Word Clouds for Efficient Document Labeling

Christin Seifert[1], Eva Ulbrich[2], and Michael Granitzer[1,2]

[1] University of Technology Graz, Austria
christin.seifert@tugraz.at
[2] Know-Center, Graz, Austria
eulbrich, mgrani@know-center.at

Abstract. In text classification the amount and quality of training data is crucial for the performance of the classifier. The generation of training data is done by human labelers - a tedious and time-consuming work. We propose to use condensed representations of text documents instead of the full-text document to reduce the labeling time for single documents. These condensed representations are key sentences and key phrases and can be generated in a fully unsupervised way. The key phrases are presented in a layout similar to a tag cloud. In a user study with 37 participants we evaluated whether document labeling with these condensed representations can be done faster and equally accurate by the human labelers. Our evaluation shows that the users labeled word clouds twice as fast but as accurately as full-text documents. While further investigations for different classification tasks are necessary, this insight could potentially reduce costs for the labeling process of text documents.

Keywords: text classification, visualization, user interface, word clouds, document labeling, document annotation

1 Introduction

Text classification is a common task in data mining and knowledge discovery; applications include document organization and hierarchical classification of web pages [14]. Text classification is supervised learning, i.e., the classifiers is built on the basis of a training data set. The training data consists of data items and one or more category labels for each of the data items. In general, the quality and amount of training data has great influence on the performance of the final classifier [4]. The generation of training data is usually done manually by domain experts. This means, the data items are presented to domain experts, who manually assign class labels for each item - an repetitive, time consuming work. Approaches to reduce the overall labeling time can be grouped into approaches to reduce the amount of necessary training data and approaches to reduce the time for labeling a single training data item. The former include active learning [16] and semi-supervised learning strategies [20]; an example of the latter is the "labeled feature approach" [3].

In the work presented here we follow the second route by reducing the time required to label single training items for text classification. The assumption is that the information the user needs to identify the category is hidden in some key parts of the document. Conversely most parts of the document can be considered as irrelevant, they do either not contribute information for the task of finding the correct category or even distract the user from identifying the correct category. Especially for long documents, the cognitive effort for filtering irrelevant information is high. The idea of this paper is to perform this filtering of irrelevant information automatically. We develop compressed representations of text documents and investigate whether these representations are appropriate for identifying the categories accurately, but with reduced time efforts. More specifically, we use key sentences and key phrases as compressed representations, both of which can be automatically extracted from text documents using the TextRank algorithm [11]. The extracted key phrases are presented as a cloud similar to a tag cloud using a special layout for the key words.

We perform a user evaluation to investigate whether the developed representations (key sentences and key phrases) reduce the labeling time for a single document while guaranteeing that the category is still identifiable. We compare these representations to the baseline of text documents represented as full-text.

From our user evaluation we conclude that labeling key phrases is twice as fast as labeling full-text documents while the accuracy remains the same.

The remainder of this paper is structured as follows: Section 2 discusses related work for minimizing annotation time, text summarization, and keyword layout. Section 3 explains the algorithms for extracting key sentences and phrases as well as the cloud layout for the key phrases. Section 4 and 5 describe the user evaluation followed by a discussion of the results in section 6.

2 Related Work

In the field of *Machine learning*, active learning is the most prominent approach to reduce the overall number of required training data [16]. In active learning the learning algorithm itself selects the most beneficial unlabeled item and updates its classification hypothesis using the label provided by the user [16]. Active learning aims at minimizing the number of training samples to be labeled and thus reducing the overall labeling time. However there is evidence, that (i) sequential active learning may increase the number of required training samples [13], (ii) batch-mode active learning may also require more training samples than random sampling. Furthermore, Tomanek & Olsen [18] found out in their web survey that some experts in the natural language processing community do not trust active learning to work. The research of Baldridge & Palmer [2] showed that whether active learning works can depend on the experience level of the annotator. In their experiments, expert annotators performed best with uncertainty-based active learning, while non-expert annotators achieved better results using random sampling.

While active learning minimizes the number of training documents, our goal is to minimize the time the user needs for identifying the category of a single document. Thus, active learning and our condensed text representations can be easily combined. Another work for minimizing the time required for labeling single items was done by Druck et al. [3]. The authors showed that using labeled features, i.e. single words, instead of labeled text documents resulted in better classifier accuracy given limited labeling time. However, their approach is tailored towards a specific learning algorithm which may not be the algorithm of choice for a given text classification task. In contrast to their work, our approach is classifier agnostic, we efficiently generate a set of training documents that can then be used to train any classification algorithm.

Text Summarization aims at producing a shorter version of the text while retaining the overall meaning and information content. Gupta and Lehal [7] present a review for extractive summaries from texts. Extractive summaries are a selection of meaningful document parts, while abstractive summaries are shorter rephrasing of the text. We chose to use the TextRank algorithm [11] as it allows for text summarization on two different levels of granularity by extracting (i) key sentences and (ii) key phrases.

Also the field of *Information Visualization* has to offer ideas on alternative text representations [21]. Most of the visualizations show additional aspects of the text which are not instantly accessible in full-text representations. The Word Tree [19] for example, is a application of a keyword-in-context method and visualizes word concordances. In TextArc [12] word frequencies and distributions of all words in the text are visualized. These visualizations allow to interactively investigate and explore the texts, but are neither condensing the text nor designed as topical summaries. PhraseNet [9] shows interword relations and may be considered as a condensed visualization of a text as two occurrences of the same phrase are collapsed into one node in the graph. True visual text summarizations are word clouds, such as Wordle [1], or the Document Cards visualization [17]. The latter on also resembles a normal word cloud in absence of tables or images in the documents. We use a special layout algorithm for displaying our word cloud [15], which has the following properties: (i) the words are all displayed horizontally for better readability, (ii) the most important words are in the center of the visualization, (iii) there is no line-by-line alignment of the single words. We think that this special layout best resembles the nature of the extracted key phrases: There is a relation between the extracted key phrases because they originate from the same text, but the nature of the relation is unclear and the information of the sequence is lost.

3 Methodology

This section presents the methodology to evaluate the effect of different text representations on manual labeling speed and accuracy.

Figure 1 gives an overview of our methodology. Starting from text documents (on the left) three different paths for generating the three different text

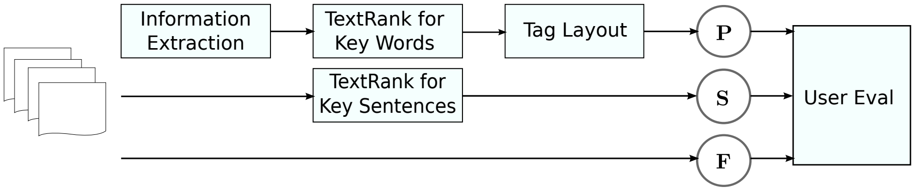


Fig. 1: Overview of the methodology

representation forms are shown. In this paper we use the word “condition” as a synonym for the text representation form, because each text representation form resembles a condition in our user evaluation. The three different conditions are denoted as **F** for full-text, **S** for key sentences (and named entities) and **P** for key phrases. In the following subsections the steps to generate the key phrases and key sentences are explained in detail. The full-text conditions serves as baseline to which we compare the users’ labeling accuracy.

3.1 Keyword and Key Sentence Extraction

We applied the TextRank algorithm [11] to extract key sentences and key words from a document. The TextRank algorithm is a graph-based ranking algorithm. The relevance of a node in the graph is determined by a voting mechanism. All predecessor nodes vote for a specific node, the score of a node is calculated from the scores of its predecessors. The final score for all nodes is determined by iteratively calculating the score for each node until the algorithm converges. To apply the TextRank algorithm, the documents need to be preprocessed. For pre-processing we used a standard information extraction pipeline consisting of the following steps: tokenization, stemming, stop-word removal, part-of-speech tagging and named entity extraction. The named entities of type “person” were added to the extracted key phrases and together they represent the key phrase condition **P** in the experiments.

TextRank for Key Sentence Extraction: For extracting key sentences the graph is constructed as follows: One node is created for each sentence. An edge between two nodes is created if their sentences are similar to each other. The similarity between two sentences is a function of their overlapping words, for instance the cosine similarity of the feature vectors of the sentences in a vector-space representation. On this weighted, undirected graph the graph-based ranking algorithm is applied. After the algorithm has converged, the nodes are sorted according to their score and the topmost nodes are selected.

TextRank for Keyword Extraction: For extracting keywords the graph is constructed as follows: (i) the text is split into tokens, (ii) part-of-speech tags are assigned to each token, (iii) for each token or all tokens for a specific part-of-speech tag a node is created, (iv) a link between two nodes is created if the words co-occur within a given window. On this unweighted, undirected graph, the graph-based ranking algorithm is applied. After the algorithm has converged, the nodes are sorted according to their score and the top T words are taken for post-processing. In the post-processing step, sequences of adjacent keywords are collapsed to multi-word keywords also termed key phrases.

3.2 Keyword Layout

The key phrases extracted by the TextRank algorithm may originate from any location of the source text. Two key phrases may belong to the same sentence and share the same context but they also may not. Consequently two key phrases have a relation as they are extracted from the same text but we do not know (anymore) which type of relation it is. We chose to use a layout for the key phrases and named entities that reflects this uncertainty in the relations. A line-by-line (Western reading-direction) layout would indicate either a relation in reading direction between the words, or none relation at all for people used to read tag clouds. We chose a layout algorithm from the family of tag layout algorithms described in [15], where the words are laid out in a circular manner, starting from the center-of-mass of the visualization boundary. The interesting property of this layout algorithm for our use case is that words are not aligned on a line and thus reading line-by-line is not possible. Compared to other words clouds, such as Wordle [1] the words are still easily readable, because all words are aligned horizontally.

4 User Evaluation

In the user evaluation we wanted to examine whether the text representation form (full-text, key sentences, key phrases) had an influence on the correctness of the labels assigned to the documents and the time required for labeling. Moreover we wanted to examine the influence of the potential mislabelings on different classifiers. In particular we tested the following hypotheses:

H1 The time required for labeling key phrases or key sentences is significantly less than for labeling full-text documents

H2 There is no difference in the number of correct labels between key phrases, key sentences and full-text.

H3 There is no difference in classifier accuracy when using labels generated in the key phrases, key sentences or full-text condition.

4.1 Design

We used a within-subjects design. The independent variable is the text representation form with three different levels (full-text **F** , key sentences **S** and key phrases **P**). We measured task completion time and correctness of the task (dependent variables). The task completion time is measured as the time difference

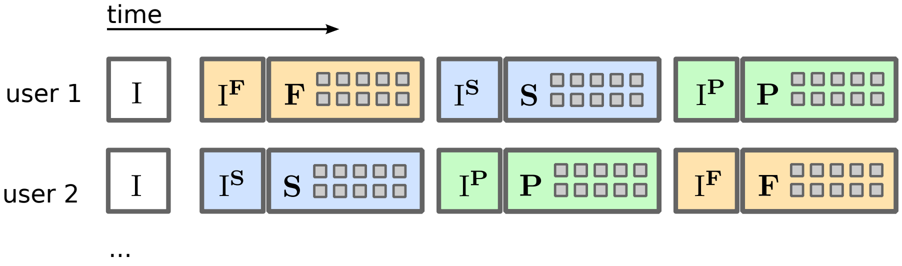


Fig. 2: Overview of the evaluation procedure, I abbreviates an Introduction step, **F** (full-text), **S** (key sentences), and **P** (key phrases) denote the different conditions.

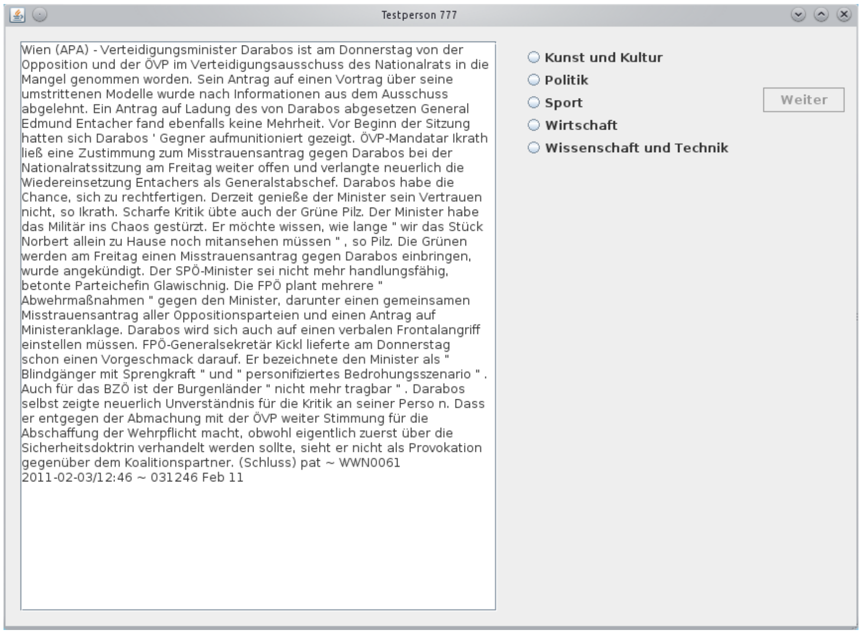
between the user first seeing the document and finishing the assignment for this document. Correctness of the task is calculated as the number of correct user labels by comparing the user labels to the ground truth of the annotated corpus.

4.2 Procedure

Figure 2 gives an overview of the evaluation procedure. For each participant, the study started with an introduction of the task and with an example document for each condition. Then the participant had time to ask questions. Thereafter the participant was asked to fill out a demographic questionnaire. Then, the three trials on the computer started. The sequence of conditions (**F**, **S** and **P**) and the documents were randomly chosen from the data set (see section 4.3 for details). For one trial (10 subsequent documents) the presentation form was the same (e.g., all documents presented as full-text). Each trial started with an introductory screen. After the participant had clicked the “OK” button, the measurements started. We measured the task completion time (the time between the two subsequent clicks on the “OK” button) and collected the labels that the participants assigned to the presented articles. For each of the three conditions, we computed the mean value for the completion time and counted the number of correct labels. Thus, for each participant $i, 1 \leq i \leq 37$ we obtained one single value for the number of correct labels l_i^c, and completion time t_i^c per condition $c \in \{\mathbf{F}, \mathbf{S}, \mathbf{P}\}$.

4.3 Test Material

We used a German news corpus from the Austrian Press Agency consisting of 27570 news articles from the year 2008. The corpus is fully labeled, i.e., each news article is annotated with one of the five classes ”economy, sports, culture, politics, science“. The articles are nearly equally distributed over the classes. The length of the articles varies between 2 and 2720 words, the average length is 247.2 words. We chose the longest articles of the corpus for our experiment, i.e. the articles longer than the 3rd quantile (> 337 words) without the statistical



Wien (APA) - Verteidigungsminister Darabos ist am Donnerstag von der Opposition und der ÖVP im Verteidigungsausschuss des Nationalrats in die Mangel genommen worden. Sein Antrag auf einen Vortrag über seine umstrittenen Modelle wurde nach Informationen aus dem Ausschuss abgelehnt. Ein Antrag auf Ladung des von Darabos abgesetzen General Edmund Entacher fand ebenfalls keine Mehrheit. Vor Beginn der Sitzung hatten sich Darabos ' Gegner aufmunitioniert gezeigt. ÖVP-Mandatar Ikrath ließ eine Zustimmung zum Misstrauensantrag gegen Darabos bei der Nationalratssitzung am Freitag weiter offen und verlangte neuerlich die Wiedereinsetzung Entachers als Generalstabschef. Darabos habe die Chance, sich zu rechtfertigen. Derzeit genieße der Minister sein Vertrauen nicht, so Ikrath. Scharfe Kritik übte auch der Grüne Pilz. Der Minister habe das Militär ins Chaos gestürzt. Er möchte wissen, wie lange " wir das Stück Norbert allein zu Hause noch mitansehen müssen " , so Pilz. Die Grünen werden am Freitag einen Misstrauensantrag gegen Darabos einbringen, wurde angekündigt. Der SPÖ-Minister sei nicht mehr handlungsfähig, betonte Parteichefin Glawischnig. Die FPÖ plant mehrere " Abwehrmaßnahmen " gegen den Minister, darunter einen gemeinsamen Misstrauensantrag aller Oppositionsparteien und einen Antrag auf Ministeranklage. Darabos wird sich auch auf einen verbalen Frontalangriff einstellen müssen. FPÖ-Generalsekretär Kickl lieferte am Donnerstag schon einen Vorgeschmack darauf. Er bezeichnete den Minister als " Blindgänger mit Sprengkraft " und " personifiziertes Bedrohungsszenario " . Auch für das BZÖ ist der Burgenländer " nicht mehr tragbar " . Darabos selbst zeigte neuerlich Unverständnis für die Kritik an seiner Perso n. Dass er entgegen der Abmachung mit der ÖVP weiter Stimmung für die Abschaffung der Wehrpflicht macht, obwohl eigentlich zuerst über die Sicherheitsdoktrin verhandelt werden sollte, sieht er nicht als Provokation gegenüber dem Koalitionspartner. (Schluss) pat ~ WWN0061
2011-02-03/12:46 ~ 031246 Feb 11

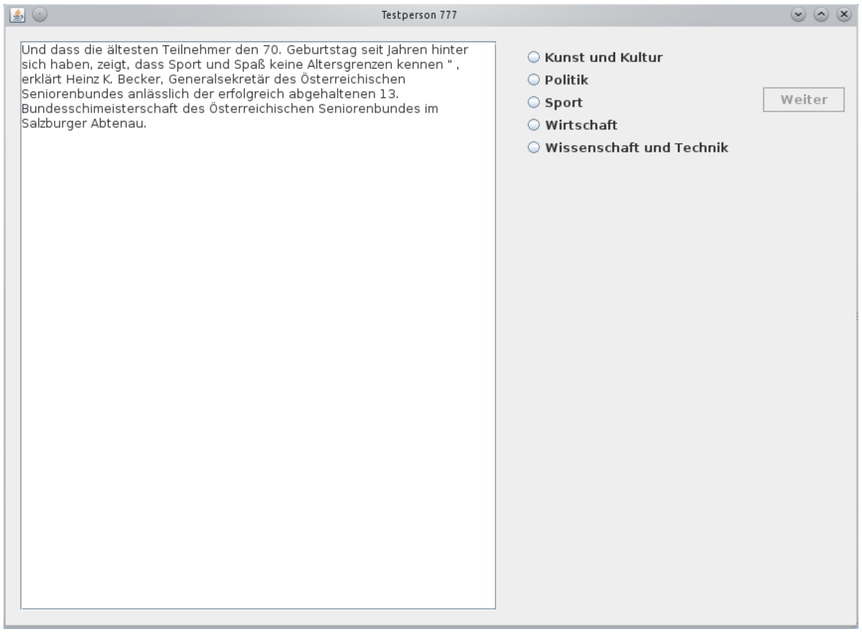
Fig. 3: Screen shots of the application for the full-text condition **F**. Data is extracted from the German test corpus.

outliers (articles with > 655 words). This leaves 6328 articles for the experiment, 1508 in class ”culture“, 1023 in ”economy“, 1409 in ”politics, 1457 in “science” and 931 in ”sports“.

For each condition a set of documents is presented to the user, we chose to take 10 documents per condition. The document set for a condition is denoted as $D^{\mathbf{F}}$, $D^{\mathbf{S}}$, $D^{\mathbf{P}}$ respectively. For a user k the sets are denoted as $D_k^{\mathbf{F}}$, $D_k^{\mathbf{S}}$, $D_k^{\mathbf{P}}$. All articles in all document sets are distinct, i.e., no users gets a document twice. For articles in set $D^{\mathbf{S}}$ key sentences, for articles in set $D^{\mathbf{P}}$ key phrases and named entities were extracted as described in section 3.1. The key sentences and the full-text were displayed in a normal text windows (see figure 3 for an full-text example and figure 4 for key sentences). The key phrases and named entities were laid out with the tag layout algorithm described in section 3.2. In order to visually separate key phrases and named entities, the key phrases were colored black and the named entities were colored blue. An example for a key phrases representation is shown in figure 5.

4.4 Participants

37 German-speaking volunteers participated in the evaluation, 18 females and 19 males. 23 of the participants were technical professionals while 14 were experts of other domains. The age of the participants ranged from 25 to 58 years (average 32.5 years).



Und dass die ältesten Teilnehmer den 70. Geburtstag seit Jahren hinter sich haben, zeigt, dass Sport und Spaß keine Altersgrenzen kennen " , erklärt Heinz K. Becker, Generalsekretär des Österreichischen Seniorenbundes anlässlich der erfolgreich abgehaltenen 13. Bundesschimeisterschaft des Österreichischen Seniorenbundes im Salzburger Abtenau.

Fig. 4: Screen shots of the application for the key sentences condition **S**. Data is extracted from the German test corpus.

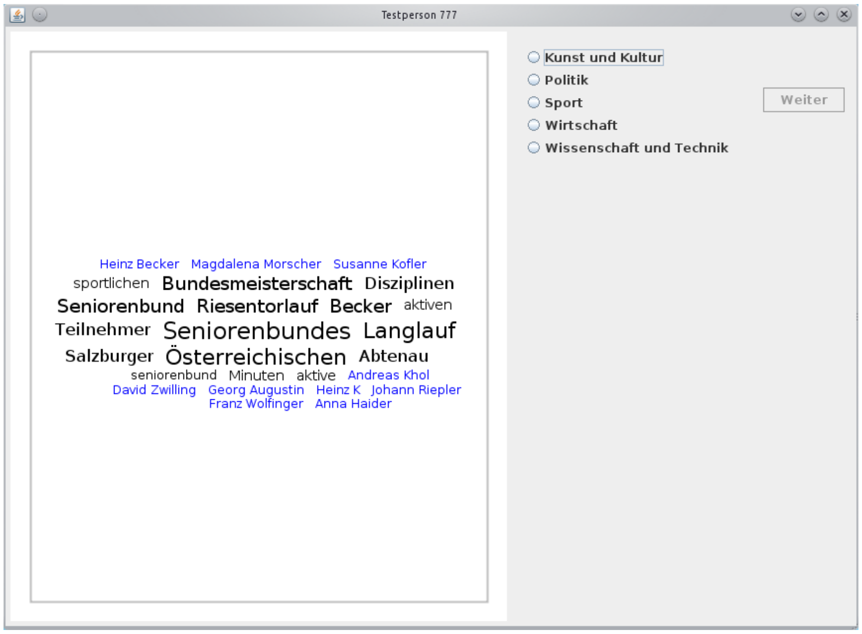


Fig. 5: Screen shots of the application for the key phrases condition **F**. Data is extracted from the German test corpus. Named entities are colored blue.

4.5 Environment

The participants were tested in a calm environment without noise distractions or additional attendees. The task was performed on a Dell Latitude e650 notebook running Windows XP Professional. The notebook was equipped with an Intel Core Duo 2.26 GHz and 3 Gb RAM. The display resolution was 1440 x 900 pixels. All users were required to use the USB mouse (and not the touch pad).

5 Results

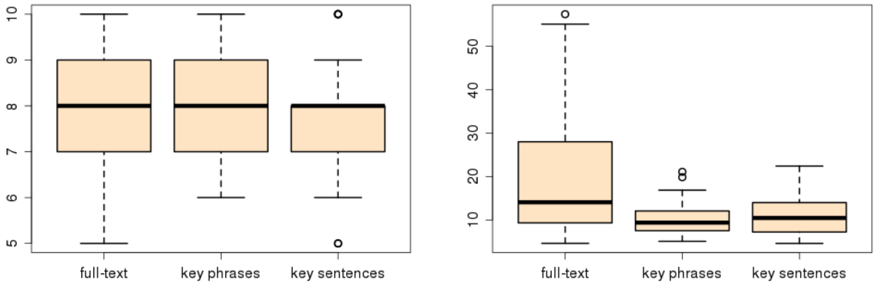
Table 1 and figure 6 summarize the measures for the number of correctly labeled examples and the task completion time. Altogether, the users assigned 290 correct labels in the full-text condition, 281 in the key sentences condition and 305 in the key phrases condition. In total 370 documents (10 documents per user, 37 users) were labeled in each condition. In the following sections we describe in detail how we tested the three hypotheses enumerated at the beginning of section 4.

Table 1: Overview of labeling time and number of correct labels (out of 10) for each condition. Values averaged over all users, showing mean and standard deviation.

	full-text	**key sentences**	**key phrases**
correct labels	7.84 ± 1.24	7.59 ± 1.38	8.24 ± 1.23
completion time [s]	19.9 ± 13.8	10.7 ± 4.4	10.4 ± 4.1

5.1 Influence on Labeling Accuracy

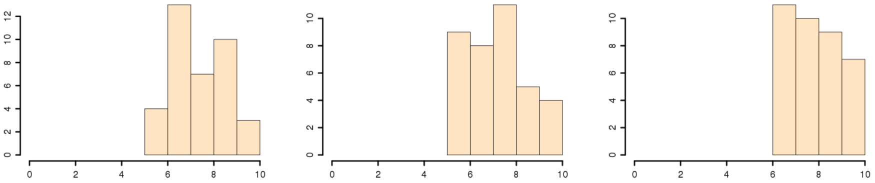
We tested whether the difference in the correct number of labels reported in table 1 are significant (Hypothesis H1). The correct number of labels is denoted as l_i^c for person i and condition c. As can be seen from the histograms of figure 7 the three variables l^f, l^s and l^p seem to be not normally distributed and thus the precondition for performing ANOVA or paired T-tests is not satisfied. However, we still tested the variables l^f, l^s and l^p for normal distribution using the Shapiro-Wilks test. All variables are not normally distributed, assuming $\alpha < .05$. Therefore, we tested on equal means with Wilcoxon rank sum test for unpaired samples. The null hypothesis for the test was that the means are equal, we set $\alpha = .05$. No difference in the mean values was found between full-text and key phrases ($W = 563$,$p = .177$) and between full-text and key sentences ($W = 754$, $p = .441$). Comparing key sentences to key phrases we found a significant difference in the mean values ($W = 504$, $p = .46$).



(a) correct labels (out of 10)

(b) task completion time in sec

Fig. 6: Box plots for task completion time and number of correct labels averaged over all users



(a) full-text

(b) key sentences

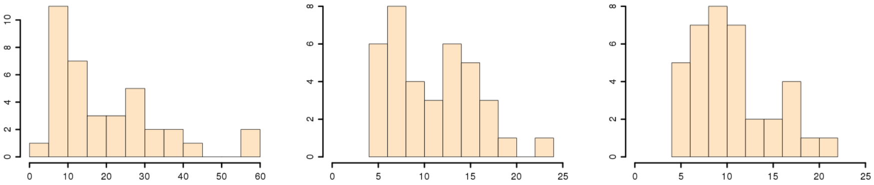
(c) key phrases

Fig. 7: Histograms of the number of correct labels averaged over all users

Summing up, we found out that users assigned significantly less correct labels when using the key sentence representation of the documents, but performed equally well with the full-text representation and the word cloud.

5.2 Influence on Labeling Time

We tested further whether the differences in task completion time reported in table 1 are significant (Hypothesis H2). The average time for labeling is denoted as t_i^c for person i and condition c. As can be seen from the histograms of figure 8 the three variables t^f, t^s and t^p seem to be not normally distributed and thus the precondition for performing ANOVA or paired T-tests is not satisfied. However, we still tested the variables t^f, t^s and t^p for normal distribution using the Shapiro-Wilks test. All variables are not normally distributed, assuming $\alpha < .05$. Therefore, we tested on equal means with Wilcoxon rank sum test for unpaired samples. The null hypothesis for the test was that the means are equal,



(a) full-text (b) key sentences (c) key phrases

Fig. 8: Histograms for the task completion times averaged over all users

we set $\alpha = .05$. No difference in the mean values was found between the full-text and key sentences ($W = 705$, $p = .830$). On the contrary, we found a significant difference comparing full-text and key phrases ($W = 956$, $p = .003$) and full-text and key sentences ($W = 982$, $p = .001$).

Summing up, we found out that users labeled the items significantly faster when using the key sentence or the key phrases representation than when using the full-text representation of the documents.

5.3 Influence on Classifier Accuracy

As reported in section 5.1 we found out that users labeled less accurately when using the key sentence representation of the text documents. We further wanted to test, whether this mislabeling would have an influence on classifiers trained on the erroneous labels (Hypothesis H3). To do so, we created two different training data sets for each condition, resulting in 6 different training data sets. Both training sets for one condition contained the documents processes by all users in this condition, one was extended by the original labels (the ground truth) and the other one was extended by the user labels. We further created an evaluation data set of 6000 randomly selected items from the data set. None of the evaluation items was contained in any of the training data sets. We trained various classifiers on both training data sets for each condition, and evaluated the trained classifiers on the evaluation data set. a_o^c denotes the accuracy of the classifier trained on original labels, a_u^c denotes the accuracy of the classifier trained on user labels for condition c. We used the following classifiers:

- bagging with decision stumps (denoted Bagging-DT) and AdaBoost with decision stumps (denoted Adaboost-DT) from the Mallet machine learning library [10]
- Naive Bayes and Hyperpipes from the WEKA machine learning library [8]
- the LibLinear library [5]
- our own implementations of the k-Nearest Neighbor classifier (denoted KNN-10 for k=10, and KNN-20 for k = 20) and the class-feature-centroid classifier [6] (denoted CFC)

Table 2: Classifier accuracy when trained on original labels (a_o) versus trained on user labels (a_u)

	full-text		**key sentences**		**key phrases**	
classifier	a_o^f	a_u^f	a_o^s	a_o^s	a_o^p	a_u^p
KNN-10	0.76	0.72	0.77	0.73	0.76	0.73
Bagging-DT	0.45	0.45	0.51	0.48	0.47	0.45
LibLin	0.80	0.74	0.80	0.76	0.79	0.74
KNN-20	0.75	0.71	0.76	0.73	0.76	0.72
Adaboost-DT	*0.36*	*0.41*	0.39	0.38	0.33	0.31
NaiveBayes	0.81	0.77	0.78	0.76	0.79	0.76
CFC, b=2.3	0.78	0.73	0.78	0.73	0.78	0.72
Hyperpipes	0.78	0.72	0.77	0.71	0.77	0.67

Table 2 reports the accuracy of the classifiers on the evaluation data set. Not surprisingly, the accuracy of the classifier trained on user labels was lower in nearly every case than when trained on the original (ground truth) labels. This is because the ground truth was labeled by domain experts and we did not explicitly communicate the rules for assignin an article to a specific category. Thus, for the boundary articles, i.e., news about a politician attending a sports event, the decision whether the article belongs to category ”sports“ or ”politics” was subjective. Because all articles were randomly selected and aligned to the three conditions this effect is likely to occur equally often in all conditions. The one exception is the Adaboost classifier in the full-text condition. However, this is also the classifier that performs worst for this classification task.

Table 3: Comparing original labels and user labels: Difference in number of correct labels and classifier accuracy (mean and standard deviation)

	full-text	**key sentences**	**key phrases**
Δcorrect labels	71	80	65
Δa	0.034 ± 0.037	0.034 ± 0.017	0.040 ± 0.022

Table 3 reports the differences in classifier accuracy averaged over all classifiers for the three conditions. When using the user-labels the accuracy decreases by less than 4% in all conditions. The difference in accuracy for the key phrases seems to be larger ($\Delta a^p = 0.040$) than for the sentence and full-text conditions ($\Delta a^s = 0.034$, $\Delta a^f = 0.034$). We investigated whether these differences are statistically significant. First we tested the variables Δa^f, Δa^s and Δa^p for normal distribution using the Shapiro-Wilks test ($\alpha = 0.05$). The two variables Δa^s and Δa^p follow a normal distribution, but Δa^f does not. This means, the preconditions for calculating ANOVA or paired t-Tests was not fulfilled. Therefore we used the Wilcoxon rank sum test for unpaired samples to compare the mean values using $\alpha = .05$. We found no significant difference between any of the conditions, the test statistics are as follows: full-text vs key phrases $W = 39$, $p = .462$, full-text vs key sentences $W = 34$, $p = .833$, key sentences vs key phrases $W = 29$, $p = .753$.

To sum up, we found no influence of the different representation forms on classifier accuracy.

6 Discussion

In this section we discuss our hypotheses outlined at the beginning of section 4 in the light of the results of the previous section. The evaluation showed that users can label key words twice as fast but with the same accuracy as full-text documents. Labeling of key sentences is fast too, but the labeling accuracy is significantly lower than in the full-text condition. This means we can accept hypotheses H1: that a compressed representation leads to faster decisions, regardless whether this decision is correct or not. Hypothesis H2 must be rejected, there is a difference in the number of correct labels when varying the representation form. More specifically, users are most accurate when using full-text or key phrases, indicating that the TextRank algorithm for keyword extraction performs well in filtering out information irrelevant for text categorization while keeping the information required to identify the category. On the contrary, the labeling accuracy for key sentences is significantly lower, indicating that key sentences are less informative on average, obviously either irrelevant or ambiguous sentences are extracted. In our experiments we found no influence of this different labeling accuracy on classifier performance confirming hypothesis H3. On the one hand this might be due to the noise tolerance of the used classifiers and the practically low amount of noise. In our experiment, it makes no difference for the classifier whether 65 or 80 out of 370 documents are labeled incorrectly. We expect this difference to become significant when the number of training items (and thus the number of mislabeled items) increases.

Summing up, our evaluation shows that: *Key phrases are a fast and accurate representation for document labeling. In fact, users labeled key phrases twice as fast and as accurately as full-text documents.*

7 Conclusion and Future Work

We investigated two different condensed representations of text, key phrases and key sentences, for the purpose of faster document labeling. Both representation forms can be generated in a fully automatic way. In a user evaluation we compared the labeling accuracy and time of the users when using these condensed representations to the baseline, the full-text representation of the texts. Our evaluation shows that the users labeled key phrases twice as fast but as accurately as full-text documents. This finding points toward a feasible way to decrease the time and cost for the generation of training data. Word clouds for labeling can be easily combined with other approaches such as active learning.

Further experiments are necessary to investigate the benefit for other classification tasks. Directions of experiments include: different languages (other than German), hierarchical and/or multi-label classification problems. Further, the process of extracting the condensed information (keyword extraction) as well as the presentation (number of keywords to show, layout algorithm) can be varied. During the user evaluation we got the impression, that different users used different reading patterns ranging from sequential word-by-word reading to scanning. We plan an eye-tracking study to investigate to which extend the reading patterns influence the efficiency of the word cloud representation. Following this direction, an application can then implement a combined or adaptive user interface: the initial representation is the word cloud, once the user feels that the presented information is insufficient to identify the label she can request the full-text article.

Acknowledgement

The Know-Center is funded within the Austrian COMET Program under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

1. Wordle - Beautiful Word Clouds. www.wordle.net, http://www.wordle.net, accessed: 2011-04-25
2. Baldridge, J., Palmer, A.: How well does active learning actually work?: Time-based evaluation of cost-reduction strategies for language documentation. In: Proc. of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 296–305. Association for Computational Linguistics, Morristown, NJ, USA (2009)
3. Druck, G., Mann, G., McCallum, A.: Learning from labeled features using generalized expectation criteria. In: SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 595–602. ACM, New York, NY, USA (2008), http://portal.acm.org/citation.cfm?id=1390436#
4. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification (2nd Edition). Wiley-Interscience, 2 edn. (November 2000)
5. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. J. Mach. Learn. Res. 9, 1871–1874 (2008)
6. Guan, H., Zhou, J., Guo, M.: A class-feature-centroid classifier for text categorization. In: Proc. of the International conference on World Wide Web (WWW). pp. 201–210. ACM, New York, NY, USA (2009)
7. Gupta, V., Lehal, G.: A survey of text summarization extractive techniques. Journal of Emerging Technologies in Web Intelligence 2(3) (2010), http://ojs.academypublisher.com/index.php/jetwi/article/view/0203258268

8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. 11, 10–18 (November 2009), http://doi.acm.org/10.1145/1656274.1656278
9. van Ham, F., Wattenberg, M., Viegas, F.B.: Mapping text with phrase nets. IEEE Transactions on Visualization and Computer Graphics 15, 1169–1176 (November 2009), http://dx.doi.org/10.1109/TVCG.2009.165
10. McCallum, A.K.: Mallet: A machine learning for language toolkit (2002), http://mallet.cs.umass.edu
11. Mihalcea, R., Tarau, P.: Textrank: Bringing order into texts. In: Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain (2004), http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Mihalcea.pdf
12. Paley, W.: Textarc: Showing word frequency and distribution in text. In: Proceedings of IEEE Symposium on Information Visualization, Poster Compendium. IEEE CS Press (2002)
13. Schein, A.I., Ungar, L.H.: Active learning for logistic regression: an evaluation. Mach. Learn. 68(3), 235–265 (2007)
14. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002), citeseer.ist.psu.edu/sebastiani02machine.html
15. Seifert, C., Kump, B., Kienreich, W., Granitzer, G., Granitzer, M.: On the beauty and usability of tag clouds. In: Proceedings of the 12th International Conference on Information Visualisation (IV). pp. 17–25. IEEE Computer Society, Los Alamitos, CA, USA (July 2008)
16. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2010), http://pages.cs.wisc.edu/~bsettles/active-learning
17. Strobelt, H., Oelke, D., Rohrdantz, C., Stoffel, A., Keim, D.A., Deussen, O.: Document cards: A top trumps visualization for documents. IEEE Transactions on Visualization and Computer Graphics 15, 1145–1152 (2009)
18. Tomanek, K., Olsson, F.: A web survey on the use of active learning to support annotation of text data. In: Proc. of the NAACL Workshop on Active Learning for Natural Language Processing (HLT). pp. 45–48. Association for Computational Linguistics, Morristown, NJ, USA (2009)
19. Wattenberg, M., Viégas, F.B.: The word tree, an interactive visual concordance. IEEE Transactions on Visualization and Computer Graphics 14, 1221–1228 (November 2008), http://portal.acm.org/citation.cfm?id=1477066.1477418
20. Zhu, X.: Semi-supervised learning literature survey. Tech. Rep. 1530, Computer Sciences, University of Wisconsin (2008), http://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf
21. Šilić, A., Bašić, B.: Visualization of text streams: A survey. In: Setchi, R., Jordanov, I., Howlett, R., Jain, L. (eds.) Knowledge-Based and Intelligent Information and Engineering Systems, Lecture Notes in Computer Science, vol. 6277, pp. 31–43. Springer Berlin / Heidelberg (2010)