Structural Differentiae of Text Types. A Quantitative Model

Olga Pustylnikov and Alexander Mehler

Faculty of Linguistics and Literature Study,
University of Bielefeld, Germany
Olga.Pustylnikov@uni-bielefeld.de
Alexander.Mehler@uni-bielefeld.de

Abstract. The categorization of natural language texts is a well established research field in computational and quantitative linguistics (Joachims 2002). In the majority of cases, the vector space model is used in terms of a *bag of words* approach. That is, lexical features are extracted from input texts in order to train some categorization model and, thus, to attribute, for example, authorship or topic categories. Parallel to these approaches there has been some effort in performing text categorization not in terms of lexical, but of structural features of document structure. More specifically, quantitative text characteristics have been computed in order to derive a sort of structural text signature which nevertheless allows reliable text categorizations (Kelih & Grzybek 2005; Pieper 1975). This *"bag of features"* approach regains attention when it comes to categorizing websites and other document types whose structure is far away from the simplicity of tree-like structures. Here we present a novel approach to structural classifiers which systematically computes structural signatures of documents. In summary, we present a text categorization algorithm which in the absence of any lexical features nevertheless performs a remarkably good classification even if the classes are thematically defined.

Keywords

QUANTITATIVE LINGUISTICS, TEXT CATEGORIZATION, FEATURE SELECTION

1 Introduction

An alternative way to categorize documents apart from the well established "*bag of words*" approach is to categorize by means of structural features. This approach functions in absence of any lexical information utilizing quantitative characteristics of documents computed from the logical document structure.[1]

[1] See also Mehler et al. (2006).

That means that markers like content words are completely disregarded. Features like distributions of sections, paragraphs, sentence length etc. are considered instead.

Capturing structural properties to build a classifier assumes that given category separations are reflected by structural differences. According to Biber (1995) we can expect that functional differences correlate with structural and formal representations of text types. This may explain good overall results in terms of *F-Measure*[2]. However, the *F-Measure* gives no information about the quality of the investigated categories. That is, no a prior knowledge about the suitability of the categories for representing homogenous classes and for applying them in machine learning tasks is provided. Since natural language categories e.g. in form of web documents or other textual units arise not necessarily with a well defined structural representation available it is important to know how the classifier behaves dealing with such categories.

Here, we investigate a large number of existing categories, thematic classes or *rubrics* taken from a 10 years newspaper corpus of *Süddeutsche Zeitung* (SZ 2004) whereas a rubric represents a recurrent part of the newspaper like `sports´ or `tv-news´. We test systematically their goodness in a structural classifier framework asking more specifically for a maximal subset of all rubrics which gives an *F-Measure* above a predefined *cut-off* $c \in [0, 1]$ (e.g. $c = 0.9$). We evaluate the classifier in the way allowing to exclude possible drawbacks with respect to:

- the categorization model used (here SVM[3] and Cluster Analysis),[4]
- the text representation model used (here the *bag of features* approach) and
- the structural homogeneity of categories used.

The first point relates to distinguishing supervised and unsupervised learning. That is, we perform these sorts of learning although we do not systematically evaluate them comparatively with respect to all possible parameters. Rather, we investigate the potential of our features evaluating them with respect to both scenarios. The representation format (vector representation) is restricted by the model used (e.g. SVM). Thus, we concentrate on the third point and apply an *iterative categorization procedure* (ICP)[5] to explore the structural suitability of categories. In summary, our experiments have twofold goals:

1. to study given categories using the ICP in order to filter out structurally inconsistent types and
2. to make judgements about the structural classifier's behavior dealing with categories of different size and quality levels.

[2] The harmonic mean of *precision* and *recall* is used here to measure the overall success of the classification

[3] Support Vector Machines.

[4] Supervised vs. unsupervised respectively.

[5] See sec. 4.

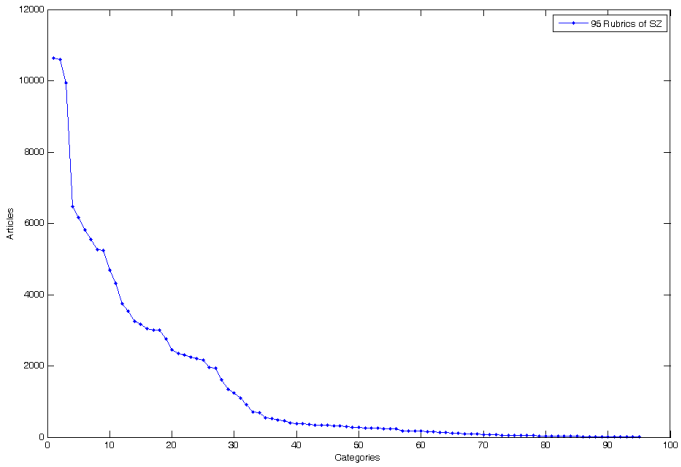

Fig. 1. Categories/Articles-Distribution of 95 Rubrics of SZ.

2 Category Selection

The 10 years corpus of the SZ used in the present study contains 95 different rubrics. The frequency distribution of these rubrics shows an enormous inequality for the whole set (See Figure 1). In order to minimize the calculation effort we reduce the initial set of 95 rubrics to a smaller subset according to the following criteria.

1. First, we compute the mean μ and the standard deviation σ for the whole set.
2. Second, we pick out all rubrics R with the cardinality $|R|$ (the number of examples within the corpus) ranging between the interval:

$$\mu - \sigma/2 < |R| < \mu + \sigma/2$$

This selection method allows to specify a window around the mean value of all documents leaving out the unusual cases.[6] Thus, the resulting subset of 68 categories is selected.

3 The Evaluation Procedure

The data representation format for the subset of rubrics uses a vector representation (*bag of features* approach) where each document is represented by a

[6] The method is taken from Bock (1974). Rieger (1989) uses it to identify above-average agglomeration steps in the clustering framework. Gleim et al. (2007) successfully applied the method to develop quality filters for wiki articles.

feature vector.[7] The vectors are calculated as structural signatures of the underlying documents. To avoid drawbacks (See Sec. 1) caused by the evaluation method in use, we compare three different categorization scenarios:

1. Supervised scenario by means of SVM-light[8],
2. Unsupervised scenario in terms of Cluster Analysis and
3. Finally, a baseline experiment based on random clustering.

Consider an input corpus K and a set of categories $\mathbb{C}$ with the number of categories $|\mathbb{C}| = n$. Then we proceed as follows to evaluate our various learning scenarios:

- For the supervised case we train a binary classifier by treating the negative examples of a category $C_i \in \mathbb{C}$ as $K \setminus [C_i]$ and the positive examples as a subset $[C_i] \subseteq K$. The subsets C_i are in this experiment pairwise disjunct and we define $\mathbb{L} = \{[C_i]|C_i \in \mathbb{C}\}$ as a partition of positive and negative examples of C_i.
  Classification results are obtained in terms of *precision* and *recall*. We calculate the *F-score* for a class C_i in the following way:

$$F_i = \frac{2}{\frac{1}{\text{recall}_i} + \frac{1}{\text{precision}_i}}$$

  In the next step we compute the weighted mean for all categories of the partition $\mathbb{L}$ in order to judge about the overall separability of given text types using the *F-Measure*:

$$\text{F-Measure}(\mathbb{L}) = \sum_{i}^{n} \frac{|C_i|}{|K|} F_i$$

- In the case of unsupervised experiments we approach as follows: The unsupervised procedure evaluates different variants of Cluster Analysis (hierarchical, k-means) trying out several linkage possibilities (complete, single, average, weighted) in order to achieve the best performance. Similar to the supervised case best clustering results are presented in terms of *F-Measure* values.
- Finally, the random baseline is calculated by preserving the original category sizes and by mapping articles randomly to them. Results of random clustering help to check the success of both learning scenarios. Thus, clusterings close to the random baseline indicate either a failure of the cluster algorithm or that the separability of the text types can't be well separated by structure.

In summary, we check the performance of structural signatures within two learning scenarios – supervised and unsupervised – and compare the results

[7] See Mehler et al. (2007) for a formalization of this approach.
[8] Joachims (2002).

with the random clustering baseline. Next Section describes the *incremental categorization procedure* (ICP) to investigate the structural homogeneity of categories.

4 Exploring the Structural Homogeneity of Text Types by means of the Iterative Categorisation Procedure (ICP)

In this Section we return to the question mentioned at the beginning. Given a *cut-off* $c \in [0,1]$ (e.g. $c = 0.9$) we ask for the maximal subset of rubrics allowing to achieve an *F-Measure* value $F > c$. Decreasing the *cut-off* c successively we get a rank ordering of rubrics ranging from the best contributors to the worst ones. The ICP allows to determine a result set of maximal size n with the maximal internal homogeneity compared to all candidate sets in question. Starting with a given set of input categories to be learned we proceed as follows:

1. **Start:** Select a seed category $C \in A$ and set $A_1 = \{C\}$. The rank r of C equals $r(C) = 1$. Now repeat:
2. **Iteration** $(i > 1)$**:** Let $B = A \setminus A_{i-1}$. Select the category $C \in B$ which when added to A_{i-1} maximizes the F-Measure value among all candidate extensions of A_{i-1} by means of single categories of B. Set $A_i = A_{i-1} \cup \{C\}$ and $r(C) = i$.
3. **Break off:** The iteration algorithm terminates if either
   i) $A \setminus A_i = \emptyset$ or
   ii) the F-Measure value of A_i is smaller than a predefined cut-off or
   iii) the F-Measure value of A_i is smaller than the one of the operative baseline.

   If none of these stop conditions holds repeat step (2).

The kind of ranking described here is more informative than the *F-Measure* value alone. That is, the *F-Measure* gives global information about the overall separability of categories. The ICP in contrast, provides additional local information about the weights of single categories with respect to the overall performance. This information allows to check the suitability of single categories to serve as structural prototypes. Knowledge about the homogeneity of each category provides a deeper insight into the possibilities of our approach.

In the next Section the rankings of the ICP applied to supervised and unsupervised learning and compared with the random clustering baseline are presented. In order to exclude a dependence of the structural approach on one of the learning methods, we also apply the *best-of-unsupervised-ranking* to the supervised scenario and compare the outcomes. That means, we use exactly the same range having performed best in the unsupervised experiment for SVM learning.

Category Set	Number
Total	95
Selected Initial Set	68
Unsupervised	55
Supervised	16
Unsupervised ∩ Supervised	14

Table 1. Corpus Formation (by Categories).

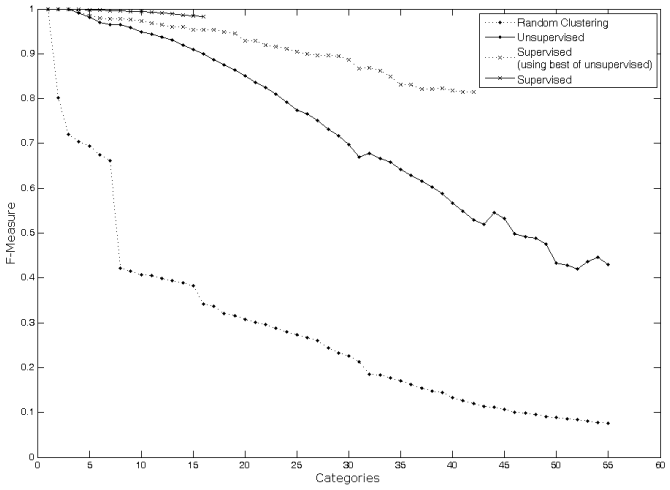

Fig. 2. The F-Measure Results of All Experiments

5 Results

Table 1 gives an overview about the categories used. From the total number of 95 rubrics 68 were selected using the selection method described in Section 2, 55 were considered in unsupervised, 16 in supervised experiments. The common subset used in both cases consists of 14 categories.

The Y-axis of Figure 2 represents the *F-Measure* values and the X-axis the rank order of categories iteratively added to the seed set. The supervised scenario (upper curve) performs best ranging around the value of 1.0. The values of the unsupervised case decrease more rapidly (the third curve from above). The unsupervised best-of-ranking categorized with the supervised method (second curve from above) lies between the best results of the two methods. The lower curve represents the results of random clustering.

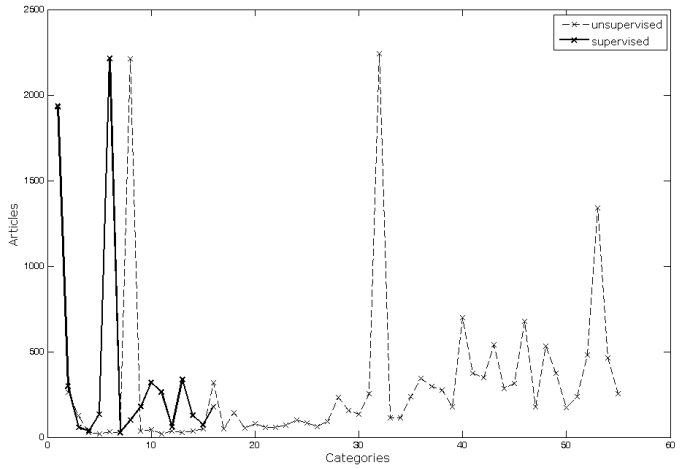

Fig. 3. Categories/Articles-Distribution of Sets used in Supervised/Unsupervised Experiments.

6 Discussion

According to Figure 2 we can see, that all *F-Measure* results lie high above the baseline of random clustering. All the subsets are well separated by their document structure which indicates a potential of structure-based categorizations. The point here was to observe the decrease of the *F-Measure* value while adding new categories.

The supervised method shows the best results remaining stable with a growing number of additional categories. The unsupervised method shows a more rapid decrease but is less time consuming. Cluster Analysis succeeds to rank 55 rubrics whereas SVM-light ranks only 16 within the same time span.

In order to compare the performance of both methods (supervised vs. unsupervised) more precisely we ran the supervised categorization based on the *best-off-ranking* of the unsupervised case. The resulting curve remains longer stable than the unsupervised one. Since the order and the features of categories are equal, the resulting difference indicates an overall better accuracy of SVM compared to Cluster Analysis.

One assumption for the success of the structural classifier was that the performance may depend on the article size, that is, on the representativeness of a category. To account for this, we compared the category size of the *best-off-rankings* of both experiments. Figure 3 shows a high variability in size, which indicates that the size factor does not influence the classifier.

7 Conclusion

In this paper we presented experiments which shed light on the possibilities of a classifier operating with structural signatures of text types. More specifically, we investigated the ability of the classifier to deal with a large number of natural language categories of different size and quality. The *best-off-rankings* showed that different evaluation methods (supervised/unsupervised) prefer different combinations of categories to achieve the best separation. Furthermore, we could see that the overall difference in performance of two methods depends rather on the method used than on the combination of categories.

Another interesting finding is that the structural classifier seems not to depend on category size allowing a good categorization of small, less representative categories. That fact motivates to use logical document (or any other kind of) structure for machine learning tasks and to extend the framework to more demanding tasks, when it comes to deal with, e.g., web documents.

References

ALTMANN, G. (1988): *Wiederholungen in Texten.* Brockmeyer, Bochum.

BIBER, D. (1995): *Dimensions of Register Variation: A Cross-Linguistic Comparison.* University Press, Cambridge.

BOCK, H.H. (1974): *Automatische Klassifikation. Theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten (Cluster-Analyse).* Vandenhoeck & Ruprecht, Göttingen.

GLEIM, R.; MEHLER, A.; DEHMER, M.; PUSTYLNIKOV, O. (2007): Isles Through the Category Forest — Utilising the Wikipedia Category System for Corpus Building in Machine Learning. In: *WEBIST '07, WIA(2).* Barcelona, Spain, 142-149.

JOACHIMS, T. (2002): *Learning to classify text using support vector machines.* Kluwer, Boston/Dordrecht/London.

KELIH, E.; GRZYBEK, P. (2005): Satzlänge: Definitionen, Häufigkeiten, Modelle (Am Beispiel slowenischer Prosatexte). In: *LDV-Forum 20(2), 31-51.*

MEHLER, A.; GEIBEL, P.; GLEIM, R.; HEROLD, S.; JAIN, B.; PUSTYLNIKOV, O. (2006): Much Ado About Text Content. Learning Text Types Solely by Structural Differentiae. In: *OTT'06.*

MEHLER, A.; GEIBEL, P.; PUSTYLNIKOV, O.; HEROLD, S. (2007): Structural Classifiers of Text Types. To appear in: LDV Forum.

PIEPER, U. (1975): Differenzierung von Texten nach Numerischen Kriterien. In: *Folia Linguistica VII, 61-113.*

RIEGER, B. (1989): *Unscharfe Semantik: Die empirische Analyse, quantitative Beschreibung, formale Repräsentation und prozedurale Modellierung vager Wortbedeutungen in Texten.* Peter Lang, Frankfurt a. M.

SÜDDEUTSCHER VERLAG (2004). *Süddeutsche Zeitung 1994-2003. 10 Jahre auf DVD.* München.