The Creation of an hierarchical Category Set as Gold Standard for Letterhead Elements

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Third Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

1 Introduction and Motivation

2 Used Data

The documents, which form the basis of this work, are old authorities correspondences where it comes to the storage of nuclear waste. The topics are mainly plans of emergency measures, eliminating access waters and the closure of individual chambers of the mine. At the beginning there were 71 authorities correspondences. 33 of these documents are freely available online. The remaining 38 documents have been provided as part of a research project. The freely accessible documents were written in the years 2009 to 2013. The others, as far as recognizable, are from the years 1971 to 1975. Although picture elements are included in the OCR output, they aren't used in this work. Twelve of the documents were removed because they effectively did not contain any information content. Thus were still 59 documents left with a total of 1239 text areas. In addition, seven empty text areas were removed. Thus remained in 1232 text areas for the following steps. The resulting in the OCR process documents were available in HTML format and has been brought in a CSV format for the next steps which contains the following information for each text area:

- Document name and index the text area within the document
- identifier of the text area assigned classification
- position of the text area in the document and its width
- by the OCR process recognized text

3 Creation of the hierarchical Category Set

The hierarchical category system extends the seven basic classes, which were used by a project partner for the first classification. The hierarchical structure also offers the possibility that instances of individual classes can be summarized in its superclass. In the following work, the documents were processed by two raters. After careful review of the documents by both raters, five subclasses were added to the first hierarchical category system. Thus, the classification model, which was used for the next steps emerged. To generate the training data each text area of the 59 documents were individually classified by the raters. In order not mutually influence each other in the award of the label, the classification of the two raters was done independently after a brief discussion about the meaning of each individual classes. In order to verify if and how well the classification matches of both raters, their agreement was calculated by Cohen's Kappa for each evaluation run. In the first evaluation run a κ value of 0.6993 was achieved.

Nevertheless, it became clear that the hierarchical category system was not yet well enough adapted to the available data. Many of the predetermined classes were not at all or only very rarely assigned. This applies to all subcategories of the third party for example. The third party referred to any personal or address information within a document, which can not be assigned to the actual sender or recipient. There were not only deleted subcategories but also added new ones. This was in the category "(main) text"the case. This was assigned in 41.23 % of the text areas. Therefore, this category has been split into the two sub-categories “content“ and “structural elements“.

The second review cycle revealed that the addition of the two sub-categories “content“ and “structural elements“ to the category “(main) text“

simplifies the assignment of many text areas. Text areas that were assigned as generic "letterhead element" (i.e. the root category) because they only contained a single character and were used to structure the text, now can be assigned to the category "structural element". This meant that almost 40 % of the previously generically as "letterhead element" marked text areas could be assigned to another category. Overall, the classification of the raters differed after the second pass on the entire data 324 times. This shows that even before the subsequent merging of the data was an agreement of almost 74 % and a κ 0.7007 from among the raters. Compared to the previous run thus a minimum increase of κ is observed. In Bakeman et al. (1997) is shown that a better value for κ is obtained by a larger number of classes. That in this evaluation run now, despite a lower number of classes (37 instead of 47) an increase of κ is observed, therefore, indicates a better match of the raters.

In order to create a unique category assignment for this data on the basis of the final hierarchical category system, the text areas whose assignments were different in the two raters, were checked again together and a category was chosen. Overall, the decisions, which of the two views was followed out very evenly distributed. On the results of the merge you can see that only the upper categories of the category system were not assigned. It follows that is at least one example of the training data available for each of the categories.

4 Classification Experiments

5 Future Work

Acknowledgments

References

Bakeman, R., Quera, V., McArthur, D. und Robinson, B. F.: *Detecting Sequential Patterns and Determining Their Reliability With Fallible Observers*, in: *Psychological Methods*, Vol. 2, S. 357–370, 1997.

Autoren, erster Vorname ausgeschrieben restlichen mit . Personen mit , getrennt. Jahr. *Titel*. Verlag, Verlagsort, Verlagsland. Zeitschriftsnummer(?):Seitevon-bis.