GermEval 2019 Task 1 – Shared task on hierarchical classification of blurbs

Rami Aly, Steffen Remus, and Chris Biemann

Language Technology Group, Universität Hamburg

May 26, 2019

1 Introduction

Hierarchical multi-label classification (HMC) of Blurbs is the task of classifying multiple labels for a short descriptive text, where each label is part of an underlying hierarchy of categories. The increasing amount of available digital documents and the need for more and finer grained categories calls for a new, more robust and sophisticated text classification methods. Large datasets often incorporate a hierarchy for which can be used to categorize information of documents on different levels of specificity. The traditional multi-class text classifcation approach is thoroughly researched, however, with the increase of available data and the necessity of more specific hierarchies and since traditional approaches fail to generalize adequately, the need for more robust and sophisticated classification methods increases.

With this task we aim to foster research within this context. This task is focusing on classifying German books into their respective hierarchically structured writing genres using short advertisement texts (Blurbs) and further meta information such as author, page number, release date, etc.

GermEval is a series of shared task evaluation campaigns that focus on Natural Language Processing for the German language. The workshop of this shared task will be held in conjunction with the Conference on Natural Language Processing KONVENS[1] 2019 in Erlangen/Nürnberg.

2 Contents and Data Format

The dataset consists of information on German books (blurb, genres, title, author, URL, ISBN, date of publication), crawled from the Random House page. The rights to all blurbs and further meta information belong to Random House. The date of publication normally represents the publication date of the particular version of the book. Genres that capture properties which do not rely on content but on the shape or form of a book were removed. Since every ISBN and URL appears exactly once, each blurb can be considered unique. However, in some cases, book blurbs may appear similar, for example, if a book is part of a series. Due to the publishing and crawling process further anomalies can not be excluded. Known anomalies are missing author and incorrect publication date. Furthermore, multiple blurbs of the same book (e.g. different editions) cannot be assumed to be the same, although the author and title is identical. However,

[1] `https://dgfs.de/de/cl/konvens.html`

only around 1% of all blurbs are affected by these anomalies.

This dataset follows the policies as described in the RCV1 dataset by Lewis et al. (2004). We adapt RCV1's properties, which have been explained by its authors in detail and refer to their description. The **minimum code policy** requires the assignment of at least one category to each document of the collection. The **hierarchy policy** ensures that every ancestor of a document's label is assigned as well.

Data format of the training, dev and test dataset is the same: It uses xml tags as a structuring method.
<book> </book>: The scope of one book
<title> </title>: Title of the book the blurb is taken from
<body> </body>: Contains the actual blurb
<copyright> </copyright>: Statement regarding copyrights
<categories> </categories>: Contains all extracted categories of a book
<category> </category>: Contains the actual genre and all ancestor genres
<topic d="n"></topic> Name of topic. It is on level n of the hierarchy. Most specific label is marked with *label = True.*
<author> </author>: Author of the book
<published> </published>: Date of publication
<isbn> </isbn>: Book ISBN.
<url> </url>: URL of the page the blurb was taken from.

An example entry of the resulting dataset is shown below:

```
<book date="2019-01-04" xml:lang="de">
<title>Blueberry Summer</title>
<body>In neuer Sommer beginnt für Rory und Isabel -- mit einer kleinen
   Neuerung: Rory ist fest mit Connor Rule zusammen und deshalb als
   Hausgast in den Hamptons. Und genau das bringt Komplikationen mit sich,
   denn irgendwie scheint Connor ein Problem damit zu haben, dass Rory
   nicht mehr für seine Familie arbeitet. Isabel dagegen arbeitet zur
   Überraschung aller als Kellnerin, um einen süssen Typen zu
   beeindrucken -- irgendwie muss sie ja über ihre Affäre mit Mike
   inwegkommen. Das klappt ganz gut, bis Rory auf Isabels Neuen
   trifft ... Und Isabel wieder auf Mike.</body>
<copyright>(c) Verlagsgruppe Random House GmbH</copyright>
<categories>
<category>
<topic d="0">Kinderbuch & Jugendbuch </topic>
<topic d="1" label="True">Liebe, Beziehung und Freundschaft</topic>
</category>
<category>
<topic d="0">Kinderbuch & Jugendbuch </topic>
<topic d="1" label="True">Echtes Leben, Realistischer Roman</topic>
</category>
</categories>
<authors>Joanna Philbin</authors>
<published>2015-02-09</published>
<isbn>9780451457998</isbn>
<url>https://www.randomhouse.de/Taschenbuch/Blueberry-Summer/Joanna-Philbin
/cbj-Jugendbuecher/e455949.rhd%0A/</url>
</book>
```

Listing 1: Example entry of a book in the dataset

Additionally, a file that contains only the hiararchy in form of parent-child relationships is provided.

3 Quantitative characteristics

The datset is split in the ratio of 70%, 10% and 20% for train, validation and test respectively. An overview of the complete dataset can be seen in Table 1.

Dataset	BGC-DE
Number of samples	20,784
Average length of blurb	94.67
Total number of classes	343 (8, 93, 242 on level 1,2, 3 resp.)

Table 1: Quantitative characteristics of the dataset.

It is important to note that the most specific category of a book does not have to be a leaf. For instance, the most specific class of a book could be *Romane & Erzählungen*, although *Romane & Erzählungen* has further children categories, such as *Romanbiographien*. Also, the label distribution is highly unbalanced.

4 License

The copyright to all blurbs belongs to **Verlagsgruppe Random House GmbH**, its licensors, vendors and/or its content providers since the blurbs were obtained through the Random House website[2]. The blurbs serve promotional/public purposes and permission has been granted by Penguin Random House to share this dataset. The dataset is shared under the CC BY-NC 4.0 license[3], which allows copying and redistributing the dataset as long as appropriate credit is given, especially to Penguin Random House and its content providers. In addition to indicating changes to the dataset, you may not use the dataset for commercial purposes.

5 Contact

For further inquires or questions regarding the shared task, send an e-mail to remus@informatik.uni-hamburg.de.

References

Lewis, D. D., Y. Yang, T. G. Rose, and F. Li (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research 5*(Apr), 361–397.

[2] `https://www.randomhouse.de/`

[3] `(https://creativecommons.org/licenses/by-nc/4.0/)`