Investigating Word Correlation at Different Scopes – a Latent-Concept Approach

Gregor Heinrich GREGOR@ARBYLON.NET
Arbylon, Nußbaumallee 25, 64297 Darmstadt, Germany;
Fraunhofer Insitute for Computer Graphics (IGD), Fraunhoferstraße 5, 64283 Darmstadt, Germany

Jörg Kindermann JOERG.KINDERMANN@AIS.FRAUNHOFER.DE
Fraunhofer Institute for Autonomous Intelligent Systems, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

Codrina Lauth CODRINA.LAUTH@AIS.FRAUNHOFER.DE
Fraunhofer Institute for Autonomous Intelligent Systems, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

Gerhard Paaß GERHARD.PAASS@AIS.FRAUNHOFER.DE
Fraunhofer Institute for Autonomous Intelligent Systems, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

Javier Sanchez Monzon JAVIER.SANCHEZ-MONZON@AIS.FRAUNHOFER.DE
Fraunhofer Institute for Autonomous Intelligent Systems, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

Abstract

This paper presents work in progress on clustering methods that identify semantic concepts in a document collection. These methods are based on the observation that semantically related words occur close together. We investigate the size of neighborhood which should be taken into account for this purpose: sentences or documents. We further investigate how local co-occurrence affects the clustering quality by including word bigrams as additional terms. We apply two different latent-concept models, probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA), to a corpus of German news stories. The resulting soft clusterings are compared with a given a priori classification of documents using an information-based distance metric. Preliminary results show that this cluster distance was smaller using (1) entire documents (compared to combinations of documents and sentences), as well as (2) combinations of unigrams and bigrams (compared to exclusive use of unigrams or bigrams).

Appearing in *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

1. Introduction

With the advent of powerful computers and the proliferation of the Web, the automatic extraction of information from text documents became possible at a large scale. Besides classifying text into pre-specified categories, *text mining* may be used to find meaningful partitions of the content feature space without any preclassification. Both approaches provide powerful tools to support the task of ontology creation from unstructured text for one of the most promising extensions to the current Web, the "Semantic Web".

In this paper, we investigate text mining methods that are able to group the words in documents into a large number of semantic concepts in an automatic way. This approach is based on the observation that words related to a semantic concept have the tendency to occur "close" to each other. Often all words of a document are considered to be close in this sense, assuming that the whole document is devoted to a specific topic. In many cases, however, a shift of concepts occurs within a document. Therefore it is not clear what notion of closeness automatic clustering procedures for detecting semantic concepts should use: bigrams, sentences, paragraphs, documents or something else.

Our work focuses on the question whether the extraction of concepts yields better results if they are based on short-range correlations (sentences) or long-range correlations (whole documents). Additionally,

we generalize terms considered for concept extraction to "word n-gram combinations", namely word unigrams, word bigrams and combinations thereof. Bigrams can be considered to capture correlation at a very short distances, and we investigate the influence of this correlation on concept extraction, as well.

We extract concepts from documents with two different approaches that characterize a concept by a distribution over the words belonging to the concept: probabilistic latent semantic indexing (PLSI) and Latent Dirchlet allocation (LDA). We generate concepts using these approaches based on the words co-occuring within longer ranges (whole documents), shorter ranges (sentences), or both. We compare the resulting concepts by a cluster similarity metric. This allows to determine whether the derived concepts on the document level are better able to reconstruct the given topic categories of the documents or not.

In the following section, we take a closer look at long- and short-range word correlations from a linguistic perspective. Section 3 will give an overview of latent concept extraction methods and a metric to compare clusters. In Section 4, we describe our data, the experimental setups and the results. In the final section, the results are discussed and summarized.

2. Long and Short-Range Word Correlations from a Linguistic Perspective

We know from the linguistic perspective that the meaning of words does not come only from the word itself, but especially from its usage in a certain context, e.g., sentence, paragraph or document level have to be taken all into account. "The major principle is that the unit of meaning is not located at the level of the word, but at the level of elementary sentences." [1].

The major challenge for the natural language processing lies in defining the right patterns for extracting the meaning of words in different contexts, especially in identifying semantic relations between words. In terms of representation of meaning in context, we speak here in the abstract form about "*concepts*" (or "*topics*"). In this paper, we employ the distributional statistic's approach of representing concepts/topics by correlated words within longer ranges (whole documents) or shorter ranges (elementary sentences). Salton and McGill already stated that "the distribution of words in a document is related to its topics" [2].

Statistical language modelling is concerned with modelling the semantic coherence between words in context, independent of the size of the context. Window models are insufficient. The classical "bag of words models" capture short distance correlations quite well, but long distance correlations of words, e.g., global sentence features, synonymy, polysemy are insufficiently solved or even not considered.

3. Latent Concept Extraction

The key property of latent concepts is that they tend to describe text content closer to meaning than literal terms. This can be used to resolve linguistic phenomena like polysemy and synonymy, which depict problems in many text processing fields, such as information retrieval or ontology generation from text data. An intuitive explanation is that a text corpus has an underlying structure of latent concepts, which is imagined to be obscured by "word-choice" noise.

There have been several approaches to implement extraction of latent concepts from text. One of the first was the Latent Semantic Analysis (LSA) approach by Deerwester et al. [3], which used singular value decomposition (SVD) to extract latent components from a term-document occurrence data. If a group of terms often occurs together in a document, it is represented in one or more of these latent concepts, which corresponds to a linear combination of terms. Empirically, this approach was quite successful, but the Gaussian noise assumption in the SVD-based approach is suboptimal with respect to the true statistical behaviour of term frequency data [4, 5].

3.1. Probabilistic Latent Semantic Analysis

To resolve the shortcomings of the initial LSA approach, Hofmann [4] developed a probabilistic model for document corpora. It assumes the existence of a latent variable z with k different values and specifies the following generative model for the words w in a document d_i, i=1, . . . , n_d:

$$p(w|d_i) = \sum_{r=1}^{k} p(w|z{=}r)p(z{=}r|d_i)\,. \qquad (1)$$

Therefore, the probability that a word occurs in document d_i is a mixture distribution of the $p(w|z{=}r)$. The $n^{(d_i)}$ different words of document d_i are generated independently. Fig. 1 (a) shows the corresponding Bayesian network.

Each value $z{=}r$ describes a latent concept. Obviously, the terms in document d_i may stem from several latent concepts, and for each document there is a distribution $p(z|d_i)$ over the concepts involved. This distribution characterizes the concepts in the document and can be

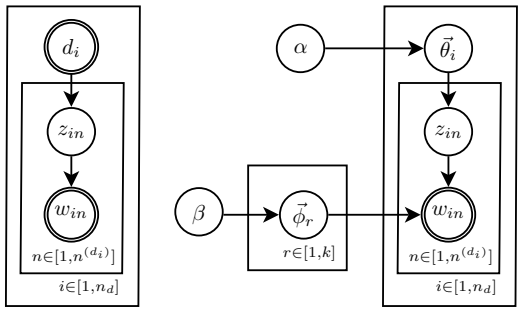

Figure 1. Bayesian networks of (a) PLSA and (b) LDA

considered a soft clustering of the documents. From a dual view, terms with high probabilities $p(w|z{=}r)$ with respect to a latent concept $z{=}r$ can be considered a soft clustering of terms, which, for instance, could include terms with similar meanings or synonyms. On the other hand, the same term w may be associated with different concepts, which may happen if it is polysemic, i.e., has different meanings.

Although this model is a vast simplification, it leads to meaningful concepts, as terms that often occur together have a high probability $p(w|z{=}r)$ with respect to some concept. Hofmann has shown experimentally that PLSA achieves a higher perplexity reduction than LSA [4].

3.1.1. PLSA with Overlapping Units

As discussed in Section 2, a document may be grouped into different units, e.g., paragraphs and sentences. We assume that within the subunits the distribution of latent concepts may be varying. Let $s_{i,1}, \ldots, s_{i,m_i}$ be the sentences of document d_i. In this paper, we consider sentences to be generated using the same latent concepts as in the whole document. Then we may assume the following generative model:

$$
\begin{aligned}
&p(w|d_i, s_{il}) \\
&\quad = \sum_{r=1}^{k} p(w|z{=}r)p(z{=}r|d_i, s_{il}) && (2)\\
&\quad = \sum_{r=1}^{k} p(w|z{=}r)(p(z{=}r|d_i)\gamma + \\
&\qquad\qquad p(z{=}r|s_{il})(1-\gamma)) && (3)\\
&\quad = \gamma \sum_{r=1}^{k} p(w|z{=}r)p(z{=}r|d_i) + \\
&\qquad (1-\gamma) \sum_{r=1}^{k} p(w|z{=}r)p(z{=}r|s_{il})\,. && (4)
\end{aligned}
$$

This model posits that the number of documents, the number of words in the documents, the number of sentences in a document as well as their lengths are independent of the contents (these assumptions may be relaxed). Fig. 2 (a) shows an equivalent Bayesian network, which compared to Fig. 1 (a) has an additional sentence plate and models word w_{ln} with sentence-based indices running over the $n_s^{(d_i)}$ sentences in a document d_i.

If a new word in document d_i is generated, we first decide with probability γ whether it is generated according to the document-specific latent class distribution $p(z{=}r|d_i)$ or the sentence-specific latent class distribution $p(z{=}r|s_{il})$. (In the Bayesian network, a choice variable c_{ln} is introduced for this.) Subsequently, the selected latent class distribution is used to generate a latent class $z{=}r$, and finally a new word is generated according to $p(w|z{=}r)$.

The document-specific latent class distribution $p(z{=}r|d_i)$ takes into account long-distance correlations between all terms in a document, whereas the sentence-specific latent class distribution $p(z{=}r|s_{il})$ only reflects the terms of a single sentence and hence short-distance correlations.

If we set $\gamma{=}1$, then we are back at the usual PLSA-Model. With $\gamma{=}0$ we have a sentence model. We can estimate γ from the data and thus determine the relative influence of short- and long-distance correlations. In this paper, we fix γ to 0.5 and compare the relative quality of clusterings under this assumption. This reduces to a version of PLSA, where weights for single words are used. Later we may estimate γ to determine the importance of short- and long-range correlations.

3.2. Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) by Blei et al. [6] extends the generative model of PLSA by defining the concept-specific multinomial term distributions $p(w|z)$ and document-specific mixture weights $p(z|d)$ as random variables themselves, following a Bayesian approach.

More specifically, LDA defines a generative model that includes Dirichlet-distributed priors over the masses of the multinomials $p(w|z)$ and $p(z|d)$. Fig. 1 (b) shows the corresponding Bayesian network. For the generation of document mixture weights, the multinomial $p(z|d_i)$, which in PLSA is an empirical distribution conditioned on the index d_i, becomes a distribution $p(z|\vec{\theta}_i)$, conditioned on a vector of parameters $\vec{\theta}_i$, which are sampled from a Dirichlet distribution $p(\vec{\theta}_i|\alpha)$ with hyperparameter α. Generating such a document-specific concept mixture from a prior also allows to estimate the concepts of previously unknown documents after training, which is not directly possible in PLSA.

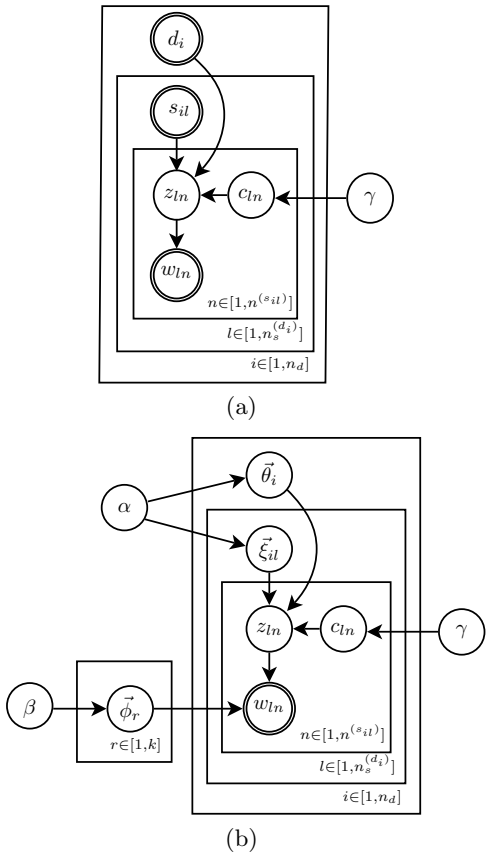

Figure 2. Bayesian networks of (a) PLSA and (b) LDA with overlapping units.

In a similar manner, for $p(w|z{=}r)$ a term distribution $p(w|z{=}r,\vec{\phi}_r)$ is introduced with a parameter vector $\vec{\phi}_r$ for each concept $z{=}r$, sampled from a Dirichlet distribution $p(\vec{\phi}_r|\beta)$ with hyperparameter β (see [6] for details). If the hyperparameters α and β are not trained, appropriate choice of values allows parameter smoothing to avoid overfitting.

On the down-side, exact inference is generally intractable in LDA, and several approximate inference algorithms have been proposed, e.g., mean-field variational EM [6], expectation propagation [7], and Gibbs sampling [8].

For overlapping units, similar considerations apply as described above for PLSA. That is, an initial generation step is introduced that decides with probability γ that a word is chosen from the latent class distribution specific to document d_i, $\vec{\theta}_i$, and with probability $1-\gamma$ that it is chosen from the distribution specific to sentence s_{il}, for which we have introduced the sentence-specific parameter vector $\vec{\xi}_{il}$. Fig. 2 (b) shows a Bayesian network equivalent to this approach.

3.3. A Measure for Comparing Clusterings

There are a number of different measures to compare clusterings on documents. In our case, we have soft clusterings where for each document there exists a distribution of clusters or concepts. Virtually all criteria for comparing clusterings can be described using the so-called *confusion matrix*. For each document, it records which clusters occur together and averages this over the corpus.

In this paper, we adopt the cluster distance measure proposed by Meila [9]. Assume we have documents $d_1,\ldots,d_n$. A soft clustering C assigns to each document d_i a distribution $p(c{=}r|d_i)$ for $r{=}1,\ldots,k$. If we have a second clustering $\tilde{C}$, we have a new distribution $p(\tilde{c}{=}r|d_i)$ for $r{=}1,\ldots,\tilde{k}$. Note that both clusterings may have different number k and $\tilde{k}$ of clusters.

If the clusterings are very similar, there will be pairs of clusters that will often occur together. On the other hand, if both clusterings are independent, the pairs of clusters $c{=}r$ and $\tilde{c}{=}s$ will appear with probability $p(c{=}r)p(\tilde{c}{=}s)$. Therefore, we may determine the Kullback-Leibler divergence between this "independent" distribution and the actual distribution $p(c{=}r,\tilde{c}{=}s)$. This is just the mutual information between the random variables induced by the clusterings

$$I(C,\tilde{C}) = \sum_{r=1}^{k}\sum_{s=1}^{\tilde{k}} p(c{=}r,\tilde{c}{=}s)\log\frac{p(c{=}r,\tilde{c}{=}s)}{p(c{=}r)p(\tilde{c}{=}s)} \quad (5)$$

We may determine the required probabilities by averaging over the distributions of clusters in documents

$$p(c{=}r) = \frac{1}{n}\sum_{i=1}^{n} p(c{=}r|d_i)$$

$$p(c{=}r,\tilde{c}{=}s) = \frac{1}{n}\sum_{i=1}^{n} p(c{=}r,\tilde{c}{=}s|d_i)$$

The following properties are well-known: $I(C,\tilde{C}){=}I(\tilde{C},C)\geq 0$; if $I(C,\tilde{C}){=}0$ then both clusterings are independent; $I(C,\tilde{C})\leq\min\{H(C),H(\tilde{C})\}$ where $H(C){=}-\sum_{r=1}^{k}p(r)\log p(r)$ is the entropy; $I(C,\tilde{C}){=}H(C){=}H(\tilde{C})$ if the two clusterings are equal. To arrive at a *cluster distance* measure Meila defines

$$D(C,\tilde{C}) = H(C)+H(\tilde{C})-2I(C,\tilde{C}) \quad (6)$$

and shows that $D(C,\tilde{C})$ has a property of a metric: non-negativity $D(C,\tilde{C})\geq 0$ and $D(C,\tilde{C}){=}0$ iff $C{=}\tilde{C}$; symmetry: $D(C,\tilde{C}){=}D(\tilde{C},C)$;and the triangle inequality: $D(C,\tilde{C})+D(\tilde{C},B)\geq D(C,B)$. Meila [9] discusses the properties of this metric and compares it to other approaches.

4. Experiments

We applied the algorithm to newswire articles from the Deutsche Presseagentur (dpa) from the year 2000. A stop word lists was generated from a statistical analysis of the corpus and manual extraction of frequent terms that were neutral to document topics. After removing about 500 stopwords as well as words appearing up to three times there remained about 140000 different words. We did not use stemming or lemmatization.

We varied the experiments with respect to several conditions:

- The number of documents used: either 10000, 20000, 50000 and the complete dpa-corpus with 229500 documents.
- The number of latent concepts, k: 20, 50, 100 and 200.
- The terms: only word unigrams or word unigrams combined with bigrams. We did not investigate experiments with only bigrams yet.
- Combinations of experiments with only documents, documents and sentences and only sentences.
- The algorithm used for the determination of latent concepts: PLSA or LDA.

For PLSA, the algorithm was stopped if the likelihood decreased on a crossvalidation sample. For LDA, an approach based on a Gibbs sampler has been taken, similar to the one described in [8]. We did not train the hyperparameters but rather varied them a priori and chose values that yielded good results for a variety of experimental conditions (α=0.01 and β=0.5). LDA parameter estimation is stopped after the sampler has reached a stationary distribution. An additional convergence criterion for both PLSA and LDA is the stability of the cluster distance taken over different training iterations. The computation time depends on the number of factors and the corpus size and varied for both PLSA and LDA between 30 minutes and about 30 hours.

The documents of the dpa corpus are pre-categorized according to the IPTC category code involving about 300 categories. Note that more than one category could be assigned to a document. Giving each of the categories the same weight the classification inherently defined a distribution for each document, which could be compared to the distribution of latent concepts generated by PLSA or LDA using the cluster metric.

Table 1. Selected Latent Concepts for the dpa Corpus

Group Name	items
political parties	CDU Partei Kohl Aufklärung Schäuble Zeitung Union Krise Wahrheit Affäre Christdemokraten Glaubwürdigkeit Konsequenzen
soccer prem. league	FC SC München Borussia SV VfL Kickers SpVgg Uhr A Köln Bochum Freiburg VfB Eintracht Bayern Hamburger Bayern+München
police, accident	Polizei verletzt schwer Auto Unfall Fahrer Angaben schwer+verletzt Menschen Wagen Verletzungen Lawine Mann vier Meter Strasse
Tchechnia	Rebellen russischen Grosny russische Tschetschenien Truppen Kaukasus Moskau Angaben Interfax tschetschenischen Agentur
politics in Hesse	FDP Koch Hessen CDU Koalition Gerhardt Wagner Liberalen hessischen Westerwelle Wolfgang Roland+Koch Wolfgang+Gerhardt
wheather	Grad Temperaturen Regen Schnee Süden Norden Sonne Wetter Wolken Deutschland zwischen Nacht Wetterdienst Wind
politics in Croatia	Parlament Partei Stimmen Mehrheit Wahlen Wahl Opposition Kroatien Präsident Parlamentswahlen Mesic Abstimmung HDZ
green party	Grünen Parteitag Atomausstieg Trittin Grüne Partei Trennung Mandat Ausstieg Amt Röstel Jahren Müller Radcke Koalition
Russian politics	Russland Putin Moskau russischen russische Jelzin Wladimir Tschetschenien Russlands Wladimir+Putin Kreml Boris Präsidenten
police in schools	Polizei Schulen Schüler Täter Polizisten Schule Tat Lehrer erschossen Beamten Mann Polizist Beamte verletzt Waffe

Hence we can define $p(c{=}r|d_i)$ for the classification as a multinomial with

$$p(c{=}r|d_i) = \begin{cases} 1/n_c^{(d_i)} & \text{if } d_i \text{ in IPTC category } c{=}r, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $n_c^{(d_i)}$ is the number of IPTC categories associated with d_i. The estimated distribution of concepts is directly derived from the conditional probabilities $p(\tilde{c}{=}s|d_i)$.

Table 2. Results with k latent classes for 18400 dpa documents from Jan 2000. The experiments are named according to their configuration including documents (D), sentences (S), and use of unigrams (U) and bigrams (B).

Experiment	k	Cluster Dist.	Mutual Inf.
PLSA			
jan20ubd1	20	4.74	1.06
jan20ubd2	20	4.75	1.06
jan20ubds	20	5.77	0.58
jan20ubs1	20	6.61	0.17
jan20ubs2	20	6.62	0.17
jan100ud	100	4.89	0.95
jan100us	100	6.62	0.13
jan100ubd	100	4.86	0.97
jan100ubds	100	5.98	0.44
jan100ubs	100	6.61	0.14
jan200ubd	200	5.39	1.05
jan200ubds	200	6.98	0.43
jan200ubs	200	7.25	0.16
LDA			
jan20ubd1	20	4.27	1.39
jan20ubd2	20	4.28	1.39
jan20ubds	20	4.86	1.18
jan20ubs	20	6.55	0.38
jan20ud1	20	4.32	1.50
jan20ud2	20	4.32	1.49
jan20uds	20	5.21	1.07
jan20us	20	5.66	0.86
jan100ubd	100	4.33	1.52
jan100ubds	100	5.09	1.30
jan100ubs	100	6.66	0.62
jan100ud	100	4.47	1.71
jan100uds	100	5.18	1.34
jan100us	100	6.04	0.94
jan200ubd	200	4.61	1.39
jan200ud	200	4.64	1.71

To illustrate the resulting concepts, we list the terms for different concepts sorted according decreasing probability $p(w|z)$ in Table 1. Because of space restrictions, only a randomly selected part of the results can be reproduced. In the first column, we print an English description of the latent concept. In the experiment, 200 latent classes were generated by PLSA for 50k documents. The terms connected by "+" are bigrams. As can be seen, the classes are rather focused and describe a content concept quite well. The latent concepts generated by LDA are similar (e.g. the "soccer" concept: Ball, Fußball, Ergebnisse, Bundesliga, FC, Trainer, Team, Spiel,...).

Table 2 contains experimental results for PLSA and LDA on dpa documents for one month, January 2000. This relatively constrained subcorpus (18.4k documents) with dpa news only from month January 2000 has served as input basis for comparing the first testbed results between LDA and PLSI. A few experiments have been repeated to explore the variability resulting from random starting seeds. It turns out that the variation is small.

As can be seen, the lowest distance to the prior classification is achieved for latent concepts which were computed with respect to the whole document. If the latent concepts are also computed with respect to sentences, the mutual information drops and the cluster distance is increased. Latent concepts computed with respect to sentences exhibit the worst relation with the a priori categories.

With respect to the number of latent concepts, the results are relatively insensitive. Especially the mutual information is relatively constant, varying from 1.06, 0.95, 0.97, to 1.05 as the number of concepts grows from 20 to 200 in PLSA, while LDA sometimes shows a larger variation, with higher mutual information. The highest mutual information, 1.7137, was reached for LDA in the experiment jan200ud with unigrams and documents, which is almost equal to the value for jan200ubd with additional bigrams, 1.7085.

The comparison between unigrams and bigrams is not conclusive. Concepts based on combined unigrams and bigrams seem to have a small advantage with a cluster distance of 4.86 compared to 4.89 with unigrams in PLSA and 4.27 compared to 4.32 with LDA (experiments jan100ubd and jan100ud). With respect to cluster distance, LDA seems to perform generally better in our experiments.

The results for the whole corpus in Table 3 support the results discussed above. Again the use of sentences lead to a deterioration of results. The largest mutual information, 1.58, was reached for 200 latent concepts, compared to a value of 1.44 for 100 latent concepts.

5. Conclusion and Outlook

In this paper, we describe work in progress to investigate the influence of word correlations at different scopes on the quality of concepts derived using different soft-clustering methods. We showed that the procedures are able to process large corpora with up to 229500 documents and generate up to 200 latent concepts. It turned out that the inclusion of short-range scopes to determine concepts did not improve the quality of concepts if their distance to the prior classifica-

Table 3. PLSA results with k latent classes for n documents extracted from the dpa corpus of the year 2000.

Experiment	k	Cluster Dist.	Mutual Inf.
PLSA results using 10000, 20000 and 50000 documents			
10000_100D	100	5.48	1.48
10000_100DS	100	7.18	0.68
10000_100S	100	8.07	0.24
20000_100D	100	5.58	1.44
20000_200D	200	5.99	1.58
20000_100DS	100	7.33	0.61
20000_200DS	200	7.99	0.63
20000_100S	100	8.14	0.21
50000_200D	200	6.36	1.42
50000_200S	200	8.79	0.24
PLSA results using all documents			
ALL20UDS	20	4.48	0.39
ALL20UBD	20	3.75	0.74
ALL100UBDS	100	6.17	0.35
ALL100UDS	100	6.14	0.37
ALL100UBD	100	5.17	0.82
ALL100UD	100	5.16	0.82
ALL200UBD	200	5.90	0.80
ALL200UD	200	5.91	0.79

tion is used as a quality criterion. These results can be observed for PLSA as well as LDA, where LDA yielded generally better lower distances and higher mutual information values, i.e., performed better compared to PLSA.

One reason for the higher distance of estimated concepts compared to the prior categorization might be that the prior categorization refers to entire documents while concepts derived from sentences may refer to more local concepts, although the actual comparison only is done for concept distributions of the whole document. We will further investigate this aspect. Another possible reason may be the increased number of degrees of freedom, as the latent class distributions for sentences involve a number of parameters for each sentence, which might lead to overfitting. Currently we are investigating ways to tackle this problem by using a prior distribution in a Bayesian framework.

Another finding is the slightly better distance between prior and estimated clusterings when using bigram data in addition to unigram data. Correlation at shortest scope therefore seems to be useful for concept extraction.

Future work will take a closer look at other sub-document settings, such as bigrams as exclusive term units or paragraphs and other logical document partitions as additional document units in addition to sentences. We will also investigate the dynamic behavior of topics over sub-document structures, which is crucial for the development of retrieval systems that use document structure. Another goal to investigate correlations in the extraction of topic hierarchies [10].

References

[1] Gross, M.: The lexicon-grammar: Application to french. In Asher, R.E., ed.: Encyclopedia of Language and Linguistics, Pergamon Press, London (1994) 2195–2203

[2] Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill (1983)

[3] Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. Journal of the American Society of Information Science **41** (1990) 391–407

[4] Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning Journal **41** (2001) 177–196

[5] Jansche, M.: Parametric models of linguistic count data. In: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 41), Sapporo, Japan (2003) 288–295

[6] Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. In: Advances in Neural Information Processing Systems 14, Cambridge, MA, MIT Press (2002)

[7] Minka, T.: A family of algorithms for approximate Bayesian inference. Phd thesis, MIT (2001)

[8] Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences **101** (2004) 5228–5235

[9] Meila, M.: Comparing clusterings. Technical Report Technical Report 418, Department of Statistics, University of Washington (2002)

[10] Blei, D., Griffiths, T., Jordan, M., Tenenbaum, J.: Hierarchical topic models and the nested Chinese restaurant process. In: Advances in Neural Information Processing Systems 16, Cambridge, MA, MIT Press (2004)