Extracting Topical Phrases from Clinical Documents

Yulan He
School of Engineering and Applied Science
Aston University, UK
y.he@cantab.net

Abstract

In clinical documents, medical terms are often expressed in multi-word phrases. Traditional topic modelling approaches relying on the "bag-of-words" assumption are not effective in extracting topic themes from clinical documents. This paper proposes to first extract medical phrases using an off-the-shelf tool for medical concept mention extraction, and then train a topic model which takes a hierarchy of Pitman-Yor processes as prior for modelling the generation of phrases of arbitrary length. Experimental results on patients' discharge summaries show that the proposed approach outperforms the state-of-the-art topical phrase extraction model on both perplexity and topic coherence measure and finds more interpretable topics.

Introduction

Topics models such as Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) have been extensively used to automatically discover topical themes from text documents. However, most of the topic models rely on the bag-of-words assumption by ignoring word orders and the extracted topics are often listed as a sequence of word unigrams which could hinder the interpretability of the discovered topics. In clinical documents, medical terms are often expressed in multi-word phrases, for example, "blood glucose" and "white blood cell". These two phrases, if split into unigrams, would lose their original semantic meanings. Also, they might be grouped under the same topic by the unigram topic models because of the common word "blood". For this reason, simply splitting documents into word unigrams would generate ambiguous and spurious topics (Arnold and Speier 2012).

There are in general three categories of approaches to topical phrase extraction. The first one is to pre-process documents by extracting phrases, either based on some statistical measures for word collocation terms detection or through frequent pattern mining, and then run traditional topic models such as LDA on the "bag-of-phrases". Since each extracted phrase is considered as a single term, the resulting vocabulary is significantly enlarged which leads to more sparse data. The topical phrase mining method (ToPMine) (El-Kishky et al. 2014) also operates LDA on the "bag-of-phrases". But it decomposes each phrase and imposes a constraint in LDA inference that all the constituent words within the same phrase should be assigned with the same topic. However, their method cannot detect less frequent phrases sharing common lower-order n-grams, for example, drugs with different dosage levels such as "ciprofloxacin 500 mg" and "ciprofloxacin 250 mg". The second category of approaches is to extract topical phrases as a post-processing step to unigram topic models (Blei and Lafferty 2009; Danilevsky et al. 2014). Such approaches assume that words which are simultaneously labeled with the same topic many times can be grouped as a phrase. However, in unigram topic models, words within the same phrase may not be assigned with the same topic. Moreover, as topic models are typically run on data with stopwords removed, the post-process approaches would have a difficulty in recognising phrases which contain stopwords such as "short of breath". The last category of approaches combine phrase boundary detection with topic inference into a unified model. Examples include the Topic N-Gram (TNG) model (Wang, McCallum, and Wei 2007), the phrase-discovering topic model (PDLDA) (Lindsey, Headden III, and Stipicevic 2012) and the N-gram Topic Segmentation model (Jameel and Lam 2013). But the models are computationally too complex and the topical phrases detected often have lower quality.

In this paper, we propose a new approach which lies in between the aforementioned categories 1 and 3 of topical phrase extraction approaches. We first extract the medical phrases from patient's discharge summaries using an off-the-shelf medical concept mention extraction tool. The phrases extracted are guaranteed to be of high quality and are also clinical-relevant. Then we learn a topic model which takes a hierarchy of Pitman-Yor Processes (PYPs) as priors. This allows the capture of n-grams of arbitrary length naturally by taking into account word orders within phrases. As will be shown in our experimental results, the proposed approach outperforms the other topical phrase models in terms of both perplexity and topic coherence measure and generates more interpretable topics.

We proceed to describe related work in topic modelling of clinical documents and phrase discovery topic models. We then discuss how we extract medical phrases and present our proposed Topical Phrase Model (TPM) followed by experimental results. Finally, we conclude the paper.

Related Work

Topic Modelling of Clinical Documents

Arnold and William (2012) proposed a topic model that captures temporal topic patterns in an individual patient's medical record. In such a model, each patient has his or her own timeline consisting of a subset of topics. In each of the patients' clinical reports, the hidden topics not only generate observed words, but also generate the timestamp associated with each report. Paul and Dredze (2013) used a "three-dimensional" LDA variant to jointly model combinations of drug (marijuana, salvia, etc.), aspect (effects, chemistry, etc.) and route of administration (smoking, oral, etc.) for generating extractive summaries about drug usage from the web. In both models, documents are modelled as finite mixtures over an underlying set of latent topics which are inferred from word co-occurrence patterns with word order ignored. As such, some of the extracted topics are not necessarily clinical relevant and often post-processing is required in order to filter out clinically irrelevant topics. For example, Zeng et al. (2006) proposed to identify topics relevant to biology based on calculating the mutual information between the topics and the controlled vocabulary of the Gene Ontology (GO) terms tagged in biomedical documents.

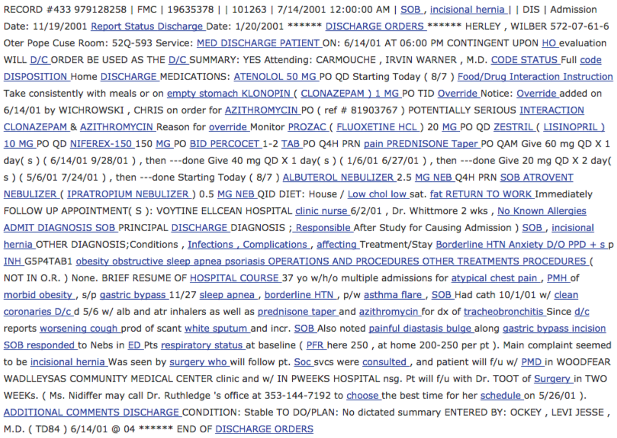
Other approaches performed pre-processing on clinical documents before applying the LDA model. For example, Yu et al. (2013) first extracted noun phrases from clinical documents and then run LDA on the "bag-of-noun-phrases". Lehman et al. (2014) extracted a set of UMLS codes from each patient's hospital discharge summary. They then trained a topic model on documents containing unordered sets of UMLS codes. The resulting topics have been shown more easier to interpret. However, they restricted the UMLS codes to three categories only, either disease, symptom, or finding, and ignore other UMLS codes and words. As such, the scope of study is limited.

Phrase Discovery Topic Models

The early topic model which goes beyond bag-of-words is the Bigram Topic Model (BTM) (Wallach 2006) in which each word is generated from the distribution over words for the context defined by a latent topic and the previous word. Wang et al. (2007) proposed a Topical N-Gram (TNG) model which extended BTM by introducing a switch variable at each word position to signal either the start of a new n-gram or a continuation of a previously identified n-gram. In TNG however, words within a n-gram do not share the same topic. Post-processing is required to take the topic of the final word in a n-gram as the topic of the entire n-gram. Also, in TNG, the topic-specific bigram distributions do not share probability mass with their constituent unigram distributions. To overcome these drawbacks, Lindsay et al. (2012) proposed a phrase-discovering topic model (PDLDA) which used the hierarchical Pitman-Yor process (HPYP) priors for the topic-word matrix. Jameel and Lam (2013) proposed a topic segmentation model which can detect topics at the segment level and at the n-gram word level. All these models essentially involve an additional step in learning phrase boundaries apart from topic detection. Performing phrase segmentation and topic detection simultaneously is computationally more expensive. Also, phrases detected in this way often have lower quality.

More recently, El-Kishky et al. (2014) proposed a topical phrase mining method called TopMine which consists of two steps. It first discover phrases from text using a method similar to frequent pattern mining commonly used in association rule mining, and then train a LDA model on the "bag-of-phrases" input under the constraint that words in the same phrase should be assigned with the same topic. Our proposed approach also follows a two-step process in which we first extract medical phrases and then train a topical phrase model on the "bag-of-phrases" input. However, unlike ToPMine which does not explicitly model the generation of n-grams, our model naturally captures n-grams with arbitrary length through the use of HPYPs. Also, as phrase detection is separated from topic inference, our approach is computationally less complicated compared to PDLDA or the N-gram Topic Segmentation model.

Extracting Medical Phrases



RECORD #433 979128258 | FMC | 19635378 | | 101263 | 7/14/2001 12:00:00 AM | SOB , incisional hernia | | DIS | Admission Date: 11/19/2001 Report Status Discharge Date: 1/20/2001 ****** DISCHARGE ORDERS ****** HERLEY , WILBER 572-07-61-6 Oter Pope Cuse Room: 52Q-593 Service: MED DISCHARGE PATIENT ON: 6/14/01 AT 06:00 PM CONTINGENT UPON HO evaluation WILL D/C ORDER BE USED AS THE D/C SUMMARY: YES Attending: CARMOUCHE , IRVIN WARNER , M.D. CODE STATUS Full code DISPOSITION Home DISCHARGE MEDICATIONS: ATENOLOL 50 MG PO QD Starting Today (8/7) Food/Drug Interaction Instruction Take consistently with meals or on empty stomach KLONOPIN (CLONAZEPAM) 1 MG PO TID Override Notice: Override added on 6/14/01 by WICHROWSKI , CHRIS on order for AZITHROMYCIN PO (ref # 81903767) POTENTIALLY SERIOUS INTERACTION CLONAZEPAM & AZITHROMYCIN Reason for override Monitor PROZAC (FLUOXETINE HCL) 20 MG PO QD ZESTRIL (LISINOPRIL) 10 MG PO QD NIFEREX-150 150 MG PO BID PERCOCET 1-2 TAB PO Q4H PRN pain PREDNISONE Taper PO QAM Give 60 mg QD X 1 day(s) (6/14/01 9/28/01) , then ---done Give 40 mg QD X 1 day(s) (1/6/01 6/27/01) , then ---done Give 20 mg QD X 2 day(s) (5/6/01 7/24/01) , then ---done Starting Today (8/7) ALBUTEROL NEBULIZER 2.5 MG NEB Q4H PRN SOB ATROVENT NEBULIZER (IPRATROPIUM NEBULIZER) 0.5 MG NEB QID DIET: House / Low chol low sat. fat RETURN TO WORK Immediately FOLLOW UP APPOINTMENT(S): VOYTINE ELLCEAN HOSPITAL clinic nurse 6/2/01 , Dr. Whittmore 2 wks , No Known Allergies ADMIT DIAGNOSIS SOB PRINCIPAL DISCHARGE DIAGNOSIS ; Responsible After Study for Causing Admission) SOB , incisional hernia OTHER DIAGNOSIS;Conditions , Infections , Complications , affecting Treatment/Stay Borderline HTN Anxiety D/O PPD + s p INH G5P4TAB1 obesity obstructive sleep apnea psoriasis OPERATIONS AND PROCEDURES OTHER TREATMENTS PROCEDURES (NOT IN O.R.) None. BRIEF RESUME OF HOSPITAL COURSE 37 yo w/h/o multiple admissions for atypical chest pain , PMH of morbid obesity , s/p gastric bypass 11/27 sleep apnea , borderline HTN , p/w asthma flare , SOB Had cath 10/1/01 w/ clean coronaries D/c d 5/6 w/ alb and atr inhalers as well as prednisone taper and azithromycin for dx of tracheobronchitis Since d/c reports worsening cough prod of scant white sputum and incr. SOB Also noted painful diastasis bulge along gastric bypass incision SOB responded to Nebs in ED Pts respiratory status at baseline (PFR here 250 , at home 200-250 per pt). Main complaint seemed to be incisional hernia Was seen by surgery who will follow pt. Soc svcs were consulted , and patient will f/u w/ PMD in WOODFEAR WADLLEYSAS COMMUNITY MEDICAL CENTER clinic and w/ IN PWEEKS HOSPITAL nsg. Pt will f/u with Dr. TOOT of Surgery in TWO WEEKs. (Ms. Nidiffer may call Dr. Ruthledge 's office at 353-144-7192 to choose the best time for her schedule on 5/26/01). ADDITIONAL COMMENTS DISCHARGE CONDITION: Stable TO DO/PLAN: No dictated summary ENTERED BY: OCKEY , LEVI JESSE , M.D. (TD84) 6/14/01 @ 04 ****** END OF DISCHARGE ORDERS

Figure 1: An example of medical term extraction result.

We extract medical phrases from text using a medical term extraction system built upon an open source toolkit called MedTagger[1], which combines machine learning and knowledge bases to identify medical concept mentions in clinical text. MedTagger assigns each extracted medical concept mention to one of the 15 semantic groups which are further classified into 133 subgroups. MedTagger has been shown constantly achieving an F-score of over 0.84 at various NLP challenges on the medical concept mention extraction task (Liu et al. 2012). An advantage of MedTagger is that it also performs concept mention normalisation. For example, a medical term "SOB" extracted can be identified as the medical phrase "short of breath". In cases where an extracted term can be mapped to multiple possible medical phrases by MedTagger, we simply take the first identified phrase as the

[1] http://www.ohnlp.org/index.php/MedTagger

normalised medical term. Figure 1 shows an example output generated by the medical phrase extraction system where detected medical terms/phrases are highlighted in blue colour. Clicking on any medical term brings up a dialog box showing the annotation results including the corresponding normalised phrase, its mapped semantic group, etc.

Topical Phrase Model (TPM)

Once the medical phrases have been identified, each document consists of a mixed set of unigram words and phrases. We propose in this section a Topical Phrase Model (TPM) which is able to learn hidden topics from text and model the generation of n-grams. Since TPM is built upon the Hierarchical Pitman-Yor Process (HPYP), we first describe HPYP before discussing the details of TPM.

Hierarchical Pitman-Yor Process (HPYP)

In previous work (Teh 2006), the hierarchical Pitman-Yor process language model (HPYLM) has been shown to recover exactly the formulation of interpolated Kneser-Ney (Chen and Goodman 1999), one of the best smoothing methods for n-gram language models. In HPYLM, a unigram word distribution $G_\emptyset$ is first generated from the PYP as:

$$G_\emptyset \sim \text{PYP}(a_0, b_0, G_0), \tag{1}$$

where G_0 is a uniform distribution over a fixed vocabulary $\mathcal{W}$ of V words, $G_0(w) = \frac{1}{V}; \forall w \in \mathcal{W}$, a_0 is the discount parameter, b_0 is called concentration parameter which controls the amount of variability of $G_\emptyset$ around the prior G_0. Then bigram word distributions $\{G_u\}_{u \in \mathcal{W}}$ is generated using $G_\emptyset$ as the base distribution, $G_u \sim \text{PYP}(a_1, b_1, G_\emptyset)$. Trigram word distributions $\{G_{uv}\}_{(u,v) \in \mathcal{W}^2}$ can then be successively generated from $G_{uv} \sim \text{PYP}(a_2, b_2, G_u)$. This process continues until the context length reaches $n-1$. In general, given a context $\boldsymbol{u}$ consisting of a sequence of up to $n-1$ words, the distribution over the current word w is generated by

$$G_{\boldsymbol{u}} \sim \text{PYP}(a_{n-1}, b_{n-1}, G_{\boldsymbol{u}'}) \tag{2}$$

where $\boldsymbol{u}'$ is the suffix of $\boldsymbol{u}$ consisting of all but the first word.

The generative process of drawing words from the prior is analogous to the generalised Chinese Restaurant Process (CRP) (Pitman 2002) where a restaurant corresponds to each $G_{\boldsymbol{u}}$ which has an infinite number of tables and each of which has infinite seating capacity. Each table is served a dish chosen from the base distribution $G_{\boldsymbol{u}'}$ (i.e., a distinct value drawn from the base distribution $G_{\boldsymbol{u}'}$). A sequence of customers corresponding to words drawn from $G_{\boldsymbol{u}}$ arrives in the restaurant. The first customer sits at the first table; the $(n+1)$th customer chooses an occupied table in proportional to the number of customers already sitting down and share the dish with other customers, or chooses a new table in proportional to some constant parameter and order a dish from the base distribution. Choosing a dish is equivalent to sending the new table as a proxy customer to the parent restaurant in a recursive manner. This process repeats until the proxy customer chooses to sit in an existing table or there is no more parent restaurant.

In HPYLM, the probability of a word following context $\boldsymbol{u}$ given the seating arrangement is:

$$P_{\boldsymbol{u}}(w|\Lambda) = \frac{c_{\boldsymbol{u}w} - at_{\boldsymbol{u}w}}{c_{\boldsymbol{u}\cdot} + b} + \frac{at_{\boldsymbol{u}\cdot} + b}{c_{\boldsymbol{u}\cdot} + b} P_{\boldsymbol{u}'}(w|\Lambda), \tag{3}$$

where $c_{\boldsymbol{u}w}$ is the number of customers having dish w in restaurant $\boldsymbol{u}$, $t_{\boldsymbol{u}w}$ is the number of tables serving w in restaurant $\boldsymbol{u}$ and Λ denotes the current seating arrangement. A dot is used to indicate marginal counts (i.e., $c_{\boldsymbol{u}\cdot} = \sum_w c_{\boldsymbol{u}w}$ and $t_{\boldsymbol{u}\cdot} = \sum_w t_{\boldsymbol{u}w}$). For the global base distribution, the predictive probability is $P_\emptyset(w|\Lambda) = G_\emptyset(w)$. Equation 3 corresponds to interpolated Kneser-Ney which estimates the probability of word w following context $\boldsymbol{u}$ by discounting the true count $c_{\boldsymbol{u}w}$ by a fixed amount $at_{\boldsymbol{u}w}$ and interpolates the estimated probability of word w with lower order m-gram probabilities (Teh 2006).

Generative Process of TPM

We consider a Bayesian nonparametric version of topic model where a topic is assigned to each word token from a document-specific multinominal distribution and a word is generated from a topic-specific distribution taking the PYP as priors. Moreover, a HPYP process is used which allows the modelling of n-gram word sequences of arbitrary length. Compared to the Dirichlet prior commonly used in LDA, using PYP as priors is more appropriate to deal with natural language since PYP can capture the fact that words in natural language follows a power law.

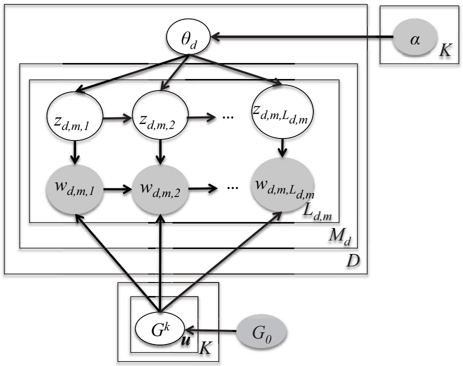


Figure 2: Topical Phrase Model (TPM).

Our proposed Topical Phrase Model is illustrated in Figure 2 and the generative process is described below:

- For each topic $k \in \{1, .., K\}$
  - First generate a unigram word distribution, $G^k_\emptyset \sim \text{PYP}(a_0, b_0, G_0)$
  - Then given a context $\boldsymbol{u}$ consisting of a sequence of up to $n-1$ words, generate a n-gram word distribution, $G^k_{\boldsymbol{u}} \sim \text{PYP}(a_{n-1}, b_{n-1}, G^k_{\boldsymbol{u}'})$, where $\boldsymbol{u}'$ denotes the parent context of $\boldsymbol{u}$.
- For each document $d \in \{1, .., D\}$
  - Choose a topic distribution $\theta_d \sim \text{Dirichlet}(\alpha)$

- For each phrase $m \in \{1,..,M_d\}$, and for each word within a phrase $i \in \{1,..,L_{d,m}\}$
  - If it is the first word in the phrase, i.e., $l = 1$
    - Select a topic $z_{d,m,l} \sim \text{Discrete}(\theta_d)$
    - Draw a word $w_{d,m,l} \sim \text{Discrete}(G_{\emptyset}^{z_{d,m,l}})$
  - Else
    - Set $z_{d,m,l} = z_{d,m,l-1}$
    - Draw a word $w_{d,m,l}|\boldsymbol{u} \sim \text{Discrete}(G_{\boldsymbol{u}}^{z_{d,m,l}})$

Note that in the plate diagram shown in Figure 2, we have explicitly shown the phrase plate that a document d contains a total of M_d phrases and each phrase m contains $L_{d,m}$ words. In our implementation of the Gibbs sampling procedure of TPM, we still loop over every single word from a total of N_d words in document d. For each word w_{di} at position i of document d, we use a switch variable x_{di} to indicate whether the word w_{di} should be concatenated with the previous word $w_{d,i-1}$ to form a multi-term phrase. If $x_{di} = 0$, then w_{di} is either the first word of a multi-term phrase or a single word by its own, and a topic z_{di} is sampled from a topic-specific multinomial distribution and a word is drawn from a topic-specific unigram distribution. If $x_{di} = 1$, then the word w_{di} is part of a multi-term phrase, and its topic z_{di} is taken to be the same as $z_{d,i-1}$ and a word is drawn from a topic-specific distribution conditioned on its context $\boldsymbol{u}$ which includes all the previous words in the phrase. Since we have already identified all the phrases as has been discussed in the previous section, the switch variable $\mathbf{x}$ is observed and does not need to be sampled from data.

In TPM, we place PYP as a prior distribution over word probabilities. For each topic $k \in 1,..,K$, we can generate n-gram word distributions by the PYP, $G_{\boldsymbol{u}}^k \sim \text{PYP}(a_{n-1}, b_{n-1}, G_{\boldsymbol{u}'}^k)$.

To build the Gibbs sampling algorithm, we first derive the joint distribution over words $\boldsymbol{w}$, topic assignments $\boldsymbol{z}$, table configuration $\boldsymbol{r}$. Let $\Omega = \{a_0, b_0, \alpha\}$, we have:

$$P(\boldsymbol{z},\boldsymbol{w},\boldsymbol{r}|\Omega) = \prod_{k=1}^{K} P(G_{\boldsymbol{u}}^k|a_0,b_0,G_0) \prod_{d=1}^{D} P(\theta_d|\alpha)$$
$$\prod_{i=1}^{N_d} P(z_{di}|\theta_d) P(w_{di}|G_{\boldsymbol{u}}^{z_{di}}). \quad (4)$$

If we omit Ω for clarity and let the index $y = (d;i)$ and the subscript $\setminus y$ denote a quantity that excludes counts in word position i of document d, we have:

$$P(z_y = k|\boldsymbol{z}_{\setminus y}, \boldsymbol{w}_{\setminus y}, \boldsymbol{r}_{\setminus y}) \propto P(z_y = k|\boldsymbol{z}_{\setminus y}) P_{\boldsymbol{u}}^k(w_y|\boldsymbol{w}_{\setminus y}, \boldsymbol{r}_{\setminus y}),$$

where $P(z_y = k|\boldsymbol{z}_{\setminus y}) = \frac{C_{d,k\setminus y}+\alpha_k}{C_{d\setminus y}+\sum_z \alpha_z}$ if the yth word is a unigram or the first word in a phrase. Here, $C_{d,k}$ is the number of times topic label k being assigned to some word tokens in document d. If the yth word is part of a multi-term phrase, we simply take $z_y = z_{y-1}$ and do not sample a topic.

As we aim to capture n-grams under each topic, there are a hierarchy of PYP distributions which model word context of different length for each topic. Let Λ denote the current seating arrangement, the generation of next word w from $G_{\boldsymbol{u}}^k$ can be computed recursively as:

$$P_{\boldsymbol{u}}^k(w|\Lambda) = \frac{c_{\boldsymbol{u}w}^k - a_{n-1}t_{\boldsymbol{u}w}^k}{c_{\boldsymbol{u}\cdot}^k + b_{n-1}} + \frac{a_{n-1}t_{\boldsymbol{u}\cdot}^k + b_{n-1}}{c_{\boldsymbol{u}\cdot}^k + b_{n-1}} P_{\boldsymbol{u}'}^k(w|\Lambda).$$

For a word $w \in \mathcal{W}$, the context $\boldsymbol{u}$ consists of a sequence of $n-1$ words, $\boldsymbol{u} \in \mathcal{W}^{n-1}$, and $\boldsymbol{u}'$ is the context consisting of all words in $\boldsymbol{u}$ except the first one. We use $c_{\boldsymbol{u}w}^k$ to denote the number of customers eating dish w in restaurant $\boldsymbol{u}$ owned by k (i.e., the number of occurrences of w following $\boldsymbol{u}$ in topic k), $t_{\boldsymbol{u}w}^k$ to denote the number of tables serving dish w in restaurant $\boldsymbol{u}$ owned by k, $c_{\boldsymbol{u}\cdot}^k = \sum_w c_{\boldsymbol{u}w}^k$ and $t_{\boldsymbol{u}\cdot}^k = \sum_w t_{\boldsymbol{u}w}^k$. For a unigram w,

$$P_{\emptyset}^k(w|\Lambda) = \frac{c_{\emptyset w}^k - a_0 t_{\emptyset w}^k}{c_{\emptyset\cdot}^k + b_0} + \frac{a_0 t_{\emptyset\cdot}^k + b_0}{c_{\emptyset\cdot}^k + b_0}\frac{1}{V},$$

where $\emptyset$ denotes no previous word and V is the vocabulary size.

PDLDA also used HPYP priors. However, PDLDA only involves a single hierarchical n-gram language model with each topic considered as part of context in $\boldsymbol{u}$, i.e., for the current word w_i, its context is defined as $\boldsymbol{u} =< z_i, w_{i-1}, w_{i-2}, ..., w_{i-n+1} >$ which consists of both topic and the preceding $n-1$ words. Nevertheless, topics are not part of the word vocabulary. Our proposed TPM instead generates a separate n-gram language model for each topic. Also, the hyperparameters a_{n-1} and b_{n-1} are set to different values for each context length n (different depth in HPYP) by auxiliary variable sampling (Teh et al. 2006), but are shared across all topics.

Experiments

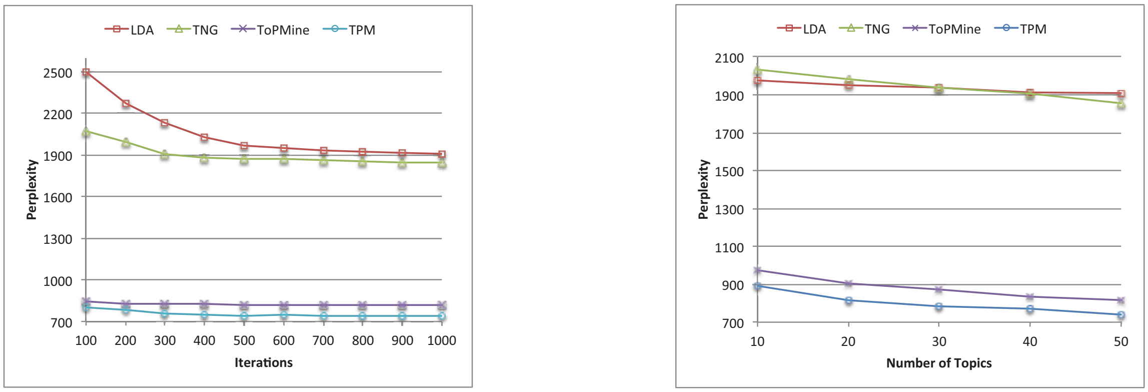
We use the clinical record data released as part of the i2b2 Natural Language Processing Challenges for Clinical Records (Uzuner et al. 2010). The data contains a total of 1,243 de-identified discharge summaries. The original challenge focused on the identification of medications, their dosages, modes (routes) of administration, frequencies, durations, and reasons for administration in discharge summaries. In our experiments here, we focused on extracting topical phrases from discharge summaries.

Each document is pre-processed to remove common stopwords and clinical stopwords such as "status report", "discharge", "Dr." etc. We use our medical phrase extraction system built upon MedTagger to identify phrases from documents. We did not perform stemming. The total number of word tokens is 791,097. The vocabulary size is 7,738 in the "bag-of-words" representation, and 32,893 in the "bag-of-phrases" representation. It can be seen that if considering each phrase as a single token, the vocabulary size is significantly enlarged.

We train TPM with a maximum of 1,000 Gibbs sampling iterations and stop if the total log-likelihood converges. We optimise all the hyperparameters including α and a_{n-1}, b_{n-1} for different context length n in HPYP every 50 iterations. We compare our proposed approach with the following baselines:

- LDA. We use the MALLET[2] implementation of the LDA model to extract topics from our pre-processed data where medical phrases have already been identified and are

[2] http://mallet.cs.umass.edu/



(a) Perplexity vs. Gibbs sampling iterations.

(b) Perplexity vs. different topic numbers.

Figure 3: A comparison of perplexity values of various topic models.

treated as single tokens. For all the hyperparameters, we use the default settings and perform optimisation every 50 Gibbs sampling iterations.

- TNG (Wang, McCallum, and Wei 2007). The Topical N-Gram model is used to simultaneously detect n-grams and infer topics. Again, the MALLET implementation of TNG is used with the default hyperparameter settings.
- ToPMine (El-Kishky et al. 2014). This approach first extracts phrases using a method similar to frequent pattern mining commonly used in association rule mining and then train a modified LDA model on the "bag-of phrases" input. It has been shown outperforming a number of phrase discovery topic models including TNG and PDLDA.

Perplexity

Perplexity has been commonly used in evaluating topic models' prediction ability on unseen data. It is defined as the reciprocal geometric mean of the likelihood of a test corpus. Lower perplexity implies better predictiveness, and hence a better model. We use 10% of the data as a held-out set and compare how different models perform in predicting the held-out set. Figure 3(a) shows the perplexity values versus Gibbs sampling iterations when the topic number is set to 50. It can be observed that TNG has better perplexity values compared to LDA. Both ToPMine and TPM perform significantly better than TNG and LDA with much lower perplexity values. We also vary the number of topics and observe a general trend that perplexity values decrease with the increasing number of topics for all the models as shown in Figure 3(b). TNG and LDA perform similarly while ToPMine and TPM achieve much lower perplexities with the best performance given by TPM.

Topical Coherence

Various topic coherence measures have been proposed to evaluate topics regarding their understandability. It has been recently reported in (Röder, Both, and Hinneburg 2015) through an extensive study that a new coherence measure based on a combination of some known approaches gives the best results in terms of approximating the human ratings of topic interpretability, outperforming all the other existing topic coherence measures including the widely used measure based on the pointwise mutual information (PMI) of all word pairs in the given top topic words (Newman et al. 2010). In particular, the new coherence measure retrieves cooccurrence counts for the given words from Wikipedia using a sliding window of size 110. For each of the top n words in a given topic, the normalised PMI value with respect to every other top words is calculated based on the cooccurrence counts. Thus, each top word is represented as a vector of normalised PMI values. The coherence measure of each topic is the arithmetic mean of the cosine similarity measurement of all vector pairs.

We report in Figure 4 the topic coherence measure calculated on the top 10 words/phrases of each topic based on the method proposed in (Röder, Both, and Hinneburg 2015). It can be observed that in general the coherence value increases with the increasing number of topics. TNG has the worst topic coherence values compared to the other models. ToPMine only slightly outperforms LDA. TPM consistently gives superior performance over all the other models for all the topic settings.

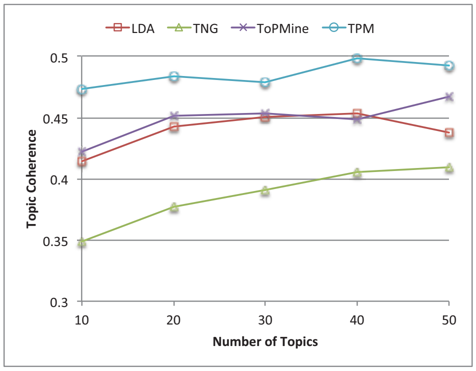


Figure 4: Topic coherence measure vs. number of topics.

Topic 1	Topic 2	Topic 3	Topic 4
LDA			
swallowing	infection	surgery	chest_pain
transferred	vancomycin	repair	aspirin
intubated	antibiotics	wound	heart
intensive_care	culture	signs	ischemia
speech	fever	removed	myocardial_ischemia
wean	intervertebral	nasogastric_tube	hypertension
extubated	culture_blood	diets	set
tube_feeds	fluid	female	lipitor
arrest	levaquin	hospital_course	coronary_artery_disease
aspiration	gentamicin	diverticulitis	emergency_department
ToPMine			
heart failure	chest pain	chest x ray	alert override override added
heart rate	shortness of breath	chest x ray showed	override notice override added
heart transplant	nausea and vomiting	physical therapy	reason for override aware
rate control	dyspnea on exertion	neck supple	previous override information
heart sounds	increased shortness of breath	showed no evidence	po ref potentially serious interactions
heart anatomy	pain control	chest x ray result	reason for override aware
distant heart sounds	substernal chest pain	chest x ray revealed	reason for override
respiratory distress	chest pressure	head and neck	alert override override added on
filled pressure	back pain	motor vehicle accident	order for coumadin
respiratory failure	atypical chest pain	neck supple no adenopathy	reason for override will monitor
TPM			
restrictive lung disease	right coronary artery	ciprofloxacin 500 mg	dressing changes
pleuritic chest pain	left upper extremity	levofloxacin 500 mg	superficial femoral artery
ferrous sulfate 325 mg	systemic vascular resistance	levofloxacin 250 mg	great toe
dyspnea on exertion	systolic ejection murmur	ciprofloxacin 250 mg	right fourth toe
vq scan	pulmonary vascular resistance	metronidazole 500 mg	vancomycin 250 mg
breath chest pain	flash pulmonary edema	chronic urinary tract infection	plastic surgery
interstitial lung disease	shortness of breath	white blood cell count	cellulitis of right foot
morbid obesity	transesophageal echocardiogram	benign prostatic hyperplasia	amputation of right foot
arterial blood gas	left internal mammary artery	recurrent urinary tract	split thickness skin graft
pulmonary function tests	coronary artery disease	irbesartan 150 mg	bone and bone

Table 1: Topic examples extracted from 50-topic runs. Each column shows the top 10 words/phrases ordered by likelihood.

Qualitative Evaluation Results

We list in Table 1 some example topics extracted from the 50-topic run. Since TNG has the lowest topic coherence scores compared to all the other models, we do not list the TNG topics due to the space constraint.

It can be observed that topic words listed under LDA topics are still dominated by unigrams. This is not surprising since LDA was simply operated on the "bag-of-phrases". The occurrence frequencies of most phrases are usually much lower than those of unigram words. As such, only a few of them appear in the top 10 words for each topic. ToPMine extracts phrases based on frequent pattern mining. It tends to group phrases sharing common words to form a topic. For example, most top words in Topic 1 has a common word "heart". While for Topic 3 and 4, the common word shared among most top words is "x-ray" and "override", respectively. TPM, on the contrary, is able to detect topics comprising a diverse range of words. More interestingly, TPM can detect symptom, diagnosis method and medication for certain diseases. For example, Topic 1 is about "lung disease". The top words include the disease name ("restrictive lung disease", "interstitial lung disease"), symptom ("pleuritic chest pain", "dyspnea on exertion", "morbid obesity"), diagnosis method ("vq scan", "arterial blood gas", "pulmonary function tests") and possible medication ("ferrous sulfate 325 mg"). Also, some topics show drugs with different dosages. For example, Topic 3 includes the antibiotics, Ciprofloxacin and Levofloxacin, with different dosages, which are both used to treat urinary tract infection. Detecting topics which consist of phrases at such a fine granularity level would not be possible with LDA run on "bag-of-phrases" or other topic models without explicitly modelling the generation of n-grams.

Conclusions

In this paper, we have proposed a new approach which first detects high quality phrases and then trains a topic model which explicitly models the generation of n-grams of arbitrary length. Compared to existing methods relying on frequent pattern mining for phrase identification, our approach can detect less frequent topical phrases but sharing common lower-order n-grams, for example, the same drug with different dosage levels. Also, since phrase detection is separated from topic inference, our approach is computationally less complicated compared to models which need to detect phrase boundaries and infer topics simultaneously. The experimental results show that our approach outperforms the state-of-the-art method in both perplexity and topic interpretability.

Acknowledgments

The author would like to thank Udochukwu Orizu for implementing a medical phrase extraction system based on MedTagger. The work is partly funded by the Innovate UK under the grant number 101779.

References

Arnold, C., and Speier, W. 2012. A topic model of clinical reports. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1031–1032.

Blei, D. M., and Lafferty, J. D. 2009. Visualizing topics with multi-word expressions. *arXiv preprint arXiv:0907.1013*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of machine Learning research* 3:993–1022.

Chen, S. F., and Goodman, J. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language* 13(4):359–393.

Danilevsky, M.; Wang, C.; Desai, N.; Ren, X.; Guo, J.; and Han, J. 2014. Automatic construction and ranking of topical keyphrases on collections of short documents. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*.

El-Kishky, A.; Song, Y.; Wang, C.; Voss, C. R.; and Han, J. 2014. Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment* 8(3):305–316.

Jameel, S., and Lam, W. 2013. An unsupervised topic segmentation model incorporating word order. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR)*, 203–212.

Lehman, L.-w.; Long, W.; Saeed, M.; and Mark, R. 2014. Latent topic discovery of clinical concepts from hospital discharge summaries of a heterogeneous patient cohort. In *Proceedings of the 36th IEEE International Conference on Engineering in Medicine and Biology Society (EMBC)*, 1773–1776.

Lindsey, R. V.; Headden III, W. P.; and Stipicevic, M. J. 2012. A phrase-discovering topic model using hierarchical pitman-yor processes. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 214–222.

Liu, H.; Wu, S. T.; Li, D.; Jonnalagadda, S.; Sohn, S.; Wagholikar, K.; Haug, P. J.; Huff, S. M.; and Chute, C. G. 2012. Towards a semantic lexicon for clinical natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2012, 568–576.

Newman, D.; Lau, J. H.; Grieser, K.; and Baldwin, T. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 100–108.

Paul, M. J., and Dredze, M. 2013. Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 168–178.

Pitman, J. 2002. Combinatorial stochastic processes. Technical report, Department of Statistics, University of California at Berkeley.

Röder, M.; Both, A.; and Hinneburg, A. 2015. Exploring the space of topic coherence measures. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining (WSDM)*, 399–408.

Teh, Y. W.; Jordan, M. I.; Beal, M. J.; and Blei, D. M. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association* 101(476).

Teh, Y. W. 2006. A hierarchical bayesian language model based on Pitman-Yor processes. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, 985–992.

Uzuner, Ö.; Solti, I.; Xia, F.; and Cadag, E. 2010. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association* 17(5):519–523.

Wallach, H. M. 2006. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning (ICML)*, 977–984.

Wang, X.; McCallum, A.; and Wei, X. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM)*, 697–702.

Yu, Z.; Johnson, T. R.; and Kavuluru, R. 2013. Phrase based topic modeling for semantic information processing in biomedicine. In *Proceedings of the 12th IEEE International Conference on Machine Learning and Applications (ICMLA)*, volume 1, 440–445.

Zheng, B.; McLean, D. C.; and Lu, X. 2006. Identifying biological concepts from a protein-related corpus with a probabilistic topic model. *BMC bioinformatics* 7(1):58.