Tagging and Linking Web Forum Posts

Su Nam Kim, Li Wang and **Timothy Baldwin**
Dept of Computer Science and Software Engineering
University of Melbourne, Australia
sunamkim@gmail.com, li.wang.d@gmail.com, tb@ldwin.net

Abstract

We propose a method for annotating post-to-post discourse structure in online user forum data, in the hopes of improving troubleshooting-oriented information access. We introduce the tasks of: (1) post classification, based on a novel dialogue act tag set; and (2) link classification. We also introduce three feature sets (structural features, post context features and semantic features) and experiment with three discriminative learners (maximum entropy, SVM-HMM and CRF). We achieve above-baseline results for both dialogue act and link classification, with interesting divergences in which feature sets perform well over the two sub-tasks, and go on to perform preliminary investigation of the interaction between post tagging and linking.

1 Introduction

With the advent of Web 2.0, there has been an explosion of web authorship from individuals of all walks of life. Notably, social networks, blogs and web user forums have entered the mainstream of modern-day society, creating both new opportunities and challenges for organisations seeking to engage with clients or users of any description. One area of particular interest is web-based user support, e.g. to aid a user in purchasing a gift for a friend, or advising a customer on how to configure a newly-acquired wireless router. While such interactions traditionally took place on an individual basis, leading to considerable redundancy for frequently-arising requests or problems, user forums support near-real-time user interaction in the form of a targeted thread made up of individual user posts. Additionally, they have the potential for perpetual logging to allow other users to benefit from them. This in turn facilitates "support sharing"—i.e. the ability for users to look over the logs of past support interactions to determine whether there is a documented, immediately-applicable solution to their current problem—on a scale previously unimaginable. This research is targeted at this task of enhanced support sharing, in the form of text mining over troubleshooting-oriented web user forum data (Baldwin et al., to appear).

One facet of our proposed strategy for enhancing information access to troubleshooting-oriented web user forum data is to preprocess threads to uncover the "content structure" of the thread, in the form of its post-to-post discourse structure. Specifically, we identify which earlier post(s) a given post responds to (linking) and in what manner (tagging), in an amalgam of dialogue act tagging (Stolcke et al., 2000) and coherence-based discourse analysis (Carlson et al., 2001; Wolf and Gibson, 2005). The reason we do this is gauge the relative role/import of individual posts, to index and weight component terms accordingly, ultimately in an attempt to enhance information access. Evidence to suggest that this structure can enhance information retrieval effectiveness comes from Xi et al. (2004) and Seo et al. (2009) (see Section 2).

To illustrate the task, consider the thread from the CNET forum shown in Figure 1, made up of 5 posts (*Post 1, ..., Post 5*) with 4 distinct participants (*A, B, C, D*). In the first post, *A* initiates the thread by requesting assistance in creating a web form. In response, *B* proposes a Javascript-based solution (i.e. responds to the first post with a proposed solution), and *C* proposes an independent solution based on .NET (i.e. also responds to the first post with a proposed solution). Next, *A* responds to *C*'s post asking for details of how to include this in a web page (i.e. responds to the third post asking for clarification), and in the final post, *D* proposes a different solution again (i.e. responds to the first post with a different solution again).

HTML Input Code - CNET Coding & scripting Forums

User *A* Post 1	**HTML Input Code** ... Please can someone tell me how to create an input box that asks the user to enter their ID, and then allows them to press go. It will then redirect to the page ...
User *B* Post 2	**Re: html input code** Part 1: create a form with a text field. See ... Part 2: give it a Javascript action ...
User *C* Post 3	**asp.net c# video** Ive prepared for you video.link click ...
User *A* Post 4	**Thank You!** Thanks a lot for that ... I have Microsoft Visual Studio 6, what program should I do this in? Lastly, how do I actually include this in my site?...
User *D* Post 5	**A little more help** ... You would simply do it this way: ... You could also just ... An example of this is:...

Figure 1: Snippeted posts in a CNET thread

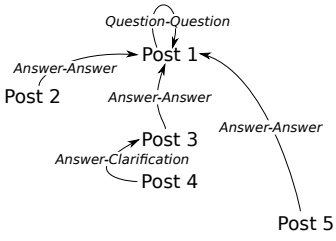

Figure 2: Post links and dialogue act labels for the example thread in Figure 1

In this, we therefore end up with a tree-based dependency link structure, with each post (other than the initial post) relating back to a unique preceding post via a range of link types, as indicated in Figure 2. Note, however, that more generally, it is possible for a post to link to multiple preceding posts (e.g. refuting one proposed solution, and proposing a different solution to the problem in the initial post).

Our primary contributions in this paper are: (1) a novel post label set for post structure in web forum data, and associated dataset; and (2) a series of results for post dependency linking and labelling, which achieve strong results for the respective tasks.

2 Related Work

Related work exists in the broad fields of dialogue processing, discourse analysis and information retrieval, and can be broken down into the following tasks: (1) dialogue act tagging; (2) discourse "disentanglement"; (3) community question answering; and (4) newsgroup/user forum search.

Dialogue act (DA) tagging is a means of capturing the function of a given utterance relative to an encompassing discourse, and has been proposed variously as a means of enhancing dialogue summarisation (Murray et al., 2006), and tracking commitments and promises in email (Cohen et al., 2004; Lampert et al., 2008), as well as being shown to improve speech recognition accuracy (Stolcke et al., 2000). A wide range of DA tag sets have been proposed, usually customised to a particular medium such as speech dialogue (Stolcke et al., 2000; Shriberg et al., 2004), task-focused email (Cohen et al., 2004; Wang et al., 2007; Lampert et al., 2008) or instant messaging (Ivanovic, 2008). The most immediately relevant DA-based work we are aware of is that of Xi et al. (2004), who proposed a 5-way classification for newsgroup data (including QUESTION and AGREEMENT/AMMENDMENT), but did not present any results based on the tagset.

A range of supervised models have been applied to DA classification, including graphical models (Ji and Bilmes, 2005), kernel methods (Wang et al., 2007), dependency networks (Carvalho and Cohen, 2005), transformation-based learning (Samuel et al., 1998), maxent models (Ang et al., 2005) and HMMs (Ivanovic, 2008). There is some contention about the import of context in DA classification, with the prevailing view being that context aids classification (Carvalho and Cohen, 2005; Ang et al., 2005; Ji and Bilmes, 2005), but also evidence to suggest that strictly local modelling is superior (Ries, 1999; Serafin and Di Eugenio, 2004).

In this work, we draw on existing work (esp. Xi et al. (2004)) in proposing a novel DA tag set customised to the analysis of troubleshooting-oriented web user forums (Section 3), and compare a range of text classification and structured classification methods for post-level DA classification.

Discourse disentanglement is the process of automatically identifying coherent sub-discourses in a single thread (in the context of user forums/mailing lists), chat session (in the context of IRC chat data: Elsner and Charniak (2008)), system interaction (in the context of HCI: Lemon et al. (2002)) or document (Wolf and Gibson, 2005). The exact definition of what constitutes a sub-discourse varies across domains, but for our purposes, entails an attempt to resolve the informa-

tion need of the initiator by a particular approach; if there are competing approaches proposed in a single thread, multiple sub-discourses will necessarily arise. The data structure used to represent the disentangled discourse varies from a simple connected sub-graph (Elsner and Charniak, 2008), to a stack/tree (Grosz and Sidner, 1986; Lemon et al., 2002; Seo et al., 2009), to a full directed acyclic graph (DAG: Rosé et al. (1995), Wolf and Gibson (2005), Schuth et al. (2007)). Disentanglement has been carried out via analysis of direct citation/user name references (Schuth et al., 2007; Seo et al., 2009), topic modelling (Lin et al., 2009), and clustering over content-based features for pairs of posts, optionally incorporating various constraints on post recency (Elsner and Charniak, 2008; Wang et al., 2008; Seo et al., 2009).

In this work, we follow Rosé et al. (1995) and Wolf and Gibson (2005) in adopting a DAG representation of discourse structure, and draw on the wide set of features used in discourse entanglement to model coherence.

Community question answering (cQA) is the task of identifying question–answer pairs in a given thread, e.g. for the purposes of thread summarisation (Shrestha and McKeown, 2004) or automated compilation of resources akin to Yahoo! Answers. cQA has been applied to both mailing list and user forum threads, conventionally based on question classification, followed by ranking of candidate answers relative to each question (Shrestha and McKeown, 2004; Ding et al., 2008; Cong et al., 2008; Cao et al., 2009). The task is somewhat peripheral to our work, but relevant in that it involves the implicit tagging of certain posts as containing questions/answers, as well as linking the posts together. Once again, we draw on the features used in cQA in this research.

There has been a spike of recent interest in **newsgroup/user forum search**. Xi et al. (2004) proposed a structured information retrieval (IR) model for newsgroup search, based on author features, thread structure (based on the tree defined by the reply-to structure), thread "topology" features and content-based features, and used a supervised ranking method to improve over a baseline IR system. Elsas and Carbonell (2009) — building on earlier work on blog search (Elsas et al., 2008) — proposed a probabilistic IR approach which ranks user forum threads relative to selected posts in the overall thread, and again demonstrated the superiority of this method over a model which ignores thread structure. Finally, Seo et al. (2009) automatically derived thread structure from user forum threads, and demonstrated that the IR effectiveness over the "threaded" structure was superior to that using a monolithic document representation.

The observations and results of Xi et al. (2004) and Seo et al. (2009) that threading information (or in our case "disentangled" DAG structure) enhances IR effectiveness is a core motivator for this research.

3 Post Label Set

Our post label set contains 12 categories, intended to capture the typical interactions that take place in troubleshooting-oriented threads on technical forums. There are 2 super-categories (QUESTION, ANSWER) and 3 singleton classes (RESOLUTION, REPRODUCTION, and OTHER). QUESTION, in turn, contains 4 sub-classes (QUESTION, ADD, CONFIRMATION, CORRECTION), while ANSWER contains 5 sub-classes (ANSWER, ADD, CONFIRMATION, CORRECTION, and OBJECTION), partially mirroring the sub-structure of QUESTION. We represent the amalgam of a super- and sub-class as QUESTION-ADD, for example.

All tags other than QUESTION-QUESTION and OTHER are relational, i.e. relate a given post to a unique earlier post. A given post can potentially be labelled with multiple tags (e.g. confirm details of a proposed solution, in addition to providing extra details of the problem), although, based on the strictly chronological ordering of posts in threads, a post can only link to posts earlier in the thread (and can also not cross thread boundaries). Additionally, the link structure is assumed to be transitive, in that if post A links to post B and post B to post C, post A is implicitly linked to post C. As such, an explicit link from post A to post C should exist only in the case that the link between them is not inferrable transitively.

Detailed definitions of each post tag are given below. Note that **initiator** refers to the user who started the thread with the first post.

QUESTION-QUESTION (Q-Q): the post contains a new question, independent of the thread context that precedes it. In general, QUESTION-QUESTION is reserved for the first post in a given thread.

QUESTION-ADD (Q-ADD): the post supple-

ments a question by providing additional information, or asking a follow-up question.

QUESTION-CONFIRMATION (Q-CONF): the post points out error(s) in a question without correcting them, or confirms details of the question.

QUESTION-CORRECTION (Q-CORR): the post corrects error(s) in a question.

ANSWER-ANSWER (A-A): the post proposes an answer to a question.

ANSWER-ADD (A-ADD): the post supplements an answer by providing additional information.

ANSWER-CONFIRMATION (A-CONF): the post points out error(s) in an answer without correcting them, or confirms details of the answer.

ANSWER-CORRECTION (A-CORR): the post corrects error(s) in an answer.

ANSWER-OBJECTION (A-OBJ): the post objects to an answer on experiential or theoretical grounds (e.g. *It won't work.*).

RESOLUTION (RES): the post confirms that an answer works, on the basis of implementing it.

REPRODUCTION (REP): the post either: (1) confirms that the same problem is being experienced (by a non-initiator, e.g. *I'm seeing the same thing.*); or (2) confirms that the answer should work.

OTHER (OTHER): the post does not belong to any of the above classes.

4 Feature Description

In this section, we describe our post feature representation, in the form of four feature types.

4.1 Lexical features

As our first feature type, we use simple lexical features, in the form of unigram and bigram tokens contained within a given post (without stopping). We also POS tagged and lemmatised the posts, postfixing the lemmatised token with its POS tag (using Lingua::EN::Tagger and `morpha` (Minnen et al., 2001)). Finally, we bin together the counts for each token, and represent it via its raw frequency.

4.2 Structural features

The identity of the post author, and position of the post within the thread, can be indicators of the post/link structure of a given post. We represent the post author as a simple binary feature indicating whether s/he is the thread initiator, and the post position via its relative position in the thread (as a ratio, relative to the total number of posts).

4.3 Post context features

As mentioned in Section 2, post context has generally (but not always) been shown to enhance the classification accuracy of DA tagging tasks, in the form of Markov features providing predicted post labels for previous posts, or more simply, post-to-post similarity. We experiment with a range of post context features, all of which are compatible with features both from the same label set as that being classified (e.g. link features for link classification), as well as features from a second label set (e.g. DA label features for link classification).

Previous Post: There is a strong prior for posts to link to their immediately preceding post (as observed for 79.9% of the data in our dataset), and also strong sequentiality in our post label set (e.g. a post following a Q-Q is most likely to be an A-A). As such, we represent the predicted post label of the immediately preceding post, as a first-order Markov feature, as well as a binary feature to indicate whether the author of the previous post also authored the current post.

Previous Post from Same Author: A given user tends to author posts of the same basic type (e.g. QUESTION or ANSWER) in a given thread, and pairings such as A-A and A-CONF from a given author are very rare. To capture this observation, we look to see if the author of the current post has posted earlier in the thread, and if so, include the label and relative location (in posts) of their most recent previous post.

Full History: As a final option, we include the predictions for all posts $P_1, ..., P_{i-1}$ preceding the current post P_i.

4.4 Semantic features

We tested four semantic features based on post content and title.

Title Similarity: For forums such as CNET which include titles for individual posts (as represented in Figure 1), a post having the same or similar title as a previous post is often a strong indicator that it responds to that post. This both provides a strong indicator of which post a given post responds (links) to, and can aid in DA tagging. We use simple cosine similarity to find the post with the most-similar title, and represent its relative location to the current post.

Post Similarity: Posts of the same general type tend to have similar content and be linked. For example, A-A and A-ADD posts tend to share content. We capture this by identifying the post with most-similar content based on cosine similarity, and represent its relative location to the current post.

Post Characteristics: We separately represent the number of question marks, exclamation marks and URLs in the current post. In general, question marks occur in QUESTION and CONFIRMATION posts, exclamation marks occur in RES and OBJECTION posts, and URLs occur in A-A and A-ADD posts.

User Profile: Some authors tend to answer questions more, while others tend to ask more questions. We capture the class priors for the author of the current post by the distribution of post labels in their posts in the training data.

5 Experimental Setup

As our dataset, we collected 320 threads containing a total of 1,332 posts from the Operating System, Software, Hardware, and Web Development sub-forums of CNET.[1]

The annotation of post labels and links was carried by two annotators in a custom-built web interface which supported multiple labels and links for a given post. For posts with multilabels, we used a modified version of Cohen's Kappa, which returned κ values of 0.59 and 0.78 for the post label and link annotations, respectively. Any disagreements in labelling were resolved through adjudication.

Of the 1332 posts, 65 posts have multiple labels (which possibly link to a common post) and 22 posts link to two different links. The majority post label in the dataset is A-A (40.30%).

We built machine learners using a conventional Maximum Entropy (ME) learner,[2] as well as two structural learners, namely: (1) SVM-HMMs (Joachims et al., 2009), as implemented in SVM-struct[3], with a linear kernel; and (2) conditional random fields (CRFs) using CRF++.[4] SVM-HMMs and CRFs have been successfully applied to a range of sequential tagging tasks such as syllabification (Bartlett et al., 2009), chunk parsing (Sha and Pereira, 2003) and word segmentation (Zhao et al., 2006). Both are discriminative models which capture structural dependencies, which is highly desirable in terms of modelling sequential preferences between post labels (e.g. A-CONF typically following a A-A). SVM-HMM has the additional advantage of scaling to large numbers of features (namely the lexical features). As such, we only experiment with lexical features for SVM-HMM and ME.

All of our evaluation is based on stratified 10-fold cross-validation, stratifying at the thread level to ensure that if a given post is contained in the test data for a given iteration, all other posts in that same thread are also in the test data (or more pertinently, not in the training data). We evaluate using micro-averaged precision, recall and F-score ($\beta = 1$). We test the statistical significance of all above-baseline results using randomised estimation ($p < 0.05$; Yeh (2000)), and present all such results in bold in our results tables.

In our experiments, we first look at the post classification task in isolation (i.e. we predict which labels to associate with each post, underspecifying which posts those labels relate to). We then move on to look at the link classification task, again in isolation (i.e. we predict which previous posts each post links to, underspecifying the nature of the link). Finally, we perform preliminary investigation of the joint task of DA and link classification, by incorporating DA class features into the link classification task.

6 DA Classification Results

Our first experiment is based on post-level dialogue act (DA) classification, ignoring link structure in the first instance. That is, we predict the labels on edges emanating from each post in the DAG representation of the post structure, without

[1] http://forums.cnet.com/?tag=TOCleftColumn.0

[2] http://maxent.sourceforge.net/

[3] http://www.cs.cornell.edu/People/tj/svm_light/svm_hmm.html

[4] http://crfpp.sourceforge.net/

Features	CRF	SVM-HMM	ME
Lexical	—	.566	.410
Structural	**.742**	.638	**.723**

Table 1: DA classification F-score with lexical and structural features (above-baseline results in **bold**)

Features	CRF	SVM-HMM	ME
Struct+R	**.740**	.640	.632
Struct+A	**.742**	**.676**	**.693**
Struct+F	**.744**	.641	.577
Struct+RA	.397	.636	**.665**
Struct+AF	.405	.642	.586

Table 2: DA classification F-score with structural and DA-based post context features (R = "Previous Post", A = "Previous Post from Same Author", and F = "Full History"; above-baseline results in **bold**)

specifying the edge destination. Returning to our example in Figure 2, e.g., the gold-standard classification for Post 1 would be Q-Q, Post 2 would be A-A, etc.

As a baseline for DA classification, simple majority voting attains an F-score of 0.403, based on the A-A class. A more realistic baseline, however, is a position-conditioned variant, where the first post is always classified as Q-Q, and all subsequent posts are classified as A-A, achieving an F-score of 0.641.

6.1 Lexical and structural features

First, we experiment with lexical and structural features (recalling that we are unable to scale the CRF model to full lexical features). Lexical features produce below-baseline performance, while simple structural features immediately lead to an improvement over the baseline for CRF and ME.

The reason for the poor performance with lexical features is that our dataset contains only around 1300 posts, each of which is less than 100 words in length on average. The models are simply unable to generalise over this small amount of data, and in the case of SVM-HMM, the presence of lexical features, if anything, appears to obscure the structured nature of the labelling task (i.e. the classifier is unable to learn the simple heuristic used by the modified majority class baseline).

The success of the structural features, on the other hand, points to the presence of predictable sequences of post labels in the data. That SVM-HMM is unable to achieve baseline performance with structural features is slightly troubling.

6.2 Post context features

Next, we test the two post context features: *Previous Post* (P) and *Previous Post from Same Author* (A). Given the success of structural features, we retain these in our experiments. Note that the labels used in the post context are those which are interactively learned by that model for the previous posts.

Table 2 presents the results for structural features combined with DA-based post context; we do not present any combinations of Previous Post and Full History, as Full History includes the Previous Post.

Comparing back to the original results using only the structural results, we can observe that Previous Post from Same Author and Full History (A and F, resp., in the table) lead to a slight increment in F-score for both CRF and SVM-HMM, but degrade the performance of ME. Previous Post leads to either a marginal improvement, or a drop in results, most noticeably for ME. It is slightly surprising that the CRF should benefit from context features at all, given that it is optimising over the full tag sequence, but the impact is relatively localised, and when all sets of context features are used, the combined weight of noisy features appears to swamp the learner, leading to a sharp degradation in F-score.

6.3 Semantic features

We next investigate the relative impact of the semantic features, once again including structural features in all experiments. Table 3 presents the F-score using the different combinations of semantic features.

Similarly to the post context features, the semantic features produced slight increments over the structural features in isolation, especially for CRF and ME. For the first time, SVM-HMM achieved above-baseline results, when incorporating title similarity and post characteristics. Of the individual semantic features, title and post similarity appear to be the best performers. Slightly disappointingly, the combination of semantic features generally led to a degradation in F-score, almost certainly due to data sparseness. The best overall result was achieved with CRF, incorporat-

Features	CRF	SVM-HMM	ME
Struct+T	**.751**	.636	**.660**
Struct+P	**.747**	.636	**.662**
Struct+C	**.738**	.587	.630
Struct+U	**.722**	.564	.620
Struct+TP	**.740**	.627	**.720**
Struct+TC	**.744**	**.646**	.589
Struct+TU	**.738**	.600	.609
Struct+PC	**.745**	.630	.583
Struct+PU	**.736**	.626	.605
Struct+CU	**.730**	.599	.619
Struct+TPC	**.739**	.622	.580
Struct+TPU	**.729**	.613	.6120
Struct+TCU	**.750**	.611	.6120
Struct+PCU	**.738**	.616	.614
Struct+TPCU	**.737**	.619	.605

Table 3: DA classification F-score with semantic features (T = "Title Similarity", P = "Post Similarity", C = "Post Characteristics", and U = "User Profile"; above-baseline results in **bold**)

Features	R	A	RA
Struct+T	**.649**	**.649**	**.649**
Struct+P	**.737**	**.736**	**.742**
Struct+C	**.741**	**.741**	**.742**
Struct+U	**.745**	**.742**	**.737**
Struct+TP	**.645**	**.656**	**.658**
Struct+TC	.383	.402	.408
Struct+TU	**.650**	**.652**	**.652**
Struct+PC	**.730**	**.743**	**.753**
Struct+PU	.232	.232	.286
Struct+CU	**.719**	.471	**.710**
Struct+TPC	.498	.469	.579
Struct+TPU	.248	.232	.248
Struct+TCU	.388	.377	.380
Struct+PCU	.231	.231	.261
Struct+TPCU	.231	.231	.231

Table 4: DA classification F-score for CRF with different combinations of post context features and semantic features (R = "Previous Post", and A = "Previous Post from Same Author"; T = "Title Similarity", P = "Post Similarity", C = "Post Characteristics", and U = "User Profile"; above-baseline results in **bold**)

ing structural features and title similarity, at an F-score of 0.751.

To further explore the interaction between post context and semantic features, we built CRF classifiers for different combinations of post context and semantic features, and present the results in Table 4.[5] We achieved moderate gains in F-score, with all post context features, in combination with structural features, post similarity and post characteristics achieving an F-score of 0.753, slightly higher than the best result achieved for just structural and post context features.

It is important to refer back to the results for lexical features (comparable to what would have been achieved with a standard text categorisation approach to the task), and observe that we have achieved far higher F-scores using features customised to user forum data. It is also important to reflect that post context (in terms of the features and the structured classification results of CRF) appears to markedly improve our results, contrasting with the results of Ries (1999) and Serafin and Di Eugenio (2004).

[5]We omit the results for Full History post context for reasons of space, but there is relatively little deviation from the numbers presented.

7 Link Classification Results

Our second experiment is based on link classification in isolation. Here, we predict unlabelled edges, e.g. in Figure 2, the gold-standard classification for Post 1 would be NULL, Post 2 would be Post 1, Post 3 would be Post 1, etc.

Note that the initial post cannot link to any other post, and also that the second post always links to the first post. As this is a hard constraint on the data, and these posts simply act to inflate the overall numbers, we exclude all first and second posts from our evaluation of link classification.

We experimented with a range of baselines as presented in Table 5, but found that the best performer by far was the simple heuristic of linking each post (except for the initial post) to its immediately preceding post. This leads to an F-score of 0.631, comparable to that for the post classification task.

7.1 Lexical and structural features

Once again, we started by exploring the effectiveness of lexical and structural features using the three learners, as detailed in Table 6.

Similarly to the results for post classification,

Baseline	Prec	Rec	F-score
Previous post	.641	.622	**.631**
First post	.278	.269	.274
Title similarity	.311	.301	.306
Post similarity	.255	.247	.251

Table 5: Baselines for link classification

Features	CRF	SVM-HMM	ME
Lexical	—	.154	.274
Structural	.446	.220	.478

Table 6: Link classification F-score with lexical and structural features (above-baseline results in **bold**)

structural features are more effective than lexical features for link classification, but this time, neither feature set approaches the baseline F-score for any of the learners. Once again, the results for SVM-HMM are well below those for the other two learners.

7.2 Post context features

Next, we experiment with link-based post context features, in combination with the structural features, as the results were found to be consistently better when combined with the structural features (despite the below-baseline performance of the structural features in this case). The link-based post context features in all cases are generated using the CRF with structural features from Table 6. As before, we do not present any combinations of Previous Post and Full History, as Full History includes the Previous Post

As seen in Table 9, here, for the first time, we achieve an above-baseline result for link classification, for SVM and ME based on Previous Post from Same Author in isolation, and also sometimes in combination with the other feature sets. The results for CRF also improve, but not to a level of statistical significance over the baseline. Similarly to the results for DA classification, the results for CRF drop appreciably when we combine feature sets.

7.3 Semantic features

Finally, we experiment with semantic features, once again in combination with structural features. The results are presented in Table 8.

The results for semantic features largely mirror those for post classification: small improvements are observed for title similarity with CRF, but otherwise, the results degrade across the board, and the combination of different feature sets compounds this effect.

Features	CRF	SVM-HMM	ME
Struct+R	.234	.605	.618
Struct+A	.365	**.665**	**.665**
Struct+F	.624	**.648**	.615
Struct+RA	.230	.615	**.661**
Struct+AF	.359	**.663**	.621

Table 7: Link classification F-score with structural and link-based post context features (R = "Previous Post", A = "Previous Post from Same Author", and F = "Full History"; above-baseline results in **bold**)

Features	CRF	SVM-HMM	ME
Struct+T	.464	.223	.477
Struct+P	.433	.198	.453
Struct+C	.438	.213	.419
Struct+U	.407	.160	.376
Struct+TP	.459	.194	.491
Struct+TC	.449	.229	.404
Struct+TU	.456	.174	.353
Struct+PC	.422	.152	.387
Struct+PU	.439	.166	.349
Struct+CU	.397	.178	.366
Struct+TPC	.449	.185	.418
Struct+TPU	.449	.160	.365
Struct+TCU	.459	.185	.358
Struct+PCU	.439	.161	.358
Struct+TPCU	.443	.163	.365

Table 8: Link classification F-score with semantic features (T = "Title Similarity", P = "Post Similarity", C = "Post Characteristics", and U = "User Profile"; above-baseline results in **bold**)

The best overall result achieved for link classification is thus the 0.743 for CRF with the structural and post context features.

We additionally experimented with combinations of features as for post classification, but were unable to improve on this result.

7.4 Link Classification using DA Features

Ultimately, we require both DA and link classification of each post, which is possible by combining the outputs of the component classifiers described

Features	CRF	SVM-HMM	ME
Struct+R	.586	.352	.430
Struct+A	.591	.278	.568
Struct+F	**.704**	.477	.546
Struct+RA	.637	.384	.551
Struct+AF	**.743**	.527	.603

Table 9: Link classification F-score with structural and post-based post context features (R = "Previous Post", A = "Previous Post from Same Author", and F = "Full History"; above-baseline results in **bold**)

above, by rolling the two tasks into a single classification task, or alternatively by looking to joint modelling methods. As a preliminary step in this direction, and means of exploring the interaction between the two tasks, we repeat the experiment based on post context features from above (see Section 7.2), but rather than using link-based post context, we use DA-based post context.

As can be seen in Table 9, the results for SVM-HMM and ME drop appreciably as compared to the results using link-based post context in Table 9, while the results for CRF jump to the highest level achieved for the task for all three learners. The effect can be ascribed to the ability of CRF to natively model the (bidirectional) link classification history in the process of performing structured learning, and the newly-introduced post features complementing the link classification task.

8 Discussion and Future Work

Ultimately, we require both DA and link classification of each post, which is possible in (at least) the following three ways: (1) by combining the outputs of the component classifiers described above; (2) by rolling the two tasks into a single classification task; or (3) by looking to joint modelling methods. Our results in Section 7.4 are suggestive of the empirical potential of performing the two tasks jointly, which we hope to explore in future work.

One puzzling effect observed in our experiments was the generally poor results for SVM. Error analysis indicates that the classifier was heavily biased towards the high-frequency classes, e.g. classifying all posts as either Q-Q or A-A for DA classification. The classifications for the other two learners were much more evenly spread across the different classes.

CRF was limited in that it was unable to capture lexical features, but ultimately, lexical features were found to be considerably less effective than structural and post context features for both tasks, and the ability of the CRF to optimise the post labelling over the full sequence of posts in a thread more than compensated for this shortcoming. Having said this, there is more work to be done exploring synergies between the different feature sets, especially for DA classification where all feature sets were found to produce above-baseline results.

Another possible direction for future research is to explore the impact of inter-post time on link structure, based on the observation that follow-up posts from the initiator tend to be temporally adjacent to posts they respond to with relatively short time intervals, while posts from non-initiators which are well spaced out tend not to respond to one another. Combining this with profiling of the cross-thread behaviour of individual forum participants (Weimer et al., 2007; Lui and Baldwin, 2009), and formal modelling of "forum behaviour" is also a promising line of research, taking the lead from the work of Götz et al. (2009), inter alia.

9 Conclusion

In this work, we have proposed a method for analysing post-to-post discourse structure in online user forum data, in the form of post linking and dialogue act tagging. We introduced three feature sets: structural features, post context features and semantic features. We experimented with three learners (maximum entropy, SVM-HMM and CRF), and established that CRF is the superior approach to the task, achieving above-baseline results for both post and link classification. We also demonstrated the complementarity of the proposed feature sets, especially for the post classification task, and carried out a preliminary exploration of the interaction between the linking and dialogue act tagging tasks.

Acknowledgements

This research was supported in part by funding from Microsoft Research Asia.

References

Jeremy Ang, Yang Liu, and Elizabeth Shriberg. 2005. Automatic dialog act segmentation and classifica-

tion in multiparty meetings. In *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, pages 1061–1064, Philadelphia, USA.

Timothy Baldwin, David Martinez, Richard Penman, Su Nam Kim, Marco Lui, Li Wang, and Andrew MacKinlay. to appear. Intelligent Linux information access by data mining: the ILIAD project. In *Proceedings of the NAACL 2010 Workshop on Computational Linguistics in a World of Social Media: #SocialMedia*, Los Angeles, USA.

Susan Bartlett, Grzegorz Kondrak, and Colin Cherry. 2009. On the syllabification of phonemes. In *Proceedings of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies 2009 (NAACL HLT 2009)*, pages 308–316, Boulder, USA.

Xin Cao, Gao Cong, Bin Cui, Christian S. Jensen, and Ce Zhang. 2009. The use of categorization information in language models for question retrieval. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 265–274, Hong Kong, China.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, pages 1–10, Aalborg, Denmark. Association for Computational Linguistics Morristown, NJ, USA.

Vitor R. Carvalho and William W. Cohen. 2005. On the collective classification of email "speech acts". In *Proceedings of 28th International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, pages 345–352.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to classify email into "speech acts". In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 309–316, Barcelona, Spain.

Gao Cong, Long Wang, Chin-Yew Lin, Young-In Song, and Yueheng Sun. 2008. Finding question-answer pairs from online forums. In *Proceedings of 31st International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*, pages 467–474, Singapore.

Shilin Ding, Gao Cong, Chin-Yew Lin, and Xiaoyan Zhu. 2008. Using conditional random fields to extract context and answers of questions from online forums. In *Proceedings of the 46th Annual Meeting of the ACL: HLT (ACL 2008)*, pages 710–718, Columbus, USA.

Jonathan L. Elsas and Jaime G. Carbonell. 2009. 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)*, pages 714–715, Boston, USA.

Jonathan L. Elsas, Jaime Arguello, Jamie Callan, and Jaime G. Carbonell. 2008. Retrieval and feedback models for blog feed search. In *Proceedings of 31st International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*, pages 347–354, Singapore.

Micha Elsner and Eugene Charniak. 2008. You talking to me? a corpus and algorithm for conversation disentanglement. In *Proceedings of the 46th Annual Meeting of the ACL: HLT (ACL 2008)*, pages 834–842, Columbus, USA.

Michaela Götz, Jure Leskovec, Mary McGlohon, and Christos Faloutsos. 2009. Modeling blog dynamics. In *Proceedings of the Third International Conference on Weblogs and Social Media (ICWSM 2009)*, pages 26–33, San Jose, USA.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intention and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Edward Ivanovic. 2008. Automatic instant messaging dialogue using statistical models and dialogue acts. Master's thesis, University of Melbourne.

Gang Ji and Jeff Bilmes. 2005. Dialog act tagging using graphical models. In *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, pages 33–36, Philadelphia, USA.

Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu. 2009. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59.

Andrew Lampert, Robert Dale, and Cécile Paris. 2008. The nature of requests and commitments in email messages. In *Proceedings of the AAAI 2008 Workshop on Enhanced Messaging*, pages 42–47, Chicago, USA.

Oliver Lemon, Alex Gruenstein, and Stanley Peters. 2002. Collaborative activities and multitasking in dialogue systems. *Traitement Automatique des Langues (TAL), Special Issue on Dialogue*, 43(2):131–154.

Chen Lin, Jiang-Ming Yang, Rui Cai, Xin-Jing Wang, Wei Wang, and Lei Zhang. 2009. Modeling semantics and structure of discussion threads. In *Proceedings of the 18th International Conference on the World Wide Web (WWW 2009)*, pages 1103–1104, Madrid, Spain.

Marco Lui and Timothy Baldwin. 2009. You are what you post: User-level features in threaded discourse. In *Proceedings of the Fourteenth Australasian Document Computing Symposium (ADCS 2009)*, Sydney, Australia.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore. 2006. Incorporating speaker and discourse features into speech summarization. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 367–374.

Klaus Ries. 1999. HMM and neural network based speech act detection. In *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-99)*, pages 497–500, Phoenix, USA.

Carolyn Penstein Rosé, Barbara Di Eugenio, Lori S. Levin, and Carol Van Ess-Dykema. 1995. Discourse processing of dialogues with multiple threads. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 31–38, Cambridge, USA.

Ken Samuel, Carbeery Sandra Carberry, and K. Vijay-Shanker. 1998. Dialogue act tagging with transformation-based learning. In *Proceedings of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98)*, pages 1150–1156, Montreal, Canada.

Anne Schuth, Maarten Marx, and Maarten de Rijke. 2007. Extracting the discussion structure in comments on news-articles. In *Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management*, pages 97–104, Lisboa, Portugal.

Jangwon Seo, W. Bruce Croft, and David A. Smith. 2009. Online community search using thread structure. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 1907–1910, Hong Kong, China.

Riccardo Serafin and Barbara Di Eugenio. 2004. FLSA: Extending latent semantic analysis with features for dialogue act classification. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pages 692–699, Barcelona, Spain.

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of the 3rd International Conference on Human Language Technology Research and 4th Annual Meeting of the NAACL (HLT-NAACL 2003)*, pages 213–220, Edmonton, Canada.

Lokesh Shrestha and Kathleen McKeown. 2004. Detection of question-answer pairs in email conversations. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 889–895, Geneva, Switzerland.

Elinzabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 97–100, Cambridge, USA.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Pail Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339–373.

Yi-Chia Wang, Mahesh Joshi, and Carolyn Rosé. 2007. A feature based approach to leveraging context for classifying newsgroup style discussion segments. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions (ACL 2007)*, pages 73–76, Prague, Czech Republic.

Yi-Chia Wang, Mahesh Joshi, William W. Cohen, and Carolyn Rosé. 2008. Recovering implicit thread structure in newsgroup style conversations. In *Proceedings of the Second International Conference on Weblogs and Social Media (ICWSM 2008)*, pages 152–160, Seattle, USA.

Markus Weimer, Iryna Gurevych, and Max Mühlhäuser. 2007. Automatically assessing the post quality in online discussions on software. In *Proceedings of the 45th Annual Meeting of the ACL: Interactive Poster and Demonstration Sessions*, pages 125–128, Prague, Czech Republic.

Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287.

Wensi Xi, Jesper Lind, and Eric Brill. 2004. Learning effective ranking functions for newsgroup search. In *Proceedings of 27th International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 394–401. Sheffield, UK.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 947–953, Saarbrücken, Germany.

Hai Zhao, Chang-Ning Huang, and Mu Li. 2006. An improved Chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 162–165. Sydney, Australia.