Bridging the Language Gap: Topic Adaptation for Documents with Different Technicality

Shuang Hong Yang **Steven P. Crain** **Hongyuan Zha**
College of Computing, Georgia Institute of Technology, Atlanta, GA 30318
{shy, s.crain, zha}@gatech.edu

Abstract

The language-gap, for example between low-literacy laypersons and highly-technical expert documents, is a fundamental barrier for cross-domain knowledge transfer. This paper seeks to close the gap at the thematic level via *topic adaptation*, i.e., adjusting the topical structures for cross-domain documents according to a domain factor such as technicality. We present a probabilistic model for this purpose based on joint modeling of topic and technicality. The proposed τLDA model explicitly encodes the interplay between topic and technicality hierarchies, providing an effective topic-level bridge between lay and expert documents. We demonstrate the usefulness of τLDA with an application to consumer medical informatics.

1 Introduction

Although knowledge access is easier today than ever with the availability of numerous information sources on the Internet, transferring knowledge across domains remains a critical challenge. Particularly, transferring expert knowledge to lay users is hampered by the fundamental *language-gap* – lay users do not have enough literacy to understand expert jargons and terminologies; likewise, experts might be unfamiliar with the slang words to best popularize their expertise to, or precisely capture the inquiries of, a common audience.

Existing research (Can & Baykal, 2007; Zeng et al., 2006) attempts to close the gap at word-level by exploiting shallow word-correlations based on machine translation techniques, e.g., by augmenting or substituting the words in a lay document with a bundle of technical words that are statistically or semantically similar to the original text's content. These approaches are not entirely satisfactory because the translation is neither interpretable nor organized. They also turn to confuse different semantic themes as documents are assumed to be topically homogeneous throughout the corpus, which is, however, generally not true (Blei et al., 2003). In this paper, we attempt to close the gap at a deeper thematic level with a *topic bridge* that connects different domains semantically. We propose *topic adaptation*, a framework that adapts the underlying topical structures (rather than content words) according to a domain factor (e.g., time, sentiment, technicality) while the topics are discovered from cross-domain texts.

Topic adaptation naturally arises in cross-domain topic modeling. Probabilistic topic models (Blei et al., 2003; Griffiths & Steyvers, 2004) interpret a document d as a mixture θ over a set of topical bases (multinomial distributions) β. A basic assumption in existing topic models is that all the documents within a corpus are drawn using the same *shared topical structures* (i.e., a single β). While this assumption works well for texts from a single domain, it is undesirable for texts from multiple domains where a significant language gap usually arises. For example, while lay texts in the topic "cancer" are dominated by common words like "cancer", "tumor" and "abnormal", an experts' knowledge base would favor technical words like "neoplasm", "carcinoma" and "metastasis". Naively learning the two domains with a single set of topical bases will inevitable lead to models with unacceptable fitting bias and in turn harm the quality of the extracted topics. Therefore, it is imperative to retain related but not identical topical bases β for each domain and capture the correlations among βs so that topics are both coherent within each domain and consistent across domains according to the changes of a domain factor (e.g., technicality). We refer to this problem as *topic adaptation* (TA).

The learning task involved in TA is challenging. On

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

the one hand, separately learning topical bases β locally from each domain corpus will lose the topical correspondence among domains — there is no guarantee that the k-th topic learned from lay texts is thematically relevant to the k-th topic of expert documents; On the other hand, simultaneously learning multiple βs from the joint corpus requires decomposing text contents into multiple sets of word occurrence patterns (βs), which is intractable due to the lack of appropriate supervision – we only observe the words in each text but not their technicality stamps (i.e., the degree of a word being technical). Of course, the task would be greatly eased if the quantity of technicality, ideally for each word in the vocabulary, could be assigned a priori. However, manual annotation is often too expensive (e.g., vocabulary is huge, and dominated power-law by rare words that could easily exceed any individual's scope of literacy) and unreliable (e.g., any annotator could easily bias toward her own interest areas) to be practical.

In this paper, we present the *topic-adapted latent Dirichlet allocation* (τLDA) for topic adaptation from cross-domain documents. The τLDA model devises a technicality-hierarchy in parallel to the topic-hierarchy of LDA, and encodes the interplay between the two hierarchies in the generative process. It leverages a mild supervision (the per-domain technicality stamps) to guide *cross-domain consistency*, making sure topics be adaptive to technicality changes. Moreover, it retains domain-specific topic bases βs to ensure *in-domain coherence*, which is efficiently parameterized via a two-mode mixture. We derive efficient inference and learning algorithm for τLDA based on variational Bayesian methods and evaluate it with an application to consumer medical informatics.

2 Related work

The language discrepancy between low-literacy laypeople and expert-produced documents has been widely recognized as a fundamental barrier for cross-domain knowledge transfer. (Zeng et al., 2002; Schwartzberg et al, 2005) observed that the language gap substantially degrades the performance of medical information services. (Uijttenbroek et al., 2008) reported a similar challenge in legal informatics. Conventional efforts (Kripalani et al., 2006; Bickmore et al., 2009) take a very manual approach, e.g., by educating clinicians and manually constructing communication scripts tailored for patients. Recently, several researches attempted to close this gap via word-level machine translation (Can & Baykal, 2007; Zeng et al., 2006). In contrast, we attempt to bridge domains at the topic level to capture the deeper thematic correlation among domains, with add-on benefits such as readily interpretable results (e.g. the topic and technicality structures offer a comprehensible organization of texts for browsing or summarization).

Topic models have been established for cross-domain texts, for example, the *cross-language topic models* (Zhao & Xing, 2006; Mimno et al., 2009; Boyd-Graber & Blei, 2009). These works, however, are fundamentally different from ours: in their setting, topics are multinomials over different vocabularies; whereas in ours, topics are different multinomials over the same vocabulary. In essence, we are addressing the subtle variations within a language, which are arguably (more) challenging. These models are also limited in applications as they require a corpus containing approximately parallel documents. Perhaps most relevant to our work is the *dynamic topical model* (Blei & Lafferty, 2006; Wang et al., 2008), which learns drifted topics from time-evolving domains. Although such *topic drifting* is a special type of *topic adaptation*, the assumptions for *time factor* (for example, causality and Markov assumptions) are less suitable for other domain factors such as technicality.

Technicality is an important factor of natural language text, yet (surprisingly) rarely explored in topic modeling. A noticeable exception is the recently proposed *hierarchical topic model* or hLDA (Blei et al., 2010), which extracts a tree structure for learned topics. Although the depth of each topic in the hLDA topic tree roughly reflects the degree of its specificity, the only guidance for learning the tree is a nonparametric prior (i.e., the nested Chinese restaurant process), which admits a plausible *monotonic constraint* for topics: high-specificity topics are always contained by lower ones. Because the regulation imposed by the nCRP prior is rather weak and diminishing quickly as observations increase, and the monotone assumption could be inaccurate, the technicality quantified by hLDA is usually unsatisfactory. Furthermore, hLDA is not applicable to topic adaptation for cross-domain documents as it models a single topic structure β.

Our prior work (Crain et al., 2010) established a model for extracting topic structures for different dialects of a language (slang, common and technical) using per-word technicality features. In this paper, we further examine technicality in a continuous range $\tau \in [0, 1]$ and explore how the topic structure β evolves according to technicality τ. We do so by extracting a family of *aligned topic structures* $\{\beta(\tau)|\tau \in [0, 1]\}$, where for any topic ID k the multinomials $\beta_k(\tau_1)$ and $\beta_k(\tau_2)$ are semantically about the same theme (with different technicalities). Such aligned topic structures serve as a *topic-level bridge* which allow us to safely assess the similarity (e.g., match query with documents in IR) of two documents, d_1 and d_1, solely in terms of the corresponding topic memberships θ_1 and θ_2 with-

out worrying how different their technicalities are or how they differ in BOW representations. The model is also practically appealing as it requires only mild corpus-level (rather than word-level) supervision.

3 Topic adaptation via topic-adapted latent Dirichlet allocation

Assume we have a family of domains $\{\mathcal{D}(\tau) : \tau \in [0,1]\}$, distinguished from each other with distinct values of a domain factor[1] τ. Without loss of generality, we assume τ is a continuous variable in the range [0,1]. A domain is a collection of documents with the same τ, $\mathcal{D}(\tau) = \{(d, \tau_d) : \tau_d = \tau\}$, and a document is a finite sequence of words $d = w_1 w_2 \dots w_n$. Our goal in topic adaptation is to infer the topical structures and guarantee its in-domain coherence and cross-domain consistency for a given multi-domain corpus $\mathcal{D} = \bigcup_{\tau \in [0,1]} \mathcal{D}(\tau)$.

As in the *latent Dirichlet allocation* (LDA) model (Blei et al., 2003), we decompose the document-word co-occurrence matrix in terms of document-specific topic mixtures θ_d and a set of topical bases β (multinomial distributions). However, instead of a single common set of bases as in LDA, we retain domain-specific topic bases $\{\beta(\tau) : \tau \in [0,1]\}$, which requires the nontrivial task of learning a functional family of multinomials. For simplicity, we adopt a two-mode mixture to efficiently parameterize $\beta(\tau)$. Particularly, we assume any $\beta(\tau)$ is a mixture of two extremes $\beta^0 = \beta(0)$ and $\beta^1 = \beta(1)$, that is: $\beta(\tau) = (1-\tau)\beta^0 + \tau\beta^1$.

We establish a probabilistic generative model for topic adaptation. The key innovations are as follows: (1) we assume a hierarchy for technicality, in parallel to the LDA topic hierarchy; (2) we model the interplays between the topic and technicality hierarchies at the latent level; and (3) we let each word sampling be conditioned on both latent topic and latent technicality assignments. Specifically, the *topic-adapted latent Dirichlet allocation* (τLDA) model assume the following generation process for each document-technicality pair, (d, τ_d), in the joint corpus $\mathcal{D}$:

1. Draw topic mixture $\theta \sim \text{Dir}(\alpha)$
2. For each topic, draw topic-level technicality $\pi_k \sim \text{Beta}(\lambda_{1k}, \lambda_{2k})$
3. For each word:
   a) Choose a topic assignment $z_n \sim \text{Mult}(\theta)$;
   b) Choose domain (i.e., technicality) assignment $t_n \sim \text{Bernoulli}(\pi_{z_n})$;
   c) Sample word $w_n \sim \text{Mult}(\beta^{t_n}_{z_n})$;
4. Generate document technicality $\tau \sim p(\tau | t_{1:N}, \omega)$.

[1]Hereafter, the domain factor τ always refers to *technicality*, although other factors such as *sentiment* and *time* might be equally applicable.

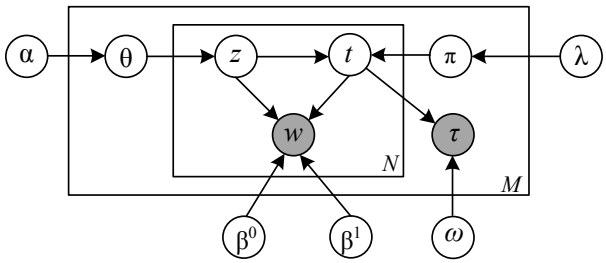

Figure 1: The topic-adapted latent Dirichlet allocation (τLDA) model for cross-domain texts.

In the model, we assume the number of topics, K, is a priori specified and fixed (in practice, it could be determined via Bayesian model comparison (Griffiths & Steyvers, 2004)). As in the plain LDA model, the per-document topic mixture θ is drawn from a K-dimensional Dirichlet distribution $\text{Dir}(\alpha)$, and the per-word topic assignment z is from a discrete distribution conditioned on θ, i.e., $\text{Mult}(\theta)$. In parallel to this topic hierarchy, we also model a technicality hierarchy. The per-document technicality π is a K-vector, with each entry π_k specifying the degree of technicality for each topic; each π_k is drawn independently from a Beta distribution[2] $\text{Beta}(\lambda_{1k}, \lambda_{2k})$. The per-word technicality assignment t is a binary scalar, it is generated conditioned on both π and z from $\text{Bernoulli}(\pi_z)$. Basically, when the i-th topic is sampled (i.e., $z_i = 1$) and its technicality π_i is given, t further specifies a domain – from which of the two extreme domains the topic is sampled (e.g., whether the "cancer" topic is talked about in layman or expert domain). Thereafter, a word is sampled conditioned on both topic and domain (technicality) assignment from a multinomial distribution $\text{Mult}(\beta^t_z)$, where $\beta^t = (\beta^0)^{1-t}(\beta^1)^t$; both β^0 and β^1 are $K \times V$ matrices, where V is the size of the vocabulary. Finally, the technicality stamp associated with each document is modeled as a response variable generated conditioned on all the technicality assignments: $p(\tau | \omega^\top \bar{y})$, where $\bar{y} = \frac{1}{N}\sum_{n=1}^N y_n$, $y_n = t_n z_n$ is a topic-aware code of t_n. For now, we use a cosine regression model that enjoys the best interpretability (Yang et al., 2010); other models will be explored later. Particularly, we assume $p(\tau | \omega^\top \bar{y}) = \frac{1}{Z}\exp(\tau \omega^\top \bar{y})$, a degraded log-linear model with constant partition $Z = \int_0^1 \exp(\tau \omega^\top \bar{y}) d\tau = \text{const}$. This model leads to regression by maximizing the Frobenius inner product between the model prediction and the ground-truth: $\omega = \arg\max \langle \tau_{1:M}, \omega^\top \bar{y}_{1:M} \rangle$.

The overall model is depicted with a graphical representation in Figure 1. For each (d, τ) pair, the joint

[2]In our implementation, we assume $\lambda_{1k} + \lambda_{2k} = \text{const}$ for all k to further reduce parameters.

distribution is given by:

$$\begin{aligned} P_d =& p(\theta, \pi, z_{1:N}, t_{1:N}, w_{1:N}, \tau | \alpha, \lambda, \beta^0, \beta^1, \omega) \\ =& p(\theta|\alpha) p(\tau|\omega^\top \bar{y}) \prod_{k=1}^K p(\pi_k|\lambda_{\cdot k}) \\ & \prod_{n=1}^N p(z_n|\theta) p(t_n|\pi_{z_n}) p(w_n|\beta^{t_n}_{z_n}). \end{aligned} \quad (1)$$

4 Inference and learning

Both parameter estimation and inferential tasks in τLDA involve the intractable computation of marginalizing P_d over the latent variables. In this section, we derive approximate algorithms based on variational methods (Jaakkola & Jordan, 2000).

4.1 Variational approximation

We lower bound the log likelihood by applying the mean-field variational approximation:

$$\begin{aligned} & \log p(d, \tau | \alpha, \lambda, \beta^0, \beta^1, \omega) = \log \int_{\theta,\pi} \sum_{z,t} P_d \\ =& \mathcal{L}(\gamma, \Phi, \eta, \mu) + KL(q||p) \approx \max_{\gamma,\Phi,\eta,\mu} \mathcal{L}(\gamma, \Phi, \eta, \mu), \end{aligned}$$

where the posterior $p(\theta, \pi, \mathbf{z}, \mathbf{t}|d, \tau, \alpha, \lambda, \beta, \omega)$ is approximated by a variational distribution q. Here, we assume a fully-factorized distribution (per document) on the latent variables:

$$\begin{aligned} & q(\theta, \pi, z_{1:N}, t_{1:N} | \gamma, \phi, \eta, \mu) \\ =& \mathrm{Dir}(\theta|\gamma) \prod_{k=1}^K \mathrm{Beta}(\pi_k|\eta_{\cdot k}) \prod_{n=1}^N \mathrm{Mult}(z_n|\phi_n) \mathrm{Ber}(t_n|\mu_n) \end{aligned}$$

Denote $\mathcal{H}_q$ the entropy of q, ℓ the operator $\log p(\cdot)$, the variational lower bound (variational free energy) of the log likelihood, $\mathcal{L}$, is given by:

$$\begin{aligned} \mathcal{L} =& \mathbb{E}_q[\ell(\theta|\alpha)] + \sum_{n=1}^N \mathbb{E}_q[\ell(z_n|\theta)] + \sum_{k=1}^K \mathbb{E}_q[\ell(\pi_k|\lambda_{\cdot k})] + \mathcal{H}_q \\ & + \sum_{n=1}^N \left(\mathbb{E}_q[\ell(t_n|\pi_{z_n})] + \mathbb{E}_q[\ell(w_n|\beta^{t_N}_{z_n})] \right) + \mathbb{E}_q[\ell(\tau|\omega^\top \bar{y})]. \end{aligned}$$

The terms in the first line are similar to those in LDA. The terms in the second line are given (in order) by:

$$\begin{aligned} \mathbb{E}_q[\ell_{[t]}] &= \sum_{k=1}^K \phi_{nk} \left(\mu_n \Psi(\eta_{1k}) + (1-\mu_n)\Psi(\eta_{2k}) - \Psi(\eta_{1k}+\eta_{2k}) \right) \\ \mathbb{E}_q[\ell_{[w]}] &= \sum_{k=1}^K \phi_{nk} \left(\mu_n \log \beta^1_{kv} + (1-\mu_n) \log \beta^0_{kv} \right) \quad (2) \\ \mathbb{E}_q[\ell_{[\tau]}] &= \frac{1}{N} \sum\nolimits_{n=1}^N \sum_{k=1}^K \tau \mu_n \omega_k \phi_{nk} \end{aligned}$$

where v is the index of w_n in the vocabulary. By setting the derivatives of $\tilde{\mathcal{L}}$ (the Lagrangian relaxation of $\mathcal{L}$) w.r.t. the variational parameters to zero, we obtain the following coordinate ascent algorithm:

$$\gamma_k = \alpha_k + \sum\nolimits_{n=1}^N \phi_{nk} \quad (3)$$

$$\eta_{1k} = \lambda_{1k} + \sum\nolimits_{n=1}^N \phi_{nk} \mu_n \quad (4)$$

$$\eta_{2k} = \lambda_{2k} + \sum\nolimits_{n=1}^N \phi_{nk} (1-\mu_n) \quad (5)$$

$$\begin{aligned} \phi_{nk} \propto & \ (\beta^0_{kv})^{1-\mu_n} (\beta^1_{kv})^{\mu_n} \exp\{\Psi(\gamma_k) - \Psi(\eta_{1k}+\eta_{2k}) \\ & + \mu_n \Psi(\eta_{1k}) + (1-\mu_n)\Psi(\eta_{2k}) + b_k \mu_n \} \end{aligned} \quad (6)$$

$$\mu_n = \zeta\Big(\sum_{k=1}^K \phi_{nk} \big(\Psi(\eta_{1k}) - \Psi(\eta_{2k}) + \log \frac{\beta^1_{kv}}{\beta^0_{kv}} + b_k\big)\Big) \quad (7)$$

where $\Psi(\cdot)$ is the digamma function, the logistic mapping $\zeta(x) = \frac{1}{1+\exp(-x)}$, and b_k is a supervision bias due to the response model. For the cosine regression model, we have $b_k = \frac{1}{N}\tau\omega_k$.

These formulas are intuitively interpretable. We observe that the per-word topic distribution, ϕ, is learned as a result of negotiation between the two extreme domains. This can be seen by rewriting Eqn.(6) as $\phi_{nk} \propto (\phi^0_{nk})^{1-\mu_n} (\phi^1_{nk})^{\mu_n}$, where $\phi^i_{nk} = E_q[z_{nk}|t_n = i]$, $i = 0$ or 1. Particularly, each word occurrence is split according to its technicality into two parts, μ_n and $1-\mu_n$; then ϕ^0 and ϕ^1 are inferred individually in each domain conditioned on topic, domain, word as well as technicality samplings; and finally, the two domains negotiate with each other and output the combined results ϕ_{nk}. Another interesting observation is how the algorithm assigns technicality for each word. It uses a *logistic regression* model, where the per-topic regressors (consisting of three parts: the prior contrast $\Psi(\eta_{1k}) - \Psi(\eta_{2k})$, the domain contrast $\log \beta^1_{kv}/\beta^0_{kv}$, and the supervision bias b_k) are weighted by per-word topic distribution ϕ and then mapped by a logistic function.

We finally note that our variational inference algorithm for τLDA is efficient enough. From Eqn.(3-7), it requires $O(KN)$ operations per iteration, which is the same complexity as LDA.

4.2 Parameter estimation

The parameters of τLDA are learned by maximizing the evidence lower bound:

$$\max L = \sum_{m=1}^M \mathcal{L}_m(\alpha, \lambda, \beta, \omega; \gamma_m, \phi_m, \eta_m, \mu_m).$$

This two-layer optimization involves two groups of parameters, corresponding to τLDA and its variational model respectively. Optimizing alternatively between these two groups leads to a Variational Expectation Maximization (VEM) algorithm, where the E-step corresponds to applying variational approximation (i.e.,

Table 1: The cross-domain corpus is a combination of document collections from five domains.

Domains	Yahoo!	PubMed	MeSH	CDC	WebMD
τ	0.0	1.0	1.0	0.7	0.3
#doc	74226	161637	25588	192258	275620

Eqn.(3-7)) to each observation (d_m, τ_m) in the corpus and the M-step maximizes L with respect to the model parameters. Particularly, for the topic bases, we have:

$$\begin{aligned}\beta^0_{kv} &\propto \sum_{m,n,k} (1-\mu_{mn})\phi_{mnk} w^v_{mn},\\ \beta^1_{kv} &\propto \sum_{m,n,k} \mu_{mn}\phi_{mnk} w^v_{mn},\end{aligned} \tag{8}$$

where $w^v_{mn} = 1(w_{mn} = v)$, $1(\cdot)$ is the indicator function. From Eqn.(8), we see again that each word occurrence is split according to its technicality into two part, μ and $1-\mu$, which contribute to the two extreme topic bases β^1 and β^0 respectively.

Then, both the topic and technicality mixture priors, α and λ, are solved (independently) by Newton-Raphson procedures conditioned on values of γ and η respectively. And finally, the response parameter, ω is learned by maximizing the conditional likelihood:

$$\begin{aligned}&\max_{\omega} \mathbb{E}_q[\log p(\tau_{1:M}|\bar{y}_{1:M}, \omega)] = \langle \tau_{1:M}, \omega^\top \mathbb{E}_q[\bar{y}_{1:M}]\rangle,\\ &\text{s.t.} : ||\tau_{1:M}|| = ||\omega^\top \mathbb{E}_q[\bar{y}_{1:M}]||,\end{aligned} \tag{9}$$

where we pose a constraint to eliminate the scale freedom of ω. Based on the Karush-Kuhn-Tucker optimality of Eqn.(9), we derive a very simple close-form solution for ω:

$$\hat{\omega} = \bar{h}/||\bar{h}||_A,$$

where $\bar{h} = \frac{1}{M}\sum_m \tau_m \mathbb{E}_q[\bar{y}_m]$, $\mathbb{E}_q[\bar{y}] = \frac{1}{N}\sum_n \mu_n \phi_{nk}$, and $A = \mathbb{E}_q[\bar{y}_{1:M}]\mathbb{E}_q[\bar{y}_{1:M}]^\top / ||\tau_{1:M}||^2$, $||x||_A = \sqrt{x^\top A x}$ denotes the A-weighted l_2-norm.

4.3 Technicality analysis

Here, we derive empirical Bayesian methods to quantify technicality at different granularities. The first task is to predict the *document technicality*, which enables *domain identification* (Yang et al., 2009; 2010). For a given document d, we first run variational inference on d, then, we have: $\hat{\tau} = \omega^\top \bar{y}$. Note that the terms involving the supervision bias should be removed[3] in variational inference as τ is unobserved for incoming documents.

At a more compact level, *topic technicality* directly reflects the specificity of each topic, similar to the node-depth in the hLDA topic tree (Blei et al., 2010). Here, we have:

$$\hat{\pi}_k = \mathbb{E}_d(\mathbb{E}_q[\pi_{mk}|d_m]) = \frac{1}{M}\sum_{m=1}^{M} \frac{\eta_{m,1k}}{\eta_{m,1k} + \eta_{m,2k}}. \tag{10}$$

Finally, *word technicality* analysis provides a function mapping for the vocabulary: $t(v) : \mathcal{V} \to [0, 1]$, which quantifies the relative specificity of a word w.r.t a targeted expertise-intensive domain and also the relative difficulty for a lay user to grasp. Again, we use empirical Bayesian:

$$\begin{aligned}\hat{t}_v &= \mathbb{E}_d(\mathbb{E}_w[\mathbb{E}_q\{t_{mn}|w_{mn} = v\}|d_m])\\ &= \sum_{m,n} w^v_{mn}\mu_{mn} / \sum_{mn} w^v_{mn}.\end{aligned}$$

5 Experiments

In this section, we apply τLDA to medical documents. We wish to find how a same topic is expressed differently in lay and expert languages, and how topics are shifted according to domain technicality.

Data As shown in Table 1, our corpus is a combination of documents collected from five different domains. The *Yahoo!* subset is a collection of user questions and corresponding answers from the health category of *Yahoo! QA* (`answer.yahoo.com`), representing lay domain labeled with lowest technicality (τ=0). The *PubMed* (medical journal articles from `www.pubmedcentral.nih.gov`) and *MeSH* (medical subject descriptors from `www.nlm.nih.gov/mesh`), in the other extreme, represent expert domain with highest technicality (τ=1). In between, *WebMD* (documents crawled from `www.webmd.com`) represents mildly non-technical domain (τ=0.3), and *CDC* (crawled from `www.cdc.gov`) mildly technical domain (τ=0.7)[4]. Note that these coarsely-assigned per-domain technicality labels, $\{\tau\}$, are the only pieces of supervision information we used for topic adaptation in τLDA.

Results The language gap leads to a substantial discrepancy of word usages between different domains, making it difficult to maintain a global vocabulary that is effectively balanced across domains (Otherwise, the

[3]For the cosine regression model, set τ_d=0.5 $\forall d$; for LR and LAD, set $b_k = 0$.

[4]WebMD is intended to be generally comprehensible, yet it contains substantially more technical words than Yahoo!. CDC is intended for both medical experts and public readers, more technical than average.

Table 2: Example topics found by τLDA: each topic is shown by the top-ten words in both layman domain (β^0) and expert domain (β^1); the top row indicates the technicality of each topic.

#1: $\pi = 0.06$		#2: $\pi = 0.15$		#3: $\pi = 0.18$		#4: $\pi = 0.19$		#5: $\pi = 0.26$		#6: $\pi = 0.54$	
β^0	β^1	β^0	β^1	β^0	β^1	β^0	β^1	β^0	β^1	β^0	β^1
who	protein	problem	activ	better	nucleic	how	relat	treatment	gene	recommend	data
ask	associ	risk	chemic	below	structur	think	program	weight	analysi	medicin	method
much	immunolog	you	process	she	same	you	report	your	determin	not	import
bodi	psycholog	your	therapi	children	inhibitor	someth	previous	food	blood	tell	deriv
you	purif	fill	substanc	farther	genom	femal	web	profession	model	littl	depart
not	virolog	thought	poison	abl	possibl	googl	various	fda	amino	past	diagnost
eat	enzymolog	skin	conserv	print	pcr	mmwr	file	health	enzym	test	measur
period	induc	anyth	organ	transmiss	express	histori	databas	diet	biosynthesi	quit	complet
sometim	parasitolog	face	virus	season	chromosom	partner	establish	fat	signal	social	design
agre	patholog	regular	cell	treatment	dna	websit	analys	dose	yeast	progress	generat

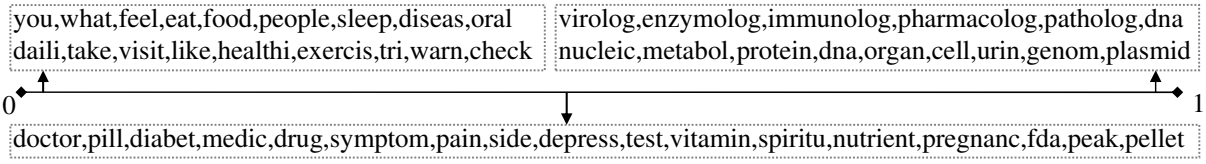

Figure 2: Example words with low-, medium- and high- technicalities.

vocabulary could be extremely skewed such that the majority of words come from lay domains). To this end, we first select terms (after stemming and stopword removal) locally from each domain based on DF (document frequency) scores, and then interleave the sub-selection round-robin to form the global vocabulary (over 10K words).

An important issue for implementing τLDA is how to initialize β's. Although reasonable results are obtained by totally random initialization, we find that a simple pre-feeding initialization procedure leads to substantial performance improvement. Particularly, we first profile the technicality for each word by using the empirical average technicality of the training documents containing the word, i.e. $\hat{\mu}_v = \sum_{\{d:f_d^v>0\}} f_d^v \tau_d$, where f_d^v denotes the term frequency of word v in document d; we then train a plain LDA model and compute the initial β's using Eqn.(8), where the pre-feeded word technicalities $\hat{\mu}$ are used in place of μ.

We train the τLDA model on a 60% subset of randomly sampled documents, and test on the rest. The results are averaged over 5 repeats. The variational algorithm is efficient: for each iteration, τLDA takes (on average) 7.6 more time than LDA (the LDA-C implementation) to converge.

Table 2 shows an intuitive view of six example topics found by τLDA. For each topic, we list the top-10 most probable words for lay (i.e., high values of β^0) and expert (i.e., high β^1) domains respectively. These results reveal a notable language gap between the two domains – almost all the representative words for lay domain are commonly-used or even slang words, in contrast, most words on the expert-domain side are highly-technical medical terminology (for example, words suffixed by "-ology"). The language gap is even evident when the same topic is concerned, indicating that laypeople and experts interpret differently even the same ideas.

As a reference, the technicality of each topic is also shown in the top row of Table 2. A very interesting observation is that there is no highly-technical topic — the maximum technicality for topics are around 0.5. In essence, this indicates that the language gap is asymmetric: experts can occasionally talk about topics of lay interests (but in a language mixing common words and their jargon), but laypeople are unlikely to be interested in expert's highly-technical topics. The absence of highly-technical topics in lay domain makes the corresponding word-occurrence pattern too submissive (infrequent) in the overall corpus to be captured by the model. To validate this hypothesis, we plot the topic distance between domains $D(\beta_k^0||\beta_k^1)$ (the Jensen-Shannon divergence) as a function of topic technicality π_k in Figure 3(a). As expected, we see an evident negative correlation between D and π, suggesting that the closer π_k is to the middle, the more β_k^0 and β_k^1 are overlapped. Also note that the technicality of a topic k does not depend on the word-distribution of β_k^0 or β_k^1, but rather on how much probability (i.e., relative frequency, see also Eq.(10)) a topic is present in technical than lay domain. Therefore, although the first topic in Table 2 seemingly covers the most technical words in its β^1, it has the lowest technicality – it clusters very common word in β^0 and very-technical word in β^1, but the former appears far more frequent than the latter.

β^0 and β^1 are two different distributions over the same vocabulary. Because different words have different intrinsic frequencies (e.g. technical words are less frequent), a better way to understand the learned topics might be to label each topic with most representative terms based on foreground-background contrast (e.g. by selecting words with highest ratio scores β^τ/β, where β is a background multinomial regardless of domains, $\tau \in \{0, 1\}$). According to this analysis[5], the two sets of topic bases are nicely aligned. Here, for ease of comparison with results of existing topic models, we comply with conventional topic labeling standard and demonstrate (in Table 2) each topic with most frequent words (i.e. solely according to foreground multinomial). Even from this somewhat naive analysis, we can still see that, except the first one, the two topical structures are approximately aligned. For example, topic #2 is about beauty and health, #3 birth and heredity, #4 medical records, #5 diet and weight control, #6 diagnosis and laboratory, etc. This observation indicates that, for a given k, β^0_k and β^1_k are roughly talking about the same topic. Such aligned topical bases are the key to cross-domain knowledge transfer. They fundamentally provide a topical bridge between lay domain and expert domain such that (1) documents from different domains can be mapped to the same simplex space $\mathcal{S} = \{\theta : ||\theta||_1 = 1, \theta_k \geqslant 0\}$, and (2) the distances between θ's precisely captures the semantic similarity between documents, no matter they are from same or different domains.

To quantitatively evaluate the quality of topic alignment, we perform an information retrieval task based on the topic mixture θ learned by τLDA. Our evaluation is confined by the availability of labeled data. As a preliminary test, we use a small number of lay documents as queries to retrieve technical documents. The results are manually graded on a 4-point scale ranging from 0 (irrelevant) to 4 (relevant). Based on this small data set of 25 queries with 100 documents per query [6], we report the performance in terms of the *normalized Discounted Cumulative Gain* on the top-five results (*nDCG@5*). The *nDCG@5* for τLDA is as high as 0.51 – a huge improvement over 0.38 of LDA based retrieval model (Wei & Croft, 2006).

τLDA also provide a simple mechanism to quantify technicality for words, which was previous achieved only by sophisticated models (e.g., hLDA). A rough view of word technicality learned by τLDA is given in Figure 2. We see that most results reasonably coincide with human intuition.

[5] `www.cc.gatech.edu/~syang46/Topic_Label.50.txt`

[6] Labeled data for this task is extremely expensive as it requires annotators with moderate medical knowledge. We are working on collecting more labeled data from paid annotators and extending this experiment.

It would be interesting to examine the relationship between the topic bases learned by LDA (i.e., β) and those by τLDA (β^0 and β^1). For this purpose, we use an element-wise interpolation: $\beta_{kv} = x_{kv}\beta^1_{kv} + (1 - x_{kv})\beta^0_{kv}$ or $\log\beta_{kv} = x_{kv}\log\beta^1_{kv} + (1 - x_{kv})\log\beta^0_{kv}$, and examine the distribution of the interpolation coefficient x. We find that the interpolations are distributed quite diversely: (1) a majority of β entries (about 68.3%, see also Figure 3(d)) are within the convex span of β^0 and β^1 (i.e., $x \in [0, 1]$), the rest 31.7% are not; (2) while the distribution peaks around $x = 0$ and $x = 1$, there is no singe dominant x that could fit all the entries well; (3) the correlation between $[x_{kv}]_{1:V}$ and $[t_v]_{1:V}$ is very low (in the range $[-0.05, 0.1]$), hence using a global per-word technicality function t_v to assist plain LDA (as in the initialization procedure) could not work either. These observations indicate that the interaction with technicality has fundamentally changed the topic structure so that from β to (β^0, β^1) is non-trivially a nonlinear decomposition. From another perspective, the observations also suggest that no *single* topic structure β is able to interpret a cross-domain corpus adequately well, hence, models with a unanimous β such as LDA will inevitably lead to substantial learning bias if brutally applied to this scenario.

We also examined the TF difference (i.e., $\beta^1_{kv} - \beta^0_{kv}$ or $\log\beta^1_{kv} - \log\beta^0_{kv}$) between domains as a function of word-technicality t_v. Figure 33(b&c) shows the relationship for an example topic k, where the differences are normalized to [-1,1]. We see that the topic structures are nicely consistent with technicality: technical words are more frequent in technical domain than in lay domain, and vice versa.

We finally report the prediction performance of τLDA. Our first evaluation is based on the test-set log likelihood (Wallach et al., 2009; Chang et al., 2009), a commonly used measure for topic models. We compare[7] τLDA with LDA and its supervised version (sLDA, (Blei & Mcauliffe, 2008)). The results are shown in Figure 3(e). We see that τLDA significantly outperforms both LDA and sLDA. This observation suggests that, by retaining topic bases for each domain, τLDA is more suitable for cross-domain topic learning than the other two competitors, which learn a single structure β for all the domains. We then apply τLDA to domain identification, i.e., to predict technicality τ_d for an unseen document d. Considering that our labeling of τ is very coarse (piecewise constant) and that precisely quantifying the degree of technicality for each domain is usually impractical in practice, this task re-

[7] We use fold-in evaluation, e.g. for τLDA: $p(w_n|\mathcal{D}) = \sum_k \hat{\phi}_k(\hat{\mu}_n\hat{\beta}^1_{kv} + (1 - \hat{\mu}_n)\hat{\beta}^0_{kv})$, where v is the ID of w_n in vocabulary. This comparison is fair across different models.

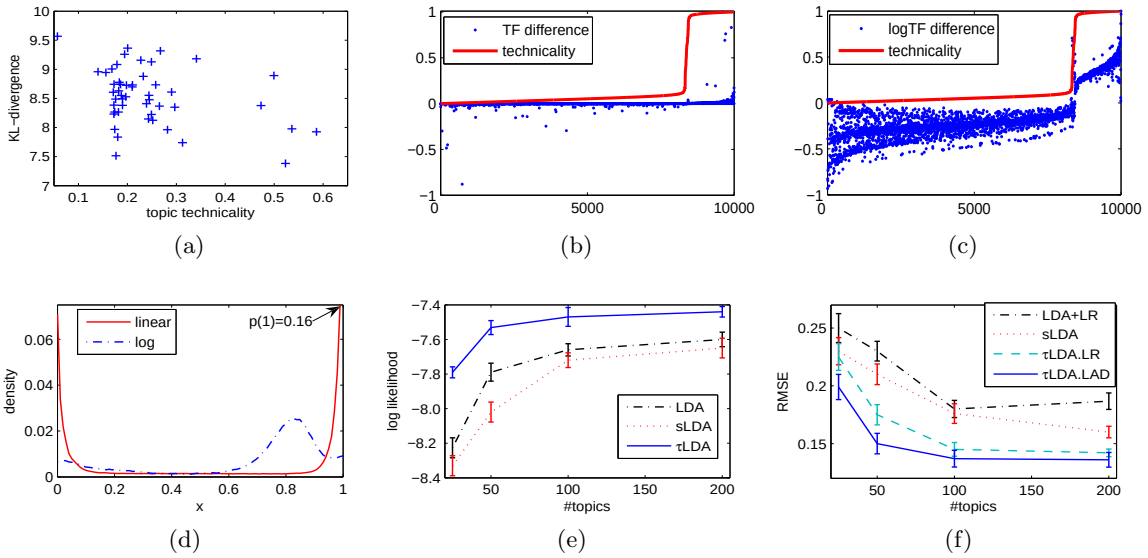

Figure 3: (a) topic variation vs. topic-technicality; (b-c) TF variation vs. word-technicality; (d) τLDA topics and LDA topic interpolation; (e) test-set predictive likelihood; (f) domain identification accuracy.

quires a model capable of handling noisy data. The cosine regression model is too sensitive to noise to fulfill this purpose. Here, we consider two other response models. The first one is linear regression (LR):

$$p(\tau|\omega^\top \bar{y}) = \mathcal{N}(\tau|\omega^\top \bar{y}, \sigma^2);$$

The other is least absolute deviation (LAD):

$$p(\tau|\omega^\top \bar{y}) = \mathfrak{L}(\tau|\omega^\top \bar{y}, \delta),$$

where $\mathfrak{L}$ denotes the Laplacian distribution. For these two models, the inference algorithms are almost the same as that of cosine regression except that the supervision bias b_k is different. In particular, for LR:

$$b_k = \frac{1}{N\sigma^2}\tau\omega_k - \frac{1}{2N^2\sigma^2}[\sum_{i\neq n}\sum_j \omega_j\omega_k\phi_{ij}\mu_m + \omega_k^2].$$

For LAD, we have:

$$b_k = \text{sign}(\tau - \mathbb{E}_q[\bar{y}])\frac{\omega_k}{N\delta}, \text{ where } \mathbb{E}_q[\bar{y}] = \frac{1}{N}\sum_{nk}\omega_k\mu_n\phi_{nk}.$$

Similarly, the learning procedure is different only in estimating ω. Specifically, for the LR model:

$$\hat{\omega} = (Y^\top Y)^{-1}YT \text{ where } Y = \bar{y}_{1:M}, T = \tau_{1:M}.$$

The LAD regression leads to an *iterative reweighted least square* algorithm, which iterative updates:

$$\Lambda^{\text{new}} = \text{diag}(\hat{\omega}^{\text{old}\top} Y),$$

$$\hat{\omega}^{\text{new}} = (Y^\top \Lambda^{\text{new}} Y)^{-1}Y^\top \Lambda^{\text{new}} T.$$

The results are reported in Figure 3(f) with the *root mean squared error* (RMSE) as evaluation metric. The performance of τLDA with cosine regression is much worse than the others (RMSE>0.3) and is therefore omitted. We can see that, although sLDA is worse than LDA in terms of predictive log-likelihood, it obtains better technicality prediction than LDA; yet, the two τLDA variants consistently outperform both LDA and sLDA (over 20% improvements). Also, less surprisingly, the LAD version of τLDA obtains significantly better performance than the LR variant as the former is more robust to noise.

6 Conclusion

We presented a generative model to learn related topic structures for documents from multiple domains. The τLDA model encodes both topic and domain factor (e.g., technicality) hierarchies as well as the interactions between them, providing an effective way to discover topic structures that are coherent within each domain and consistent among domains. The model offers a topic-level bridge for cross-domain knowledge transfer as demonstrated in eHealth tasks.

Today's personalized information services (e.g. Web 2.0) call for machine learning algorithms that are capable of capturing such subtle cognitive aspects of users (e.g. interests, capability, literacy, expertise, learning style) from their contextual texts and in turn adapting services accordingly. The τLDA offers a promising startpoint for learning user's literacy and expertise. It would be interesting to explore how other cognitive aspects of a user can be captured based on the texts she crafted/read. We plan to extend τLDA and the generative models for supervised classification and disambiguation (Yang et al., 2009; 2010) for this purpose.

Acknowledgement

The authors would like to thank Yu Jiao (@ORNL) and the anonymous reviewers for helpful comments. Part of this work is supported by NSF #IIS-1049694 and a grant from Hewlett-Packard.

References

Bickmore, T. W., Pfeifer, L. M.,and Jack, B. W. Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In *CHI' 2009*, pp. 1265–1274, 2009.

Blei, David M. and Lafferty, John D. Dynamic topic models. In *ICML' 2006*, pp. 113–120, 2006.

Blei, David M. and Mcauliffe, Jon D. Supervised topic models. In *Advances in Neural Information Processing Systems 21*, pp. 121–128, 2008.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Blei, David M., Griffiths, Thomas L., and Jordan, Michael I. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):1–30, 2010.

Boyd-Graber, Jordan and Blei, David M. Multilingual topic models for unaligned text. In *UAI' 2009*, pp. 75–82. 2009.

Can, Aysu Betin and Baykal, Nazife. Medicoport: A medical search engine for all. *Computer Methods and Programs in Biomedicine*, 86(1):73–86, 2007.

Chang, Jonathan, Boyd-Graber, Jordan, Gerrish, Sean, Wang, Chong, and Blei, David. Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems 22*, pp. 288–296, 2009.

Crain, Steven P., Yang, Shuang-Hong, and Zha, Hongyuan. Dialect Topic Modeling for Improved Consumer Medical Search. *Annual Symposium of American Medical Informatics Association*, 2010.

Griffiths, Thomas L. and Steyvers, Mark. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, 2004.

Jaakkola, Tommi S. and Jordan, Michael I. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

Kripalani, S., Jacobson, K. L., Brown, S., Manning, K., Rask, K. J., and Jacobson, T. A. Development and implementation of a health literacy training program for medical residents. *Medical Education Online*, 11, 2006.

Mimno, David, Wallach, Hanna M., Naradowsky, Jason, Smith, David A., and McCallum, Andrew. Polylingual topic models. In *EMNLP' 2009*, pp. 880–889, 2009.

Schwartzberg, Joanne G., VanGeest, Jonathan B. and Wang, Claire C. *Understanding health literacy: implications for medicine and public health*. 2005.

Uijttenbroek, E. M., Lodder, A. R., Klein, M., Wildeboer, G. R., Steenbergen, W., Sie, R., Huygen, P.E. M., and Harmelen, F. Retrieval of case law to provide layman with information about liability: Preliminary results of the best-project. In *Computable Models of the Law: Languages, Dialogues, Games, Ontologies*, pp. 291–311. Springer, 2008.

Wallach, Hanna M., Murray, Iain, Salakhutdinov, Ruslan, and Mimno, David. Evaluation methods for topic models. In *ICML' 2009*, pp. 1105–1112, New York, NY, USA, 2009. ACM.

Wang, Chong, Blei, David, and Heckerman, David. Continuous time dynamic topic models. In *UAI' 2008*, 2008.

Wei, Xing and Croft, W. Bruce. Lda-based document models for ad-hoc retrieval. In *SIGIR' 2006*, pp. 178–185, 2006.

Yang, S.H., Zha, H., and Hu, B.-G. Dirichlet-bernoulli alignment: A generative model for multi-class multi-label multi-instance corpora. In *Advances in Neural Information Processing Systems 22*.

Yang, S.H., Bian, J., and Zha, H. Hybrid generative/discriminative learning for automatic image annotation. In *UAI' 2010*.

Zeng, Q., Kogan, S., Ash, N., Greenes, R. A., and Boxwala, A. A. Characteristics of consumer terminology for health information retrieval. *Methods of Information in Medicine*, 41(4):289–298, 2002.

Zeng, Q., Crowell, J., Plovnick, R. M., Kim, E., and nd Emily Dibble, L.. Assisting consumer health information retrieval with query recommendations. *Journal of the American Medical Informatics Association*, 13(1):80–90, 2006.

Zhao, Bing and Xing, Eric P. Bitam: bilingual topic admixture models for word alignment. In *ACL' 2006*, pp. 969–976, 2006.