - Look-up performs best: 76%
  - Database trained tagger slightly worse: 69%
  - Generic tagger performs worst: 33%

Correcting Wrong Column Errors

- Aim: Detecting and correcting values that are entered into the wrong column of the database
- Method:
  - train a classifier to predict appropriate column for text string based on typographical features and token overlap
  - signal potential error in the database if the predicted column deviates from the original column
- Results:
  - detection recall up to 100%
  - detection precision around 20% (high enough for semi-automatic error correction)
  - correction accuracy: up to 91% (promising)

This work is carried out as part of the MITCH project (Mining for Information in Texts from the Cultural Heritage), a joint research project between Tilburg University and Naturalis, the Dutch National Museum for Natural History.

http://ilk.uvt.nl/mitch