Topic Modeling for Conference Analytics

Pengfei Liu[1]*, Shoaib Jameel*[1]*, Wai Lam*[1]*, Bin Ma*[2]*, Helen Meng*[1]

[1]Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Human Language Technology Department,
Institute for Infocomm Research, Singapore
{pfliu, msjameel, wlam, hmmeng}@se.cuhk.edu.hk, mabin@i2r.a-star.edu.sg

Abstract

This work presents our attempt to understand the research topics that characterize the papers submitted to a conference, by using topic modeling and data visualization techniques. We infer the latent topics from the abstracts of all the papers submitted to Interspeech2014 by means of Latent Dirichlet Allocation. Per-topic word distributions thus obtained are visualized through word clouds. We also compare the automatically inferred topics against the expert-defined topics (also known as *tracks* for Interspeech2014). The comparison is based on an information retrieval framework, where we use each latent topic as a query and each track as a document. For each latent topic, we retrieve a ranked list of tracks scored by the degree of word overlap. Each latent topic is associated with the top-scoring track. This analytic procedure was applied to all submissions to Interspeech2014 and sheds some interesting light in terms of providing an overview of topic categorization in the conference, popular versus unpopular topics, emerging topics and topic compositions. Such insights are potentially valuable for understanding the technical content of a field and planning the future development of its conference(s).

Index Terms: topic modeling, conference analytics, information retrieval

1. Introduction

Academic conferences such as INTERSPEECH usually call for papers with a list of expert-designed research tracks and subtracks. These tracks are useful for the organization of a technical conference where authors can submit their papers to their preferred tracks, reviewers can choose papers to review from their preferred tracks, and readers can search for papers by filtering tracks. Conference organizers may be interested in questions such as, *Are the list of tracks is representative and diverse enough to cover all major research topics of the field? Are the tracks described well? Do the tracks match the research topics in the submitted papers as they vary from year to year? Are authors submitting their papers to the relevant tracks? If not, then are the track names ambiguous? Alternatively, are the track descriptions confusing?*

To answer these questions properly, we present a new task named *topic-track matching*, which matches latent topics inferred from the content (e.g., abstract) of submitted papers with a list of expert-designed conference tracks. The task can be helpful to the conference organizers in the following ways: (1) Help the organizers monitor whether authors are submitting their papers to the appropriate track; (2) Help the organizers monitor whether a track is ambiguously described; (3) Analyze the popularity of the tracks and merge less popular tracks with other related tracks for future conferences; (4) Help the organizers observe the popularity of topics in their field, which may engender new tracks for emerging and popular topics.

Probabilistic topic models [1, 2, 3, 4] such as Latent Dirichlet Allocation (LDA) [1] are statistical models that find patterns of words or underlying latent topics from a large collection of documents, which have been widely used in the past to study academic conferences [5, 6, 7, 4, 8]. For example, Blei et al. [6] presented a hierarchical latent Dirichlet allocation (hLDA) model based on a nested Chinese restaurant process, and demonstrated a 3-level topic hierarchy estimated from 1,717 paper abstracts from NIPS; Hall et al. [7] studied about how ideas in linguistics conferences had changed over time, using the LDA model to detect temporal changes in the ideas from the research papers of many linguistic conferences such as ACL, COLING, etc.

In this paper, we propose to apply topic modeling techniques for the *topic-track matching* task. We investigate the task under an *information retrieval* framework [9], with each latent topic as a query and each track as a document for retrieval. We apply the LDA model on abstracts of all the submitted papers from the conference organizers of Interspeech2014 to infer K latent topics and match each topic with the most similar track, which has the highest F-score calculated by counting the overlapping words between *top-ranking words* based on decreasing probability of a topic and *key words* of a track. The outputs of the task are matches between topics and tracks, word cloud visualization of topics and distribution analysis of matching results to give suggestions for future conference organization. Our experimental results show that the topic-track matching task enables the detection of mismatches between inferred latent topics and expert-designed tracks. This can help suggest which track descriptions may be revised. Our results can also facilitate the analysis of the popularity of tracks for better future conference organization.

2. Approach

2.1. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic topic model for collections of discrete data such as text corpora. Following the notations in [1], given the hyper-parameters α and β, LDA defines the probability of a corpus D with M documents, as illustrated in formula (1), where θ_d is a topic mixture for document d, w_{dn} is the nth word from document d, and z_{dn} is the latent topic assignment for the word w_{dn} given θ_d.

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn},\beta) \right) d\theta_d \quad (1)$$

In the generative process of LDA, each document is generated by first sampling a document-specific topic proportion θ_d from a Dirichlet distribution, and then drawing each word from a topic-specific Multinomial distribution $p(w_{dn}|z_{dn},\beta)$. The model generates a low-dimensional representation of data, consisting of a word distribution of $P(w|z)$, which states the probability of a word w belonging to a topic z and a topic distribution in a document $P(z|d)$, which specifies the mixture of topics in a document d. Our interest is on $P(w|z)$ as we will match the words in topics with the words in expert-designed tracks of an academic conference. The LDA model can be estimated by several algorithms such as the variational Bayes algorithm by Blei et al. [1], the expectation propagation algorithm by Minka et al. [10] and the collapsed Gibbs sampling algorithm by Griffiths and Steyvers in [5] and so on.

2.2. Topic-Track Matching

Conceptually, we assume a 1:N relationship between track and topic, and we define a match as a pair between a topic and its *top-one* similar track based on F-score. We tackle the *topic-track matching* problem under an information retrieval framework, with each latent topic as a *query* and each track as a *document*. The latent topics are obtained by applying LDA on the paper abstracts and we represent each topic by choosing its top $N = 200$ words based on the decreasing probability of each word. This parameter is set empirically to cover approximately 80% of the probability space of the words in each topic. Each *document* typically consists of 20-50 words pre-processed from the corresponding track description.[1]

Our topic-track matching system is illustrated in Figure 1. We first applied the same pre-processing step to both *Conference Tracks* and *Paper Abstracts*. Then, we applied LDA to get the latent topics with a list of *top words* in descending order of probability $P(w|z)$, which are queries for retrieving the tracks represented with a set of *key words*.

For each *query* (topic), we match it with the *document* (track) which has the highest F-score obtained by calculating their overlapping words. As the F-score measure used in *Text::Similarity*[2] for pair-wise similarity of files or strings, we calculated F-score by first counting the number of matching words between the key words of a track (W_k) and the top words (W_t) of a topic, and then computing *Precision*, *Recall* and *F-score* accordingly, as shown in formulas (2), (3) and (4).

$$\text{Precision} = \frac{|W_k \cap W_t|}{|W_k|} \quad (2)$$

$$\text{Recall} = \frac{|W_k \cap W_t|}{|W_t|} \quad (3)$$

$$\text{F-score} = \frac{|W_k \cap W_t|}{\frac{|W_k|+|W_t|}{2}} \quad (4)$$

2.3. System Implementation

We developed our matching system in Java which includes dataset pre-processing, F-score calculation and topic-track matching, as well as a LDA component from the MALLET [11] toolkit which implements the collapsed Gibbs sampling algorithm for model inference. Specifically, we used the Java class *ParallelTopicModel* in MALLET, a parallel threaded implementation of LDA, whose detailed algorithm is described in [12, 13]. We share the source code of our matching system at `https://github.com/ppfliu/conference-topic`.

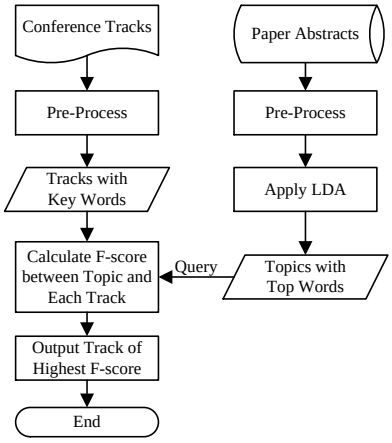


Figure 1: The Topic-Track Matching System.

3. Experiments

3.1. Corpus

We conducted experiments on the paper abstracts of 12 main tracks set by the conference organizers of Interspeech2014[3]. In the dataset, the total number of paper abstracts is 1,078, and the total number of words is 101,312. The average number of words in a paper abstract is about 94. After stopword removal and stemming, the vocabulary contains 5,732 unique words.

3.2. Experimental Settings

Preprocessing: For the step of pre-processing, we kept only content words and stemmed each word by morphology (i.e., computing the base form of English words by removing inflections such as noun plurals, pronoun case and verb endings). We lowercased all words and removed stop words.

Hyperparameters of α and β: We then applied the LDA model on the pre-processed dataset, with the hyper-parameters optimized using the hyper-parameter sampling algorithm implemented in MALLET.

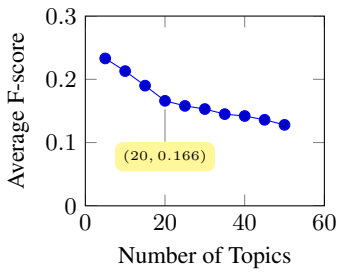
Number of Topics: We empirically set the number of latent topics as 20, which is higher than the number of tracks (12) because there may exist some topics not described in tracks by conference organizers but included in a number of papers. This number is close to 22 given by the the Hierarchical Direchlet Process (HDP) model [14, 15], which takes a nonparametric Bayesian approach to find the number of latent topics automatically [14, 3, 4]. We also verified the number of topics by showing the average matching F-score under different number of topics for Interspeech2014, in Figure 2a, which shows that the number of topics should not be too high to lead to a low average F-score.

Number of Iterations: We tune the number of iterations by comparing the *training perplexity* [1] of the LDA model on the

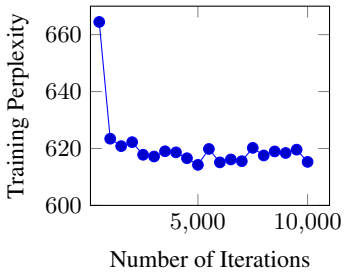
[1] We also tried to apply LDA on the tracks directly to infer their top covered topics, which are however not distinguishable among tracks.

[2] https://metacpan.org/pod/Text::Similarity

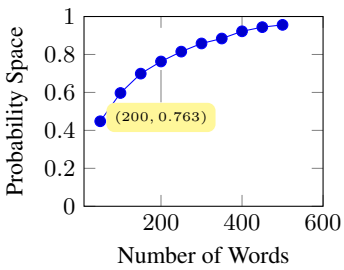
[3] http://www.interspeech2014.org/public.php?page=conference_areas.html



(a) Average matching F-score by different number of topics.



(b) Training perplexity with 20 topics under different number of iterations.



(c) Probability space covered by different number of top words in 20 topics.

Figure 2: Experimental Settings on *Number of Topics*, *Iterations* and *Top Words* in Each Topic for Interspeech2014.

Speech Synthesis	*Language Model*	*Speaker Identification*	*Spoken Language*	*Speech Prosody*	*Signal Processing*	*Neural Network*	*Method Algorithm*
speech	language	speaker	system	tone	vocal	network	method
synthesis	model	system	keyword	pitch	tract	dnn	algorithm
base	datum	vector	search	prosodic	source	neural	sparse
voice	word	recognition	term	speech	voice	train	dictionary
quality	lm	verification	speak	word	frequency	deep	function
system	gram	performance	base	duration	method	feature	propose
hmm	resource	show	detection	syllable	signal	recognition	matrix
synthetic	cross	identification	word	mandarin	excitation	layer	base
conversion	multilingual	variability	language	stress	formant	task	nmf
unit	domain	base	query	lexical	pitch	system	source

Table 1: Top ten words from some selected topics obtained using the LDA model on our dataset. The words appear in decreasing order of probabilities. The first entry in each column is a title given by a domain expert in speech area.

whole dataset under different number of iterations, as shown in Figure 2b. We can see that 10,000 iterations is enough to lead the perplexity to be stable. Therefore, we conducted all the experiments with the number of iterations as 10,000, which is also feasible for a relatively small dataset like Interspeech2014.

Number of Top Words: We chose top 200 words in the order of descending probability to represent each topic. This number is chosen empirically, which covers approximately 80% (0.763) of the probability space of each topic. Setting the number of topics as 20 and the number of iterations as 10000, we plot the relationship between the average probability space over the topics and the number of top words.

3.3. Topical Words

In Table 1, we analyzed the LDA results by presenting the top 10 words sorted by decreasing probability for 8 topics out of 20 topics inferred from the dataset. We chose the top 10 words because these words can provide sufficient detail to convey the subject of a topic, and distinguish one topic from another [16]. We also named each topic with a title manually in the table header which helps us gain an overview of the research topics in Interspeech2014, such as *Speech Recognition*, *Speech Synthesis*, *Language Modeling*, etc.

3.4. Visualization of Topic-Track Matching

We present two illustrations of topic-track matching (Again, topics are automatically derived and tracks are expert defined.). The first example is between the topic of *Neural Network* (See Table 1) and *Track 7* (See Table 2), to which there are 177 submitted papers. Figure 3 shows a word cloud of the topic *Neural Network* by the tool Wordle[4] and the matching words with Track 7, where words with larger font have higher probability

[4]http://www.wordle.net/

Figure 3: Word Cloud of the Topic *Neural Network*: it matches with Track 7 (See Table 2) with the F-score of 0.138 and the matching words are: *acoustic asr conversational cross deep discriminative extraction feature level model network neural process recognition robustness speech train*.

Figure 4: Word Cloud of the Topic *Speech Synthesis*: the matching F-score with Track 6 (See Table 2) is 0.130 and the matching words are: *analysis conversion evaluation generation method model modification parametric process prosody quality speech statistical synthesis text voice*.

in the topic while the layout and color are randomly set for visualization. We observe in Figure 3 that *deep neural network* has become a popular topic for Track 7 in *speech recognition*.

The other matching example is between the topic *Speech Synthesis* (See Table 1) and *Track 6* (See Table 2), which has 105 submitted papers. Figure 4 shows a word cloud of the topic and its matching words with Track 6, which presents the areas of *speech synthesis* and *voice conversion* and the *HMM* statistical model.

6: Speech Synthesis and Spoken Language Generation	7: Speech Recognition - Signal Processing, Acoustic Modeling, Robustness, and Adaptation
6.1 Grapheme-to-phoneme conversion for synthesis	7.1 Feature extraction and low-level feature modeling for ASR
6.2 Text processing for speech synthesis (text normalization, syntactic and semantic analysis)	7.2 Prosodic features and models
6.3 Segmental-level and/or concatenative synthesis	7.3 Robustness against noise, reverberation
6.4 Signal processing/statistical model for synthesis	7.4 Far field and microphone array speech recognition
6.5 Speech synthesis paradigms and methods, silence speech, articulatory synthesis,,parametric synthesis etc.	7.5 Speaker normalization (e.g., VTLN)
6.6 Prosody modeling and generation	7.6 Deep neural network
...	...

Table 2: Descriptions of Track 6 and Track 7 from Interspeech2014.

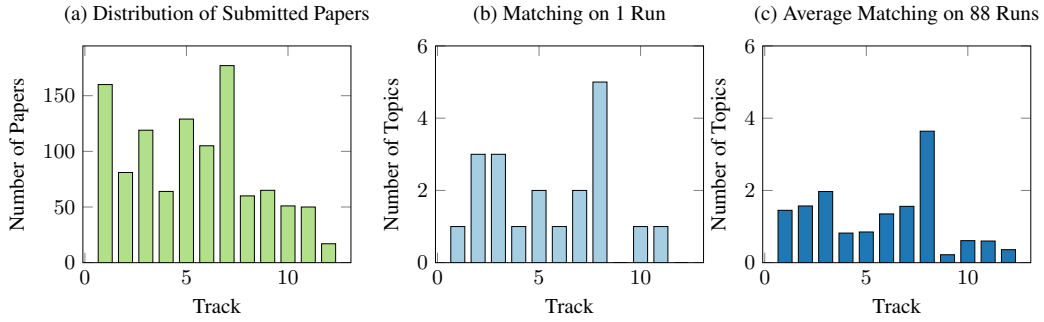


Figure 5: Distribution of Submitted Papers, and Matches between Latent Topics and Tracks in Interspeech2014.

3.5. Analysis of Matches between Latent Topics and Tracks

Figure 5(a) shows the paper submissions across the 12 Interspeech2014 tracks, as selected by the authors. Figure 5(b) shows the number of latent topics matched to each track according to highest F-scores. This example is based on one LDA run (i.e., applying LDA on the dataset once for 10, 000 iterations) with 20 topics and 200 top words.

It is known that different runs of the LDA model may give slightly different results. Hence, we conducted 88 runs shown in Figure 5(c), where we varied the number of latent topics from 10 to 20 and the number of top words for each topic as (50, 100, 150, 200). We conducted the whole experiment twice, and had a total of 88 ($88 = 11 \times 4 \times 2$) runs of LDA on the dataset. Our observations from Figure 5 include:

(1) Track 1 (*Speech Perception and Production*) has 160 submitted papers. However, it has fewer matching topics than expected, based on the proportion of submitted papers. This may indicate that some authors had submitted their papers to other tracks, which suggests that *Track 1* may need revision. A similar situation also exists for Track 5 (*Speaker and Language Identification*) with 129 submitted papers.

(2) Track 3 (*Analysis of Speech and Audio Signals*) shows a high number of submitted papers in (a) and similar high matches in both (b) and (c). This may indicate that Track 3 is important and matches well with the submitted papers.

(3) Subfigures of (b) and (c) show similar matching distributions with slight differences due to some randomness involved in the collapsed Gibbs sampling algorithm of LDA.

(4) The matches in both (b) and (c) for Track 7 and Track 8 (*Speech Recognition - Architecture, Search & Linguistic Components*) show reversed distributions compared with the paper submissions in (a). Quite a number of authors submitted their papers to Track 7 while our LDA-based matching system assigned the papers to Track 8 through analyzing their abstracts. Track 7 and 8 may be revised for better categorization.

(5) Some tracks are weakly matched with latent topics. Track 9 (*LVCSR and Its Applications, Technologies and Systems for New Applications*) has 65 submitted papers (see Figure5(a)), but no match in 5(b) and only 19 matches (1.4%) in 5(c). Track 10 (*Spoken Language Processing - Dialogue, Summarization, Understanding*), Track 11 (*Spoken Language Processing - Translation, Info Retrieval*) and Track 12 (*Spoken Language Evaluation, Standardization and Resources*) also have few submissions and low topic matches. These tracks may be revised for future conference organization.

4. Conclusions and Future Work

In this paper, we propose a new task of *topic-track matching* which applies LDA on paper abstracts and matches inferred latent topics with expert-designed conference tracks. We investigated the task under an information retrieval framework, with each topic as a *query* and each track as a *document*. For each topic, we retrieved a ranked list of tracks by calculating their word overlapping score and chose the *top-scoring* track as the matching *document*. We analyzed the matches to obtain trends which may be helpful for future conference organization. Experiments on Interspeech2014 show that the new task and our LDA-based method can facilitate future conference organization by visualizing popular topics, e.g., *deep neural networks*, and identifying the popular and less popular tracks based on topic matches.

One interesting future direction is to learn hierarchical topic structures from conference papers, which not only help researchers know current research topics but also facilitate future conference organizers to derive better tracks and sub-tracks.

5. Acknowledgements

This work is affiliated with the Stanley Ho Big Data Decision Analytics Research Center of The Chinese University of Hong Kong.

6. References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[2] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook of latent semantic analysis*, vol. 427, no. 7, pp. 424–440, 2007.

[3] D. M. Blei and J. D. Lafferty, "Topic models," *Text mining: classification, clustering, and applications*, vol. 10, no. 71, p. 34, 2009.

[4] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[5] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[6] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical topic models and the nested chinese restaurant process," *NIPS*, vol. 16, p. 17, 2004.

[7] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the history of ideas using topic models," in *EMNLP*, 2008, pp. 363–371.

[8] A. Ahmed and E. P. Xing, "Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream," *arXiv preprint arXiv:1203.3463*, 2012.

[9] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge University Press, 2008, vol. 1.

[10] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," in *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 2002, pp. 352–359.

[11] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit," 2002, http://mallet.cs.umass.edu.

[12] D. Newman, A. Asuncion, P. Smyth, and M. Welling, "Distributed algorithms for topic models," *JMLR*, vol. 10, 2009.

[13] L. Yao, D. Mimno, and A. McCallum, "Efficient methods for topic model inference on streaming document collections," in *ACM SIGKDD*, 2009, pp. 937–946.

[14] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the american statistical association*, vol. 101, no. 476, 2006.

[15] G. Heinrich, "Infinite LDA implementing the HDP with minimum code complexity," *Technical note, Feb*, vol. 170, 2011.

[16] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *NAACL*, 2010, pp. 100–108.