The Utility of Discourse Structure in Forum Thread Retrieval

Li Wang,[♠♡] Su Nam Kim,[♢] and Timothy Baldwin[♠♡]

♠ Department of Computing and Information Systems, The University of Melbourne
♡ NICTA Victoria Research Laboratory
♢ Faculty of Information Technology, Monash University
li@liwang.info, sunamkim@gmail.com, tb@ldwin.net

Abstract. Web user forums are a valuable means for users to resolve specific information needs, both interactively for the participants and statically for users who search/browse over historical thread data. However, the complex structure of forum threads can make it difficult for users to extract relevant information. Information retrieval (IR) over forum threads is one important way to obtain useful information on questions asked by others. In this paper, we investigate the task of IR over web user forums by utilising the discourse structure of forum threads. Experimental results show that exploiting the characteristics of discourse structure of forum threads can benefit IR, when compared to previously-published results.

Keywords: Discourse Structure, Web User Forum, Information Retrieval, Social Media, Dialogue Act

1 Introduction

Web user forums (or simply "forums") are online platforms for people to discuss information and obtain information via a text-based threaded discourse, generally in a pre-determined domain (e.g. IT support or DSLR cameras). With the advent of Web 2.0, there has been an explosion of web authorship in this area, and forums are now widely used in various areas such as customer support, community development, interactive reporting and online eduction. In addition to providing the means to interactively participate in discussions or obtain/provide answers to questions, the vast volumes of data contained in forums make them a valuable resource for "support sharing", i.e. looking over records of past user interactions to potentially find an immediately applicable solution to a current problem. On the one hand, more and more answers to questions over a wide range of domains are becoming available on forums; on the other hand, it is becoming harder and harder to extract and access relevant information due to the sheer scale and diversity of the data.

One potential way to enhance information access and support sharing in forums is to improve information retrieval (IR) effectiveness over forum threads. To this end, Elsas [1] amassed a forum dataset for forum thread retrieval and

Wang, Li, Su Nam Kim and Timothy Baldwin (to appear) The Utility of Discourse Structure in Forum Thread Retrieval, In Proceedings of the Ninth Asian Information Retrieval Societies Conference (AIRS 2013), Singapore.

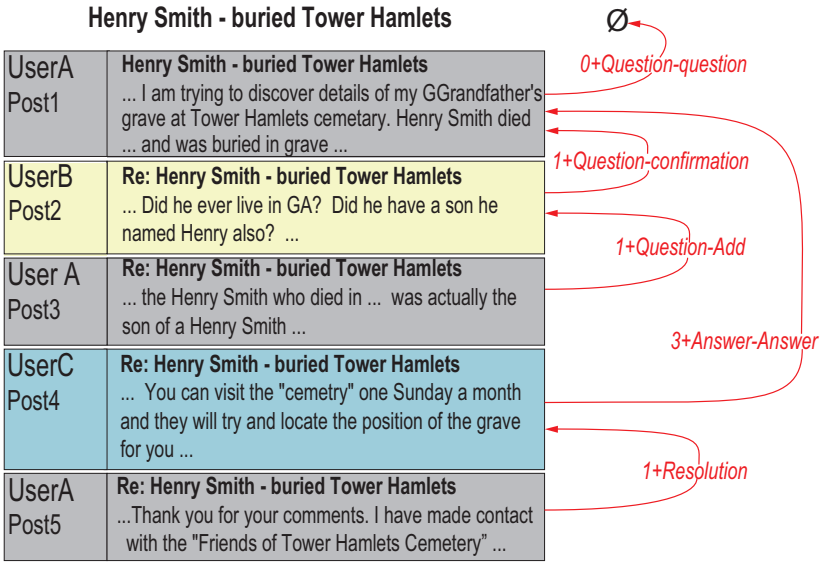

Fig. 1. A snippeted and annotated `Ancestry` thread.

conducted initial experiments. We build on this earlier work, in exploring the hypothesis that incorporating thread discourse structure [2, 3] into the IR model can improve retrieval effectiveness.

The discourse structure of a thread is modelled as a rooted directed acyclic graph (DAG), with the posts in the thread represented as nodes in the DAG. The reply-to relations between posts take the form of directed edges (Links) between nodes in the DAG, and dialogue acts (DAs) are used to label the edges. For the purposes of illustration, we use an annotated example thread from Elsas' `Ancestry` dataset [1], made up of 5 posts from 3 distinct participants, as shown in Fig. 1. In this example, UserA initiates the thread with a question (DA = Question-question) in the first post, seeking information about his/her great-grandfather. In response, UserB asks for more details about the question (DA = Question-confirmation). Then UserA responds to UserB to add extra information to his/her original question (DA = Question-add). Finally, UserC proposes a solution to the original question (DA = Answer-answer), and UserA confirms that this answer is correct (DA = Resolution). It should be noted that the discourse structure of most threads actually takes the form of a tree, as shown in Fig. 1. However, in some rare cases, a given post can reply to two or more previous posts, producing a DAG structure.

Specifically in this paper, we automatically infer the thread discourse structure of a target forum dataset by using a discourse parser that is trained over out-of-domain annotated data. We then incorporate information derived from this thread discourse structure into a state-of-the-art IR model for forum retrieval,

and find that thread discourse structure can, indeed, benefit thread retrieval. We also investigate the reason behind the improvements.

2 Related Work

As far as we are aware, there has been very little IR work that is specifically targeted at web user forum data. The most closely-related work is that performed by Elsas [1], on which this work is directly based; we describe the relevant details of Elsas' work later in this paper. Other closely-related work was done by Seo et al. [4], in improving thread retrieval by automatically inferring thread structure and incorporating it into the retrieval model. They explore different thread document representations, such as at the thread-level (i.e. concatenate all the posts in a thread into a single document), pair-level (i.e. treat each pair of posts as a document), dialogue-level (i.e. treat each sub-thread in a thread as a document), and combinations of these. They show that using the linking structure of threads boosts thread retrieval effectiveness. Elsas and Carbonell [5] conducted preliminary research on thread retrieval and also showed that thread structure is useful in thread ranking. Additionally, they found that message/post selection can contribute to thread retrieval.

Research on thread discourse structure analysis and classification over user forums has gained in momentum in recent years. Fortuna et al. [6] defined 5 post-level dialogue acts to describe the levels of agreement (i.e. agreement, disagreement, insult) and identify questions and answers (i.e. question and answer) in forum posts. Xi et al. [7] defined 5 prevalent types of post-level dialogue acts in forum threads. This set of dialogue acts was then adapted and extended by us in earlier work [2] to describe possible types of posts in troubleshooting-oriented online forums. Specifically, we devised a post-level dialogue act set and annotated a set of threads from `forums.cnet.com`. In this work, we proposed a set of novel features, which they applied to the separate tasks of post link classification and dialogue act classification. We later applied the same basic methodology to dialogue act classification over one-on-one live chat data with provided message dependencies [8], demonstrating the generalisability of the original method. In both cases, however, we tackled only a single task, either link classification (optionally given dialogue act tags) or dialogue act classification, but never the two together.

In later work, we delved into the task of thread discourse structure parsing further [3]. We used the same features as [2], but different parsing approaches. Specifically, we approached thread discourse structure parsing as a joint link and dialogue act classification task, using conditional random fields [9] and dependency parsing [10]. We also demonstrated that our discourse structure parsing method was able to perform equally well over partial threads as complete threads, by experimenting with "in situ" classification of evolving threads.

There has also been research focusing on particular types of dialogue acts, such as question–answer pairs in emails [11] and forum threads [12], question–context–answer in forum threads [12], initiation–response pairs (e.g. question–

answer, assessment–agreement, and blame–denial) in forum threads [13], as well as request and commitment in emails [14].

Thread discourse structure can be used to facilitate different tasks in web user forums. For example, we demonstrated that the information extracted from thread discourse structure can be used to improve Solvedness (i.e. whether the problem presented in the thread is solved or not) classification of forum threads [15]. Additionally, threading information has been shown to enhance retrieval effectiveness for post-level retrieval [7, 4], thread-level retrieval [4, 5], sentence-level shallow information extraction [16], and near-duplicate thread detection [17]. Moreover, Wang and Rose [13] demonstrated that initiation–response pairs (e.g. question–answer, assessment–agreement, and blame–denial) from online forums have the potential to enhance thread summarisation and automatically generate knowledge bases for Community Question Answering (cQA) services such as Yahoo! Answers. Furthermore, Kim et al. [18] showed that dialogue acts can be used to classify student online discussions in web-enhanced courses. Specifically, they use dialogue acts to identify discussion threads that may have unanswered questions and need the attention of an instructor.

3 Dataset Description

3.1 The `Ancestry` Forum Dataset

The Ancestry.com Forum Dataset (`Ancestry`) was created by Jonathan Elsas and Ancestry.com, a website which supports historical genealogical research. The `Ancestry` dataset contains a full snapshot of the Ancestry.com online forum (`boards.ancestry.com`) from December 1995 to July 2010. The dataset includes 22,054,728 posts spanning 9,040,958 threads, from 165,358 sub-forums. The total number of unique users is 3,775,670. The `Ancestry` dataset is presented at the post-level, and information associated with each post includes: post identifiers, the subforum name, thread identifier, author name/identifier, timestamp (at the day level), URL of the original post, post title and post body. The inter-post link structure of each thread, in terms of the reply-to structure generated by users when posting to the thread, are also provided.

The `Ancestry` dataset also comes with a selected set of 191 queries from Ancestry.com's query log, and pairwise preference relevance judgements for each query over the Ancestry.com forum data.

To create the pairwise preference relevance judgements annotation, a document pool is simulated as the first step. Firstly, `Indri` (`lemurproject.org`), `Terrier` (`terrier.org`), `Zettair` (`www.seg.rmit.edu.au/zettair`) and Ancestry.com's ranked boolean system are applied over the whole dataset to produce post rankings, with each ranking containing 1000 posts. Then, three aggregation methods, namely `Mean`, `Max` and Pseudo-Cluster Selection (`PCS`) [19], are used to convert each post ranking to a thread ranking. Lastly, the document pool is created by combining the top 100 threads of each thread ranking. The document pool contains 374 unique threads per query on average.

Relevance assessment is conducted by Ancestry.com, by collecting document-pair preferences [20]. This approach presents side-by-side document pairs (L, R) and collects judgements: L is preferred to R, R is preferred to L, L and R are duplicates, L is bad or R is bad. During the assessment process, a document pair selection algorithm, which is described in detail in [1], is used to reduce the number of assessments.

Out of the 191 queries, 50 queries were first selected for a pilot assessment, with each query annotated by two assessors. The results of the pilot assessment were analysed and used as a guide to set the parameters of the document pair selection algorithm, as well as adjust assessor training and assessment guidelines. Then, each of the remaining 141 queries was assessed by one assessor, with the adjusted parameters of the pair selection algorithm.

3.2 The CNET Forum Dataset

The CNET forum dataset of Kim et al. [2][1] contains 1332 annotated posts spanning 315 threads, collected from the Operating System, Software, Hardware and Web Development sub-forums of CNET.[2] Each post is labelled with one or more links (including the possibility of null-links, where the post doesn't link to any other post), and each link is labelled with a dialogue act. The dialogue act set is made up of 5 super-categories: Question, Answer, Resolution (confirmation of the question being resolved), Reproduction (external confirmation of a proposed solution working) and Other. The Question category contains 4 sub-classes: question, add, confirmation and correction. Similarly, the Answer category contains 5 sub-classes: answer, add, confirmation, correction and objection. For example, the label Question-add signifies the Question superclass and add subclass, i.e. addition of extra information to a question.

3.3 The ILIAD Forum Dataset

The ILIAD (Improved Linux Information Access by Data Mining) dataset [21] contains 1158 posts spanning 250 threads, collected from Linuxquestions[3] and Debian mailing lists.[4] We hand-annotated the discourse structure of the ILIAD dataset [15], based on a slightly modified version of the dialogue act set from our earlier work [2]. As part of this annotation, we proposed an additional Question-information dialogue act, for posts which provide information in non-troubleshooting threads. We also slightly adjusted the definition of the Resolution dialogue act. For full details of the ILIAD dataset and the annotations over it, see [21] and [15], respectively.

[1] Available from http://www.csse.unimelb.edu.au/research/lt/resources/conll2010-thread/

[2] http://forums.cnet.com/

[3] http://www.linuxquestions.org

[4] http://lists.debian.org/completeindex.html

IR Systems	
System	Configuration Used
`Indri`	Bag-of-words (`BoW`) queries
`Indri`	Dependence Model (`DM`) queries [23], with suggested model weights
`Indri`	Fielded query with linear combination
`Indri`	Fielded query with loglinear combination
`Terrier`	*PL2* with default parameters
`Terrier`	*InL2* with default parameters
`Zettair`	Default Okapi BM25 ranking algorithm
`Ancestry.com`	The ranked boolean system used by Ancestry.com
Aggregation Methods	
Name	Description
Mean	Thread score is the mean of retrieved posts' scores
Max	Thread score is the max score of the retrieved posts
Pseudo-Cluster Selection (`PCS`)[19]	Thread score is the geometric mean of the top-k retrieved posts' scores ($k = 5$ is used)

Table 1. Summary of Elsas' [1] experimental setup.

4 Pairwise Preference Evaluation

As explained in Section 3.1, the relevance judgements in the `Ancestry` dataset are pairwise preferences, rather than traditional absolute preferences (judgements). As analogues to absolute evaluation measures such as Precision at a cutoff ($P@k$) and Average Precision (AP), Elsas [1] uses Precision of Preferences at a cutoff ($ppref@k$) and a modified version of Average Precision of Preferences ($mAPpref$), which was originally proposed by Carterette [22]. $ppref@k$ represents the proportion of correctly ordered preferences to ordered preferences, where at least one document/thread in the pair is ranked above k. $mAPpref$ is the average of $ppref$ values over the ranks (i.e. k) of all documents which have ever been preferred to any other documents. While $ppref$ used by Elsas [1] is unchanged, the original $APpref$ proposed by Carterette [22] is the average of $ppref$ values over the ranks (i.e. k) at which the recall of preferences ($rpref$) increases. $rpref$ is the proportion of correctly ordered preferences to the total number of preferences made by assessors.

For comparability, the primary evaluation metrics used in this paper are $ppref@10$ and $mAPpref$, based on the evaluation script provided by Elsas [1].[5]

5 Baseline Systems

Elsas [1] conducted a series of IR experiments over the `Ancestry` dataset, using 4 retrieval systems with various configurations. The retrieval was done at the post-level, and 3 different aggregation methods were used to convert the post-level

[5] Available at `https://github.com/jelsas/Pairwise-Preference-Evaluation`

System	Aggregation Method	*mAPpref*		*ppref*@10	
		Orginal	Reproduced	Orginal	Reproduced
Indri-BoW	Mean	.542	.533	.492	.501
	Max	.599	.591	.561	.556
	PCS	.656	.650	.640	.633
Indri-DM	Mean	.549	.536	.506	.510
	Max	.608	.597	.571	.568
	PCS	**.661**	**.657**	**.646**	**.664**

Table 2. Elsas' [1] IR results (Original) and our reproduced results (Reproduced) over the Ancestry dataset. Retrieval is performed at the post-level, and evaluation is conducted at the thread-level. Three aggregation methods are used for each system to transform post-level scores to thread-level scores. The best results for each column are **bold-faced**.

rankings to thread-level rankings. A summary of the retrieval systems with the configurations used, as well as the aggregation methods, is presented in Table 1.

According to the experiments of Elsas [1], Indri with bag-of-words (BoW) and dependence model (DM: [23]) query formulation perform the best; our experiments support this conclusion. The DM used is a full dependency variant of a Markov Random Field, which assumes that all query terms are in some way dependent on each other. It considers the BoW representation (with weight 0.8) of the whole query, as well as ordered representation (with weight 0.1) and unordered representation (with weight 0.1) of the subsets of the query.

We tried to reproduce the results presented in [1] using Indri-BoW and Indri-DM for post-level retrieval with three different aggregation methods: Mean, Max and Pseudo-Cluster Selection (PCS). Our experimental results are displayed alongside the results reported in [1] in Table 2. Although there are slight differences between our results and Elsas' [1] results, the overall results are comparable. Because Indri-DM with PCS (Indri-DM-PCS) obtains the best results for both *mAPpref* and *ppref*@10, it will be used as our baseline IR method.

Following the work of Seo et al. [4], we also experimented with retrieval based on contexts of differing size, such as the thread-level, pair-level, dialogue-level, and various combinations of these. None of these experiments resulted in better results than the Indri-DM-PCS baseline, and the results are omitted from the paper.

6 Discourse Structure Parsing for Thread Retrieval

It is not practical for us to manually annotate the discourse structure of the whole Ancestry dataset nor just the portion of the dataset retrieved by the different IR systems. Rather, we opt to use automatically-predicted discourse structure. To build a discourse parser for Ancestry threads, we randomly selected and annotated 50 threads from the whole dataset to use for parameter tuning.

Train dataset setup	LD	Link	DA
Ancestry	.513	**.842**	.530
CNET	.359	.681	.435
ILIAD	.529	.801	**.569**
Ancestry+CNET	.427	.711	.501
Ancestry+ILIAD	**.539**	.827	**.569**
CNET+ILIAD	.406	.688	.478
Ancestry+CNET+ILIAD	.488	.730	.563

Table 3. Discourse structure parsing F-scores by applying CRFSGD with Initiator feature using the Combine approach over different training dataset setups. (The best result for each column is **bold-faced**.)

Discourse structure parsing, as discussed in [3], can be addressed in several ways. If a structured classification approach, such as a conditional random field (CRF), is used, we can either classify the links (Link) and dialogue act (DA) separately and compose them afterwards (denoted as Composition), or classify the combined Link and DA (e.g. treat 0+Question-question as a single label) directly (denoted as Combine). Another approach is to treat discourse parsing as a dependency parsing problem. Dependency parsing [24] is the task of automatically predicting the dependency structure of a token sequence, in the form of binary asymmetric dependency relations with dependency types. The joint classification task of Link and DA is a natural fit for dependency parsing, in that the task is intrinsically one of inferring labelled dependencies between posts.

For discourse parsing, we follow our earlier work [3]. All experiments were carried out based on stratified 10-fold cross-validation, stratifying at the thread level to ensure that all posts from a given thread occur in a single fold. Additionally, we augment the training data with the CNET and ILIAD datasets. The results are evaluated using post-level micro-averaged F-score ($\beta = 1$). All three discourse parsing methods described above were tested in our experiments, using CRFSGD [25] and MaltParser [10]. For features, we experimented with all the features proposed in our earlier work [3], as well as many of our own features. We found that using CRFSGD with a simple feature indicating whether a post's author is the initiator of the thread and the Combine approach achieves the highest Link and DA joint (LD) F-scores, as shown in Table 3. Because the availability of annotated discourse structure data cannot always be assumed, we decided to use only out-of-domain data to train the discourse parsers. Therefore, only the configurations of CNET, ILIAD and CNET+ILIAD are used in later experiments.

7 Augment Thread Retrieval with Discourse Structure

The basic idea of using the discourse structure to enhance existing IR systems is to use either links (Links) or dialogue acts (DAs) to modify the document ranking. For example, in the framework of Pseudo-Cluster Selection (PCS), one could imagine that a retrieved Answer-answer (i.e. an independent answer to a

question) post should be weighted higher than Other posts (including irrelevant posts), and thus contribute more to the thread ranking score. Under this assumption, we examined all the correctly predicted instances from the parsers described in Section 6 over our `Ancestry` development set, and found that the correctly predicted set only contains 5 dialogue acts, namely: Question-question (Qq), Question-add (Qadd), Answer-answer (Aa), Answer-add (Aadd), and Resolution (Res). Therefore, only predictions for these 5 dialogue acts are considered. Build on the `Indri-DM-PCS` system, our system (`Indri-DM-LD`) modifies the post-level rankings based on the predicted DA types of the posts. If a post's predicted DA type belongs to the selected DA subset (`DASubset`), it is considered to be more important than other posts and its score is increased/promoted by a certain factor. In addition to the 5 dialogue acts (`DAs+ALL`), we experimented with omitting one DA at a time (e.g. `DAs`–Qq = the five DAs minus Question-question predictions), to gauge the impact of each DA on the overall results.

Furthermore, in the model of `PCS`, one crucial parameter is the k which governs the number of retrieved posts that are used to calculate the thread-level ranking scores. Because of the potential interaction between this parameter k and our DA promotion model `Indri-DM-LD`, we also examined the effect of k in the baseline system `Indri-DM-PCS` as well as in our system `Indri-DM-LD`. We found that while $k = 5$ produces the best results for `Indri-DM-PCS`, $k = 4$ is the best setting for our `Indri-DM-LD` system. All experimental results reported in this paper are based on these respective k settings.

Table 4 presents the *mAPpref* / *ppref*@10 results for our `Indri-DM-LD` system with different `DASubset` configurations and promotion factors (i.e. 30%–70%). We test for statistical significance over the `Indri-DM-PCS` baseline with the two-tailed t-test ($p < 0.05$).

From Table 4 we can see that our system outperforms the `Indri-DM-PCS` baseline system (*mAPpref* = .657 and *ppref*@10 = .664) in most cases, demonstrating the superiority of our method. Our best results (*mAPpref* = .674 and *ppref*@10 = .678) are achieved using the combined `CNET` and `ILIAD` datasets for discourse parser training, the `DASubset` of `DAs`–Qq, and a `DA` promotion factor of 50%. The intuition behind Question-question posts not warranting promotion is that they contain question and not answer data, and are less likely to contain information relevant to the resolution of a query. It is important to reinforce that the discourse structure information used in these experiments was derived automatically based on out-of-domain data.

To investigate the mechanics behind our system, we conducted error analysis over `Indri-DM-PCS` vs. `Indri-DM-LD`. In one case, there are two threads, namely Thread1 and Thread2, which relate to Query 38 (*jacob lazarus; great synagogue, dukes place, london*). In the gold-standard annotation, Thread1 is preferred to Thread2. The posts retrieved by `Indri-DM` system are posts 3, 4 and 9 for Thread1 and posts 2, 7 and 12 for Thread2. Under the `Indri-DM-PCS` baseline system, Thread2 is ranked higher than Thread1. However, with `Indri-DM-LD` and `DAs`–Qq, the correct ordering of Thread1 and Thread2 is predicted, as the DA of post 12 in Thread2 is Question-question while the DA of all other posts is

DA training	DASubset	*mAPpref*					*ppref*@10				
		30%	40%	50%	60%	70%	30%	40%	50%	60%	70%
CNET	DAs +ALL	**.667**	**.668**	**.668**	**.669**	**.670**	.668	.673	.672	.664	.664
	–Qq	**.670**	**.673**	**.673**	**.674**	**.674**	**.674**	.673	**.678**	.671	.666
	–Qadd	**.667**	**.669**	**.670**	**.670**	**.671**	.667	.673	.673	.665	.666
	–Aa	.656	.655	.654	.654	.654	.660	.659	.658	.660	.657
	–Aadd	**.667**	**.668**	**.668**	**.669**	**.670**	.668	.673	.671	.664	.664
	–Res	**.666**	**.667**	**.667**	**.668**	**.670**	.666	.669	.669	.661	.661
ILIAD	DAs +ALL	**.666**	**.668**	**.668**	**.669**	**.669**	.668	.673	.671	.664	.664
	–Qq	**.670**	**.673**	**.673**	**.674**	**.674**	.672	.673	.673	.666	.666
	–Qadd	**.667**	**.668**	**.669**	**.671**	**.671**	.666	.668	.671	.667	.668
	–Aa	**.666**	**.666**	**.667**	**.667**	**.668**	.670	.669	.669	.668	.668
	–Aadd	.661	.661	.661	.659	.658	.660	.661	.660	.657	.657
	–Res	**.666**	**.668**	**.668**	**.669**	**.669**	.668	.673	.672	.663	.663
CNET+ILIAD	DAs +ALL	**.667**	**.668**	**.668**	**.669**	**.670**	.669	.673	.672	.663	.665
	–Qq	**.670**	**.673**	**.674**	**.674**	**.674**	**.674**	.673	**.678**	.671	.671
	–Qadd	**.667**	**.669**	**.669**	**.670**	**.671**	.664	.669	.672	.668	.663
	–Aa	.657	.655	.655	.654	.654	.661	.659	.658	.660	.657
	–Aadd	**.667**	**.668**	**.668**	**.669**	**.670**	.669	.673	.671	.663	.664
	–Res	**.666**	**.668**	**.667**	**.669**	**.670**	.667	.671	.669	.662	.663

Table 4. The *mAPpref*/*ppref*@10 scores from Indri-DM-LD when training the discourse parser over different training data sets (CNET, ILIAD or CNET+ILIAD), and with different promotion factors for the selected DAs; **boldface** signifies a better result than the Indri-DM-PCS baseline at a level of statistical significance ($p < 0.05$).

in DAs–Qq. As a consequence, the relative promotion of Thread1 is greater than Thread2, and the correct ranking is derived.

During our experiments, we demonstrated that making use of discourse structure of forum threads can boost retrieval effectiveness. As an alternative to full discourse parsing, we experimented with simply promoting all non-first posts (under the assumption that first posts are most likely to be Question-question posts). The best results achieved for this simple method are $mAPpref = .667$ and $ppref@10 = .670$. Although the *mAPpref* score is significantly better than the baseline, the *ppref*@10 is not (and both results are slightly below the best results achieved with discourse parsing, of $mAPpref = .674$ and $ppref@10 = .678$). Nevertheless it shows the potential of using a lighter-weight version of discourse structure to improve IR effectiveness. We will explore this line of research further in future work.

8 Conclusion

In this research, we have explored the hypothesis that IR over forum threads can be improved by incorporating thread discourse structure in the form of a rooted DAG over posts, with edges labelled with dialogue acts. When compared to

previous research conducted over the `Ancestry` dataset, we achieved significantly better results using automatically-predicted thread discourse structure.

In future work, we plan to firstly investigate more ways to capture thread discourse structure information. Furthermore, we intend to look into means of exploiting the structural information of threads for the purpose of IR, and their interaction with thread discourse structure. For example, the same dialogue act of Answer-answer may contribute to the thread ranking differently if it appears at different positions in a thread (e.g. second post vs. last post).

Acknowledgements

NICTA is funded by the Australian government as represented by Department of Broadband, Communication and Digital Economy, and the Australian Research Council through the ICT centre of Excellence programme.

References

1. Elsas, J.: The Ancestry.com forum dataset. available at `http://www.cs.cmu.edu/~jelsas/data/ancestry.com/Ancestry_TR.pdf` (2011) CMU LTI Tech Report CMU-LTI-017.
2. Kim, S.N., Wang, L., Baldwin, T.: Tagging and linking web forum posts. In: Proceedings of the 14th Conference on Computational Natural Language Learning (CoNLL-2010), Uppsala, Sweden (2010) 192–202
3. Wang, L., Lui, M., Kim, S.N., Nivre, J., Baldwin, T.: Predicting thread discourse structure over technical web forums. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK (2011) 13–25
4. Seo, J., Croft, W.B., Smith, D.A.: Online community search using thread structure. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009), Hong Kong, China (2009) 1907–1910
5. Elsas, J.L., Carbonell, J.G.: It pays to be picky: An evaluation of thread retrieval in online forums. In: Proceedings of 32nd International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09), Boston, USA (2009) 714–715
6. Fortuna, B., Rodrigues, E.M., Milic-Frayling, N.: Improving the classification of newsgroup messages through social network analysis. In: Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM 2007), Lisbon, Portugal (2007) 877–880
7. Xi, W., Lind, J., Brill, E.: Learning effective ranking functions for newsgroup search. In: Proceedings of 27th International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004), Sheffield, UK (2004) 394–401
8. Kim, S.N., Cavedon, L., Baldwin, T.: Classifying dialogue acts in one-on-one live chats. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010), Boston, USA (2010) 862–871
9. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, Williamstown, USA (2001) 282–289

10. Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., Marsi, E.: MaltParser: A language-independent system for data-driven dependency parsing. Natural Language Engineering **13**(02) (2007) 95–135
11. Shrestha, L., McKeown, K.: Detection of question-answer pairs in email conversations. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), Geneva, Switzerland (2004) 889–895
12. Cong, G., Wang, L., Lin, C.Y., Song, Y.I., Sun, Y.: Finding question-answer pairs from online forums. In: Proceedings of 31st International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08), Singapore (2008) 467–474
13. Wang, Y.C., Rosé, C.P.: Making conversational structure explicit: identification of initiation-response pairs within online discussions. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010). (2010) 673–676
14. Lampert, A., Dale, R., Paris, C.: Detecting emails containing requests for action. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010), Los Angeles, California (2010) 984–992
15. Wang, L., Kim, S.N., Baldwin, T.: The utility of discourse structure in identifying resolved threads in technical user forums. In: Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), Mumbai, India (2012) 2739 – 2756
16. Sondhi, P., Gupta, M., Zhai, C., Hockenmaier, J.: Shallow information extraction from medical forum data. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Posters Volume, Beijing, China (2010) 1158–1166
17. Muthmann, K., Barczyński, W.M., Brauer, F., Löser, A.: Near-duplicate detection for web-forums. In: Proceedings of the 2009 International Database Engineering & Applications Symposium (IDEAS 2009), Cetraro, Italy (2009) 142–151
18. Kim, J., Chern, G., Feng, D., Shaw, E., Hovy, E.: Mining and assessing discussions on the web through speech act analysis. In: Proceedings of the ISWC'06 Workshop on Web Content Mining with Human Language Technologies, Athens, USA (2006)
19. Seo, J., Croft, W.B., Smith, D.A.: Online community search using conversational structures. Information Retrieval **14**(6) (2011) 547–571
20. Carterette, B., Bennett, P.N., Chickering, D.M., Dumais, S.T.: Here or there: Preference judgments for relevance. In: Proceedings of the IR research, 30th European conference on Advances in information retrieval (ECIR'08), Glasgow, UK (2008) 16–27
21. Baldwin, T., Martinez, D., Penman, R.B.: Automatic thread classification for Linux user forum information access. In: Proceedings of the 12th Australasian Document Computing Symposium (ADCS 2007), Melbourne, Australia (2007) 72–79
22. Carterette, B., Bennett, P.N.: Evaluation measures for preference judgments. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08), Singapore (2008) 685–686
23. Metzler, D., Croft, W.B.: A markov random field model for term dependencies. In: +SIGIR2005, Salvador, Brazil (2005) 472–479
24. Kübler, S., McDonald, R., Nivre, J.: Dependency parsing. Synthesis Lectures on Human Language Technologies **2**(1) (2009) 1–127
25. Bottou, L.: CRFSGD software. `http://leon.bottou.org/projects/sgd` (2011)