Much Ado About Text Content. Learning Text Types Solely by Structural Differentiae

— Extended Abstract —

Alexander Mehler¹, Peter Geibel², Rüdiger Gleim¹, Sebastian Herold², Brijnesh-Johannes Jain¹, and Olga Pustylnikov¹

¹ Bielefeld University
{Alexander.Mehler,Ruediger.Gleim,Olga.Pustylnikov}@uni-bielefeld.de
² University of Osnabrück
{pgeibel,sherold}@uos.de
³ Technical University of Berlin
jbj@cs.tu-berlin.de

Abstract.

1 Introduction

In text linguistics and related disciplines, textual units are distinguished in terms of their structure, content and function [6]. From a text technological point of view we may add the layout of texts apart from their logical document structure as a further starting-point of differentiation [3, 26]. These four reference points open different perspectives on text classification which are not necessarily orthogonal. Generally speaking, text classification belongs to explorative data analysis which aims at automatically dividing sets of objects into classes of maximum internal homogeneity and external heterogeneity. According to Bock [5], any such approach is related to at least five questions:

1. What do classes of objects denote?
2. Which differentiae [31] determine the membership of objects to classes?
3. What is the underlying (e.g. hierarchical, partitive) structure of classification?
4. Into which relations (e.g. subordination, overlapping) do the classes enter?
5. By which criteria can we identify good classifications?

In this paper, we deal with classifying texts into classes which denote text types whose textual instances serve more or less homogeneous functions. Other than mainstream approaches to text classification, which rely on the vector space model [30] or some of its descendants [2] and, thus, on content-related lexical features, we solely refer to structural differentiae, that is, to patterns of text structure as determinants of class membership. Further, we suppose that text types span a type hierarchy based on the type-subtype relation [31]. Thus, although we admit that class membership is fuzzy so that overlapping classes are
inevitable, we suppose a non-overlapping type system structured into a rooted tree – whether solely based on functional or additional on, e.g., content- or media-based criteria [1]. What regards criteria of goodness of classification, we perform a classical supervised categorization experiment [30] based on cross-validation as a method of model selection [11]. That is, we perform a categorization experiment in which for all training and test cases class membership is known \textit{ex ante}. In summary, we perform a supervised experiment of text classification in order to learn functionally grounded text types where membership to these types is solely based on structural criteria.

2 The Quantitative, Distributional Approach

In order to make the present approach comparable to its rivals, we need to relate the concept of a \textit{text type} to related notions in text linguistics. Further, we have to clarify the relation of \textit{structure} and \textit{function}, that is, the relation by which the function of a text can be predicted by its structure. In order to do that we generally define a \textit{document class} as a class of textual or hypertextual units which share class constitutive differentiae. Varying reference points of clarifying the ontological status of these differentiae lead to different notions of a document class. If we focus, for example, on functional or situative criteria of class membership, we deal with so called \textit{genres} [19, 36] and \textit{registers} [4, 10], respectively. By analogy with these distinctions, we speak of \textit{hypertext sorts}, \textit{digital genres} and \textit{web genres} in the case of classes of hypertextual units [9, 16, 25]. Generally speaking, if we consider the composition of classes in terms of their \textit{extension}, that is, from the point of view of enumerating their elements, we deal with \textit{sorts} of documents – e.g. \textit{text sorts} in the sense of Heinemann [13]. If in contrast to this, class membership is defined in \textit{intensional} terms, we deal with \textit{text patterns} [13] or \textit{superstructures} [34] as prototypical representations of class members, whose expectation-driven production/reception they support. The basic idea of all these approaches is that the structure and shape of (hyper-)textual units vary, though not deterministically dependent on the communicative situation or function they manifest. If we focus on structure, abstracting from shape or layout, we deal with what is called the \textit{logical document structure} [26, 27].

The various notions of document classes as, for example, \textit{genre} and \textit{register} are specific to certain linguistic theories. In order to use a general terminology without preferring any such theory, we speak – by analogy with object oriented modeling [28] – of document \textit{classes} as intensional specifications in contrast to their \textit{class extensions}. Thus, a document is an element of a certain class extension while it is an instance or a member of the corresponding class. As far as we focus on textual units, disregarding hypertextual ones, we speak more narrowly of text classes. Finally, instead of the term \textit{class} we will use its synonym \textit{type} [32] and will, therefore, speak of \textit{text types}.

Now we can explain the relation of text structure and function in more detail. The basic idea to predict the function of a text based on the patterns it instantiates comes from the quantitative, distributional approach [4]. Accord-
To this approach, distributional patterns vary dependent on the functions of the texts in which they are observed [4]. Starting from the weak contextual hypothesis of Miller & Charles [23] which says that the similarity of the contextual representations of words contributes to their semantic similarity, one might formulate that different text structures reflect differences in function as far as they are confirmed by significantly many texts and thus are recognizable as recurrent text patterns. In the reversal case one might say that texts with similar functions tend to have similar structures. With a focus on registers, Biber [4, 59] puts this as follows: “[…] preferred linguistic forms of a register are those that are best suited functionally to the situational demands of the variety […].”

As the relation of structure and function is not deterministic (neither can we deterministically infer a unique function based on observing some text pattern, nor is the same function always manifested by the same pattern), the task of learning text types by exploring structural patterns is all but not trivial.

In the present paper, we focus on the logical document structure as a source of feature selection while we disregard layout as well as any content indicating lexical units of the texts to be classified. Since we focus on text types, we leave out hypertext and, especially, web-based documents. Furthermore, since we aim at modelling text types, we go beyond classical approaches to text categorization which learn classifiers in order to enumerate class members without any effort in interpreting these classifiers as representations of text patterns. Rather, we perform our experiments as a preliminary step towards learning classifiers as representations of text patterns which determine membership to text types.

In summary, the present paper aims at a terminological ontology of text types by example of the area of press communication. As far as the text classifiers being learnt allow deriving representations of text patterns which, in turn, allow computing the similarity of texts with respect to these patterns, we contribute to a prototype ontology in the sense of Sowa [31].

3 The Experimental Setting

The main hypothesis of our approach is that structure-based classification is a serious alternative to approaches based on lexical feature vectors. Thus, we expect

\begin{table}
\centering
\begin{tabular}{|c|c|}
\hline
ID & Category & \# \\
\hline
1 & Company of the Day & 2,213 \\
2 & Table of Contents & 1,933 \\
3 & Weekly Chronicle & 553 \\
4 & Dates & 176 \\
5 & University News & 73 \\
6 & Culture Events & 49 \\
7 & Questions and Answers & 20 \\
\hline
\end{tabular}
\caption{The corpus of rubrics of the German newspaper Süddeutsche Zeitung.}
\end{table}
a high selectivity of structure-sensitive features which in the case of function-
ally delimitable text types are at least significant as lexical, content-indicating
features. In order to support this hypothesis, we process two corpora:

- As a proof of concept, we perform SVM-based classification on 5,015 texts
  which belong to a corpus of seven so called rubrics (i.e. the categories to be
  learned; cf. Table 1). These are text types which according to visual inspec-
tion are evidently separable by the structure of their instances. Taken these
rubrics as categories to be learnt, we expect a high selectivity of structure-
based classifiers.

- In order to expose our approach to a more demanding experiment, we per-
form the same experiment by means of a corpus of all 95 rubrics of a ten
years release of the Süddeutsche Zeitung [33]. Since the sample sizes of
these rubrics vary drastically (from 2 texts of the rubric poll to 10,991 texts
of the rubric Interview) an since we do not control the selection of rubrics
by inspection of their instances, we expect a (moderate) break-in of our
classifiers.

In recent experiments [22], we have studied the selectivity of structural fea-
tures in an unsupervised learning scenario. That is, we did not perform any su-
pervised training. Our findings have shown that unsupervised structure-oriented
classification performs well above the baseline of random category assignment.
However, these experiments have also shown that there is room for improve-
ment. As a complementary approach, we now tackle the question what “golden
standard” can be achieved by using supervised methods of machine learning. In
this context, we develop two structure-oriented SVM-based methods:

- SVM learning is based on the use of a kernel that operates on pairs of
  examples [35]. Unlike many text mining approaches that only consider a
  vector-based representation of the documents [7], we want to focus on their
  structural properties. The theory of SVMs, however, ensures that kernels can
also be defined for structured data like DOM-trees describing texts [29, 12,
15]. A large class of structure kernels is formed by the so-called convolution
kernels [12] including the one defined by Collins and Duffy [8] for labeled,
ordered trees. This tree kernel has previously been applied to structure based
classification of sentences described by their parse trees [8, 24]. In order to
classify whole texts, we apply an extended version of this kernel to the DOM-
trees describing their structures. This new kernel allows a variable number
of descendants for some tree nodes (corresponding to particular XML-tags).
We will compare the performance of our generalized tree kernel with an
approach [15] that combines SVMs and a graph similarity measure, which is
also suitable for DOM-trees.

- In order to have a baseline of a structure-oriented classifier, we build a feature
  vector model whose dimensions solely code structural features of the input
text. This approach is inspired by Liiv & Tuldava [18] who cluster texts by
reference to structural variables as well as by Köhler [17] who develops a
reference system of structural variables.
Finally, we are interested in the question whether a combination of structure- and content-oriented classifiers perform better than any of these approaches in isolation. Consequently, we also develop integrated, that is, structure- and content-based classifiers.

These three approaches operate on input texts in conformance with an element name-adapted version of XCES [14]. That is, we automatically preprocess input texts by means of the Text Miner system [20] which, amongst others, annotates the input text’s logical document structure down to the level of lexical tokens (by disregarding sentence structure). This is illustrated in Figure (1) which shows an outline of an input document and the kind of ordered rooted tree derived thereof as the proper input of feature selection.

4 Conclusion

By exploring the information value of document structure as a source of feature selection, we contribute to the area of structure mining. This is a valuable task as more and more documents are online whose structure is an invaluable resource of web mining. Thus, the present paper does more than just presenting yet another
text classifier. Rather, it is a very first step towards the far-reaching goal of developing a prototype ontology of document types in the fast growing field of web-based information processing.
Bibliography


