Statistical modeling of medical indexing processes for biomedical knowledge information discovery from text

Markus Bundschus
Institute for Computer Science
University of Munich
Oettingenstr. 67
80538 Munich, Germany
bundschu@dbs.ifi.lmu.de

Mathaeus Dejori
Integrated Data Systems Dep.
Siemens Corporate Research
755 College Road East
Princeton, NJ 08540, USA
mathaeus.dejori@siemens.com

Shipeng Yu
CAD & Knowledge Solutions
Siemens Medical Solutions
51 Valley Stream Parkway
Malvern, PA 19355, USA
shipeng.yu@siemens.com

Volker Tresp
Information &
Communications, IC4
Siemens CT
Otto-Hahn-Ring 6
81739 Munich, Germany
volker.tresp@siemens.com

Hans-Peter Kriegel
Institute for Computer Science
University of Munich
Oettingenstr. 67
80538 Munich, Germany
kriegel@dbs.ifi.lmu.de

ABSTRACT

The overwhelming amount of published literature in the biomedical domain and the growing number of collaborations across scientific disciplines results in an increasing topical complexity of research articles. This represents an immense challenge for efficient biomedical knowledge discovery from text. We present a new graphical model, the so-called TOPIC-CONCEPT MODEL, which extends the basic Latent Dirichlet Allocation framework and reflects the generative process of indexing a PubMed abstract with terminological concepts from an ontology. The generative model captures the latent topic structure of documents by learning the statistical dependencies between words, topics and MeSH (Medical Subject Headings) concepts. A number of important tasks for biomedical knowledge discovery can be solved with the here introduced model. We provide results for the extraction of the hidden topic-concept structure from a large medical text collection, the identification of the most likely topics given a specific MeSH concept, and the extraction of statistical relationships between MeSH concepts and words. Moreover, we apply the introduced generative model to a challenging multi-label classification task. A benchmark with several classification methods on two independent data sets proves our method to be competitive.

Keywords

Document Modeling, topic modeling, multi-label classification, ontologies

BIOKDD '08 Las Vegas, NV, USA

1. INTRODUCTION

In the last decade, powerful new biomedical research tools and methods have been developed, resulting in an unprecedented increase of biomedical data and literature. High-throughput experiments, such as DNA microarrays or protein arrays, produce large quantities of high-quality data, leading to an explosion of scientific articles published in this field. Thus, automated extraction of useful information from large document collections has become an increasingly important research area [12, 11]. To ensure an efficient access to this steadily increasing source of bibliographic information, it is required to efficiently index incoming articles, i. e. to label unstructured free text with a structured machine readable annotation. Articles selected for inclusion in PubMed[1], for example, are indexed with concepts from the Medical Subject Headings[2] (MeSH) thesaurus to facilitate later retrieval. This additional meta information provides a rich source of knowledge, which can be exploited for biomedical knowledge discovery and data mining tasks and this is the focus of this work.

Recently, powerful techniques such as Probabilistic Latent Semantic Analysis (PLSA) [15] or Latent Dirichlet Allocation (LDA) [7] have been proposed for automated extraction of useful information from large document collections. Applications include automatic topic extraction, query answering, document summarization, and trend analysis. Generative statistical models such as the above mentioned ones, have been proven effective in addressing these problems. In general, the following advantages of topic models are highlighted in the context of document modeling: First, topics can be extracted in a complete unsupervised fashion, requiring no initial labeling of the topics. Second, the resulting representation of topics for a document collection is interpretable and last but not least, each document is usually expressed by a mixture of topics, thus capturing the topic combinations that arise in documents [15, 7, 14]. In the

[1] http://www.ncbi.nlm.nih.gov/pubmed/
[2] http://www.nlm.nih.gov/mesh/

biomedical domain, the classical LDA model has been applied to the task of finding life span related genes from the Caenorhabditis Genetic Center Bibliography [5] and to the task of identifying biological concepts from a protein-related corpus [33]. Depending on the addressed generative process, the LDA framework has been extended e. g. to model the dependencies between authors, topics and documents [30] or the dependencies between author and recipients [20]. Further approaches include the modeling of images and their corresponding captions [6] as well as the modeling of dependencies between topics and named entities [25].

In this paper, we introduce another extension of the LDA framework, the so-called Topic-Concept (TC) model, to resemble the generative process of creating an indexed PubMed abstract. The approach simultaneously models the way how the document is generated as well as the way how the document is subsequently indexed with MeSH concepts (see figure 1 for a comparison with the classical LDA approach). We refer to MeSH as a terminological ontology, where relations are partially described as subtype-supertype relations and where the concepts are described by concept labels or synonyms [2].

By modeling the indexing process of PubMed abstracts, we can answer a range of important queries for knowledge discovery about the content of biomedical text collections. With such a model, we can provide a bird's eye view of biomedical topics discussed in a large document collection associated with prominent MeSH concepts (i. e. uncovering the hidden topic-concept structure in a biomedical text collection). In contrast to the classical LDA, this results in a richer representation of topics, since topics are not solely represented by their most likely words. Instead, topics in the TC model are, in addition to the words, associated with their most likely MeSH terms (see section 3.2.1). Furthermore, we can identify several types of statistical relationships between different classes of document entities (i. e. words, MeSH concepts and topics). We provide results for identifying statistical relationships between concepts and words based on the topics (see section 3.2.2). Another interesting use case we consider, is the estimation of the most likely topics given a MeSH concept. This results in a fast overview over the topics in which a specific MeSH term is most likely to be involved (see section 3.2.2). Last but not least, we can use the TC model for multi-label classification. To validate the predictive power of the here presented model, we apply our generative method to a challenging multi-label classification problem with 108 classes. A benchmark on two independent corpora against (1) a multi-label naive Bayes classifier, (2) a method currently used by the National Library of Medicine (NLM) and (3) a state-of-the-art multi-label support vector machine (SVM) shows encouraging results.

The remainder of the paper is organized as follows: In Section 2 we describe the extension of the classical LDA towards the TC model. Section 3.1 describes the experimental setup. Afterwards results are presented and a concluding discussion is given.

2. METHODS

In the following we will describe two generative models, the first simulating the process of document generation alone and the second simulating both the process of document generation and the process of document indexing. Let

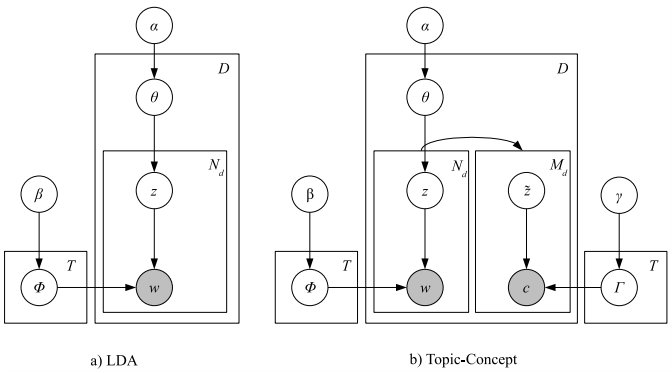


Figure 1: Graphical model for a) LDA and b) Concept-LDA in plate notation. Shaded nodes represent observed random variables, unshaded nodes represent latent random variables.

$\mathbf{D} = \{d_1, d_2, ..., d_D\}$ be a set of documents, where D denotes the number of documents in the corpus. A document d is represented by a vector of N_d words, w_d, where each word w_i is chosen from a vocabulary of size N. In the second model, a document d is additionally described by a vector of M_d MeSH concepts c_d, where each concept c_i is chosen from a set of MeSH concepts of size M. The collection of D documents is defined by $\mathbf{D} = \{(w_1, c_1), ..., (w_D, c_D)\}$.

2.1 Classical Latent Dirichlet Allocation (LDA) model

The Latent Dirichlet Allocation model (LDA) is based upon the idea that the probability distribution over words in a document can be expressed as a mixture of topics, where each topic is expressed as a mixture of words [7]. In LDA, the generation of a document collection is modeled as a three step process. First, for each document, a distribution over topics is sampled from a Dirichlet distribution. Second, for each word in the document, a single topic is chosen according to this distribution. Finally, a word is sampled from a multinomial distribution over words specific to the sampled topic. The hierarchical Bayesian model shown (using plate notation) in Figure 1(a) depicts this generative process. θ represents the document-specific multinomial distribution over T topics, each being drawn independently from a symmetric Dirichlet prior α. Φ denotes the multinomial distribution over N vocabulary items for each of T topics being drawn independently from a symmetric Dirichlet prior β. For each of the N_d words w in document d, z denotes the topic responsible for generating that word, drawn from θ, and w is the word itself, drawn from the topic distribution Φ conditioned on z. According to the graphical model representation, the probability distribution over N vocabulary items for the generation of word w_i within a given document is specified as

$$p(w_i) = \sum_{t=1}^{T} p(w_i|z_i = t)p(z_i = t) \quad (1)$$

where $z_i = t$ represents the assignment of topic t to the ith word, $p(w_i|z_i = t)$ is given by the topic-word distribution Φ and $p(z_i = t)$ by the document-topic distribution θ.

Table 1: Corpora statistics for the two data sets used in this paper.

	random 50K	genetics-related
Documents	50.000	84.076
Unique Words	22.531	31.684
Total Words	2.369.616	4.293.992
Unique MeSH Main Headings	17.716	18.350
Total MeSH Main Headings	470.101	912.231

2.2 Extension to the Topic-Concept (TC) Model

The Topic-Concept model extends the LDA framework by simultaneously modeling the generative process of *document generation* and the process of *document indexing*. In addition to the three steps mentioned above, two further steps are introduced to model the process of document indexing. For each of the M_d concepts in the document a topic $\tilde{z}$ is uniformly drawn based on the topic assignments for each word in the document. Finally, each concept c is sampled from a multinomial distribution over concepts specific to the sampled topic. This generative process corresponds to the hierarchical Bayesian model shown in Figure 1(b). In this model, Γ denotes the vector of multinomial distribution over M concepts for each of T topics being drawn independently from a symmetric Dirichlet prior γ. After the generation of words, a topic $\tilde{z}$ is drawn from the document specific distribution, and a concept c is drawn from the $\tilde{z}$ specific distribution Γ. The probability distribution over M MeSH concepts for the generation of a concept c_i within a document is specified as:

$$p(c_i) = \sum_{t=1}^{T} p(c_i|\tilde{z}_i = t)p(\tilde{z}_i = t|\mathbf{z}) \tag{2}$$

where $\tilde{z}_i = t$ represents the assignment of topic t to the ith concept, $p(c_i|\tilde{z}_i = t)$ is given by the concept-topic distribution Γ. The topic for the concept is selected uniformly out off the assignments of topics in the document model, i.e., $p(\tilde{z}_i = t|\mathbf{z}) = \text{Unif}(z_1, z_2, \ldots, z_{N_d})$ leading to a coupling between both generative components.

The generative process of the Topic-Concept model is essentially the same as the Correspondence LDA model proposed in [6] with the difference that the Topic-Concept model imitates the generation of documents and their subsequent annotation, while [7] models the dependency between image regions and captions.

2.3 Learning the Topic-Concept Model from Text Collections

Estimating Φ, θ and Γ provides information about the underlying topic distribution in a corpus and the respective word and MeSH concept distributions in each document. Given the observed documents, the learning task is to infer these parameters for each document. Instead of estimating the parameters directly [16, 6] we follow the idea of [14] and estimate Φ and θ from the posterior distribution over the assignments of words to topics $p(\mathbf{w}|\mathbf{z})$. As the posterior cannot computed directly we resort to a Gibbs sampling strategy generating samples from the posterior by repeatedly drawing a topic for each observed word from its probability conditioned on all other variables. In the LDA model, the algorithm goes over all documents word by word. For each word w_i, a topic z_i is assigned by drawing from its distribution conditioned on all other variables

$$\begin{aligned} p(z_i = t|w_i = n, \mathbf{z_{-i}}, \mathbf{w_{-i}}) &\propto \\ p(w_i = n|z_i = t)p(z_i = t) &\propto \\ \frac{C^{WT}_{nt} + \beta}{\sum_{n'} C^{WT}_{n't} + N\beta} \frac{C^{DT}_{dt} + \alpha}{\sum_{t'} C^{DT}_{dt'} + T\alpha} & \end{aligned} \tag{3}$$

where $z_i = t$ represents the assignments of the ith word in a document to topic t, $w_i = n$ represents the observation that the ith word is the nth word in the lexicon, and $\mathbf{z_{-i}}$ represents all topic assignments not including the ith word. Furthermore, C^{WT}_{nt} is the number of times word n is assigned to topic t, not including the current instance, and C^{DT}_{dt} is the number of times topic t has occurred in document d, not including the current instance. Additionally, in the Topic-Concept model, the posterior $p(c|\tilde{\mathbf{z}})$ is approximated by assigning for each concept c_i, a topic $\tilde{z}_i$ from the following distribution

$$\begin{aligned} p(\tilde{z}_i = t|c_i = m, \tilde{\mathbf{z}}_\mathbf{i}, \mathbf{z_{-i}}, \mathbf{w_{-i}}) &\propto \\ p(c_i = m|\tilde{z}_i = t)p(\tilde{z}_i = t|\mathbf{z}) &\propto \\ \frac{C^{CT}_{mt} + \gamma}{\sum_{m'} C^{CT}_{m't} + M\gamma} \frac{C^{TD}_{td}}{N_d} & \end{aligned} \tag{4}$$

where $\tilde{z}_i = t$ represents the assignments of the ith concept in a document to topic t, $c_i = m$ represents the observation that the ith concept in the document is the mth concept in the lexicon, and $\mathbf{z_{-i}}$ represents all topic assignments not including the ith concept. Furthermore, C^{CT}_{mt} is the number of times concept m is assigned to topic t, not including the current instance, and C^{TD}_{td} is the number of times topic t has occurred in document d, not including the current instance.

For any single sample we can estimate Φ, θ and Γ using

$$\hat{\Phi}_{nt} = \frac{C^{WT}_{nt} + \beta}{\sum_{n'} C^{WT}_{n't} + N\beta} \tag{5}$$

$$\hat{\theta}_{dt} = \frac{C^{DT}_{dt} + \alpha}{\sum_{t'} C^{WT}_{dt'} + T\alpha} \tag{6}$$

$$\hat{\Gamma}_{mt} = \frac{C^{CT}_{mt} + \gamma}{\sum_{m'} C^{CT}_{m't} + M\gamma} \tag{7}$$

Instead of estimating the hyperparameters α, β and γ, we fix them to $50/T$, 0.001 and $1/M$ respectively in each of the experiments. The values were chosen according to [30, 14].

3. EXPERIMENTS AND RESULTS

3.1 Experimental setting

Two large PubMed corpora previously generated by [23, 24] were used in the experiments. The first data set is a collection of PubMed abstracts randomly selected from the MEDLINE 2006 baseline database provided by the NLM.

Table 2: Selected topics, learned from the genetics-related corpus ($T = 300$). For each topic the fifteen most probably words and MeSH terms are listed with their corresponding probabilities.

Topic 6				Topic 17			
Word	Prob.	Mesh Term	Prob.	Word	Prob.	Mesh Term	Prob.
ethic	0.043	Humans	0.150	viru	0.118	Humans	0.06
research	0.039	United States	0.038	viral	0.064	HIV-1	0.06
issu	0.023	Informed Consent	0.017	infect	0.058	HIV Infections	0.059
public	0.014	Ethics, Medical	0.011	hiv-1	0.047	Virus Replication	0.045
medic	0.013	Personal Autonomy	0.001	virus	0.035	RNA, Viral	0.042
health	0.013	Decision Making	0.001	hiv	0.033	Animals	0.027
moral	0.013	Ethics, Research	0.008	replic	0.033	DNA, Viral	0.027
consent	0.012	Great Britain	0.008	immunodef.	0.025	Cell-Line	0.023
practic	0.012	Human Experimentation	0.007	envelop	0.012	Genome, Viral	0.022
concern	0.011	Public Policy	0.007	aids	0.012	Viral Proteins	0.020
polici	0.001	Morals	0.007	particl	0.011	Molecular Sequence Data	0.017
conflict	0.008	Biomedical Research	0.006	capsid	0.011	Anti-HIV Agents	0.016
right	0.008	Research Subjects	0.006	host	0.011	Viral Envelope Proteins	0.013
articl	0.008	Social Justice	0.006	infecti	0.010	Drug Resistance, Viral	0.012
accept	0.008	Confidentiality	0.006	antiretrovir	0.001	Acquired Immunodef. Synd.	0.011

Topic 16				Topic 26			
Word	Prob.	Mesh Term	Prob.	Word	Prob.	Mesh Term	Prob.
phosphoryl	0.130	Phosphorylation	0.123	breast	0.372	Breast Neoplasms	0.319
kinas	0.118	Prot.-Serine-Threonine Kin.	0.075	cancer	0.323	Humans	0.120
activ	0.060	Proto-Oncogene Prot.	0.060	women	0.032	Middle Aged	0.024
akt	0.060	Proto-Oncogene Proteins c-akt	0.047	tamoxifen	0.028	Receptors, Estrogen	0.023
tyrosin	0.036	1-Phosphatidylinositol 3-Kin.	0.047	mcf-7	0.026	Tamoxifen	0.022
protein	0.029	Humans	0.043	estrogen	0.012	Antineopl. Agents, Hormon.	0.017
phosphatas	0.025	Signal Transduction	0.038	mda-mb-231	0.007	Aged	0.016
signal	0.025	Animals	0.028	adjuv	0.007	Carcinoma, Ductal, Breast	0.013
pten	0.024	Protein Kinases	0.021	statu	0.007	Chemotherapy, Adjuvant	0.013
pi3k	0.022	Tumor Suppressor Proteins	0.016	hormon	0.007	Mammography	0.012
pathwai	0.020	Phosphoric Monoester Hydrol.	0.016	tam	0.006	Breast	0.012
regul	0.018	Enzyme Activation	0.015	aromatas	0.006	Adult	0.011
serin	0.015	Cell Line, Tumor	0.014	ductal	0.006	Neoplasm Staging	0.010
inhibit	0.015	Enzyme Activation	0.001	mammari	0.006	Aromatase Inhibitors	0.009
src	0.015	Mice	0.013	postmenop.	0.005	Receptors, Progesterone	0.009

The collection consists of $D = 50.000$ abstracts, $M = 17.716$ unique MeSH main headings and $N = 22.531$ unique word stems. Word tokens from title and abstract were stemmed with a standard Porter stemmer [27] and stop words were removed using the PubMed stop word list [3]. Additionally, word stems occurring less than five times in the corpus were filtered out. Note that no filter criterion was defined for the MeSH vocabulary.

The second data set contains $D = 84.076$ PubMed abstracts, with $M = 18.350$ unique MeSH main headings and a total of $N = 31.684$ unique word stems. The same filtering steps were applied as described above. This corpus is composed of genetics-related abstracts from the MEDLINE 2005 baseline corpus. The here introduced bias towards genetics-related abstracts resulted from using NLM's Journal Descriptor Indexing Tool by applying some genetics-related filtering strategies [23]. See [23, 24] for more information about both corpora. In the following, the data sets are referred to as *random 50K* data set and *genetics-related* data set respectively. For the extraction of statistical relationships between the various document entities and for uncovering the hidden-topic concept structure, we decided to use the larger genetics-related corpus with all 18.350 MeSH main headings (see section 3.2.1 and section 3.2.2), while for the multi-label classification task, we used both corpora in a pruned setting (see next section 3.1.1).

Parameters for the Topic-Concept model were estimated by averaging samples from ten randomly-seeded runs, each running over 100 iterations, with an initial burn-in phase of 500 iterations (resulting in a total of 1.500 iterations). We found 500 iterations to be a convenient choice by observing a flattening of the log likelihood. The training time ranged from ten to fifteen hours depending on the size of the data set, the number of used MeSH concepts as well as on the predefined number of topics (run on a standard Linux PC with Opteron Dual Core processor, 2.4 GHz).

3.1.1 Multi-label classification task

In this setting, we prune each MeSH descriptor to the first level of each taxonomy-subbranch resulting in 108 unique MeSH concepts ($M = 108$). For example, if a document is indexed with *Muscular Disorders, Atrophic [C10.668.550]*, the concept is pruned to *Nervous System Diseases [C10]*. Therefore, the task is to assign at least one of the 108 classes to an unseen PubMed abstract. Note that from a machine learning point of view, this is a challenging 108 multi-label classification problem and corresponds to other state-of-the-art text classification problems such as the Reuters text classification task [19], where the number of classes is approximately the same. In the pruned setting of our task, we have on average 9.6/10.5 (random 50K/genetics-related) pruned

[3] http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#Stopwords

Table 3: Selected MeSH concepts from the *Disease* and the *Drug & Chemicals* subbranch with the 20 most probable word stems estimated based on a topic-concept model learned from the genetics-related corpus ($T = 300$). The font size of each word stem encodes its probability given the corresponding MeSH concept. The number in brackets is euqal to the number of times, the MeSH terms occurs in the corpus

Diseases

Myelodysplastic Syndromes (208)

acut aml bcr-abl blast chronic cml flt3 hematolog imatinib leukaemia leukem leukemia lymphoblast marrow mds myelodysplast myeloid patient relaps syndrom

Pulmonary Embolism (39)

activ associ case clinic diagnos diagnosi diagnost factor incid men mortal patient platelet preval protein rate risk studi women year

Drugs & Chemicals

Erythropoietin (85)

abnorm anaemia anemia caus cell defect defici disord epo erythrocyt erythroid erythropoietin g6pd hemoglobin increas model normal patient sever studi

Paclitaxel (309)

advanc agent anticanc cancer chemotherapi cisplatin combin cytotox drug effect median paclitaxel patient phase regimen respons sensit surviv toxic treatment

MeSH labels per document. Parameter estimation remains the same as mentioned in the previous paragraph.

In particular, we are interested in evaluating the classification task in a user-centered or semi-automatic scenario, where we want to recommend a set of classes for a specific document (e. g. a human indexer gets recommendations of MeSH terms for a document). Thus, we decided to follow the evaluation of [13] and average the effectiveness of the classifiers over documents rather than over categories. In addition, we weight recall over precision and use the F2-macro measure, because it reflects that human indexers will accept some inappropriate recommendations as long as the major fraction of recommended index terms will be correct [13].

3.2 Results

3.2.1 Uncovering the hidden topic-concept structure

Table 2 illustrates several different topics (out of 300) from the genetics-related corpus, obtained from a particular Gibbs sampler run after the 1.500th iteration. Each table shows the fifteen most likely word stems assigned to a specific topic and its corresponding most likely MeSH main headings. To show the descriptive power of our learned model, we chose four topics describing different aspects of biomedical research. Topic 6 is ethics-related, topic 16 is related to a special biochemical process, namely signal transduction, and the last two topics represent aspects of specific disease classes. Topic 26 represents a topic centered around breast cancer, while topic 17 refers to HIV. The model includes several other topics related to specific diseases, biochemical processes, organs and other aspects of biomedical research like e. g. Magnetic Resonance Spectroscopy. Recall that the here investigated corpus is biased towards genetics-related topics, thus, some topics can describe quite specific aspects of genetics research. More generic topics in the corpus are related to terms, common to almost all biomedical research areas including terminology, describing experimental setups or methods. In general, the extracted topics are, of course, dependent on the corpus seed. The full list of topics with corresponding word and MeSH distributions is available at `www.dbs.ifi.lmu.de/~bundschu/TCmodel_supplementary/TC_structure.txt`.

It can be seen that the word stems already provide an intuitive description of specific aspects. Furthermore, the topics gain more descriptive power by their associated MeSH concepts, providing an accurate description in structured from. Note that the standard topic models are only able to represent topics with the single word descriptions. In contrast, the TC model provides a richer representation of topics by additionally linking topics to concepts from a terminological ontology. We found that the topics obtained from different Gibbs sampling runs were relatively stable. A variability in terms of ranking of the words and MeSH terms in the topics can be observed, but overall the topics match very closely. For studies about topic stability in aspect models, please refer to [29].

3.2.2 Extraction of statistical relationships

Besides uncovering the hidden topic-concept structure, we apply the model to derive statistical relations between MeSH concepts and word stems, thus bridging the gap between natural free text and the structured semantic annotation. The derived relations could be e. g. used for improving word sense disambiguation [18]. In Table 3, four MeSH concepts from the *Disease* and the *Drug & Chemicals* subbranch and their twenty most probable word stems are shown. For each MeSH concept, the distribution over words is graphically represented by varying the font size for each word stem with respect to the probability. Given a concept c, the conditional probability for each word is estimated by $p(w|c) \propto \sum_t p(w|t)p(t|c)$, which is computed from the learned model parameters. The word distributions describe the corresponding MeSH concept in an intuitive way, capturing the topical diversity of certain MeSH concepts. Note

Table 4: Selected MeSH concepts from the *Disease* and the *Drug & Chemicals* subbranch with the three most probable topics estimated based on a topic-concept model learned from the genetics-related corpus ($T = 300$). Topics are illustrated here by the twenty most probable word stems.

MeSH term	Topic	Word stems
Myelodysplastic Syndromes (208)	Topic 46 ($p = 0.20$)	leukemia acut myeloid aml mds lymphoblast leukaemia blast leukem patient myelodysplast marrow syndrom malign flt3 bone promyelocyt hematolog mll granulocyt
	Topic 75 ($p = 0.02$)	transplant donor recipi graft stem allogen reject autolog cell immunosuppress allograft marrow surviv hematopoiet condit receiv acut gvhd engraft diseas
	Topic 25 ($p = 0.01$)	chromosom aberr transloc cytogenet delet abnorm rearrang genom karyotyp gain loss region arm breakpoint trisomi mosaic duplic cgh case imbal
Erythropoietin (85)	Topic 177 ($p = 0.30$)	defici adren anemia malaria parasit plasmodium mosquito falciparum erythrocyt cortisol erythropoietin caus g6pd insuffici adrenocort acth anaemia epo anophel develop
	Topic 14 ($p = 0.14$)	cell stem progenitor hematopoiet differenti embryon lineag hsc adult marrow bone erythroid cd34+ precursor potenti cd34 marker hematopoiesi msc self-renew
	Topic 140 ($p = 0.07$)	activ nf-kappab factor nuclear transcript express cell induc inhibit constitut ap-1 regul c-jun suppress p65 kappa curcumin transloc nfkappab c-fo

that there are many other opportunities to access statistical relations between MeSH concepts and words. One could e. g. use measurements like relative frequency or χ^2 statistics. It may be that the TC model captures relationships that can't be captured in a simpler way, but this evaluation is out of scope of the here presented work. We provide all word clouds for all MeSH terms occurring in the corpus from the *Disease* and the *Drug & Chemicals* subbranch as supplementary data (`www.dbs.ifi.lmu.de/~bundschu/TCmodel_supplementary/`).

Another important use case we consider, is the task of estimating the most likely topics given a specific MeSH term with respect to a seed corpus. This results in a fast overview over the topics in which a specific MeSH term is most likely to be involved. Table 4 shows two such examples extracted from the genetics-related corpus. Because of lack of space, we only represent the topics by the most likely word stems (the associated MeSH terms for the topics can be investigated in the supplementary file, mentioned in section 3.2.1). The first example shows the three most likely topics for the MeSH term *myelodysplastic syndromes*. Myelodysplastic syndromes, also called pre-leukemia or 'smoldering' leukemia, are diseases in which the bone marrow does not function normally and not enough blood cells are produced [26]. This fact is reflected by the most likely topic for this MeSH term (Table 4, Topic 46). Furthermore, a state-of-the-art treatment of this disease, is bone marrow transplantation. First, all of the bone marrow in the body is going to be destroyed by high-doses of chemotherapy and/or radiation therapy. Then healthy marrow is taken from a donor (i. e. another person) and is given to the patient [26]. This is described by the second most likely topic (Table 4, Topic 75). Topic 25 constitutes that Myleodysplastic syndromes have a genetic origin and that gene and chromosome aberrations are a likely cause of this disease [26].

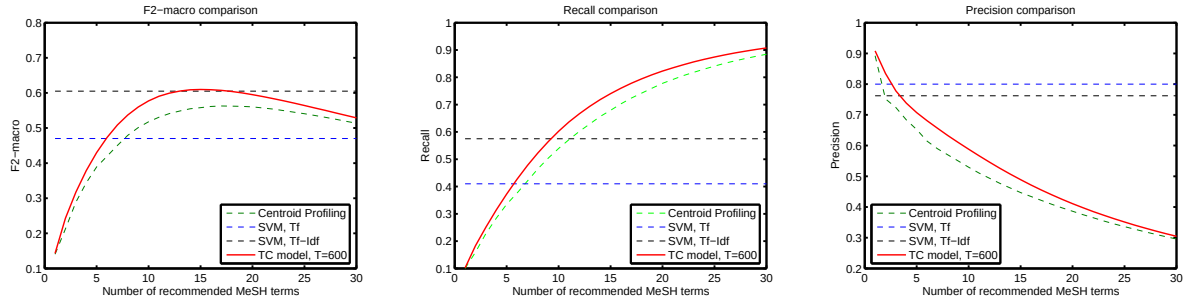
The second MeSH term in table 4, *Erythropoietin* (EPO), is a hormone which is produced by the kidney and liver. It is known to regulate red blood cell production. In the mined genetics-related corpus, the most likely topic (Table 4, Topic 177) states that erythropoietin could be used as a treatment during malaria infection [9] and this is a current issue of ongoing research [3, 31]. Erythropoietin is known to directly promote the generation of neuronal stem cells from progenitors, which is reflected by Topic 14. Last but not least, Topic 140 provides information about the gene regulatory context of EPO. NF-kappaB, e. g. , regulates EPO [8], while EPO in turn regulates expression of c-jun and AP-1 [28].

A full list of all MeSH terms and its most likely associated topics is available online. (`www.dbs.ifi.lmu.de/~bundschu/TCmodel_supplementary/mesh_associated_topics.pdf`).

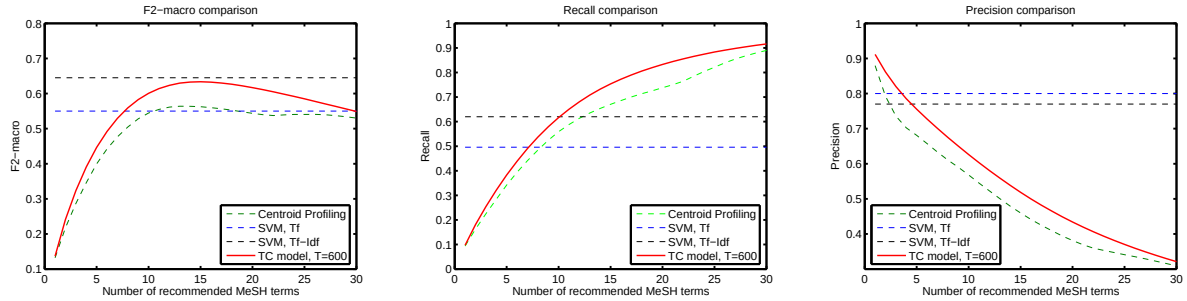
3.2.3 Multi-label classification

In what follows, we will first describe the used benchmark methods and then present the results for the multi-label classification problem with 108 classes for the genetics-related corpus and the random 50K corpus. The prediction results of the Topic-Concept model are benchmarked against a method currently used by the NLM [17], which we refer to as *centroid profiling*, a multi-label naive Bayes classifier and a multi-label SVM. For both data sets and all methods, 5-fold cross-validation was conducted.

In [17] classification is tackled by computing for each word token w_i and each class label y_m, in a training corpus, a term frequency measure $TF_{i,m} = w_{i,y_m} / \sum_{m=1}^{M} w_{i,y_m}$ with M equals to the total number of classes. Thus, $TF_{i,m}$ measures the number of times a specific word w_i co-occurs with the class label y_m, normalized by the total number of times the word w_i occurs. As a consequence, each word token in the training can be represented by a profile consisting of the term frequency distribution over all M classes. When index-



(a) *random 50K corpus*



(b) *genetics-related corpus*

Figure 2: F2-macro, recall and precision plots for discipline-based indexing. Results are plotted according to the number of top n recommended MeSH terms. In average every document has 9.6 such assignments in our experimental setting. (a) Plots for the randomly selected data set. (b) Plots for the genetics-related data set

ing a new unseen document, the centroid over all profiles for the word tokens in the test document is computed. This centroid represents the ranking of all class labels for the test document. This method was chosen, because it is currently used by the NLM in a classification task to predict so-called journal descriptors [17].

According to [22], naive Bayes classifiers are a very successful class of algorithms for learning to classify text documents. For the multi-label naive Bayes classifier, we assumed a bag of words representation like for the Topic-Concept model and trained it for each of the 108 labels. We used the popular multinomial model for naive Bayes [21].

The multi-label SVM setting was implemented according to [19]. In this setting, a linear kernel is used and the popular so-called binary method is used to adapt the SVM to a multi-label setting. It has been shown that this setting produced very competitive results on a large-scale text classification task on the RCV1 Reuters corpus [19]. LIBLINEAR, a part of the LIBSVM package [10] is used for the implementation. Two different weighting schemes are evaluated: Term frequency (Tf) as well as cosine-normalized Term frequency-inverse document frequency (Tf-Idf).

In the TC-model, the prediction of concept terms for unseen documents can be formulated as follows: Based on the word-topic and concept-topic count matrices learned from an independent data set, the likelihood of a concept c given the test document d is $p(c|d) = \sum_t p(c|t)p(t|d)$. The first probability in the sum, $p(c|t)$, is given by the learned topic-concept distribution (see Equation 7). The mixture of topics for the document $p(t|d)$ is estimated by drawing for each word token in the test document a topic based on the learned word-topic distribution $p(w|t)$ (see Equation 5). Therefore, the TC model directly predicts a ranked list of class recommendations, in contrast to the classical task of topic models in text classification problems, where they are usually used for dimensionality reduction and afterwards standard classifiers are applied [7].

We now discuss experimental results using 5-fold cross-validation. Figure 2 plots F2-macro measure, recall and precision against the number of recommended MeSH terms. Figure 2(a) shows results for the random 50K data set and Figure 2(b) for the genetics-related data set respectively. Our TC model and the centroid profiling method provide as output a ranked list of recommendations. In order to be able to compare these two methods with the other classifiers, a thresholding strategy is needed [32]. We decided to use the simple rank-based thresholding (*Rcut*) [32] and evaluate the results until a cut-off value of 30 (Recall that each document has in average 9.6 (random 50K) and 10.5 (genetics-related) MeSH entries in our experimental setting. The Topic-Concept model was trained with two different number of topics on both data sets ($T = 300$, $T = 600$ for the 50K random corpus and $T = 300$, $T = 600$ for the genetics-related corpus). For clarity, we only show the results for $T = 600$ here, since experimental validation showed

that the number of topics is not very sensitive to the overall performance. We also exclude the NB classifier from the figure for clarity (F-measure 0.58 and 0.60 for random 50K and genetics-related). In terms of F2-macro, recall and precision, the Topic-Concept model clearly outperforms the centroid profiling. The naive Bayes classifier already yields quite competitive results. Regarding F2-macro, the TC models reach their optimum at 15 returned recommendations for both data sets (0.61 (random 50K)/0.635 (genetics-related)). At a cut-off value of 15 recommendations, centroid profiling reaches a F2-macro of 0.558 for the random 50K data set (optimum at 17 recommendations with 0.562) and 0.562 for the genetics-related corpus (optimum at 13 recommendations with 0.564). Using a cut-off value which equals to the number of average MeSH assignments (rounded-up) in the two training corpora the F2-macro is for the best TC models 0.59 (random 50K) and 0.61 (genetics-related), while the centroid profiling reaches only 0.517 (random 50K) and 0.55 (genetics-related) at this cut-off value. Note that using the average number of MeSH assignments is the most simple way to determine an appropriate cut-off value. A more analytical way of determining the threshold would be to set up an independent development set for the given corpus and to maximize the F2-macro measure according to the number of recommendations. Other approaches e. g. use a default length of 25 recommended index terms [1] for unpruned MeSH recommendation. The evaluation of the multi-label SVM shows that the performance is very sensitive to the used term weighting scheme (see Figure 2). When using Tf-Idf, the SVM is approximately on par with the TC model in terms of F2-macro on both data sets (F2-macro SVM, Tf-Idf is 0.60 (random 50K) and 0.645 (genetics-related)). The SVM is clearly superior in terms of precision due to its discriminative nature. When considering recall, the TC model outperforms the SVM with Tf-Idf, effective from a cut-off value of recommended MeSH terms, which is the average number of MeSH terms in the training corpora.

4. CONCLUSION AND OUTLOOK

This study presents a new probabilistic topic model for modeling medical text indexing processes. The so-called Topic-Concept model automatically learns the relation between words, MeSH terms, documents and topics from large text corpora of PubMed abstracts. The method uses a generative probabilistic process to learn the just mentioned relationships by extracting the latent topic structure. Gibbs sampling is used to learn the Topic-Concept model.

The TC model uncovers novel information from a biomedical text corpus, including the extraction of the hidden topic-concept structure, using all occurring unique MeSH terms in the corpus (18.350 distinct MeSH terms). In contrast to standard topic models, where topics are solely represented by their most likely words, the here extracted topic-concept structure can be interpreted as a richer representation of topics by additionally linking to concepts from the MeSH thesaurus. Thus, the enriched topic representation provides important additional information from a terminological ontology. Other use cases we explore, are the extraction of statistical relationships between words and MeSH terms as well as between topics and MeSH terms. The just mentioned applications can have impact on several other closely related areas such as information retrieval or information extraction (see e. g. [25]).

The Topic-Concept model can be easily applied to text classification tasks. Even though the here proposed method is generative, the experimental evaluation on a challenging multi-label classification problem on two independent data sets with 108 class labels against discriminative methods proves our method to be competitive in terms of F2-macro and even superior in terms of recall. In contrast to most text categorization algorithms, the here proposed model provides a ranking of recommended index terms for prediction tasks. Up to now, the choice of the number of returned recommended index terms is user-defined. Using a simple cut-off value which is equal to the number of average index terms assigned in a training collection, already yields competitive results.

In the current setting, our model neglects the hierarchical property of the MeSH thesaurus. The extension of the underlying generative process for capturing the hierarchy of terminological ontologies is a matter of ongoing research. To further tune prediction performance, we are also considering an expansion of the generative Topic-Concept model to a supervised topic model for multi-label classification as lately proposed by [4] for multi-class classification problems.

Funding

The work presented here was partially funded by the German Federal Ministry of Economy and Technology (BMWi) under the THESEUS project.

Acknowledgements

The authors wish to thank Aurélie Névéol from the NLM for answering questions about her research as well as for providing the data sets.

5. REFERENCES

[1] A. R. Aronson, J. G. Mork, C. W. Gay, S. M. Humphrey, and W. J. Rogers. The nlm indexing initiative's medical text indexer. In *Medinfo 2004*, pages 268–272. IOS Press, 2004.

[2] C. Biemann. Ontology learning from text: A survey of methods. *LDV-Forum*, 20(2):75–93, 2005.

[3] A.-L. L. Bienvenu, J. Ferrandiz, K. Kaiser, C. Latour, and S. Picot. Artesunate-erythropoietin combination for murine cerebral malaria treatment. *Acta tropica*, February 2008.

[4] D. Blei and J. Mcauliffe. Supervised topic models. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA, 2008.

[5] D. M. Blei, K. Franks, M. I. Jordan, and I. S. Mian. Statistical modeling of biomedical corpora: mining the caenorhabditis genetic center bibliography for genes related to life span. *BMC Bioinformatics*, 7(1), May 2006.

[6] D. M. Blei, M. I. Jordan, J. Callan, G. Cormack, C. Clarke, D. Hawking, and A. Smeaton. Modeling annotated data. *SIGIR Forum*, (SPEC. ISS.):127–134, 2003.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[8] G. Carvalho, C. Lefaucheur, C. Cherbonnier, D. Metivier, A. Chapel, M. Pallardy, M.-F. Bourgeade, B. Charpentier, F. Hirsch, and G. Kroemer. Chemosensitization by erythropoietin through inhibition of the nf-[kappa]b rescue pathway. *Oncogene*, aop(current), December 2004.

[9] C. Casals-Pascual, R. Idro, N. Gicheru, S. Gwer, B. Kitsao, E. Gitau, R. Mwakesi, D. J. Roberts, and C. R. Newton. High levels of erythropoietin are associated with protection against neurological sequelae in african children with cerebral malaria. *Proceedings of the National Academy of Sciences*, 105(7):2634–2639, February 2008.

[10] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines, 2001.

[11] K. B. Cohen and L. Hunter. *Natural language processing and systems biology*, pages 147–174. Springer, December 2004.

[12] R. Feldman, Y. Regev, E. Hurvitz, and M. Finkelstein-Landau. Mining the biomedical literature using semantic analysis and natural language processing techniques. *Drug Discovery Today: BIOSILICO*, 1(2), May 2003.

[13] C. W. Gay, M. Kayaalp, and A. R. Aronson. Semi-automatic indexing of full text biomedical articles. In *AMIA Annu Symp Proc*, pages 271–275, Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD 20894, USA., 2005.

[14] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc Natl Acad Sci U S A*, 101 Suppl 1:5228–5235, April 2004.

[15] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, 1999.

[16] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, V42(1):177–196, January 2001.

[17] S. Humphrey, C. Lu, W. Rogers, and A. Browne. Journal descriptor indexing tool for categorizing text according to discipline or semantic type. In *AMIA Annu Symp Proc*, 2006.

[18] S. M. Humphrey, W. J. Rogers, H. Kilicoglu, D. Demner-Fushman, and T. C. Rindflesch. Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: preliminary experiment. volume 57, pages 96–113, 2006.

[19] D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[20] A. Mccallum, A. Corrada-Emmanuel, and X. Wang. Topic and role discovery in social networks. 2005.

[21] A. Mccallum and K. Nigam. A comparison of event models for naive bayes text classification, 1998.

[22] T. M. Mitchell. *Machine Learning*. McGraw-Hill Science/Engineering/Math, March 1997.

[23] A. Névéol, S. E. Shooshan, S. M. Humphrey, T. C. Rindflesch, and A. R. Aronson. Multiple approaches to fine-grained indexing of the biomedical literature. In R. B. Altman, K. A. Dunker, L. Hunter, T. Murray, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, pages 292–303. World Scientific, 2007.

[24] A. Névéol, S. E. Shooshan, J. G. Mork, and A. R. Aronson. Fine-grained indexing of the biomedical literature: Mesh subheading attachment for a medline indexing tool. In *Proc. AMIA Symp*, 2007.

[25] D. Newman, C. Chemudugunta, and P. Smyth. Statistical entity-topic models. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 680–686, New York, NY, USA, 2006. ACM Press.

[26] S. D. Nimer. Myelodysplastic syndromes. *Blood*, 111(10):4841–4851, May 2008.

[27] M. F. Porter. An algorithm for suffix stripping. pages 313–316, 1997.

[28] S. R. Seong, J. W. Lee, Y. K. Lee, T. I. Kim, D. J. Son, D. C. Moon, Y. W. Yun, d. o. . Y. Yoon, and J. T. Hong. Stimulation of cell growth by erythropoietin in raw264.7 cells: association with ap-1 activation. *Archives of pharmacal research*, 29(3):218–223, March 2006.

[29] M. Steyvers and T. Griffiths. *Probabilistic Topic Models*. 2007.

[30] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315, New York, NY, USA, 2004. ACM Press.

[31] L. Wiese, C. Hempel, M. Penkowa, N. Kirkby, and J. A. L. Kurtzhals. Recombinant human erythropoietin increases survival and reduces neuronal apoptosis in a murine model of cerebral malaria. *Malaria Journal*, 7:3+, January 2008.

[32] Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1/2):69–90, 1999.

[33] B. Zheng, D. C. Mclean, and X. Lu. Identifying biological concepts from a protein-related corpus with a probabilistic topic model. *BMC Bioinformatics*, 7, 2006.