Alignment-based Expansion of Textual Database Fields

Piroska Lendvai

ILK / Communication and Information Sciences
Tilburg University
P.O. Box 90153, 5000 LE, Tilburg
The Netherlands

Abstract. Our study describes the induction of a secondary metadata layer from textual databases in the cultural heritage domain. Metadata concept candidates are detected and extracted from complex fields of a database so that content can be linked to new, finer-grained labels. Candidate labels are mined drawing on the output of Alignment-Based Learning, an unsupervised grammatical inference algorithm, by identifying head - modifier dependency relations in the constituent hypothesis space. The extracted metadata explicitly represent hidden semantic properties, derived from syntactic properties. Candidates validated by a domain expert constitute a seed list for acquiring a partial ontology.

1 Introduction

The data model underlying the structure of a database is often manually constructed, with the risk of becoming out-of-date over time: records that are arranged according to this structure outgrow it as the number of data attributes increases when resources of various formats get merged, updated, and new concepts emerge. These tendencies sometimes result in a mix of attributes joined in an ad-hoc way in loosely defined (free text) columns of the database, typically labelled as *Special remarks*. Such columns are of lower semantic coherence, and are suboptimal for effective database querying as they may contain several identical data types, for example numbers, that implicitly describe different, perhaps idiosyncratic properties, of a record.

Consider the following example from the specialRemarks column of a museum collection database:

```
Slides MSH 1975-xviii-27/29, 1975-xix-20/25; tape recording 1975 II
B 297-304. Acquired as gift from the British Museum (Nat. Hist.),
BMNH 1975. 1348
```

If a researcher is searching this column for tape recordings of a certain year, he needs to browse through all slide identification numbers and other ID numbers as well, because retrieving numbers cannot be securely narrowed down any further than to accessing the entire field. A query would be more efficient if the various ID numbers of slides, tape recordings, registration numbers, etc. would be separately