Text Type Structure and Logical Document Structure

Hagen Langer
Justus-Liebig-Universität/
Universität Osnabrück
hagen.langer@web.de

Harald Lüngen
Justus-Liebig-Universität
Gießen, Germany
luengen@uni-giessen.de

Petra Saskia Bayerl
Justus-Liebig-Universität
Gießen, Germany
bayerl@uni-giessen.de

Abstract

Most research on automated categorization of documents has concentrated on the assignment of one or many categories to a whole text. However, new applications, e.g. in the area of the Semantic Web, require a richer and more fine-grained annotation of documents, such as detailed thematic information about the parts of a document. Hence we investigate the automatic categorization of text segments of scientific articles with XML markup into 16 topic types from a text type structure schema. A corpus of 47 linguistic articles was provided with XML markup on different annotation layers representing text type structure, logical document structure, and grammatical categories. Six different feature extraction strategies were applied to this corpus and combined in various parametrizations in different classifiers. The aim was to explore the contribution of each type of information, in particular the logical structure features, to the classification accuracy. The results suggest that some of the topic types of our hierarchy are successfully learnable, while the features from the logical structure layer had no particular impact on the results.

1 Introduction

Our project *Semantics of Generic Document Structures* is concerned with the text type structure of scientific articles and its relations to document grammars and markup. One of its goals is to explore the feasibility of an automatic categorization of text segments of scientific articles which are annotated with logical structure tags (e.g. `section`, `paragraph`, `appendix`) into topic types such as *background, researchTopic, method,* and *results* defined in a hierarchical text type schema. The schema representing the text type structure (or, thematic structure) of scientific articles is shown in Figure 1 (explained more fully in section 2). It is assumed that authors to some degree adhere to such a schema when creating a logical structure for their documents for instance by XML markup, and that therefore such markup bears clues as to the thematic structure of scientific articles.

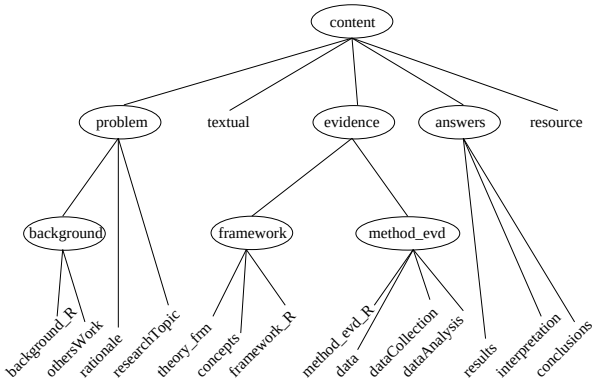

Figure 1: Text type schema

Automatic document classification is an advanced field within computational linguistics with many practical applications and has yielded a wealth of standard methods and tools over the past 10 years. Many systems have been developed on the Reuters-21578 corpus containing 21578 English newpaper articles manually categorized into 135 topic categories, e.g. (Zu et al., 2003). Besides newspaper articles, some approaches have treated the automatic categorization of (HTML) hypertexts available on the W3 into universal text topic sets such as the ones used in the LookSmart and Yahoo web directories (Dumais and Chen, 2000). An approach that focuses solely on research papers is CiteSeer (Giles et al., 1998), the online digital library of scientific articles that can be navigated via citation indices. Though the focus of CiteSeer is on the citation indices, it also provides a classification of articles from computer science into the hierarchically ordered topic set of its *computer science directory*. The bag-of-words model of document representation still prevails in automatic text categorization, but the advent of XML calls for extending this model to include logical structure features in the vector space. Yi and Sundaresan (2000) have developed a so-called semi-structured classi-

fier for semi-structured documents (e.g. XML documents) based on a document representation that combines terms with path expressions. This classifier was shown to reduce classification error rates considerably in comparison with a purely term-based classifier when run on US patent documents provided with XML markup and résumé documents in HTML. Our aim is to explore similar methods to classify thematic segments of scientific articles in XML.

2 Text type ontology for scientific articles

Previous approaches to the text type structure of scientific articles have been developed in the context of automatic text summarization. In Teufel and Moens (2002), a categorization scheme of seven "rhetorical zones" including *background, other, own, aim, textual, contrast, basis* is used for the classification of the sentences of a scientific article according to their rhetorical status and subsequently finding the most suitable ones for a summary. The schema we employ (see below) is more fine-grained, consisting of 16 categories that are hierarchically ordered (although in the present experiments we did not make use of this hierarchical order). These 16 categories refer to the dimension that is discussed under *problem structure* in (Teufel and Moens, 2002), rather than to exclusively rhetorical zones and are viewed as types of *topics*. A topic "is the semantic-pragmatic function that selects which concept of the contextual information will be extended with new information" (van Dijk, 1980, p.97). Thus while the concept *The daily newspaper Neue Zürcher Zeitung* is the topic of several sentences in the article Bühlmann (2002), it is an instance of the *topic type* `'data'` which in turn is part of the text type structure of many scientific articles. This text type structure is captured in our text type schema with 16 bottom-level topic types (Figure 1) that were obtained by evaluating articles from the disciplines of linguistics and psychology. Kando (1997) presented a similar hierarchical schema with 51 bottom-level categories, which were employed for manually annotating sentences in Japanese scientific articles. Her "text constituents" resemble our topic types, but we have aimed at sorting out functional categories such as 'Reason for...', including only purely thematic categories and keeping the number of categories lower for the experiments described in this paper.

In Figure 1, the arcs represent the *part-of* relation such that a type lower in the hierarchy is a part of the immediately dominating, more general type in terms of text type structure. The schema is supposed to represent the typical thematic structure of research papers. The order of the categories represents a canonical, expected order of topic types in a scientific article. The text type schema was initially encoded as an XML Schema grammar where topic types are represented by elements that are nested such that the XML structure reflects the structure of the text type structure tree (Figure 2).

```
<xs:element name="problem">
 <xs:complexType>
  <xs:sequence>
   <xs:element name="background" minOccurs="0">
          <xs:complexType>
             <xs:sequence>
                <xs:element name="othersWork"
                            type="xs:string"
                            minOccurs="0"/>
                <xs:element name="background_R"
                            type="xs:string"
                            minOccurs="0"/>
             </xs:sequence>
          </xs:complexType>
   </xs:element>
       ...
```

Figure 2: XML Schema grammar (extract) for the text type schema

3 Data and annotation levels

We carried out the experiments on a corpus of 47 linguistic research articles, taken from the German online journal 'Linguistik Online',[1] from the volumes 2000-2003. The selected articles came with HTML markup and have an average length of 8639 word forms, dealing with subjects as diverse as the syntax of adverbs, chat analysis, and language learning.

Taking a text-technological approach, this corpus was prepared such that all required types of information, including the target classification categories and the classification features to be extracted, are realized as XML annotations of the raw text. Thus, XML markup was provided for the thematic level, a logical structure level, and a grammatical level. As described in Bayerl et al. (2003), annotation *levels* are distinguished from annotation *layers*. An annotation level is an abstract level of information (such as the morphology and syntax levels in linguistics), originally independent of any annotation scheme. The term annotation layer, in contrast, refers to the realization of an annotation level as e.g. XML markup. There need not be a 1:1-correspondence between annotation levels and layers. As for the three annotation levels in our setting,

[1] `http://www.linguistik-online.de/`

one (the structural level) was realized as an independent layer, and two (thematic and grammatical) were realized in one single annotation layer. Each annotation layer of an article is stored in a separate file, while it is ensured that the PCDATA of each layer are identical.

3.1 Annotation of text type structure

The order of topic types in a specific scientific article may deviate from the canonical order represented in the XML schema grammar of the text type structure shown in Figure 2. Thus a flat version of the hierarchical XML schema was derived by means of an XSLT style sheet, exploiting the fact that XML schema grammars, unlike DTDs, are XML documents themselves. In the derived flat XML schema, topic types are represented as attribute values of elements called <group> and , instead of names of nested elements. Empty <group> elements represent topic types that corresponded to the nodes in the original tree of topic types, while elements correspond to leaves (terminal categories). The original hierarchical structure is still represented via the ID/IDREF attributes id and parent, similar to O'Donnell's (2000) representation of rhetorical structure trees.

For the annotation, the raw text of each article was automatically partitioned into text segments corresponding to sentences, but the annotators were allowed to modify (join or split) segments to yield proper thematic units. The problem of finding thematic boundaries other than sentence boundaries automatically (e.g. Utiyama and Isahara (2001)) is thus not addressed in this work. The annotator then provided the values of the attribute topic using the XML spy editor, choosing exactly one of the 16 terminal topic types for each segment, or alternatively the category *void_meta* for metadata such as acknowledgements. If more than one topic type could in principle be assigned, the annotators were instructed to choose the one that was most central to the argumentation. An extract from a THM annotation layer is shown in Figure 3.[2]

The two annotators were experienced in that they had received intense training as well as annotated a corpus of psychological articles according to an extended version of the schema in Figure 1 earlier (Bayerl et al., 2003). We assessed inter-rater reliability on three articles from the present linguistics corpus, which were annotated by both annotators independently according to the topic type set shown in Figure 1. (Prior to the analysis the articles were

[2]The extract, which is also shown in Figure 4, is taken from Bühlmann (2002).

```
<segment id="s75a" parent="g19"
 topic="dataAnalysis">
 Die obige Reihenfolge verändert sich etwas,
 wenn nicht die gesamte Anzahl der
 Personenbezeichnungen ausschlaggebend ist,
 sondern die Anzahl unterschiedlicher
 Personenbezeichnungen (das heisst, eine
 Personenbezeichnung wie z.B. Jugendliche,
 die acht Mal verwendet wurde, wird trotzdem
 nur einmal gezählt):
</segment>
<segment id="s76" parent="g4" topic="results">
 Im ganzen kommen in den untersuchten Artikel
 261 verschiedene Personenbezeichnungen vor.
 Davon sind über 46,7% generische Maskulina,
 und nur 31% sind Institutions- und
 Kollektivbezeichnungen. Es folgen die
 geschlechtsneutralen und -abstrakten
 Bezeichnungen mit 18,4%, und nach wie vor
 stehen die Doppelformen mit 3,8% Bezeichnungen
 am Schluss.
</segment>
```

Figure 3: THM annotation (extract)

resegmented manually so that segment boundaries were completely identical.) An average agreement of Kappa = 0.73 was reached (min: .63, max: .78), which can be interpreted as 'substantial' agreement (Landis and Koch, 1977). In order to test for annotation biases we also performed a Stuart-Maxwell Test (Stuart, 1955; Maxwell, 1970), leading to the conclusion that marginal homogeneity must be rejected on the 1% level (χ^2 = 61.24; df = 14). The McNemar Tests (McNemar, 1947) revealed that the topic types *textual*, *results*, *interpretation*, *othersWork*, and *conclusions* were the problematic categories. Subsequent log-linear analyses revealed that annotator1 systematically had assigned *background* where annotator2 had assigned *framework*. Also *interpretation* was regularly confused with *conclusions*, and *concepts* with either *background* or *othersWork* (model-fit: $\chi^2 = 173.14, df = 155, p = .15$).

3.2 Annotation of syntax and morphology

For an annotation of grammatical categories to word form tokens in our corpus, the commercial tagger Machinese Syntax by Connexor Oy was employed. This tagger is a rule-based, robust syntactic parser available for several languages and based on Constraint Grammar and Functional Dependency Grammar (Tapanainen and Järvinen, 1997). It provides morphological, surface syntactic, and functional tags for each word form and a dependency structure for sentences, and besides is able to process and output "simple" XML (that is, XML without attributes). No conflicts in terms of element overlaps can arise between our THM annotation layer

and the grammatical tagging, because all tags provided by Machinese Syntax pertain to word forms. The grammatical annotations could therefore be integrated with the THM annotations, forming the XML annotation layer that we call THMCNX. An XSLT stylesheet is applied to convert the THM annotations into attribute-free XML by integrating the information from attribute-value specifications into the names of their respective elements. After the grammatical tagging, a second stylesheet re-converts the resulting attribute-free XML representations into the original complex XML enriched by the grammatical tags. Besides, we re-formatted the original Machinese Syntax tags by omitting, merging, and renaming some of them, again using XSLT. The <cmp-head-lemma> tag (containing the lemma of the head of the present word form), for example, was derived from the original <lemma> tag, the value of which contains compound segmentation information. On the THMCNX layer, a subset of 15 grammatical tags may appear at each word form, including <pos> (part of speech), <aux> (auxiliary verb), and <num> (number feature for nominal categories).

3.3 Logical document structure annotation

Since HTML is a hybrid markup language including a mixture of structural and layout information, we chose to convert the original HTML of the corpus into XML based on the DocBook standard (Walsh and Muellner, 1999). DocBook was originally designed for technical documentation and represents a purely logical document structure, relying on style sheets to interpret the logical elements to produce a desired layout. We did not employ the whole, very large official DocBook DTD, but designed a new XML schema that defines a subset with 45 DocBook elements plus 13 additional logical elements such as `tablefootnote` and `numexample`, which appear in the annotations after the namespace prefix `log`.[3] The annotations were obtained using a perl script that provided raw DocBook annotations from the HTML markup, and the XML spy editor for validation and manually filling in elements that have no correspondences in HTML. Figure 4 shows the DocBook annotation of the extract that was also given in Figure 3.

Moreover, structural position attributes were added to each element by means of an XSLT style sheet. These '`POSINFO`' attributes make explicit the position of the element in the XML DOM tree of the document instance in an XPATH-expression as shown in Figure 5.

```
<sect2>
 ...
 <log:figure>
  <log:mediaobject>
   <xhtml:img src="buehlmann20.gif"/>
  </log:mediaobject>
 </log:figure>
 <para>Die obige Reihenfolge verändert sich
  etwas, wenn nicht die gesamte Anzahl der
  Personenbezeichnungen ausschlaggebend ist,
  sondern die Anzahl unterschiedlicher
  Personenbezeichnungen (das heisst, eine
  Personenbezeichnung wie z.B. Jugendliche, die
  acht Mal verwendet wurde, wird trotzdem nur
  einmal gezählt): Im ganzen kommen in den
  untersuchten Artikel 261 verschiedene
  Personenbezeichnungen vor. Davon sind über
  46,7% generische Maskulina, und nur 31% sind
  Institutions- und Kollektivbezeichnungen. Es
  folgen die geschlechtsneutralen und
  -abstrakten Bezeichnungen mit 18,4%, und
  nach wie vor stehen  die Doppelformen mit
  3,8% Bezeichnungen am Schluss.
  ...
 </para>
...
</sect2>
```

Figure 4: Annotation according to DocBook (extract)

```
<para POSINFO1=
  "/article[1]/sect1[4]/sect2[1]/para[10]">
  Die obige Reihenfolge verändert sich etwas,
  ...
</para>
```

Figure 5: Structural position path on the doc layer

As pointed out above, XML document structure has been exploited formerly in the automatic classification of *complete* documents, e.g. in (Yi and Sundaresan, 2000; Denoyer and Gallinari, 2003). However, we want to use XML document structure in the classification of thematic segments of documents, where the thematic segments are XML elements in the THM annotation layer. The THM and DOC layers cannot necessarily be combined in a single layer, as we had refrained from imposing the constraint that they always should be compatible, i.e. not contain overlaps. Still we had to relate element instances on the DOC layer to element instances on the THM layer.

For this purpose, we resorted to the Prolog query tool `seit.pl` developed at the University of Bielefeld in the project Sekimo[4] for the inference of re-

[3] This XML schema was designed in collaboration with the HyTex project at the University of Dortmund, `http://www.hytex.info`

[4] see (Goecke et al., 2003; Bayerl et al., 2003) and `http://www.text-technology.de/`

lations between two annotation layers of the same text. `seit.pl` infers 13 mutually exclusive relations between instances of element types on separate annotation layers on account of their shared PCDATA. In view of the application we envisaged, we defined four general relations, one of which was *Identity* and three of which were defined by the union of several more specific `seit.pl` relations:

Identity: The original identity relation from `seit.pl`.

Included: Holds if a thematic segment is properly included in a DocBook element in terms of the ranges of the respective PCDATA, i.e. is defined as the union of the original `seit.pl`-relations *included_A_in_B*, *starting_point_B* and *end_point_B*. This relation was considered to be significant because we would for example expect THM segments annotated with the topic type *interpretation* to appear within `/article[1]/sect1[5]` rather than `/article[1]/sect1[1]` elements (i.e. the fifth rather than the first `sect1` element).

Includes: Holds if a thematic segment properly includes a DocBook element in terms of the ranges of the respective PCDATA, i.e. is defined as the union of the original `seit.pl` relations *included_B_in_A, starting_point_A, end_point_A*. This relation was considered to be significant because we would for example expect logical elements such as `numexample` to be included preferably in segments labelled with the topic type *data*.

Overlap: Holds if a thematic segment properly overlaps with a DocBook element in terms of the ranges of the respective PCDATA. This relation was considered less significant because the overlapping portion of PCDATA might be very small and `seit.pl` so far does not allow for querying how large the overlapping portion actually is.

The Prolog code of `seit.pl` was modified such that it outputs XML files that contain the THM annotation layer including structural positions from the DOC layer within each segment as values of elements that indicate the relation found, cf. Figure 6.

4 Automatic text segment classification experiments

We applied different classification models, namely a KNN classifier (cf. section 4.1) and, for purposes of comparison, a simplified Rocchio classifier to text segments, in order to evaluate the feasibility of an automatic annotation of scientific articles according to our THM annotation layer. One important motivation for these experiments was to find out which kind of data representation yields the best classification accuracy, and particularly, if the combination of complementary information sources, such as bag-of-words representations of text, on the one hand, and the structural information provided by the DocBook path annotations, on the other hand, produces additional synergetic effects.

4.1 KNN classification

The basic idea of the K nearest neighbor (KNN) classification algorithm is to use already categorized examples from a training set in order to assign a category to a new object. The first step is to choose the K nearest neighbors (i.e. the K most similar objects according to some similarity metric, such as cosine) from the trainings set. In a second step the categorial information of the nearest neighbors is combined, in the simplest case, by determining the majority class.

The version of KNN classification, adopted here, uses the *Jensen-Shannon divergence* (also known as *information radius* or *iRad*) as a (dis-)similarity metric:

$$\text{iRad}(q,r) = \tfrac{1}{2}[D(q\|\tfrac{q+r}{2}) + D(r\|\tfrac{q+r}{2})]$$

$D(x\|y)$ is the Kullback-Leibler divergence (KL divergence) of probability distributions x and y:

$$D(x\|y) = \sum_{i=1}^{n} x(i)(log(x(i)) - log(y(i)))$$

iRad ranges from 0 (identity) to $2log2$ (no similarity) and requires that the compared objects are probability distributions.

Let $N_{O,C} = \{n_1, \ldots, n_m\}$ $(0 \leq m \leq K)$ be the set of those objects among the K nearest neighbors of some new object O that belong to a particular category C. Then the score assigned to the classification $O \in C$ is

$$\text{score}(O,C) = \sum_{j=1}^{m} \text{iRad}(O,n_j)^E.$$

Depending on the choice of E, one yields either a simple majority decision (if $E = 0$), a linear weighting of the iRad similarity (if $E = 1$), or a stronger emphasis on closer training examples (if $E > 1$). Actually, it turned out that very high values of E improved the classification accuracy. Finally, the KNN scores for each segment were normalized to probability distributions, in order to get comparable results for different K and E, when the KNN classifications get combined with the bigram model.

```
<segment id="s75a" topic="dataCollection">
 <included>/article[1]</included>
 <included>/article[1]/sect1[4]</included>
 <included>/article[1]/sect1[4]/sect2[1]</included>
 <includes>/article[1]/sect1[4]/sect2[1]/log:figure[5]</includes>
 <includes>/article[1]/sect1[4]/sect2[1]/log:figure[5]/log:mediaobject[1]</includes>
 <includes>/article[1]/sect1[4]/sect2[1]/log:figure[5]/log:mediaobject[1]/xhtml:img[1]</includes>
 <included>/article[1]/sect1[4]/sect2[1]/para[10]</included>
 <text>Die obige Reihenfolge verändert sich etwas, wenn nicht die gesamte Anzahl
  der Personenbezeichnungen
  ...
 </text>
</segment>
<segment id="s76" topic="results">
 <included>/article[1]</included>
 <included>/article[1]/sect1[4]</included>
 <included>/article[1]/sect1[4]/sect2[1]</included>
 <included>/article[1]/sect1[4]/sect2[1]/para[10]</included>
 <text>Im ganzen kommen in den untersuchten Artikel 261 verschiedene Personenbezeichnungen vor.
  Davon sind ...
 </text>
</segment>
```

Figure 6: Generated THMDOC layer

4.2 Bigram model

The bigram model gives the conditional probability of a topic type T_{n+1}, given its predecessor T_n.

For a sequence of segments $s_1 \dots s_m$ the total score $\tau(T, s_i)$ for the assignment of a topic type T to s_i is the product of bigram probability, given the putative predecessor topic type (i.e. the topic type T' with the highest $\tau(T', s_{i-1})$ computed in the previous step), and the normalized score of the KNN classifier. The total score of the topic type sequence is the product of its τ scores.

4.3 Information sources

In our classification experiments we used six different representations which can be viewed as different feature extraction strategies or different levels of abstraction:

- word forms (wf): a bag-of-words representation of the segment without morphological analysis; special characters (punctuation, braces, etc.) are treated as words.
- compound heads (ch): stems; in case of compounds, the head is used instead of the whole compound. These features were extracted from the THMCNX layer (cf. section 3.2).
- size (sz): number of words per segment (calculation based on the THM annotation layer, cf. section 3.1).
- DocBook paths (dbp): the segment is represented as the set of the DocBook paths which include it (the segment stand in the the *Included* relation to it as explained in section 3.3).
- selected DocBook features (df): a set of 6 DocBook features which indicate occurrences of block quotes, itemized lists, numbered examples, ordered lists, tables, and references to footnotes standing in any of the four relations listed in section 3.2.
- POS tags (pos): the distribution of part-of-speech tags of the segment taken from the THMCNX layer (cf. section 3.2).

4.4 Training and Evaluation

For each test document the bigram model and the classifier were trained with all other documents. The overall size of the data collection was 47 documents. Thus, each classifier and each bigram model has been trained on the basis of 46 documents, respectively. The total number of segments was 7330.

4.5 Results

We performed several hundred classification tests with different combinations of data representation, classification algorithm, and classifier parameter setting. Table 1 summarizes some results of these experiments. The baseline (a 'classifier' guessing always the most frequent topic type) had an accuracy of 22%.

The best combination of data representation and classifier setting achieved about 47% accuracy. In this configuration we used a mixture of the compound head representation (40%), the POS tag distribution (40%), the segment size (10%), and the selected DocBook features (10%). However, the combination of compound heads (50%) and part-of-speech tags (50%) and a similar combination in-

classifier	feature weights	K	E	accuracy classifier	accuracy classifier + bigram
most frequent	-	-	-	22.4147	-
KNN*	ch 40% pos 40% sz 10% df 10%	20	40	56.9785	-
Rocchio	ch	-	-	39.0267	-
KNN	ch 30% pos 30% dbp 40%	20	40	41.1278	41.6294
KNN	wf	20	40	38.9725	41.6429
KNN	pos	20	40	40.5314	41.9005
KNN	ch	25	40	40.4094	42.8765
KNN	ch 50% pos 50%	50	40	44.8556	45.8859
KNN	ch 50% pos 50%	13	40	44.3270	46.6179
KNN	ch 49% pos 49% dbp 2%	20	40	44.8150	46.9296
KNN	ch 40% pos 40% sz 10% df 10%	20	40	45.5063	47.0788

Table 1: Results

cluding a 2% portion of DocBook path structure features had similar results. In all experiments the KNN algorithm performed better than the simplified Rocchio algorithm. For illustrative purpose, we also included a configuration, where all other segments (i.e. including those from the *same* document) were available as training segments ('KNN*' in the second line of table 1).

The variation of classification accuracy was very high both across the topic types and across the documents. In the best configuration of our classification experiments the average segment classification accuracy per document had a range from 22% to 77%, reflecting the fact that the document collection was very heterogeneous in many respects. The topic type *resource* had an average recall of 97.56% and an average precision of 91.86%, while several other topic types, e.g. *rationale* and *dataAnalysis* were near to zero both w.r.t. precision and recall. The most frequent error was the incorrect assignment of topic type *othersWork* to segments of topic types *framework*, *concepts*, and *background*.

4.6 Discussion

The task of classifying small text segments, as opposed to whole documents, is a rather new application field for general domain-independent text categorization methods. Thus, we lack data from previous experiments to compare our own results with. Nevertheless, there are some conclusions to be drawn from our experiments.

Although the results probably suffer from limitations of our data collection (small sample size, restricted thematic domain), our main conclusion is that at least some of the topic types of our hierarchy are successfully learnable. It is, however, questionable if an overall accuracy of less than 50% is sufficient for applications that require a high reliability. Moreover, it should be emphasized that our classification experiments were carried out on the basis of manually segmented input.

The usage of structural information improved the accuracy results slightly, but the impact of this information source was clearly below our expectations. The effect of adding this kind of information was within the range of improvements which can also be achieved by fine-tuning a classifier parameter, such as K.

A somewhat surprising result was that a pure part-of-speech tag representation achieved nearly 42% accuracy in combination with the bigram model.

The usage of a bigram model improved the results in almost all configurations.

5 Conclusion

The best combination of data representation and classifier configuration included ch (40%), pos (40%), sz (10%) and df (10%), combined with a topic type bigram model, which yielded an accuracy of 47%. However, almost the same accuracy could be achieved by selecting ch and pos features only. Other test runs showed that the dbp features could not improve the results in any combination, although these features are the ones that indicate where a segment is situated in an article. An inspection of data representations revealed that, for a particular test document (i.e. text segment), the majority of training documents with an identical dpb representation are often assigned the desired topic type, but this majority is so small that many other test documents with identical dbp representation are mis-classified. An accuracy improvement might therefore be achieved by running different (local) KNN classifiers trained on different feature sets and combine their results afterwards.

More future work will focus on the inspection of categories that have a very low precision and recall (such as *rationale*) with a possible review of the text type ontology. Furthermore, we aim at testing al-

ternative algorithms (e.g. support vector machines), feature selection methods and at enlarging our training set. Besides, we will investigate the question, inhowfar our results are generalizable to scientific articles from other disciplines and languages.

References

Petra S. Bayerl, H. Lüngen, D. Goecke, A. Witt, and D. Naber. 2003. Methods for the semantic analysis of document markup. In *Proceedings of the ACM Symposium on Document Engineering (DocEng 2003)*.

Regula Bühlmann. 2002. Ehefrau Vreni haucht ihm ins Ohr... Untersuchungen zur geschlechtergerechten Sprache und zur Darstellung von Frauen in Deutschschweizer Tageszeitungen. *Linguistik Online*, 11. `http://www.linguistik-online.de`.

Ludovic Denoyer and Patrick Gallinari. 2003. Using belief networks and fisher kernels for structured document classification. In *Proceedings of the 7th European Conference on Principles and Practices of Knowledge Discovery in Databases*, Cavtat-Dubrovnik, Croatia.

Susan T. Dumais and Hao Chen. 2000. Hierarchical classification of web content. In *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval*, pages 256–263, Athens, Greece. ACM press, New York.

C. Lee Giles, Kurt Bollacker, and Steve Lawrence. 1998. CiteSeer: An automatic citation indexing system. In Ian Witten, Rob Akscyn, and Frank M. Shipman III, editors, *Digital Libraries 98 - The Third ACM Conference on Digital Libraries*, pages 89–98, Pittsburgh, PA. ACM Press.

Daniela Goecke, Daniel Naber, and Andreas Witt. 2003. Query von Multiebenen-annotierten XML-Dokumenten mit Prolog. In Uta Seewald-Heeg, editor, *Sprachtechnologie für die multilinguale Kommunikation. Beiträge der GLDV-Frühjahrstagung, Köthen 2003*, volume 5 of *Sprachwissenschaft Computerlinguistik Neue Medien*, pages 391–405, Sankt Augustin. gardez!-Verlag.

Noriko Kando. 1997. Text-level structure of research papers: Implications for text-based information processing systems. In *Proceedings of the British Computer Society Annual Colloquium of Information Retrieval Research*, pages 68–81.

J.R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

A. Maxwell. 1970. Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry*, 116:651 – 655.

Q. McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153 – 157.

A Stuart. 1955. A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 42:412 – 416.

Pasi Tapanainen and Timo Järvinen. 1997. A nonprojective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 64–71, Washington D.C. Association for Computational Linguistics.

Simone Teufel and Marc Moens. 2002. Summarizing scientfic articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Masao Utiyama and Hitoshi Isahara. 2001. A statistical model for domain-independent text segmentation. In *Meeting of the Association for Computational Linguistics*, pages 491–498.

Teun A. van Dijk. 1980. *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Norman Walsh and Leonard Muellner. 1999. *DocBook: The Definitive Guide*. O'Reilly.

Jeonghee Yi and Neel Sundaresan. 2000. A classifier for semi-structured documents. In *Proceedings of the Conference on Knowledge Discovery in Data*, pages 190–197.

Guowei Zu, Wataru Ohyama, Tetsushi Wakabayashi, and Fumitaka Kimura. 2003. Accuracy improvement of automatic text classification based on feature transformation. In Christine Vanoirbeek, C. Roisin, and Ethan Munson, editors, *Proceedings of the 2003 ACM Symposium on Document Engineering - DocEng03*, pages 118–120, Grenoble, France.