Topic-time Heatmaps for Human-in-the-loop Topic Detection and Tracking

Doug Beeferman
Center for Constructive Communication, MIT
Cambridge, MA, USA
dougb5@mit.edu

Hang Jiang
Center for Constructive Communication, MIT
Cambridge, MA, USA
hijan42@mit.edu

ABSTRACT

The essential task of Topic Detection and Tracking (TDT) is to organize a collection of news media into clusters of stories that pertain to the same real-world event. To apply TDT models to practical applications such as search engines and discovery tools, human guidance is needed to pin down the scope of an "event" for the corpus of interest. In this work in progress, we explore a human-in-the-loop method that helps users iteratively fine-tune TDT algorithms so that both the algorithms and the users themselves better understand the nature of the events. We generate a visual overview of the entire corpus, allowing the user to select regions of interest from the overview, and then ask a series of questions to affirm (or reject) that the selected documents belong to the same event. The answers to these questions supplement the training data for the event similarity model that underlies the system.

ACM Reference Format:
Doug Beeferman and Hang Jiang. 2021. Topic-time Heatmaps for Human-in-the-loop Topic Detection and Tracking. In *DaSH@KDD '21: 3rd Workshop on Data Science with Humans in the Loop: KDD, August 15-16, 2021, Virtual Conference.* ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

The Topic Detection and Tracking (TDT) task [1, 2, 4] challenges NLP researchers to organize a collection of news media into clusters of stories that pertain to the same real-world event, and to organize those events into topics. *On-line* methods [3, 18, 26], useful in alert systems, require that novel events are recognized and clusters are built in real-time as stories are processed, while *retrospective* methods [2, 27], useful for search and discovery tools, use all of the accumulated data to find the best clustering.

Human feedback is vital to both kinds of systems in order to align the notion of an "event" with the concrete goals of a deployed application. "Event" is defined broadly in the TDT guidelines – "something (non-trivial) happening at a certain place at a certain time" [27]. Recent advances in human-in-the-loop (HITL) [25] have demonstrated the effectiveness of human intervention for NLP model training [12] and deployment [6]. In terms of human-machine interaction, intuitive visualization tools based on topic models [9, 13] are developed to collect feedback from NLP non-experts. Binary or scaled user feedback [10, 15, 20] is easy to collect but sometimes oversimplify users' intentions, whereas language feedback [6, 14, 24] is more informative but also challenging for machines to interpret. The feedback can be used for incremental learning [9, 12, 21] or direct manipulation of the model [11, 14].

In this work we explore a human-in-the-loop (HITL) method that can be used to fine-tune TDT algorithms so that they capture the nature of the events of interest. In turn, the process teaches the human users about the corpus. Our high-level approach is to give the user a visual overview of the entire corpus, allow them to select regions of interest from the overview, and then ask a series of questions to affirm (or reject) that the selected documents are part of the same cluster. The answers to these questions inform how the model, described below, is retrained. Our goal is for this collaboration to enhance the efficacy of the system as both an event detector and as a sense-making tool.

2 TDT SYSTEMS

Recent TDT approaches have explored both sparse and dense features. Miranda et al. [16] proposes an online clustering method that represents documents with TF-IDF features, and demonstrate high performance on a benchmark news article data set. Building on this work, Staykovski et al. [22] adopts a BCubed metric for evaluation and compares sparse TF-IDF features with dense Doc2Vec representations, showing a sizeable improvement on the standard data set. Saravanakumar et al. [17] is the first to include BERT contextual representations for the task and achieves further improvement. Specifically, they fine-tune an entity-aware BERT model on an event similarity task with a triplet loss function. They generate triplets for each document using the batch-hard regime [7]. In each document in a mini-batch, they mark documents with the same label as positive examples and different labels as negative examples. The hardest positive (biggest positive-anchor document distance) and negative (smallest anchor-negative document distance) examples are picked per anchor document to form a triplet. The entity-aware BERT model is trained to make the embedding distance between anchor and positive documents closer than anchor and negative documents. Overall, this fine-tuning process effectively improves the contextual embedding for the overall TDT system.

Our own TDT framework (in progress) similarly mixes sparse features with dense features fine-tuned for event similarity. Instead of using hand-crafted sparse time features [16, 17, 22], we represent the document creation time with a Date2Vec embedding [8] and infuse it with an entity-aware BERT embedding with the self-attention mechanism [23] to produce a time-sensitive dense document representation. This captures interactions between topic

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
DaSH@KDD '21, 15-16 August 2021, Virtual Conference
© 2021 Association for Computing Machinery.

and time. We experiment with both online and offline triplet mining algorithms [19] to optimally train our event similarity model.

We also add a human-in-the-loop component whereby annotators can steer the system via an interactive visualization tool. Interactions with the tool are used to form triplets, which are continuously used to tune the model's representations. Here we aim to improve upon the process of adapting our system to new domains. In this paper, we will focus on this HITL component.

3 TOPIC-TIME HEATMAPS

Motivation & Procedure. When we first set out to find the events in a new document collection, we may have no labeled examples of the kinds of events we care about. Getting such golden data is expensive; clustering a large set of documents into events is time-consuming and hard to parallelize between multiple annotators [1, 2, 4], who must coordinate on how they label events. Instead, inspired by *story-story links* [5], we bootstrap our event similarity model on triplets of documents from the new data. Such triplets can be constructed by judiciously picking *pairs* of documents from the new collection, each pair annotated with whether it is from the same event or from different events. These pairwise judgments can be solicited with the guidance of annotators. The work can be distributed across a large set of annotators without the need for them to coordinate on an event naming scheme; they need only know the guidelines for how events are to be distinguished from each other in the context of the application.

Interactive Heatmap. We help annotators look for fruitful pairs in an interactive two-dimensional heatmap visualization that positions all the documents in the corpus by time (x-axis) and topic (y-axis). Here *time* refers to the date of the article, and *topic* refers to a projection of the event similarity model's representation of the document text into one dimension, grouped into M discrete buckets based on an estimate of the number of events of interest. The intensity of each cell indicates what fraction of the day's documents are mapped to that combination of topic and date under the current model. Each row is labeled with the most informative words in the text of the topic it represents.

Since the documents that constitute a news event typically have temporal and semantic locality, stories from the same event tend to be counted in the same, or in nearby, cells, and events often manifest as rectangular regions. We show heatmaps for two different domains, Twitter (Figure 1) and broadcast news (Figure 2).

User Feedback. The annotator may explore the heatmap and view a sample of the documents that are counted in each cell. If they select a region, they are shown a randomly selected pair of documents in the region and asked whether or not the documents belong to the same event. These questions are picked so as to generate useful data for the triplet training scheme described in section 3. Positive pairs (those affirmed to belong to the same event) push the representations of the even similarity model closer together for these documents, while negative pairs (those said to be from unrelated events) push them apart. Triplets are constructed when positive and negative pairs share an anchor document.

Incremental Learning. Once a set of comparisons is collected and the event similarity model is fine-tuned for the new detection task, the above process can be applied iteratively; that is, we can regenerate the heatmap according to the updated event similarity representations, and solicit more feedback from the user or annotators. Additionally, as we now have a full TDT model that can assign documents to event IDs, we can tabulate (and visualize) how well the new model addresses the cumulative human feedback collected. A flawless event similarity model would place all of the documents corresponding to an event into the same row.

4 CONCLUSION AND FUTURE WORK

In this short paper we have outlined our in-progress work on a human-in-the-loop topic detection and tracking system, and we have introduced a topic-time heatmap visualization that human annotators can use to improve both the efficacy of the system on new corpora, and their own understanding of the data. We are in the process of measuring this system's efficacy compared to established techniques for collecting event detection training data[1].

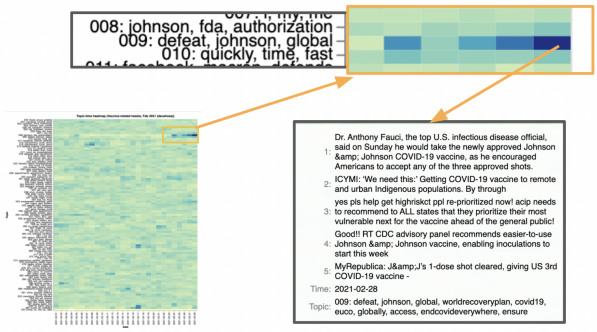

Figure 1: A topic-time heatmap (left) built from a sample of tweets from Twitter in February, 2021. The highlighted section (top) reveals discussion of the authorization of the Johnson and Johnson vaccine in the USA at the end of that month. Sample documents (right) are shown when the user examines one of the cells.

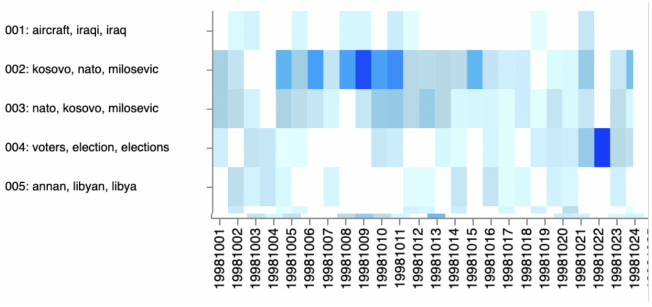

Figure 2: Part of the topic-time heatmap for the TDT3 corpus, which includes broadcast news stories starting from late 1998. The dark cells in topics 2 and 3 reveal a burst of articles related to NATO involvement in the Kosovo War.

[1]Our software for generating heatmaps from a corpus of timestamped documents is available at `https://github.com/social-machines/semsearch`

REFERENCES

[1] James Allan. 2002. Introduction to topic detection and tracking. In *Topic detection and tracking*. Springer, 1–16.

[2] James Allan, Jaime G Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic detection and tracking pilot study final report. (1998).

[3] James Allan, Ron Papka, and Victor Lavrenko. 1998. On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 37–45.

[4] Farzindar Atefeh and Wael Khreich. 2015. A survey of techniques for event detection in twitter. *Computational Intelligence* 31, 1 (2015), 132–164.

[5] Christopher Cieri. 2000. Multiple annotations of reusable data resources: Corpora for topic detection and tracking. *Actes 5ième Journées Internationales d'Analyse Statistique des Données Textuelles (JADT)* (2000).

[6] Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! *arXiv preprint arXiv:1901.05415* (2019).

[7] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).

[8] Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. 2019. Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321* (2019).

[9] Hannah Kim, Dongjin Choi, Barry Drake, Alex Endert, and Haesun Park. 2019. TopicSifter: Interactive search space reduction through targeted topic modeling. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 35–45.

[10] Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018. Can Neural Machine Translation be Improved with User Feedback? *arXiv preprint arXiv:1804.05958* (2018).

[11] Julia Kreutzer, Stefan Riezler, and Carolin Lawrence. 2020. Learning from Human Feedback: Challenges for Real-World Reinforcement Learning in NLP. *arXiv preprint arXiv:2011.02511* (2020).

[12] Varun Kumar, Alison Smith-Renner, Leah Findlater, Kevin Seppi, and Jordan Boyd-Graber. 2019. Why Didn't You Listen to Me? Comparing User Control of Human-in-the-Loop Topic Models. *arXiv preprint arXiv:1905.09864* (2019).

[13] Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies* 105 (2017), 28–42.

[14] Jiwei Li, Alexander H Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2016. Dialogue learning with human-in-the-loop. *arXiv preprint arXiv:1611.09823* (2016).

[15] Bing Liu, Gokhan Tur, Dilek Hakkani-Tur, Pararth Shah, and Larry Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. *arXiv preprint arXiv:1804.06512* (2018).

[16] Sebastiao Miranda, Artūrs Znotiņš, Shay B Cohen, and Guntis Barzdins. 2018. Multilingual clustering of streaming news. *arXiv preprint arXiv:1809.00540* (2018).

[17] Kailash Karthik Saravanakumar, Miguel Ballesteros, Muthu Kumar Chandrasekaran, and Kathleen McKeown. 2021. Event-Driven News Stream Clustering using Entity-Aware Contextual Embeddings. *arXiv preprint arXiv:2101.11059* (2021).

[18] Hassan Sayyadi, Matthew Hurst, and Alexey Maykov. 2009. Event detection and tracking in social streams. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 3.

[19] Milad Sikaroudi, Benyamin Ghojogh, Amir Safarpoor, Fakhri Karray, Mark Crowley, and Hamid R Tizhoosh. 2020. Offline versus online triplet mining based on extreme distances of histopathology patches. In *International Symposium on Visual Computing*. Springer, 333–345.

[20] Patrice Simard, David Chickering, Aparna Lakshmiratan, Denis Charles, Léon Bottou, Carlos Garcia Jurado Suarez, David Grangier, Saleema Amershi, Johan Verwey, and Jina Suh. 2014. ICE: enabling non-experts to build models interactively for large-scale lopsided problems. *arXiv preprint arXiv:1409.4814* (2014).

[21] Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system. In *23rd International Conference on Intelligent User Interfaces*. 293–304.

[22] Todor Staykovski, Alberto Barrón-Cedeno, Giovanni Da San Martino, and Preslav Nakov. 2019. Dense vs. Sparse Representations for News Stream Clustering.. In *Text2Story@ ECIR*. 47–52.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[24] Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics* 7 (2019), 387–401.

[25] Zijie J Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting humans in the natural language processing loop: A survey. *arXiv preprint arXiv:2103.04044* (2021).

[26] Jianshu Weng and Bu-Sung Lee. 2011. Event detection in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5.

[27] Yiming Yang, Tom Pierce, and Jaime Carbonell. 1998. A study of retrospective and on-line event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 28–36.