Deep Learning for Topical Words and Thematic Sentences

Jen-Tzung Chien JTCHIEN@NCTU.EDU.TW
Ying-Lan Chang YLCHANG@CHIEN.CM.NCTU.EDU.TW
Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan

Abstract

This paper presents a hierarchical theme and topic model for deep representation of sentences and words from heterogeneous documents. We extract the latent themes from sentences and simultaneously identify the latent topics from words in different sentences. In this study, we flexibly conduct structural learning according to the Bayesian nonparametrics where the numbers of themes and topics are unknown. A tree stick-breaking process is proposed to determine the theme proportions for sentence representation. Hierarchical Dirichlet process is adopted to sample the topical words of a text corpus under the same theme. In the experiments, the proposed method is evaluated to be effective for finding topical words and thematic sentences in DUC 2007 corpus.

1. Introduction

Machine learning is generally categorized into supervised learning and unsupervised learning. Supervised learning aims to find a function mapping data into their class while unsupervised learning has a broad goal of finding features and discovering the structure within the given data. There have been intensively increasing interests in developing the unsupervised learning methods and exploring the *deep architectures* with multiple layers of nonlinearities (Hinton et al., 2012). In the literature, the unsupervised learning via latent topic model has been popular for document categorization (Blei et al., 2003), speech recognition (Chien & Chueh, 2011), text segmentation (Chien & Chueh, 2012) and image analysis (Blei et al., 2010a). The latent semantic topics are learnt from a bag of words. Such topic model can capture salient themes embedded in data collection and apply for document summarization (Chang & Chien, 2009)(Chang et al., 2011). However, topic model based on latent Dirichlet allocation (LDA) (Blei et al., 2003) was constructed as a finite-dimensional mixture representation which assumed that 1) number of topics was fixed, and 2) topics were independent. The hierarchical Dirichlet process (HDP) (Teh et al., 2006) and the nested Chinese restaurant process (nCRP) (Blei et al., 2004) (Blei et al., 2010b) were proposed to relax these two assumptions. HDP is a Bayesian nonparametric extension of LDA where the representation of documents is allowed to grow structurally as more data are observed. Each word token within a document is drawn from a mixture model where the hidden topics are shared across documents. Dirichlet process (DP) is realized to find flexible data partitions or provide the nonparametric prior over number of topics for each document. The base measure for the child DPs is itself drawn from a parent DP.

In this study, we develop a hierarchical tree model for deep representation of heterogeneous documents. Each path from root node to leaf node covers from general theme to individual theme. These themes contain coherent information but in varying degrees of sharing. The brother nodes expand the diversity of themes from different sentences. This model does not only group sentences into a node in terms of its theme but also distinguish their concepts through different levels. A structural stick-breaking process is proposed to draw multiple paths and determine a sub-tree of theme proportions. We conduct deep learning and group the sentences with a diversity of themes and concepts. The number of themes and their dependency are learnt from the collected data. The words of the sentences inside a node are represented by a topic model which is drawn by DP. All the topics from different nodes are shared under a global DP. This approach is evaluated for deep learning of latent topics and themes from text documents.

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

2. Prior Works

2.1. Bayesian Nonparametrics

There have been many Bayesian nonparametric approaches developed for discovering latent features in a variety of real-world data. DP plays a crucial role in constructing mixture model with a countably infinite number of hidden variables. Bayesian inference is performed by integrating out the infinitely many parameters. HDP (Teh et al., 2006) conducts Bayesian nonparametric representation of documents or grouped data. Each document d is associated with a draw from a DP G_d, which determines how much each member of a shared set of mixture components contributes to that document. The base measure of G_d is itself drawn from a global DP G_0 which ensures that there is a set of mixtures shared across data. Each distribution G_d governs the generation of words for a document d. The strength parameter α_0 determines the proportion of a mixture in a document d. The document distribution G_d is generated by $G_0 \sim \mathrm{DP}(\gamma, H)$ and $G_d \sim \mathrm{DP}(\alpha_0, G_0)$ where γ and H denote the strength parameter and base measure, respectively. HDP is developed to represent a bag of words from a set of documents through nonparametric prior G_0.

In (Blei et al., 2004)(Blei et al., 2010b), the nCRP was proposed to conduct Bayesian nonparametric inference of topic hierarchies and learn the deeply branching trees from data collection. Using this hierarchical LDA (hLDA), each document was modeled by a path of topics along a random tree where the hierarchically-correlated topics from global topics to specific topics were extracted. In general, HDP and nCRP could be implemented by using stick-breaking process and Chinese restaurant process. The approximate inference algorithms via Markov chain Monte Carlo (MCMC) (Blei et al., 2004)(Blei et al., 2010b)(Teh et al., 2006) and variational Bayesian (Paisley et al., 2011)(Teh et al., 2007)(Wang & Blei, 2009) were developed.

2.2. Stick-Breaking Process

Stick-breaking process is designed to implement infinite mixture model according to a DP. Beta distribution is introduced to draw binary variables for stick-breaking into left segment and right segment. A random probability measure G is first drawn from a DP with base measure H using a sequence of beta variates. Using this process, a stick of unit length is partitioned at a random location. The left segment is denoted by θ_1. The right segment is further partitioned at a new location. The partitioned left segment is denoted by θ_2. We continue this process by generating the left segment θ_i and breaking the right segment at each step i. Stick-breaking depends on a random value drawn from H which is seen as center of probability measure. The distribution over sequence of proportions $\{\theta_1, \cdots, \theta_i\}$ is called GEM distribution which provides a distribution over infinite partitions of unit interval (Pitman, 2002). In (Adams et al., 2010), a tree stick-breaking process was proposed for inference of a tree structure. This method interleaved two stick-breaking procedures. The first has beta variates for depth which determine the size of a given node's partition. The second has beta variates for branch which determine the branching probabilities. Interleaving two procedures could partition the unit interval into a tree structure.

3. Deep Document Representation

In this study, a hierarchical theme and topic model (H2TM) is proposed for deep representation of text documents through Bayesian nonparametric learning.

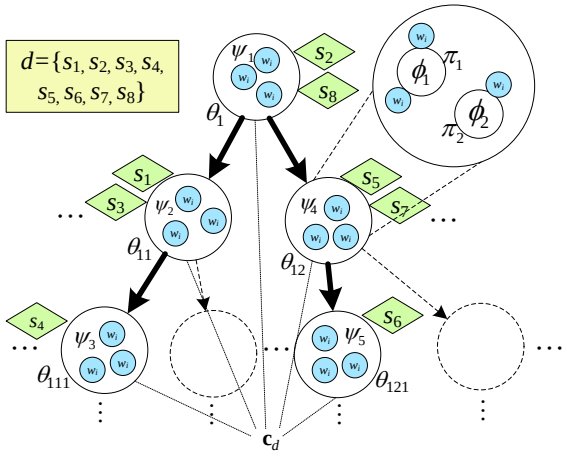

Figure 1. A tree structure for representation of words, sentences and documents. Thick arrows denote the tree paths $\mathbf{c}_d$ drawn for eight sentences of a document d. Colored rectangle, diamonds and circles denote the document, sentences and words, respectively. Each sentence s_j is assigned with a theme variable ψ_l at a tree node along tree paths with probability θ_{dl} while each word w_i in tree node is assigned with a topic variable ϕ_k with probability π_{lk}.

3.1. Model Construction

H2TM is constructed by considering the structure of a document where each document consists of a "bag of sentences" and each sentence consists of a "bag of words". Different from the infinite topic model using HDP (Teh et al., 2006) and the hierarchical

topic model using nCRP (Blei et al., 2004)(Blei et al., 2010b), we propose a new tree model for representation of a "bag of sentences" where each sentence has *variable length of words*. A two-stage procedure is developed for document representation as illustrated in Figure 1. In the first stage, each sentence s_j of a document is drawn from a *mixture of theme model* where the themes are shared for all sentences from a document collection. The theme model of a document d is composed of the themes along its corresponding tree paths $\mathbf{c}_d$. With a tree structure of themes, the unsupervised grouping of sentences into different layers is constructed. In the second stage, each word w_i of the sentences allocated in a tree node is drawn by an individual *mixture of topic model*. All topics from different nodes are drawn using a global topic model.

We assume that the words of the sentences in a tree node given topic k are conditionally independent and drawn from a topic model with infinite topics $\{\phi_k\}_{k=1}^{\infty}$. The sentences in a document given themes l are conditionally independent and drawn from a theme model with infinite themes $\{\psi_l\}_{l=1}^{\infty}$. The document-dependent theme proportions $\{\theta_{dl}\}_{l=1}^{\infty}$ and theme-dependent topic proportions $\{\pi_{lk}\}_{k=1}^{\infty}$ are introduced. Given these proportions, each word w_i is drawn from a mixture model of topics $\sum_k \pi_{lk} \cdot \phi_k$ while each sentence s_j is determined from a mixture model of themes $\sum_l \theta_{dl} \cdot \psi_l$. *Since a theme for sentences is represented by a mixture model of topics for words, we accordingly bridge the relation between themes and topics via* $\psi_l \sim \sum_k \pi_{lk} \cdot \phi_k$.

3.2. nCRP for Thematic Sentences

We construct a sentence-based tree model with infinite nodes and branches. A sentence-based nCRP (snCRP) is proposed to carry out a tree model where root node contains general theme and leaf node conveys a specific theme. Different from word-based nCRP (Blei et al., 2004)(Blei et al., 2010b) where topics along a single tree path are selected to represent all words in a document, the snCRP is developed to represent all sentences in a document based on the themes which may be from *multiple tree paths*. It is because that the variation of themes does exist in heterogeneous documents. The word-based nCRP using GEM distribution is extended to the snCRP using *tree*GEM distribution by considering sub-tree path for document representation. A tree stick-breaking process is proposed to draw a *sub-tree* and determine theme proportions for representation of all sentences in a document.

A new scenario is described as follows. There are infinite number of Chinese restaurants in a city. Each restaurant has infinite tables. A tourist visits the first (root) restaurant where each of its tables has a card showing the next restaurant which is arranged in the second layer of this tree. Such visit repeats infinitely. Each restaurant is associated with a tree layer. The restaurants in a city are organized into an infinitely-branched and infinitely-deep tree structure. H2TM is constructed by

1. For each theme l
   (a) Draw a topic model $\phi_k \sim G_0$.
   (b) Draw topic proportions $\pi_l|\lambda_0 \sim \text{DP}(\alpha_0, \lambda_0)$.
   (c) Theme model is generated by $\psi_l \sim \sum_k \pi_{lk} \cdot \phi_k$
2. For each document $d \in \{1, \cdots, D\}$
   (a) Draw sub-tree path $\mathbf{c}_d = \{c_{dj}\} \sim \text{snCRP}(\gamma_s)$.
   (b) Draw theme proportions over path $\mathbf{c}_d$ by tree stick-breaking $\theta_d|\{\alpha_s, \lambda_s\} \sim tree\text{GEM}(\alpha_s, \lambda_s)$.
   (c) For each sentence s_j
      i. Choose a theme label $z_{sj} = l|\theta_d \sim \text{Mult}(\theta_d)$.
      ii. For each word w_i
         A. Choose a topic label based on topic proportion of a theme $z_{wi} = k|\pi_l \sim \text{Mult}(\pi_l)$.
         B. Draw a word by $w_i|z_{wi} \sim \text{Mult}(\phi_{z_{wi}})$.

The hierarchical grouping of sentences is therefore obtained by a nonparametric tree model based on snCRP. Each tree node stands for a theme. A sentence s_j is determined by a theme model ψ_l.

3.3. Tree Stick-Breaking Process

Traditional GEM distribution is not suitable for characterizing a tree structure with dependencies between parent nodes and child nodes. To deal with this issue, the snCRP is fulfilled and a tree stick-breaking process is conducted to draw theme proportion along with a sub-tree path. A variety of subjects from different sentences is revealed. Each sentence is assigned by a node with theme proportion determined by all nodes in sub-tree path $\mathbf{c}_d$.

In TSB process, we draw theme proportions $\theta_d = \{\theta_{dl}\}$ for a document d subject to $\sum_{l=1}^{\infty} \theta_{dl} = 1$ based on a tree model with infinite nodes. We consider a set of a parent node and its child nodes that are connected as shown by thick arrows in Figure 2(a). Let l_a denote an ancestor node and $l_c = \{l_{a1}, l_{a2}, \cdots\}$ denote its child nodes. TSB is run for each set of nodes $\{l_a, l_c\}$ in a recursive fashion. Figures 2(a) and 2(b) illustrate how the tree structure in Figure 1 is constructed. Figure 2(b) shows how theme proportions are inferred by TSB. The theme proportion $\theta_{l_a 0}$ in child nodes denotes the initial segment of node l_a when proceeding stick-breaking process for its child nodes l_c. Here,

$\theta_0 = 1$ denotes the initial unit length, $\theta_1 = \nu_1$ denotes the first segment of stick for root node and $1 - \nu_1$ denotes the remaining segment of the stick. Given the *tree*GEM parameters $\{\alpha_s, \lambda_s\}$, the beta variable $\nu_u \sim \text{Beta}\,(\alpha_s\lambda_s, \alpha_s(1-\lambda_s))$ of a child node $l_u \in \Omega_{l_c}$ is first drawn. The probability of generating this draw is calculated by $\nu_u \prod_{v=0}^{u-1}(1-\nu_v)$. This probability is then multiplied by theme proportion θ_{l_a} of ancestor node l_a so as to find theme proportion for its child nodes. We recursively calculate theme proportion by

$$\theta_{l_u} = \theta_{l_a}\nu_u \prod_{v=0}^{u-1}(1-\nu_v), \quad \text{for } l_u \in \Omega_{l_c}. \tag{1}$$

Therefore, a tree model is constructed without limitation of tree layers and branches. We improve the efficiency of tree stick-breaking method (Adams et al., 2010) by adopting a single set of beta parameters $\{\alpha_s, \lambda_s\}$ for stick-breaking towards depth as well as branch. Using this process, we draw global theme proportions for sentences in different documents by using scaling parameter γ_s and determine sub-tree path for all sentences s_j in document d via snCRP by $\mathbf{c}_d = \{c_{dj}\} \sim \text{snCRP}(\gamma_s)$. The TSB process is performed to obtain theme proportions $\theta_d \sim \textit{tree}\text{GEM}(\alpha_s, \lambda_s)$.

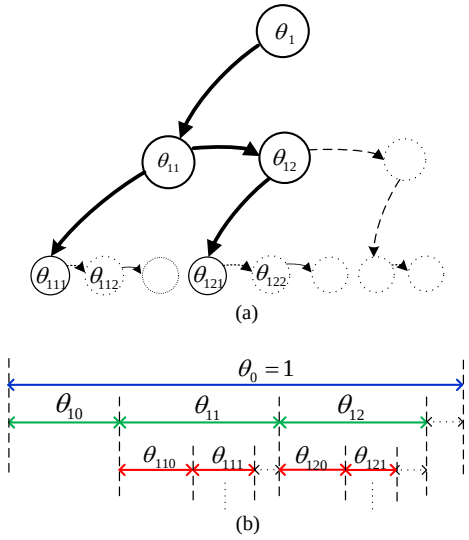

Figure 2. Illustrations for (a) tree stick-breaking process, and (b) hierarchical theme proportions.

3.4. HDP for Topical Words

After having hierarchical grouping of sentences, we treat the words corresponding to a node of theme l as grouped data and conduct HDP using the grouped data from different tree nodes. The delicate topic model is constructed to draw individual words. Importantly, each theme is represented by a mixture model of topics $\psi_l \sim \sum_k \pi_{lk} \cdot \phi_k$. HDP is applied to infer word distributions and topic proportions. The standard stick-breaking process is applied to infer topic proportions for DP mixture model based on GEM distribution. The words of a tree node corresponding to theme l is generated by

$$\begin{gathered}\lambda_0 \sim \text{GEM}(\gamma_w), \quad \pi_l \sim G(\alpha_0, \lambda_0), \quad z_{wi}|\pi_l \sim \pi_l \\ \phi_k \sim G_0, \quad w_i|z_{wi}, \{\phi_k\}_{k=1}^{\infty} \sim \text{Mult}(\phi_{z_{wi}})\end{gathered} \tag{2}$$

where λ_0 is a global prior for tree nodes, π_l is the topic proportion for theme l, ϕ_k is the kth topic, α_0 and γ_w are the strength parameters for DP. The *snCRP compound HDP* is fulfilled accordingly.

4. Bayesian Inference

The approximate Bayesian inference using Gibbs sampling is developed to infer posterior parameters or latent variables for H2TM. Each latent variable is iteratively sampled by a posterior probability with condition on the observations and all the other latent variables. We sample sub-tree path $\mathbf{c}_d = \{c_{dj}\}$ for different sentences of a document d. Each sentence s_j is grouped into a tree node with theme l which is sampled through snCRP. Each word w_i of a sentence is then assigned by the topic k according to the HDP.

4.1. Sampling Sub-Tree Path

A document is treated as "a bag of sentences" for sub-tree sampling. We iteratively sample tree paths for words $\mathbf{w}_d$ in document d consisting of sentences $\{\mathbf{w}_{dj}\}$. Sampling is performed using the posterior probability

$$\begin{aligned}p(c_{dj}|\mathbf{c}_{d(-j)}, \mathbf{w}_d, z_{sj}, \psi_l, \gamma_s) &\propto p(c_{dj}|\mathbf{c}_{d(-j)}, \gamma_s) \\ &\times p(\mathbf{w}_{dj}|\mathbf{w}_{d(-j)}, z_{sj}, \mathbf{c}_d, \psi_l)\end{aligned} \tag{3}$$

where $\mathbf{c}_{d(-j)}$ denotes the paths of all sentences in document d except sentence s_j. The notation "-" denotes self-exception. In (3), γ_s is Dirichlet prior parameter for global theme proportions. The first term in right-hand-side (RHS) calculates the probability of choosing a path for a sentence. This probability is determined by applying CRP (Blei et al., 2010b). Here, the jth sentence chooses either an occupied path h by $p(c_{dj} = h|\mathbf{c}_{d(-j)}, \gamma_s) = \frac{f_{d(c_{dj}=h)}}{f_{d\cdot}-1+\gamma_s}$ or a new path by $p(c_{dj} = \text{new}|\mathbf{c}_{d(-j)}, \gamma_s) = \frac{\gamma_s}{f_{d\cdot}-1+\gamma_s}$ where $f_{d(c_{dj}=h)}$ denotes the number of sentences in document d that are allocated along tree path h. Path h is selected for sentence $\mathbf{w}_{dj}$. The second term in RHS of (3) is calculated by referring (Blei et al., 2004)(Blei et al., 2010b).

4.2. Sampling Themes

Given the current path c_{dj} selected via snCRP by using words $\mathbf{w}_{dj}$, we sample a tree node at level ℓ or equivalently sample a theme l according to the posterior probability given current values of all other variables

$$\begin{aligned} p(z_{sj} = l|\mathbf{w}_d, \mathbf{z}_{s(-j)}, c_{dj}, \alpha_s, \lambda_s, \psi_l) \propto \\ p(z_{sj} = l|\mathbf{z}_{s(-j)}, c_{dj}, \alpha_s, \lambda_s) p(\mathbf{w}_{dj}|\mathbf{w}_{d(-j)}, \mathbf{z}_s, \psi_l) \end{aligned} \quad (4)$$

where $\mathbf{z}_s = \{z_{sj}, \mathbf{z}_{s(-j)}\}$. The number of themes is unlimited. The first term in RHS of (4) is a distribution over levels derived as an expectation of *tree*GEM which is implemented via TSB process and is calculated via a product of beta variables $\nu_u \sim \text{Beta}\,(\alpha_s\lambda_s, \alpha_s(1-\lambda_s))$ along path c_{dj}. The second term calculates the probability of sentence $\mathbf{w}_{dj}$ given the theme measure ψ_l.

4.3. Sampling Topics

According to HDP, we apply stick-breaking construction to draw topics for words in different tree nodes. We view words $\{w_{dji}\}$ of the sentences in a node with theme l as the grouped data. Topic proportions are drawn from $\text{DP}(\alpha_0, \lambda_0)$. Drawing of a topic k for word w_{dji} or w_i depends on the posterior probability

$$\begin{aligned} p(z_{wi} = k|\mathbf{w}_{dj}, \mathbf{z}_{w(-i)}, c_{dj}, \alpha_0, \lambda_0, \phi_k) \\ \propto p(z_{wi} = k|\mathbf{z}_{w(-i)}, c_{dj}, \alpha_0, \lambda_0) \\ \times p(w_{dji}|\mathbf{w}_{dj(-i)}, \mathbf{z}_w, \phi_k)\,. \end{aligned} \quad (5)$$

Calculating (5) is equivalent to estimating the topic proportion π_{lk}. The first term in RHS of (5) is a distribution derived as an expectation of GEM and is calculated via a product of beta variables using $\text{Beta}\,(\alpha_0\lambda_0, \alpha_0(1-\lambda_0))$. Given the current status of the sampler, we iteratively sample each variable conditioned on the rest variables. For each document d, the paths c_{dj}, themes l and topics k are sequentially sampled and iteratively employed to update the corresponding posterior probabilities in Gibbs sampling procedure. The true posteriors are approximated by running sufficient sampling iterations. The resulting H2TM is implemented.

5. Experiments

5.1. Experimental Setup

The experiments were conducted to evaluate the proposed H2TM for document representation. We collected the DUC (Document Understanding Conference) 2007 (http://duc.nist.gov/). In DUC 2007, there were 45 super-documents where each document contained 25-50 news articles. The number of total sentences in this dataset was 22961. The vocabulary size was 18696 after removing stop words. For simplicity, we constrained tree growing to three layers in our experiments. The initial values of three-layer H2TM were specified by $\psi_l = \phi_k = [0.05\ 0.025\ 0.0125]^T$, $\lambda_0 = \lambda_s = 0.35$, $\alpha_s = 100$ and $\gamma_s = 0.5$.

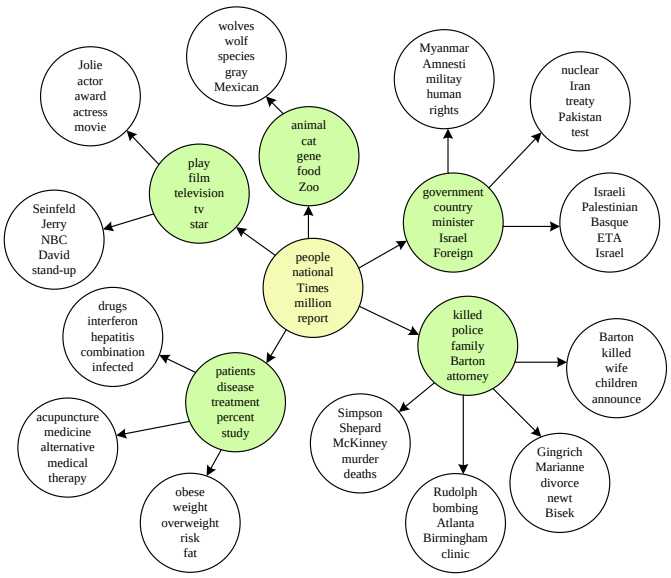

Figure 3. The tree structure of DUC 2007 showing the topical words in each theme or tree node.

5.2. Experimental Results

H2TM is presented to perform unsupervised structure learning for themes using sentences and topics using the words in different tree nodes. Each theme is a mixture model of topics. Figure 3 displays an example of three-layer tree structure which is estimated based on snCRP or *tree*GEM by using DUC sentences. For the words of all sentences allocated in tree nodes, we conduct HDP to find topic proportions corresponding to each node based on the GEM distribution. Here, five topical words are shown for tree nodes in different layers which are shaded with varying colors. The root node (yellow) contains general words while leaf nodes (white) consist of specific words. It is obvious to see semantic relationships between tree nodes in different layers along the selected five tree paths. These paths are separately related to *animal*, *television*, *disease*, *criminal* and *country*. The performance of unsupervised structural learning is obvious.

For the same tree model, Figure 4 shows the topical words as well as the *thematic sentences* along thick paths. The hierarchical clustering of sentences is illustrated. We do see that topical words and thematic sentences allocated in the same node are semantically

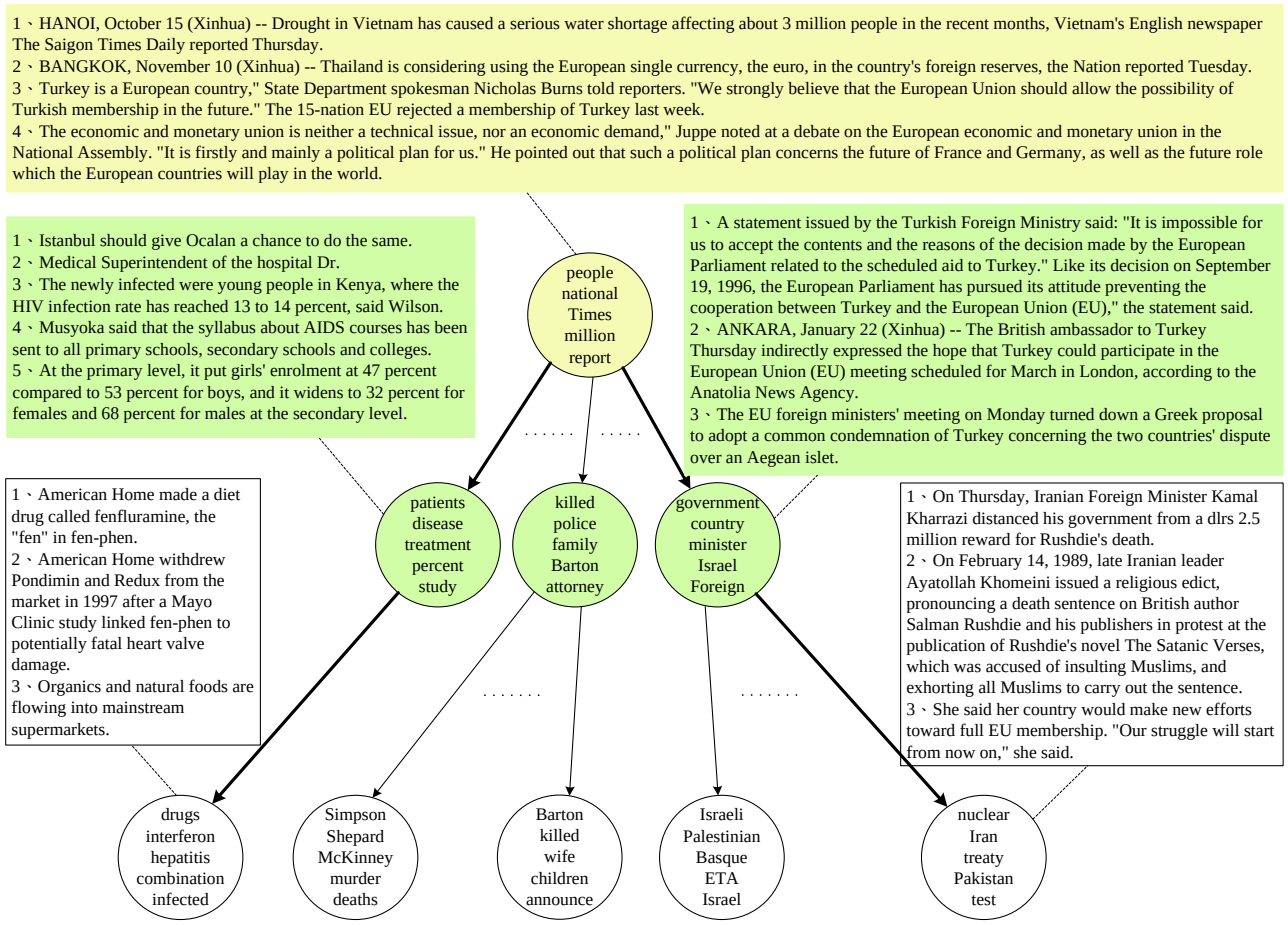

Figure 4. Hierarchical tree model of DUC 2007 showing topical words and thematic sentences.

similar in different levels. The left-hand-side (LHS) and RHS tree paths reflect different themes extracted from multiple documents. The sentences in LHS path are related to the theme on disease infection and clinic study. The sentences in lower layer are more focused on specific theme. Also, the sentences in RHS path indicate the theme on European Union and Middle East. The sentences in the lower layer are more related to Iranian Affairs. The contributions of using H2TM come from the flexible model complexity and the structural theme information which are beneficial for sentence clustering and document representation.

6. Conclusions

This paper addressed a deep learning method for document representation. A hierarchical theme model was constructed according to a sentence-level nCRP while the topic model was established through a word-level HDP. The snCRP compound HDP was proposed to build H2TM where each theme was characterized by a mixture model of topics. An organized document representation using themes in sentence level and topics in word level was proposed. We also presented a TSB process to draw sub-tree path for a heterogeneous document and built a hierarchical mixture model of themes according to the snCRP. The hierarchical clustering of sentences was implemented. The sentences were allocated in tree nodes and the corresponding words in different nodes were drawn by HDP. The proposed H2TM is a general model for unsupervised learning of different structural data. Experimental results on document representation showed that H2TM captured the latent structure in multiple documents.

References

Adams, R. P., Ghahramani, Z., and Jordan, M. I. Tree-structured stick breaking for hierarchical data. In *Advances in Neural Information Processing Systems*, 2010.

Blei, D., Carin, L., and Dunson, D. Probabilistic topic

models. *IEEE Signal Processing Magazine*, 27(6): 55–65, 2010a.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(5):993–1022, 2003.

Blei, D. M., Griffiths, T. L., Jordan, M. I., and Tenebaum, J. B. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems*, 2004.

Blei, D. M., Griffiths, T. L., and Jordan, M. I. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):7, 2010b.

Chang, Y.-L. and Chien, J.-T. Latent Dirichlet learning for document summarization. In *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, pp. 1689–1692, 2009.

Chang, Y.-L., Hung, J.-J., and Chien, J.-T. Bayesian nonparametric modeling of hierarchical topics and sentences. In *Proc. of IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2011.

Chien, J.-T. and Chueh, C.-H. Dirichlet class language models for speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 19(3): 482–495, 2011.

Chien, J.-T. and Chueh, C.-H. Topic-based hierarchical segmentation. *IEEE Transactions on Audio, Speech and Language Processing*, 20(1):55–66, 2012.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Paisley, J., Carin, L., and Blei, D. Variational inference for stick-breaking beta process priors. In *Proc. of International Conference on Machine Learning*, 2011.

Pitman, J. Poisson-Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability and Computing*, 11:501–514, 2002.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical Dirichlet process. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

Teh, Y. W., Newman, D., and Welling, M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, 2007.

Wang, C. and Blei, D. M. Variational inference for the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems*, 2009.