Unsupervised Medical Timeline Summarization of Reddit Posts

Umich EECS 598-017 Project Report, 12-09-2020

Minxue Niu
Department of Computer Science
University of Michigan
sandymn@umich.edu

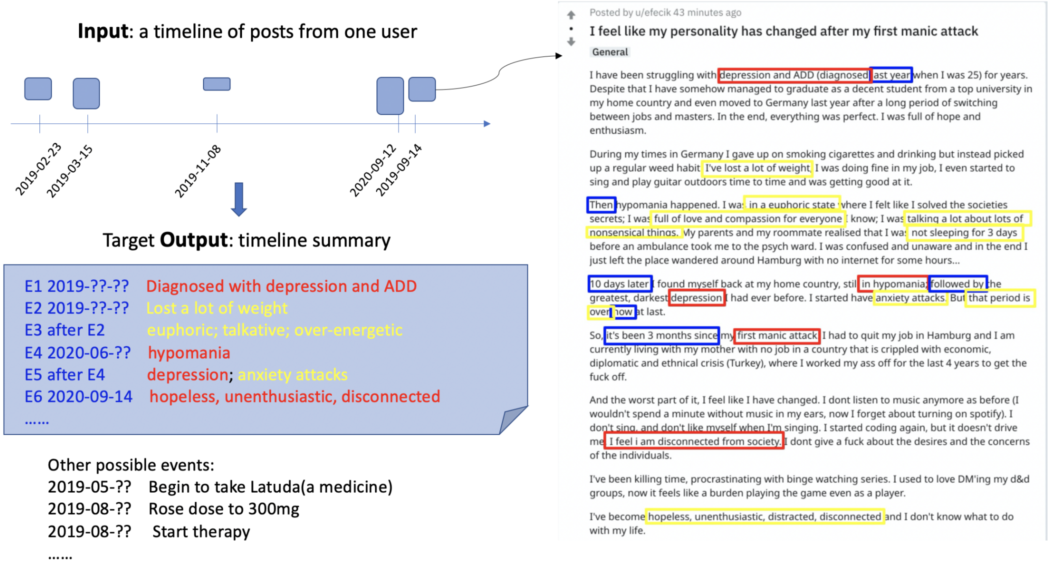
Xueming Xu
Department of Computer Science
University of Michigan
xueming@umich.edu

1 Problem Description

Diagnosis and treatment of mental disorders involve more ambiguity and subjective choices of the clinicians, compared with those of physical diseases. Symptoms differ a lot in each individual and usually don't have numerical measures, which makes it harder to do controlled experiment, comparison or causal inference. Besides, the chronical nature of mental disorders makes keeping track of mood history hard while important.

In the meantime, a lot of people post on online forums sharing their personal experience with those disorders. Reddit, for example, has subreddits for communities(patients and their loved ones) suffering from various mental disorders like depression, autism, bipolar disorder and AHDH. Among those posts, a considerable amount of them have detailed descriptions of their mood treatment and personal life histories, which is a good source for disease/treatment study.

This project aims to turn those noisy and unstructured posts into short, readable timeline summarizations in an unsupervised manner. A sample input post and output summary is shown in fig 1. Note that the output is semi-structured: each record in the format of TIME-Symptom or TIME-Treatment format, but the symptom descriptors are generated by the algorithm without constraints.


Input: a timeline of posts from one user
2019-02-23
2019-03-15
2019-11-08
2020-09-12
2019-09-14
Target **Output**: timeline summary

E1 2019-??-?? Diagnosed with depression and ADD
E2 2019-??-?? Lost a lot of weight
E3 after E2 euphoric; talkative; over-energetic
E4 2020-06-?? hypomania
E5 after E4 depression; anxiety attacks
E6 2020-09-14 hopeless, unenthusiastic, disconnected
……

Other possible events:
2019-05-?? Begin to take Latuda(a medicine)
2019-08-?? Rose dose to 300mg
2019-08-?? Start therapy
……

Posted by u/efecik 43 minutes ago

I feel like my personality has changed after my first manic attack

General

I have been struggling with depression and ADD (diagnosed last year when I was 25) for years. Despite that I have somehow managed to graduate as a decent student from a top university in my home country and even moved to Germany last year after a long period of switching between jobs and masters. In the end, everything was perfect. I was full of hope and enthusiasm.

During my times in Germany I gave up on smoking cigarettes and drinking but instead picked up a regular weed habit. I've lost a lot of weight. I was doing fine in my job, I even started to sing and play guitar outdoors time to time and was getting good at it.

Then hypomania happened. I was in a euphoric state where I felt like I solved the societies secrets; I was full of love and compassion for everyone I know; I was talking a lot about lots of nonsensical things. My parents and my roommate realised that I was not sleeping for 3 days before an ambulance took me to the psych ward. I was confused and unaware and in the end I just left the place wandered around Hamburg with no internet for some hours...

10 days later I found myself back at my home country, still in hypomania; followed by the greatest, darkest depression I had ever before. I started have anxiety attacks. But that period is over now at last.

So, it's been 3 months since my first manic attack. I had to quit my job in Hamburg and I am currently living with my mother with no job in a country that is crippled with economic, diplomatic and ethnical crisis (Turkey), where I worked my ass off for the last 4 years to get the fuck off.

And the worst part of it, I feel like I have changed. I dont listen to music anymore as before (I wouldn't spend a minute without music in my ears, now I forget about turning on spotify). I don't sing, and don't like myself when I'm singing. I started coding again, but it doesn't drive me. I feel i am disconnected from society. I dont give a fuck about the desires and the concerns of the individuals.

I've been killing time, procrastinating with binge watching series. I used to love DM'ing my d&d groups, now it feels like a burden playing the game even as a player.

I've become hopeless, unenthusiastic, distracted, disconnected and I don't know what to do with my life.

Figure 1: Task: Generate timeline summary from reddit posts

2 Related work

To the best of knowledge, no existing study works on the same problem in terms of a) uses online posts b) outputs semi-structured timeline summarization and c) is totally unsupervised. In this section, we will review some related works we introduced in our proposal, and also add some new interesting

2.1 (Supervised) Medical Time/Event Extraction

The work that has the most similar goal with ours is the SemEval-2017 Task 12: Clinical TempEval[1]. The task focuses on temporal information extraction on clinical data, where they train time/event/temporal relation extractors on carefully annotated clinical records. One sample note and annotation are shown in fig 2. Although the target output is very clean, accurate and helpful, 2 major problems limit its application in psychiatry domain: 1) Annotation too expensive. As you can see the annotations are in complicated structure and dependencies(temporal relationships rely on event/time identification), and thus are very expensive to obtain; 2) Events that can be identified are very limited and rely on domain-specific knowledge.

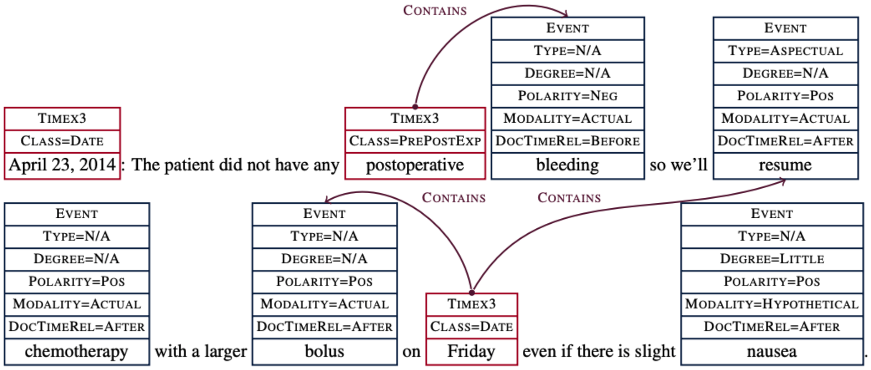


Figure 2: Example Clinical TempEval annotations

In medical domain, summarizing EHR data has been a challenge for more than a decade. An earlier study defines a set of event tags(test/medical operation/symptom/etc) markup scheme and outputs tables of tag-value pairs[2]. It especially addresses negative-event recognition - things that were mentioned but didn't actually happen. Another approach[3] generates extractive summary with concept recognition and relation detection techniques. We also find a nice survey that examines more work on summarizing EHR data[4].

There are a few attempts on summarizing medical conversations. A very relevant paper[5] generates clinician notes(similar to our targets) from transcriptions of patient-doctor conversations. Another similar work investigates abstractive summarization of patient-nurse conversation with the aim of capturing 9 predefined symptoms of interest[6].

2.2 Unsupervised Text Summarization

Most current unsupervised summarization methods are extractive. They usually pull out sentences that are deemed important by some heuristic functions such as content relevance and novelty[7]. Abstractive methods usually utilizes a self-supervised task such as important words recovery[8] or next sentence prediction[9].

2.3 Timeline Summarization

Our task is essentially timeline summarization. Timeline summarization(TLS) automatically identifies key dates of major events and provides short descriptions of what happened on these dates. To the best of our knowledge, almost all systems proposed specifically for TLS have been extractive[10, 11, 12], but only one system has been abstractive[13]. The abstractive model in Steen et al., 2019[13] was proved to outperform the previous extractive models on two publicly available datasets Crisis[14] and Timeline 17(TL17)[15]. However, the previous TLS models are not applicable to our problem for several reasons.

2.4 Social Media Mood Analysis

Another stream of work is trying to correlate social media language/activity statuses with mental status(mental disorders, or more general attributes like personality/emotion). In [16], they can approximately matching the accuracy of screening surveys by analyzing facebook language. However, we don't refer too much to these works because their labels are usually coarse and their approaches capture more sentiment and language style rather than medical details.

As said, our project differs from all of those in that it aims at unsupervised medical text summarization with timeline-format output. We believe this is a meaningful task because it does not need annotations, which is usually expensive to obtain, and the output format is semi-structured and ready for a lot of applications. However, we do expect and also find in our attempts that having accurate control over the content without supervision is very challenging.

3 Methodology

It's very hard to extract semi-structured, decently accurate information from noisy domain with end-to-end systems, and here we adopt a pipeline-based approach instead. In this section we will go over those components.

3.1 Time Expression Extraction and Normalization

Time expressions are important information, and have relatively clear format. Therefore we built a time expression(TIMEX) extractor using pattern-based matching algorithms.

We later found an exising tool HeidelTime[17], whichextracts temporal expressions from documents and normalizes them according to the TIMEX3 annotation standard. It's used by a lot in related works. We think it will be nice to have this standard time expression, and it also handles most time expressions nicely, so we decide to use this tool for time tagging. However, it tends to identify trivial expressions. For example, the sentence "I have been on therapy for 3 months now", it will identify "3 months" as a duration and "now" as a present date "now". To avoid false positives, we still use our timetagger for segmentation and sorting(more on that soon), while their time expressions for display.

3.2 Segmentation and Filtering

We first process posts into coherent segments and do a round of filtering before feeding into neural network for 1) getting rid of some irrelavant information and 2) generating segments with similar lengths and focused topics, which will help the learning process. We have described the pipeline in midterm presentation and midterm progress report, so we will only recap main ideas here.

For segmentation, we cut the post into multiple segments based on 1) natural new paragraph and 2) change of time. We also use topic tiling methods to cut posts based on the continuity of sentence similarity. However, in practice we found that either LDA-vector based or SentBERT based representations tend to reply on token similarity rather than semantics. Therefore, we only use them as a supplementary method when the segments are too long.

To further clean up the data, we performed filtering referencing [2], which explores "negative events" that refers to mentioning of events that didn't actually happen to the subject. We combine parsing(with the Stanford tool stanza) results and text matching rules to filter out 1) other subjects(e.g. "My girlfriend is bipolar...") 2) Future tense and 3) intention words ("I'm considering...", "My doctor suggested..."). We only deal with these three because they are the major "negative events", and also this part is not our focus in this project. But there are definitely more can be done in this part.

3.3 Event Summarization: Modified RMN

To summarize the events in each segment, we adapted the unsupervised RMN model that incorporates dictionary learning to generate interpretable, accurate medical trajectory[18].

After training, the model learned a descriptor matrix R with each row in the same vector space as the word embedding and corresponding to a descriptor. By finding the nearest neighbors of each row of the descriptor matrix in the word embedding space, we can interpret each descriptor with several descriptive words. Each segment on the timeline is encoded into a weight vector with dimension equal to the number of descriptors. Each entry of the weight vector is interpreted as the weight of the corresponding descriptor. We pick the descriptor with the highest weight to summarize the segment.

Here are three main modifications to the original RMN model in the paper[18]. First, instead of using Glove word embedding directly, we fine-tuned Glove on our dataset in order to include unseen vocabulary like the name of medicines. Another modification is that we concatenate the title of the post with each segment to make the input globally contextualized. Also, we changed original RNN layer to a self-attention layer motivated by the observation that topic shift between consecutive segments is sharp and abrupt.

4 Experiments

4.1 Data

We carry out our experiments in the context of bipolar disorder(BD), a chronic psychiatric illness characterized by swings of mood and energy between healthy (euthymia) and pathological states (mania or depression). The dataset contains posts(author, title, text, time) from 3 bipolar-related subreddits: #bipolar/#bipolar2/#bipolarreddit.

For post quality, we filtered out users with more than 30 or fewer than 3 posts. This finally gave us **12,861** posts across **2,459** users. The average length of posts is **1,050.36** characters. After segmentation and filtering process, we got **74653** segments remaining. We used 69365 for training and 5288 for testing. We do not split evaluation set because we only use the test set to make sure validation loss is going down, not tuning the parameters. There is no direct measure of performance for hyperparameter tuning, so for convenience we just skip that part.

4.2 Embedding Finetuning and Training

In the first round of experiments, we found that the model is unable to learn drug names because they are not in the GloVe embeddings. Subword embeddings are suggested, but we found that hard to implement because when converting back to descriptors, the embeddings only correspond to subwords makes interpretation harder.

Therefore, we finally use Mittens[19], a GloVe extension tool, to finetune the glove embeddings on our training set. As we will see, this enables the model to learn medicine descriptors, and generally more in-domain descriptors. We used R-size=30/50, learning rate=0.01, to be consistent with the paper. We tried GloVe embeddings of size 50 and 300, and we found 50-dim embeddings seem leading to cleaner descriptors.

In the R-size=30 and embedding-size = 300 setting, our model takes about 1.5h to run 50 epoch. The training process(loss) is shown in fig 3.

5 Evaluation

Ideally we should evaluate the system on some time-annotated datasets like the SemEval2012 task17 dataset we mentioned in related work, but their data distribution has strict limitations and is only accessible by group members. So we mostly did qualitative evaluation by showing samples.

In this section, we want to answer 3 questions about our system:

1. Does our model learn meaningful descriptors?
2. Do the weights on descriptors nicely correlate to segment content(thus showing an understanding of the language)?
3. Are our summaries generally more helpful than other unsupervised summarization baselines?

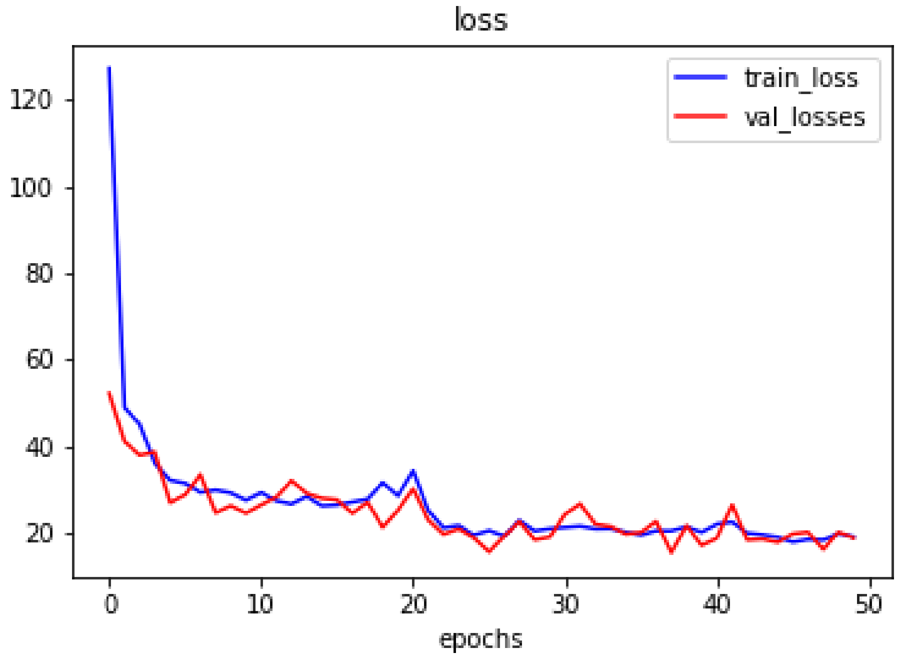


Figure 3: Training and Validation loss, timestep=50. We can see that the model converged after 20 epoches.

Descriptor Matrix Descriptor Matrix is randomly initialized and learned by back-propagation. After training, we interpret the matrix by looking printing out the nearest neighbors of each row in the word embedding.

We show the nearest word of each row, for the 50-dim and 300-dim models, respectively. (Note that these descriptors are model parameters, so we only have one set of descriptors for a set of hyper-parameters.)

50-dim: ’journey’, ’"could"’, ’suicidal’, ’provocations..’, ’suffer’, ’freckles,’, ’"everything!"’, ’counselling’, ’affect’, ’(big)pentagons.’, ’hotline.’, ’legless’, ’dose’, ’thoughts"’, ’etc.’, ’tho,’, ’topirimate,’, ’antibacterial’, ’remnant’, ’exposes’, ’purging,’, ’lifelong.’, ’deficits’, ’caff)....’, ’then..)..’, ’college-level’, ’"spiritual"’, ’impacting’, ’spins,’, ’8/9/10’]

300-dim:’ads?’, ’bleed’, ’sidekick’, ’coronavirus’, ’self-discipline’, ’"jesus...’, ’recuperation.’, ’question.**’, ’people’, ’lorazepams’, ’24-hour’, ’"discussion,"’, ’pasty’, ’seasons"?’, ’waved’, ’lightyears’, ’riddick’, ’piling.’, ’std,’, ’adderall’, ’duplicate’, ’gre’, ’bit.’, ’notified’, ’daydreamy’, ’(excuse’, ’layout’, ’dimension...’, ’licking’, ’"in-love",’

Findings:

1. GloVe embeddings might be not very clean, and finetuning process seems made that worse. A better pre-trained embedding sets can help with interpretation.

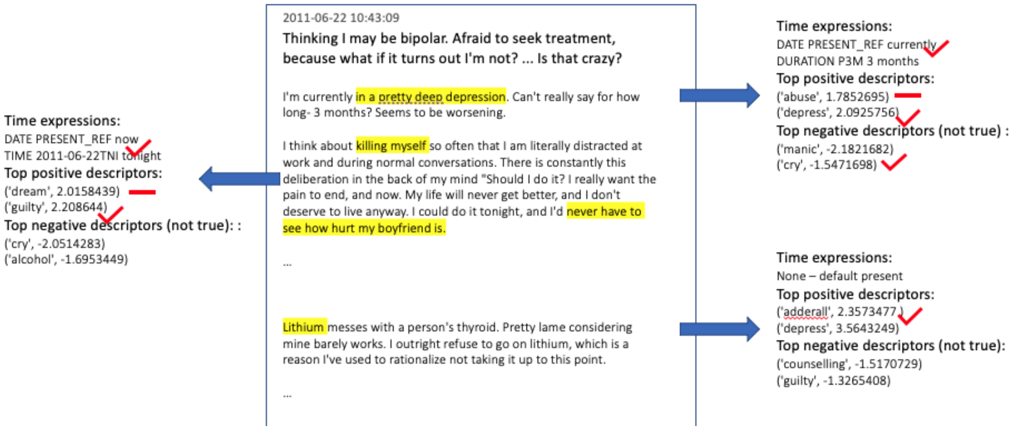
2. In general those words are in-domain, but not very representative of bipolar events.

3. In the training process, we found that in 300-dim model the word list updates slowly, with some of them not changing at all from the beginning. We think the problem is, the learned embeddings are not exactly in the original GloVe space, so even if the R matrix changes a lot the nearest neighbor will still be the same one. 50-dim descriptors tend to update faster, likely because the space is more dense.

5.1 Weights

In the following experiment, we fixed the descriptor matrix to be (18) selected top words from human annotations(on 20 posts, done by ourselves) and only let the model learn the weights. For each segment, we use different cutoffs to form machine summary sets, and calculate the precision/recall of the model, compared with human annotations(only on the 18 selected words).

A sample post and the output of our system is shown in fig 4.



2011-06-22 10:43:09

Thinking I may be bipolar. Afraid to seek treatment, because what if it turns out I'm not? ... Is that crazy?

I'm currently in a pretty deep depression. Can't really say for how long- 3 months? Seems to be worsening.

I think about killing myself so often that I am literally distracted at work and during normal conversations. There is constantly this deliberation in the back of my mind "Should I do it? I really want the pain to end, and now. My life will never get better, and I don't deserve to live anyway. I could do it tonight, and I'd never have to see how hurt my boyfriend is.

...

Lithium messes with a person's thyroid. Pretty lame considering mine barely works. I outright refuse to go on lithium, which is a reason I've used to rationalize not taking it up to this point.

...

Time expressions:
DATE PRESENT_REF now
TIME 2011-06-22TNI tonight
Top positive descriptors:
('dream', 2.0158439)
('guilty', 2.208644)
Top negative descriptors (not true): :
('cry', -2.0514283)
('alcohol', -1.6953449)

Time expressions:
DATE PRESENT_REF currently
DURATION P3M 3 months
Top positive descriptors:
('abuse', 1.7852695)
('depress', 2.0925756)
Top negative descriptors (not true) :
('manic', -2.1821682)
('cry', -1.5471698)

Time expressions:
None – default present
Top positive descriptors:
('adderall', 2.3573477)
('depress', 3.5643249)
Top negative descriptors (not true):
('counselling', -1.5170729)
('guilty', -1.3265408)

Figure 4: Sample output of our system when fixed R with 18 human-selected words and only learning weights. Word list: ’anxiety’, ’concentrate’, ’counselling’, ’cry’, ’dream’, ’emotionless’, ’happy’, ’guilty’, ’manic’, ’depress’, ’mixed’, ’lorazepams’, ’therapy’, ’worried’, ’move’, ’abuse’, ’agitated’, ’alcohol’, ’adderall’

Findings:

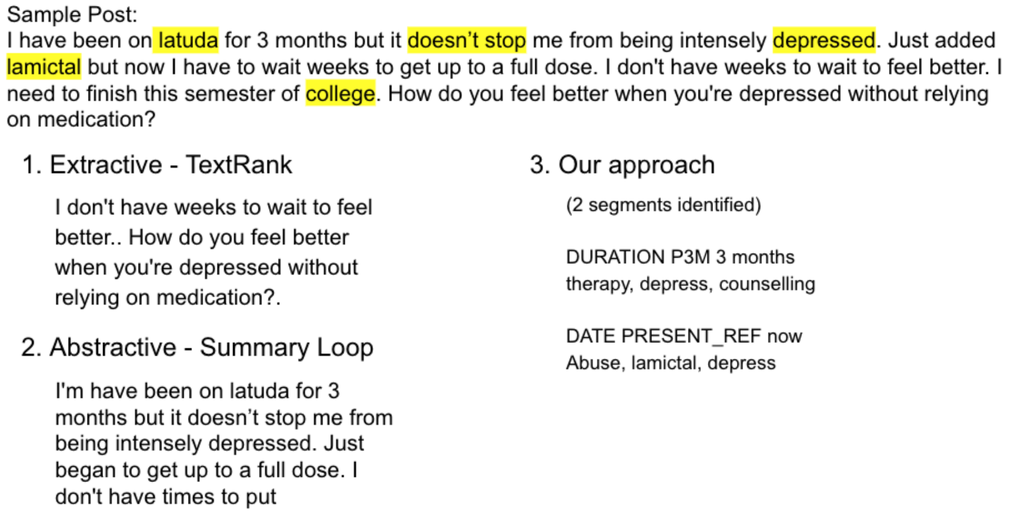
1. Segmentation and Time Expressions are decent.
2. In general the model is able to capture theme-information. For example, when medicine is mentioned, the word "adderall" will noticeably go up while it doesn’t normally show at the top. The word "depress" and "manic" usually have opposite weights, and seems our model can pick the right one in most times.
3. What triggers certain words are not clear. For example, when the post mentions suicidal thoughts and how hurt her boyfriend will be, the work "guilt" gets top weights, which is really good but in the meantime too good to be true. In general, weights go in the right direction, but the value is not accurate enough for a medical summary.
4. Longer word list would be helpful for evaluation. We just don’t have enough time to experiment the model with different lists.

5.2 Sample outputs of our model and other baselines

We compare our output with 2 unsupervised baselines, one extractive(TextRank[20]) and one abstractive(SummaryLoop[21]). As an example shown in fig 5, unfortunately none of the methods, including ours works well.

Findings:

1. TextRank sentence selection seems pretty random. The method inherently look for novel content, while for bipolar and other psychiatry timelines similar events happen a lot but are all important (e.g. depression to mania, and back to depression again; change of medications, etc.) So this method is not suitable for such tasks.
2. Summary Loop is not working(we ran the code they released). This might be because the model does not converge. The original paper reported 10 days of training, but limited by



Sample Post:
I have been on latuda for 3 months but it doesn't stop me from being intensely depressed. Just added lamictal but now I have to wait weeks to get up to a full dose. I don't have weeks to wait to feel better. I need to finish this semester of college. How do you feel better when you're depressed without relying on medication?

1. Extractive - TextRank

I don't have weeks to wait to feel better.. How do you feel better when you're depressed without relying on medication?.

2. Abstractive - Summary Loop

I'm have been on latuda for 3 months but it doesn't stop me from being intensely depressed. Just began to get up to a full dose. I don't have times to put

3. Our approach

(2 segments identified)

DURATION P3M 3 months
therapy, depress, counselling

DATE PRESENT_REF now
Abuse, lamictal, depress

Figure 5: Output from 2 baselines and out system. It a random sample from our dataset Our model fails to capture the "medication" information in the first segment.

time and resources we used the model after one day of training. And we do not train our own fluency scorer - the original one was trained with news articles. The model generally pick the first L words, with slight modifications at the beginning.

3. As to our model, the timeline structure is helpful, but the content summary is way too vague. We still believe this is a direction worth pursuing, but RMN only is not likely a solution.

6 Conclusions and Problems

6.1 Conclusion

In this project, we found that event extraction and summarization is very hard without supervision. When learning descriptors, the model mostly preserves relevant information, but not representative enough. Some human guidance like R matrix initialization/fix can be helpful, and again the model is able to capture strong signals, but far from can be actually applied. Baseline models don't perform well either, which leaves a great space for exploration on this task.

6.2 Problem: Descriptors not exactly in-domain

Looking closely into the training process, we found that it's hard to ensure the learned descriptors are in the same space with original embeddings. We inspected this by tracking the nearest GloVe word of each row vector every 10 epoches, and found that sometimes(for example when using cosine similarity instead of dot product for loss function) the words are not updated even though the R matrix changed a lot, suggesting that the R matrix is not in a dense space of GloVe embeddings. We are able to learn some meaningful descriptors, but unclear whether it's by chance or not.

6.3 Problem: Can't utilize the power of contextualized word embeddings

This is a 2018 paper and it embeds sentences by averaging word vectors, which inevitably misses a lot of sentence/segment-level semantic information. In recent 2 years transformers have renewed SOTA performance in almost all nlp tasks. It would be great to make use of their power in this summarization task. However, RMN seems not generalizable - It replies on static word embeddings to interpret the descriptors, and thus contextualized word embeddings can not be applied. We wonder if there could be ways to use transformer-like encoder-decoder model, while still preserve the interpretability of R matrix.

References

[1] Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. SemEval-2017 task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, August 2017.

[2] Eiji ARAMAKI et al. Text2table:medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the Workshop on BioNLP, ACL 2009*, pages 185–192.

[3] Jennifer Liang, Ching-Huei Tsou, and Ananya Poddar. A novel system for extractive clinical note summarization using ehr data. In *Proceedings of the Workshop on BioNLP, ACL 2009*, pages 185–192.

[4] Rimma Pivovarov and Noemie Elhadad. Automated methods for the summarization of electronic health records. In *Journal of the American Medical Informatics Association*, pages 938–947.

[5] Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. Generating SOAP notes from Doctor-Patient conversations. In *arXiv preprint arXiv:2005.01795*, 2020.

[6] Z. Liu, A. Ng, S. Lee, A. T. Aw, and N. F. Chen. Topic-aware pointer-generator networks for summarizing spoken conversations. In *arXiv preprint arXiv:1910.01335*, 2019.

[7] Akanksha Joshi, E. Fidalgoa, E. Alegrea, and Laura Fernández-Robles. Summcoder: An unsupervised framework for extractive text summarization based on deep auto-encoders. In *Expert Systems With Applications*, pages 200–215, 2019.

[8] Philippe Laban, Andrew Hsi, John Canny, and Marti A. Hearst. The summary loop: Learning to write abstractive summaries without examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[9] Peter West, Ari Holtzman, Jan Buys, and Yejin Choi. Bottlesum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle. In *arXiv preprint arXiv:1909.07405*, 2019.

[10] Kiem-Hieu Nguyen, Xavier Tannier, and Véronique Moriceau. Ranking multidocument event descriptions for building thematic timelines. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1208–1217, 2014.

[11] Hai Leong Chieu and Yoong Keok Lee. Query based event extraction along a timeline. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 425–432, 2004.

[12] William Yang Wang, Yashar Mehdad, Dragomir Radev, and Amanda Stent. A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 58–68, 2016.

[13] Julius Steen and Katja Markert. Abstractive timeline summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 21–31, 2019.

[14] Giang Tran, Mohammad Alrifai, and Eelco Herder. Timeline summarization from relevant headlines. In *European Conference on Information Retrieval*, pages 245–256. Springer, 2015.

[15] Giang Binh Tran, Mohammad Alrifai, and Dat Quoc Nguyen. Predicting relevant news events for timeline summaries. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 91–92, 2013.

[16] Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208, 2018.

[17] Jannik Strötgen and Michael Gertz. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324, 2010.

[18] Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544, 2016.

[19] Nicholas Dingwall and Christopher Potts. Mittens: an extension of glove for learning domain-specialized representations. *arXiv preprint arXiv:1803.09901*, 2018.

[20] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[21] Philippe Laban, Andrew Hsi, John Canny, and Marti A Hearst. The summary loop: Learning to write abstractive summaries without examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5135–5150, 2020.