Learning Thread Reply Structure on Patient Forums

Yunzhong Liu
Computer Science and Engineering
Arizona State University
Tempe, AZ, USA
liuyz@asu.edu

Feng Chen
Heinz College
Carnegie Mellon University
Pittsburgh, PA, USA
fchen1@cmu.edu

Yi Chen
School of Management
New Jersey Institute of Technology
Newark, NJ, USA
yi.chen@njit.edu

ABSTRACT

The thread reply structure on patient forums is important for users and automated techniques to understand the discussion content and search information effectively. However, most online patient forums only have partially labeled structures. In patient forums, the discussions by patients and caregivers contain abundance of person references, which provide strong indication of the thread reply structure. In this paper, we propose using person reference resolution, combined with a statistical machine learning model, to learn the unknown thread structure on patient forums. Our preliminary performance evaluation has verified the effectiveness of the proposed approaches.

Categories and Subject Descriptors

H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval

Keywords

Thread Reply Structure; Healthcare Informatics; Patient Forums; Person Resolution; Machine Learning.

1. INTRODUCTION

Online patient forums provide a convenient channel for patients, caregivers, and medical staff to share personal experience and support each other. A recent report shows 80% of internet users gather health information online [1], such as the Patientslikeme [1] and WebMD [2] forums. A patient forum is typically contributed by patients and their caregivers. It contains valuable information about various medical situations experienced by different patients as well as suggestions and feedbacks from others based on their experience. The fast growing content in such forums is becoming big data. A pressing challenge is to understand the semantics and to obtain knowledge from the forum data.

[1]http://www.patientslikeme.com
[2]http://exchanges.webmd.com

DARE'13, November 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2512410.2512426 ...$15.00.

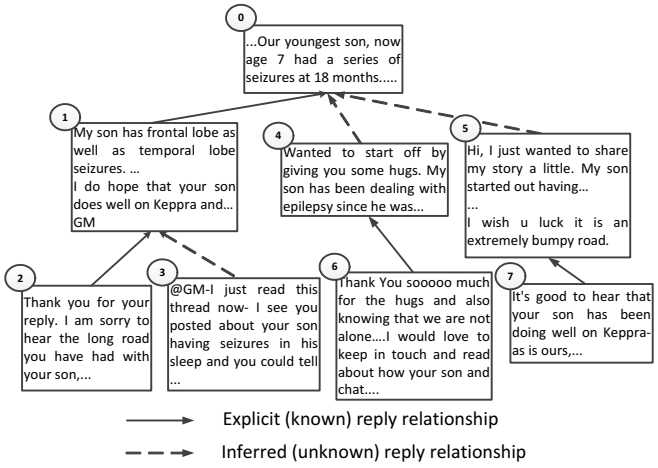

Figure 1: A simplified thread sampled from a patient forum.

The thread reply structure is important for human to understand the discussion content and for automatic methods to perform information extraction for effective search [8]. A typical forum thread consists of a sequence of posts, ordered according to the time when the post is submitted. The thread structure is often modeled as a tree, where each post is a node and has one parent to which it replies, except the first post, the root of the tree [10]. An example of a forum thread in tree representation is shown in Fig. 1, which is extracted from one thread in the epilepsy foundation forum [3].

The thread reply structure is especially important for understanding patient forums. In a patient forum, users typically describe problems, such as a symptom or a disease, or make comments in a subjective way, using personal experience as the context. A clear thread structure can help readers to easily understand who the described symptoms are related to and who is the receiver of the treatment suggestion. In [8], the thread reply structure has been used for patient-centered information extraction for effective search on healthcare forums.

However, most patient forums do not have the complete thread structure available, which means the parents of some posts are unknown. The goal of this work is to learn the complete thread reply structure. This is challenging due to the large search space as the number of candidate thread trees increases exponentially with the number of posts in a thread.

In this work, we have made two observations, namely, the characteristics of patient forums, and the prevalence of partially labeled thread structures, and then proposed our approaches for learning the complete thread reply structure.

[3]http://epilepsyfoundation.ning.com/forum

First, we observe that patient forums usually focus on the discussion of patients, and the interactions between patients and caregivers. In a patient forum, the person reference relationships are critical to understand the context, like which patient has the discussed symptoms and who has given a treatment suggestion to whom. The person reference relationship can also help learn the thread reply structure. For example, when one post replies to another, it tends to mention the person described in the parent post. Conversely, if one post mentions a person that is described in a preceding post, then this post is likely to be a child or descendant of that preceding post. According to this observation, if we can find out the person reference relationships in a person resolution process, that can provide helpful information for learning the thread reply structure.

Second, we observe that most patient forums, such as WebMD and the epilepsy forum, have abundance of partially labeled reply structures, which can be leveraged for learning a complete thread structure. There are always some post authors who have a good habit of keeping an explicit reply structure. An example of such a partially labeled thread structure is shown in Fig. 1, where the solid arrow indicates the known reply relationship while the dotted arrow indicates the unknown reply relationship. We can leverage the known reply relationships to infer the unknown reply relationships.

In this paper, we first propose using person reference resolution to learn the unknown thread structure on patient forums. Then we combine it with a statistic machine learning model for learning the complete thread structure with the partially labeled data. Our preliminary performance evaluation has verified the effectiveness of the proposed approaches.

Related Work: There is some existing work for thread structure learning on other types of forums, such as technical forums [7, 10]. The approach proposed in [10] is a supervised learning model, which requires training data that have a complete thread structure. On the other hand, the approach with unsupervised learning [7] does not leverage the partially labeled data. Besides, none of them focus on patient forums and use the abundant person reference relationships for thread structure learning.

2. PERSON RESOLUTION FOR THREAD STRUCTURE LEARNING

Person resolution (PR) is the process of identifying the same person mentioned in different contexts. In this section, we focus on how to apply person resolution for thread structure learning.

We use the inter-post and intra-thread person resolution for thread structure learning, since the thread reply structure also focuses on the relationships between posts within a thread. In other words, we only check if a pair of person mentions from two different posts in the same thread refer to the same person. Person resolution belongs to co-reference resolution. So the more general co-reference resolution can be used for person resolution. However, the current co-reference resolution systems, such as Stanford's multi-pass unsupervised deterministic co-reference resolution system [6], mainly focus on the co-reference resolution within a document, like a post. Instead, we focus on the person resolution between posts, and have defined the PR features specific to our patient forums. Similar to [6], our PR features belong to several types and each feature type has a different priority. We arrange them in the order of descending priority as follows.

PR Feature Type 1: Matching between the address in the current post content and the signature in the content of a candidate parent post. Usually, the address appears at the beginning of a post or follows some tokens like "hi", "hello", and etc., while the signature appears at the end of a post following some tokens like "thanks", "regards", and etc. We extract these features based on Name Entity Recognition (NER) [2] and some common patterns expressed by regular expressions.

PR Feature Type 2: Matching between the same role related to the same person. For example, "your daughter" in the current post can match "my/our daughter" in a candidate parent post. Here the role, "daughter", is identified using the family group semantic type by MetaMap tool in the Unified Medical Language System (UMLS) [3], while the Part-of-Speech (POS) and syntactic analysis modules in [2] are used to identify these matched phrases.

PR Feature Type 3: Matching between the second person pronoun like "you" in the current post and the first person pronoun, like "I" or "we", in a candidate parent post that tend to refer to the same person. Semantic role labeling (SRL) [4] and WordNet [5] are used for checking if they tend to refer to the same person, e.g., checking if their associated verbs are synonyms.

PR Feature Type 4: Matching between the third person pronoun in the current post and the person name or role in a candidate parent post that are consistent in gender. For example, "she" in the current post can match "Mary" in a candidate parent post.

We use these PR features to learn the unknown reply structure. Currently we have only defined the pairwise PR features with each feature involving a pair of posts. We can thus select those candidate parents for each post independently. Basically, we select out those candidate parents matched with the features of higher priority. When there are multiple candidates left at the end, we use the following two rules to break the tie. First, if one and only one candidate parent post in the candidate set is from the thread initiator, and the author of the current post is not from the initiator, then this parent post from the thread initiator will be output as the labeled parent. Here the thread initiator is the author of the first post in this thread. This rule is based on the assumption that people, except for the thread initiator, tend to reply to the person who initiates the thread. Second, if two candidates are both from the thread initiator or neither of them is from the thread initiator, then the closest candidate parent will be more preferred. This rule is based on the assumption that people tend to reply to the latest post given that all the other factors are the same.

3. LEARNING THREAD STRUCTURE WITH PR AND THREADCRF

Thread conditional random field (threadCRF), proposed in [10], has been shown to be very effective for thread structure learning. However, since threadCRF is a supervised learning model, it requires a completely labeled data set for model training before it can be applied for thread structure learning. In this section, we propose to use person resolution for generating a fully labeled training set given the partially labeled data. The generated data set can be considered as an approximation of the ground truth and used to bootstrap the supervised threadCRF model training. Then we use the trained model for re-labeling the unknown structures with the known structures as constraints.

We first materialize multiple possible thread reply structures given one partially labeled thread structure. The materialization procedure is similar to that in [9], except that we only materialize the most likely reply structures. We use the person resolution technique discussed in Section 2 to evaluate the likelihood of all possible candidate parents and only materialize the most likely candidates. In this way, we expect the materialized thread reply structures are more similar to the ground-truth reply structures, thus obtaining a

Table 1: Thread structure learning performance

	FIRST	LAST	SIM	PR	PR + ThreadCRF
Accuracy	0.444	0.429	0.361	0.583	**0.635**

more accurate training data set, which in turn can help learn a more accurate threadCRF model.

When there is no clear person resolution indication for some posts, we use some other heuristics to reduce the number of candidate parents in order to improve the model accuracy and training efficiency. In [10], three unsupervised baseline approaches are used to recover the thread reply structure: reply to the first post, reply to the last post, and reply to the post with the highest content similarity. These three baseline methods are referred as FIRST, LAST, and SIM, respectively. We also observe that, in most cases, the real parent is from these candidates. Therefore, when there are no better candidates from the person resolution perspective, we will at most materialize three candidates based on the three baselines.

With the materialized fully labeled data set, we can learn the threadCRF model parameters. We then use the learned model to re-label the unknown reply structures with the existing reply structures as constraints. Note that when training the threadCRF model, we treat all the materialized instances generated from the same thread with equal possibilities, and leave the assignment of different weights for different training instances as our future work.

4. PRELIMINARY PERFORMANCE EVALUATION

Data Set: We collected 9210 posts in 911 threads published on "Patient help patient" sub-forum in the epilepsy foundation forum. Among the 911 threads, we chose 200 threads for the experiments, which have a significant number of labeled reply relationships. Specifically, for each thread we computed the ratio of the number of known parent labels to the total number of posts. If a thread has a ratio above a threshold 0.5, then it is chosen. We observe that some threads include a large number of posts, with some parts heavily labeled while the other parts are not. If a thread does not meet the threshold as a whole, but the first i posts as a set meets the threshold, and adding any number of following posts in sequential order to that set will fail to meet the threshold, then we take the set of the first i posts as a thread. Note that these i posts still represent a thread tree, which is required for the threadCRF model learning and inference. With this threshold, we have a good size of known labels, and yet have threads that include many posts.

As discussed earlier, the goal of this work is to learn the complete thread structure with the partially labeled data. To evaluate the thread structure learning performance, those unknown reply structures, 468 unknown reply relationships in the experiment set, were manually labeled.

Results Analysis: We compare five methods: FIRST, LAST, SIM, PR, and PR + ThreadCRF, and use the accuracy of individual labels as the evaluation metric. As in [10], the accuracy of individual labels is defined as the proportion of correct labels in the whole set of predicted labels. Table 1 shows the performance comparison among the five methods for learning thread reply structures. Among the first four methods, PR achieves the best performance. When we combine PR with threadCRF, we can see that the performance is significantly improved. The performance improvement can be explained as follows: in addition to the semantic features used in PR, the original threadCRF has also incorporated some syntactic and structure features, and they jointly help to learn a more accurate thread reply structure.

5. CONCLUSION AND FUTURE WORK

In this paper, we propose to use person resolution combined with threadCRF for learning a complete thread reply structure. Our preliminary experiment evaluation shows that the proposed approaches are very promising.

In the future, our work will focus on the following directions. First, we will improve the current PR system by investigating more complex features and developing an advanced candidate thread structure selection algorithm. In particular, we may identify the person's social role, such as patient, caregiver, or doctor, and use that to help person resolution. Second, in addition to the pairwise PR features, we will also use features involving multiple pairs of posts. For example, a post may mention a person introduced in a candidate grandparent post. The candidate grandparent, parent, and the post itself thus form a possible reply path. With features involving multiple pairs of posts, a global optimization algorithm is needed to find the best thread reply structure with all the potential person reference relationships. At last, with all these PR features, we will investigate the role and importance of each feature type and their availability across multiple patient forums.

Acknowledgments

This material is based on work partially supported by NSF CAREER Award IIS-0845647, IIS-0915438, an IBM Faculty Award and a Google Research Award. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

6. REFERENCES

[1] Pew Internet: Health topics. `http://www.pewinternet.org/Reports/2011/HealthTopics.aspx`.

[2] Stanford core NLP tools. `http://nlp.stanford.edu/software/corenlp.shtml`.

[3] A. R. Aronson. Metamap: Mapping text to the UMLS metathesaurus (2006). `http://skr.nlm.nih.gov/papers/references/metamap06.pdf`.

[4] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

[5] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[6] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of ACL CoNLL-2011 Shared Task*, 2011.

[7] C. Lin, J. Yang, R. Cai, X. Wang, and W. Wang. Simultaneously modeling semantics and structure of threaded discussions: a sparse coding approach and its applications. In *Proceedings of ACM SIGIR*, 2009.

[8] Y. Liu and Y. Chen. Patient-centered information extraction for effective search on healthcare forum. In *Proceedings of International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction (SBP)*, 2013.

[9] Y. Tsuboi, H. Kashima, S. Mori, H. Oda, and Y. Matsumoto. Training conditional random fields using incomplete annotations. In *Proceedings of COLING*, 2008.

[10] H. Wang, C. Wang, C. Zhai, and J. Han. Learning online discussion structures by conditional random fields. In *Proceedings of ACM SIGIR*, 2011.