Automatic sub-post forum thread discourse structure analysis

A thesis presented
by

Liang Han (712397)

to
Master of Information Technology (Computing)
Research Project
COMP90055 (25 credits)

Supervised by
Prof Timothy Baldwin

The University of Melbourne
Melbourne, Australia
Jun 2016

Declaration

I certify that

(i) this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text.

(ii) where necessary I have received clearance for this research from the University's Ethics Committee (Approval Number: No) and have submitted all required data to the Department

(iii) the thesis is 6803 words in length (excluding text in images, table, bibliographies and appendices).

Signed: Liang Han   Date: 5 June 2016

Thesis advisor(s)
Timothy Baldwin

Author
Liang Han

Automatic sub-post forum thread discourse structure analysis

Abstract

Web forums are platforms for users to resolve information needs and seek answers for specific questions.However, complex discourse makes it harder for users to retrieve information. The sub post level discourse structure contains more discourse structure information than the higher level. In this thesis, we move beyond the common post level discourse structure and focus on the sentence level discourse analysis. We segment the CNET dataset from posts into sentences, use Conditional Random Fields to make prediction of the Link + Dialogue Act labels of sentences. We make comparison with the post level models and try to figure out whether sub post level discourse structure can resolve some of the issues in post level discourse structure. Finally, three methods, confusion matrix, correlation analysis and manual analysis, are used to perform the error analysis on the sub post level models.

Contents

Title Page i
Abstract iii
Table of Contents iv
Acknowledgments vi

1 Introduction 1
1.1 Background 1
1.2 Motivation 1
1.3 Contribution 4
1.4 Overview 4

2 Literature Review 5
2.1 Sentence boundary detection 5
2.2 Dialouge act tag set 6
2.3 Features extraction 6
2.4 Learners and Conditional Random Fields 9
2.5 General sentence level discourse analysis 10

3 Task Description 11
3.1 Dataset and preprocessing 11
3.2 Annotation 11
3.2.1 Dialogue act tag set 11
3.2.2 Annotation 12
3.2.3 Non troubleshooting thread 12
3.2.4 Multi-headedness 12
3.3 Classification setup 14
3.3.1 Learner 14
3.3.2 Features 15
3.3.3 Classification Methodology 17

4 Results And Evaluation 18
4.1 Heuristic and dummy models 18
4.2 Component-wise analysis 18

4.2.1 Comparison with post level research 20
4.3 Error Analysis 21
4.3.1 Confusion matrix 21
4.3.2 Correlation analysis 22
4.3.3 Manual analysis 23

5 Conclusion And Future Work 26
5.1 Conclusion 26
5.2 Future Directions 27

Acknowledgments

First of all, I sincerely show my thanks to my supervisor Professor Tim Baldwin. This work is wholly supervised and supported by Tim. During my work on the project, Tim not only patiently gave me clear and practical instructions about the research direction and methodology but also generously provided detailed suggestions about research techniques. The supervision of Tim was extremely helpful in improving my understanding of the problems and even some trivial aspects such as programming and writing. Without Tim, I can not learn so much about computing linguistics and can not finish the project to current extend.

Besides, I am grateful to Li Wang, graduate PhD student of Tim, for sharing his code used in the previous stage of this project. It helped me a lot in finding the technical solutions efficiently. I also appreciate Fei Liu, Yuan Li and Yiqing Zhang, all of whom are current PhD students of Department of Computing and Information Systems, for introducing Tim to me and sharing the experiences of their PhD research life. Their introduction and sharing played a crucial role when I was confused about the Master project.

Finally, I also appreciate the organisers and all the participants who took part in the oral presentation. I am sorry for not having enough time to remember your names, but your questions makes me think about my work from different perspectives and let me think of this research question more thoroughly.

Chapter 1

Introduction

1.1 Background

Web forums are online platforms where people can discuss, share and retrieve information via threaded discourse. Web forums have been widely used in many areas such as education, technical support and social community, which is valuable for users to resolve questions in a specific area. The appearance of Web 2.0 boosts the development of web forums, which improves the interactions of people obtaining and providing answers to questions but at the same time brings large volumes of data. Hence, more and more answers are available on web forums but it is becoming harder and harder to extract the information as a result of scaling and diversity of forum data.

1.2 Motivation

This research aims at enhancing information access to web forums data by investigating the discourse structure of troubleshooting-oriented forum thread level at a more granular level. Previous work in Wang *et al.* (2011) constructed the web forum post level discourse structure as rooted directed acyclic graphs. The post level discourse structure includes two types of information - inter-post links (Link) dialogue acts (DA). In this way, each post can be labeled in form of Link-DA, such as 0+Question-question and 1+Answer-answer. The author proposed three classification methods: separately classify the Links and Dialogue Acts and then compose the prediction results, jointly predict the Link-DA labels and apply dependency parsing to the problem. The experimental results shows that combined Link-DA prediction based on `CRFSGD` achieves best performance among all the approaches. Apart from the prediction of post labels, the author also explained the issues in post level discourse structure, which may interfere modelling of dependency parsing, such as multi-headedness and disconnected sub-graphs. We aim to move beyond the post

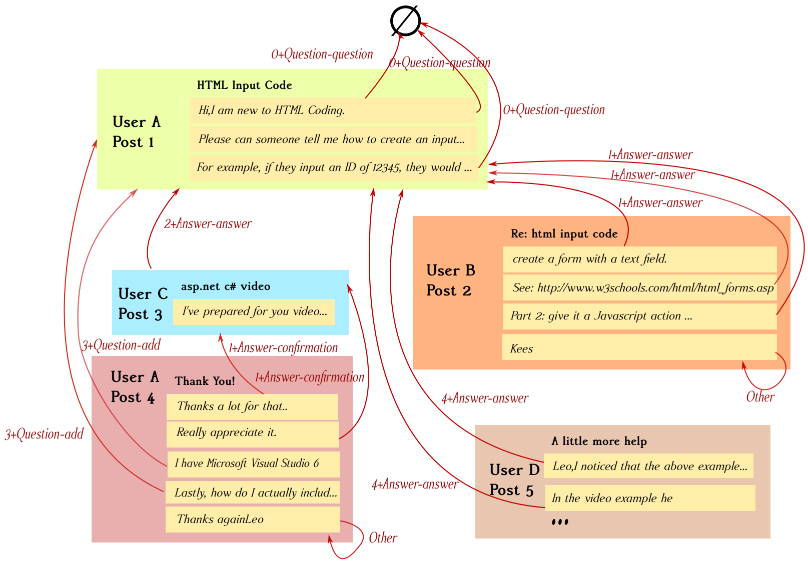


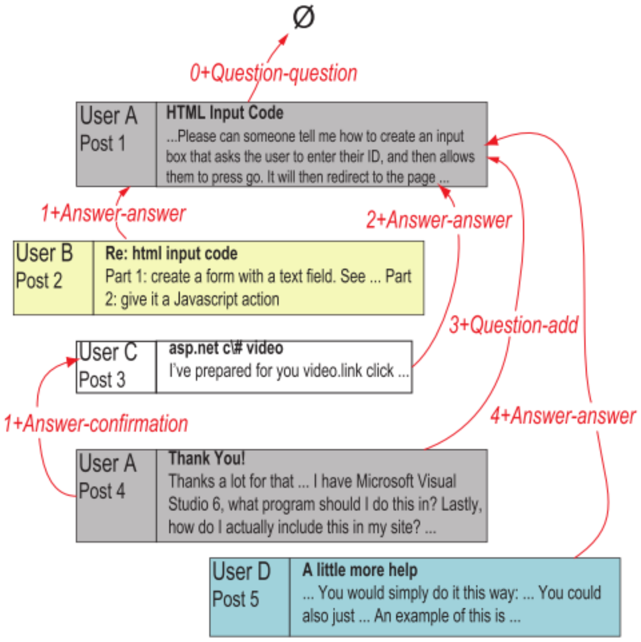
Figure 1.1: Sample thread of sentence level discourse structure

level discourse structure and investigate the prediction performance of Link-DA labels at sentence level as well as whether sub post level discourse structure can resolve the inherent issues at the higher level discourse structure.

The dataset in this research is the CNET dataset created by Kim *et al.* (2010b). CNET is an online community where people can discuss and share information about technical problems. However, the original dataset only contains threads and posts. In other words, we need to segment the posts into sentences. Based on the comparison in Read *et al.* (2012), we find the `tokenizer` tool perform best on the social media data, so we use it to segment the posts into sentences automatically initially and then correct the errors manually.

The CNET dataset has no annotation so another task is annotation. To make our research results comparable with previous work, we still use the dialogue act tag set used in Wang *et al.* (2011) which was initially proposed in Kim *et al.* (2010b). The threads in the CNET dataset are mostly troubleshooting-oriented. But some of the threads are not. In Kim *et al.* (2010b), such posts were labeled as Question-information. In the previous work by Wang *et al.* (2011), the author simply discarded the non troubleshooting-oriented treads. But in our research, we keep these threads and annotate them manually. Detailed annotation method will be explained in Section 3.2.

Figure 1.1 is a sample thread after annotation of sentence level discourse structure and Figure 1.2 is a sample thread of post level discourse structure. In Wang *et al.* (2011), the post level discourse can be constructed as a rooted directed acyclic graph, in which each link starts from a post and stops at another post. But as shown in 1.1 the threads of sentence level discourse structure can not be regarded as a rooted di-



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 ...

Re: html input code
Part 1: create a form with a text field. See ... Part 2: give it a Javascript action

asp.net c\# video
I've prepared for you video.link click ...

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? ...

A little more help
... You would simply do it this way: ... You could also just ... An example of this is ...

Figure 1.2: Sample thread of post level discourse structure

rected acyclic graph because each link in the sentence level discourse structure starts from a sentence but end at another post. As the modelling of discourse structure is different, the dependency relationship in the graph is not at the same level, dependency parsing is not feasible in this model. CRFs are conditional probability models and the label prediction of current instance is based on the probability of previous instances. Compared with dependency parser, CRFs do not rely on the dependency relationships between instances so it can be used as prediction tools in our experiments.

Additionally, we compare the sentence-level model and post-level model to see if sentence-level model can achieve some improvement in prediction and if sentence-level model can resolve the problems in post-level models, such as multi-headedness. Lastly, we perform error analysis with three methods, confusion matrix analysis, correlation analysis and manual analysis. By confusion matrix analysis, we distinguish the easily-predicted labels and hard-predicted labels. In correlation analysis, we randomly selected 100 sentences, and firstly apply Chi-square analysis to get tokens most relevant to prediction correctness. Secondly, we calculate the prediction correctness entropy corresponding to sentence positions to get the positional characteristics relevant to prediction correctness. The third technique is manual analyse. We randomly select 100 misclassified instances, define two characteristics (lexical exceptional and sequential exceptional) that may lead to the misclassification and calculate what proportion of the wrongly-predicted instances have such characteristics. For the lexically exceptional instances, we further their their lexical overlap degree and characteristics of unique tokens.

1.3 Contribution

Firstly, most of previous works focus on post level discourse structure. Our work moves beyond that and deep into sub post level.

Secondly, we make some aggregation of features proposed in previous researches and extracted new features which was not used in the past research, such as Doc2Vec feature vector.

Thirdly, we investigate issues of post level discourse structure. We investigate if more granular level discourse structure can resolve the multi-headedness issue.

Finally, we perform error analysis on the output from different points of view.

1.4 Overview

This thesis will cover both the previous research work on the discourse structure analysis and our novel experiments results and analysis at the sentence level discourse structure.

- Chapter 2 is a literature review relevant to our research topic.
- Chapter 3 is about how to set up the experiments, including data processing and annotation, classification algorithms selection.
- Chapter 4 is the experimental results and performance and error analysis.
- Chapter 5 is conclusion and future directions. We give a summary of the whole thesis and propose three possible future directions.

Chapter 2

Literature Review

This work makes further research on sub post level building on earlier work on the post level discourse structure research by Wang *et al.* (2011). In the previous work, Kim *et al.* (2010a) carried out experiments on classification of dialogue acts and Kim *et al.* (2010b) classifies post links and dialogues acts separately. Wang *et al.* (2011) composed joint-classification of post links and dialogue act tag on dataset from CNET annotated by novel tag set proposed in Kim *et al.* (2010b). As all the past works were performed on post level, we want to take a further exploration on the sub post level.

2.1 Sentence boundary detection

Our research is at sub post level and the discourse structure of original CNET dataset is at post level. One important research area is the sentence boundary detection.

Automatic sentence boundary detection was popularly used in utterance segments in conversational speech. To address this issue, automatic segmentation has been explored. Ang *et al.* (2005) used a decision tree model, a language model (lexical features), and a HMM-based combination model to automatically segment the ICSI meeting corpus. They found that while prosodic features (in the form of a decision tree) produced superior results, the combined approach gets further improvement. Finke *et al.* (1998) tried both neural network and Markov model approaches for automatic segmentation, and found that both methods could produce competitive results. Shriberg *et al.* (2000) used decision tree and hidden Markov modelling to segment the Broadcast News and Switchboard corpora, and experimented with a range of lexical and prosodic features. They found that prosodic features are especially useful, and achieved an accuracy of 96.8% on the Broadcast News transcript and 96% on the Switchboard corpus.

Read *et al.* (2012) proposed three general categories of classification methods -

Dialogue Act	Description
A	accept response
AA	acknowledge and appreciate
AC	action motivator
P	polite mechanism
QH	rhetorical question
QO	open-ended question
QR	or/or-clause question
QW	wh-question
QY	yes-no question
R	reject response
S	statement
U	uncertain response

Table 2.1: The 12 dialogue act labels defined by Jeong *et al.* (2009)

rule-based, supervised and unsupervised. The author composes comparison between 9 sentence boundary identification tools from the three categories on three corpus - Brown(Francis and Kucera, 1982), CDC(Morante and Blanco, 2012), GENIA(Kim *et al.*, 2003) and WSJ(Dridan and Oepen, 2012). In Read *et al.* (2012), 9 sentence boundary detection tools are evaluated. Among all the 9 tools, `tokenizer` method achieves better performance than other tools on social media corpus.

2.2 Dialouge act tag set

Jeong *et al.* (2009) defined a dialogue act tag set with 12 dialogue acts based on Dhillon *et al.* (2004) as shown in Table 2.1.

In the early research, Kim *et al.* (2006) proposed a tag set in Table 2.2 based on Austin (1975) and Searle (1969). They annotated a dataset from student online discussion in an Operating System course. The Kim *et al.* (2006) dialogue act tag set was used in Feng *et al.* (2006) in detecting best answer post in discussion board.

Kim *et al.* (2010b) proposed another novel post level dialogue act tag set, whose annotation method applied in Wang *et al.* (2011) as shown in Table 3.2.

2.3 Features extraction

Many kinds of features have been used in previous discourse structure analysis, including lexical features (Cong *et al.*, 2008; Kim *et al.*, 2010b), such as bag of words, structural features(Kim *et al.*, 2010b), such as relative position and semantic features (Ding *et al.*, 2008; Kim *et al.*, 2010b) such as similarity scores. Generally, lexical

Dialogue Act	Description	Positivity
Question	Question on specific problems	neutral
Announcement	Command or announcement	neutral
Complex Answer	An answer requiring a full description of procedures, reasons, etc.	neutral
Simple Answer	An answer with short phrase or words, e.g. factoid, Yes/No	neutral
Suggest	Give advices/suggestions for some problems/solutions	neutral
Elaborate	Elaborate on a previous arguments or questions	neutral
Correct	Correct a wrong answer/solution with a new one	negative
Object	Object to some argument/suggestions/solutions	negative
Criticize	Criticize an argument	negative
Support	Support others' arguments/solutions	positive
Acknowledge	Confirm or acknowledgement	positive
Compliment	Praise an argument or suggestion	positive

Table 2.2: Dialogue acts proposed by Kim *et al.* (2006)

features are less effective than other features.

Lui and Baldwin (2009) experimented with 8 features based on the previous research (Fortuna *et al.*, 2007). The 8 features are

BoW: bag of words features.

ILIAD: the features proposed by Baldwin *et al.* (2007).

WANAS: the features used by Wanas *et al.* (2008)

ILIAD-User: aggregated mean of each of ILIAD features.

WANAS-User: aggregated mean of each of WANAS features.

PostAfter: if author A has replied to at least one of author B's posts then A and B will be linked (Fortuna *et al.*, 2007).

ThreadParticipation: if two authors participated in the same thread more than a certain times then A and B will be linked Fortuna *et al.* (2007).

CommonAuthors: if thread A and thread B have more than 2 common authors then they will be linked Fortuna *et al.* (2007).

The experimental results show significant improvements when using the user level features.

Bhatia *et al.* (2012) proposed 4 features. They are

Content based features:

- whether the post quotes a previous post.
- cosine similarity between the post and the thread title, initial post and the whole thread, as three features.
- whether the post contains question marks, 5W1H words, or keywords such as same and similar, as three binary features.

Structural features:

- the absolute and relative positions of the current post in the thread, as two features.
- the total amount of words and unique words in the current post, as two features
- the total amount of words and unique words in the current post after discarding stop words and stemming, as two features.

User features:

- the total amount of posts from the post's user.
- whether the post's user is the initiator of the thread it belongs to.
- the post's user's authority score (Bhatia and Mitra, 2010).

Sentiment based features:

- whether the post contains one of the following (each represented as a binary feature): *thank*, *exclamation mark*, *did not*, or *does not*.
- the post's sentiment score calculated using `SentiStrength` (Mike *et al.*, 2010).

They observed that the content based features are the most effective and the sentiment based features are, while quite novel, the least effective.

Kim *et al.* (2010b) extracted 4 group of features which are most relevant to our previous research (Wang *et al.*, 2011). They are

Lexical features: it includes the unigram and bigram feature vector.

Structural features:

- whether the post is crated by the initiator of the thread
- the relative position of the current post

Post context features:

- the immediately preceding post predicted dialogue act
- whether the preceding post's author is also the author of current post
- the relative position of post whose author is also the author of the current post
- predicted dialogue acts of all the preceding posts

Semantic features:

- relative position of post with most similar title with the current post
- relative position of post with most similar content with the current post
- distribution of dialogue acts of the author of the current post

2.4 Learners and Conditional Random Fields

In Kim *et al.* (2010b), three classifiers were used, maximum entropy, SVM-HMM and CRFs. They find that CRFs are the most effective learner and structural features are the most effective features in predicting the dialogue acts. In prediction of link, when using dialogue act-based post contexts, the best result is achieved when applying CRFs with combination of structural and post context features.

In the web forum research, Conditional Random Fields (CRFs) was applied in many situations. Ding *et al.* (2008) focuses on the question and answer detection problem at sentence level. The author compares linear chain CRFs, skip chain-CRFs and 2D CRFs. The experimental results indicate that CRFs are better than SVMs with a polynomial kernel and decision trees. Additionally, 2D CRFs outperform linear-chain CRFs for both context and answer detection.

In Wang *et al.* (2011), CRFSGD is used as one of the classifiers to perform the joint and composite classification on the dialogue acts and inter post links.

2.5 General sentence level discourse analysis

The previous sentence level discourse structure researches mainly focus on question answer pair extraction.This task may help enrich the knowledge base of question answering services (Cong *et al.*, 2008), improve information/answer access over forum threads (Cong *et al.*, 2008) and improve thread symmetrisation (Ding *et al.*, 2008). (Ding *et al.*, 2008) aims to extract answer at sentences level.

Soricut and Marcu (2003) proposed a dependency parsing model that uses syntactically and lexically related feature set to predict discourse structures at sentence level. The research finds that syntax contributes most to the parsing of sentence level discourse structure.

Chapter 3

Task Description

This chapter introduces the dataset and corresponding preprocessing and annotation technique. Then we introduce learning algorithms and feature extraction.

3.1 Dataset and preprocessing

Our research is performed over the CNET dataset, which was created by Kim *et al.* (2010b) based on CNET forums. The original dataset only contains post-level structure. As our research is at a more granular level, we need firstly preprocess this dataset, splitting each post into sentences at sub post level.

Based on past research on sentence boundary detection (Read *et al.*, 2012), `tokenizer` performs best at sentence tokenization over social media text. Thus we firstly use tokenizer sentence boundary detection tools to automatically separate the posts into sentences automatically. As errors in the output are unavoidable, we further correct the wrongly segmented sentences manually. Table 3.1 provides an overview of CNET dataset after preprocessing.

3.2 Annotation

3.2.1 Dialogue act tag set

Annotation at the sentence level is the same as at the post level. The labels are formatted as a combination of inter-post links (Link) and dialogue acts (DA). The annotation is based on the assumption that each sentence can only link to an earlier post. The dialogue act set in this research is that proposed by Kim *et al.* (2010b) for post level analysis.

	Count
Threads	327
Posts	1371
Sentences	6991

Table 3.1: Overview of CNET dataset

3.2.2 Annotation

To make the annotation subjective, we annotate the sentences manually without referring to the post level annotation. We mainly refer to the DA tag sets proposed in Kim *et al.* (2010b) and a typical sentence label is in form of Link+DA. However, there are two minor exceptions. One is the Other, when coming across the some sentences such as signature of the author, these sentences can not labeled as the any of the dialogue acts with actual meanings but can only be regarded as Other. Besides, such sentences are also dependent on other posts. So the isolated sentences, such as the signatures, would be labelled as just Other without linking number. The other exception is the Question-information dialogue act. The annotation of such sentences will be explained in the next subsection.

3.2.3 Non troubleshooting thread

Non troubleshooting thread means the information discussed or shared within a thread is not troubleshooting-oriented. In the CNET dataset, 14 threads are not trouble shooting oriented. In the previous work (Wang *et al.*, 2011), such threads were just simply discarded. But in our work, we do not remove such threads and put them into training.

For such thread, only Question-information and Other dialogue act can be used to label the sentences, because other dialogue acts are designed for the troubleshooting thread. For all the non troubleshooting-oriented threads, we simply annotate the first sentences in the first post as 0+Question-information while other sentences as distance to initial post + Other.

3.2.4 Multi-headedness

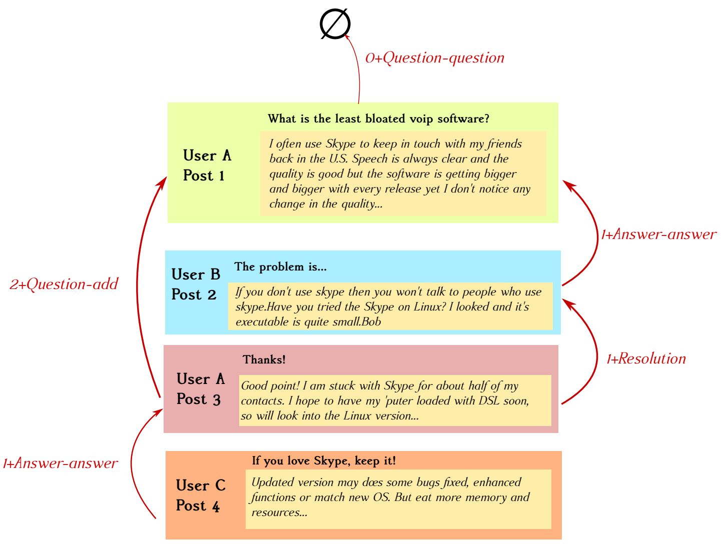
One important issue in post level discourse structure research is multi-headedness. Multi-headedness means it is possible for a given post to have multiple heads, including the possibility of multiple dependency links to the same post as defined in Wang *et al.* (2011). As shown in Figure 3.1, the Post 3 is a post with multi-headedness. As sentences are more granular, we want to investigate whether multi-headedness is alleviated through sentence level analysis.

Super-category	Sub-class	Abbreviation	Description
Question	question	QQ	the post contains a new question which is independent of the posts before it.
	add	QAd	the post provides additional information or asks a follow-up question, regarding a previous question.
	confirmation	QCn	the post confirms details or error(s) in a question.
	correction	QCr	the post corrects error(s) in a question.
	information*	QI	the posts are not troubleshooting-oriented and only provide information.
Answer	answer	AAn	the post proposes an answer to a question.
	add	AAd	the post provides additional information to an answer.
	confirmation	ACn	the post confirms details or error(s) in an answer.
	correction	ACr	the post corrects error(s) in an answer.
	objection	AO	the post objects to an answer.
Resolution	—	Res	the initiator confirms that an answer works.
Reproduction	—	Rep	a non-initiator asks a similar question, or confirms that an answer should work.
Other	—	O	the post does not belong to any of the above classes.

Table 3.2: The dialogue act (DA) tag set proposed by Kim *et al.* (2010b)

In the post level annotation, the number of posts with multi-headedness is 65, corresponding to 5% of posts. All the posts with multi-headedness have exactly 2 labels, which means no posts have more than two labels.

We find that with sentence-level annotation, the number of posts with 2 labels increases to 71 and 1 post has 3 labels after sentence level annotation. On the sentence level labels, 4 sentences have 2 labels and no sentence has more than 2 labels, which means only a small proportion (0.02%) of sentences have multi-headedness.



What is the least bloated voip software?

I often use Skype to keep in touch with my friends back in the U.S. Speech is always clear and the quality is good but the software is getting bigger and bigger with every release yet I don't notice any change in the quality...

The problem is...

If you don't use skype then you won't talk to people who use skype.Have you tried the Skype on Linux? I looked and it's executable is quite small.Bob

Thanks!

Good point! I am stuck with Skype for about half of my contacts. I hope to have my 'puter loaded with DSL soon, so will look into the Linux version...

If you love Skype, keep it!

Updated version may dœs some bugs fixed, enhanced functions or match new OS. But eat more memory and resources...

Figure 3.1: A simple thread with multi-headedness

3.3 Classification setup

3.3.1 Learner

Figure 3.2 shows a thread after annotation. We can see the sub post level discourse structure based on our annotation method is not a typical tree structure. Each node in this graph contains multi sub nodes and sub nodes are linked to a parent node. So we use a linear chain Conditional Random Field model for joint classification of both links and dialogue acts.

Conditional Random Fields (CRFs)

Normally, CRFs are applied to tasks such as part-of-speech tagging, named entity recognition and semantic role labelling, where the individual tokens are single words. In our case, sentences and threads respectively corresponds to the tokens and sentences in traditional CRF model. We perform our model based on the sequences of sentences within the threads. This way does not care about the post sequence in the thread. To take the post level sequential information into consideration, we explicitly add a position vector $header, body, tail$ to indicate current sentence's position in the post. Each element of this vector is a binary value with either 0 or 1. For example, if the position vector of current sentence is $\{1, 0, 0\}$, it means current sentence is at the first position within the post. If the position vector of current sentence is $\{1, 1, 1\}$, it means current sentence can be regarded any of header, body and tail which implicitly indicates the post only has one sentence. So the position vector can obtain the post boudnary information.

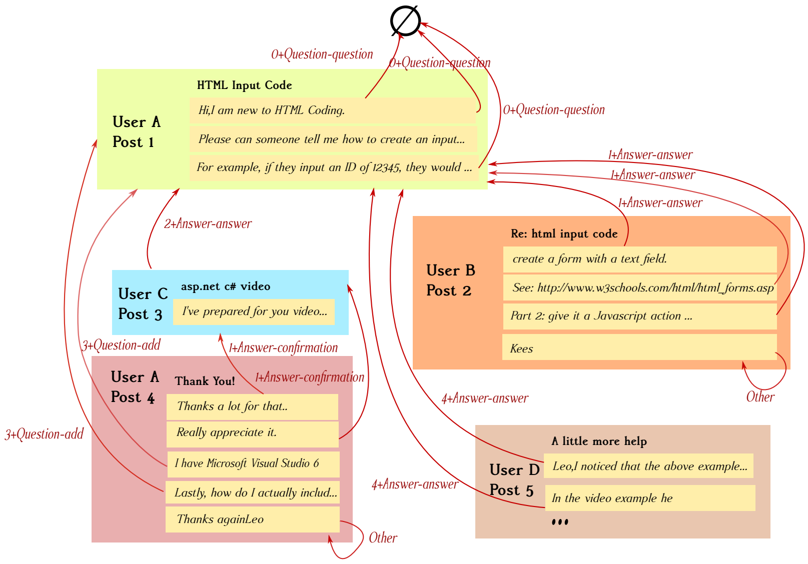


Figure 3.2: A snippeted CNET thread annotated on sentences

The CRFs package we used is `CRFSuite` (Okazaki, 2007). `CRFsuite` is an implementation of CRFs for labelling sequential data. It has the following features

- Fast training and tagging.
- Linear-chain (first-order Markov).
- State-of-the-art training methods
- An efficient file format for storing/accessing CRF models.

As our dataset have almost 7000 instances and more than 8000 features. The high dimensionality may greatly reduce the training efficiency. So the efficiency optimisation of `CRFSuite` is important in improving training speed. Besides, as we need to compare our model with the post level model, previous work of CRF model on post level is a linear chain model and uses *stochastic gradient descent*. These model characteristics are all involved or can be configured in `CRFSuite`.

3.3.2 Features

In our experiment, we extract 8375 features over 6 categories. The post level structural and linguistic features comes from the original work Wang *et al.* (2011) but aggregated into two feature set. Derived from post level structural and linguistic features, the sentence level structural and linguistic features have the same idea with post level features but are just converted into the corresponding sentence level feature sets.

Bag of words

Due to the simplicity and efficiency, we firstly extracted the bag of words feature vector. We use TreeBankWord Tokenizer and WordNet Lemmatizer from NLTK (Bird *et al.*, 2009) to segment the sentences into tokens and then use the bag of lemmas as the feature vector. Each features takes a value of either 0 (not contains this token) or 1 (contains this token).

Doc2Vec

Even though bag of words extraction is simple, the feature vector of bag of words is sparse, which suffers from data sparsity and high dimensionality. Different from a sparse feature vector of bag of words, Doc2Vec We use the Gensim (Řehůřek and Sojka, 2010) package to get the Doc2Vec features. As there are only 6991 sentences, to avoid serious dimensionality issue, the number of features is set to 300 while other configurations are set default.

Post Level Structural Features

Initiator whether the author of the current post is the initiator of the thread. It takes a value of either 0 (not the initiator) or 1 (initiator).

PostPos the relative position of parent post of the current sentence. $PostPosition \in [0, 1]$.

Post Level Linguistic Features

TitSim the relative position with the most similar title to the current post. The similarity is calculated with unweighted cosine similarity. $TitSim \in [0, 1]$.

PostSim the relative location of the post which contains the sentence most similar to the current sentence. $PostSim \in [0, 1]$.

UserProf the class distribution of the author of the current sentence. $UserProf = \{x_1, x_2, ..., x_n\}$, where n is the number of distinct classes in the training data, and x_i is the occurrence of this user's posts which have a class label of i.

Sentence Level Structural Features

SentPos the relative location of the sentence in the post. $SentPos \in [0, 1]$.

SentType a binary value vector $\{header, body, end\}$ explicitly indicates the position of the sentence in the post.

Sentence Level Linguistic Features

Punct the amount of *question marks*, *exclamation marks* and *URLs* in the current post. The three individual features which take a non-negative integer value.

SenSim the relative location in the post of the sentence most similar to the current sentence. $Sensim \in [0, 1]$

3.3.3 Classification Methodology

Our experiments were performed on 10-fold cross-validation. To make sure there is no duplicate sentences in each fold of the cross-validation. We perform the stratified cross-validation, stratifying at the thread level. In this way, we can guarantee each sentence only occurs once in each fold.

Metrics

The results are primarily evaluated using the following 5 metrics,

Sentence F1-score sentence level micro-averaged F1-score.

Post F1-score mark the post with the majority of sentence labels within the post and get the micro-averaged F1-score on post level.

Post Accuracy proportion of posts whose sentences are all correct.

Thread Sentence Accuracy proportion of threads whose sentences are all correct.

Thread Post Accuracy mark the post with the majority of sentence labels within the post and get the proportion of threads whose posts are all correct.

The post F1-score and thread post accuracy are the two major metrics used in comparison with post level models. We take majority of sentences label because we observed that more than 90% posts are consistent with its majority of sentence labels. What's more, on post level model, we can predict an individual labels for each post instead of just True or False which just indicates the correctness of prediction. So it is reasonable to use the majority of sentence labels as post label, to make it comparable with post level model.

In Wang *et al.* (2011), a thread can be treated as correctly predicted only if all the posts are correct. So we need to use the Thread Post Accuracy as we defined to get the correctness of each thread. Thus, we can use the two corresponding metrics to compare with the post level models without sacrificing much accuracy.

Chapter 4

Results And Evaluation

4.1 Heuristic and dummy models

Wang *et al.* (2011) proposed two no features involved prediction models.The first is the Heuristic model. Heuristic model takes human oracles as the label of instances. Heuristic labels is a possible prediction of the dataset which can be done by human intuitively. As the objective of machine learning is to achieve prediction performance at least better than human, the heuristic model is used as the baseline model in our research. We simply label the sentences from the first post as 0+Question-question and the sentences from the other posts as 1+Answer-answer. This annotation makes sense as the structure of the post level discourse structure is a rooted directed acyclic graph. We naively assume that all the sentences in the first post are all about questions while all the sentences in the following posts are answers to the initial questions. We get a sentence level F1 score of 0.512 and a post level F1 score of 0.507.

As CRF is a linear chain algorithms, which can also predict the current labels only based on the sequential information even without features. To distinguish the effect of individual sequential information, we perform the experiment in the data set with a CRF without any features but only based on the sequential labels of the sentences in the thread. The results are presented in Table 4.1. We can see that the dummy model is much worse than the Heuristic model on all of the metrics.

4.2 Component-wise analysis

Firstly, we compose a model with all features. From Table 4.1, we can see that sentence level F1 score and post level accuracy are slightly better than the heuristic baseline. However, the thread level accuracy with all sentences correct is much better than the heuristic baseline, which indicates that the CRF model with all features predicts more sequential and positional information of sentences in the thread. With more features, the all feature model also gets better performance than the dummy

Methods	Metrics				
	Sentence F1	Post F1	Post Acc	Thread-Sent Acc	Thread-Post Acc
Heuristic	0.512	0.507	0.275	0.011	0.300
Dummy	0.325	0.236	0.134	0.0	0.0
All Features	0.527	0.491	0.297	0.115	0.291
-BoW	0.492	0.466	0.254	0.094	0.258
-Doc2Vec	0.531	0.494	0.305	0.121	0.291
-Post Structural	0.472	0.439	0.255	0.067	0.218
-Post Linguistic	0.594	0.539	0.439	0.209	0.33
-Sentence Structural	0.503	0.471	0.256	0.091	0.261
-Sentence Linguistic	0.525	0.493	0.296	0.100	0.285

Table 4.1: Sentence/post/thread-level classification F-scores

model.

Next, we compose the component wise classification by ablating one set of features a time to see the effects of each set of features. Several interesting things can be observed on both sentence level F1 score and post level F1 score. The bag of words, post level structure features and sentence level structure features have a positive effect, as the F1 score dropped without the three sets of features. In contrast, Doc2Vec, Post Linguistic and sentence linguistic features have a negative effect on the model performance as the F1 score increases when each of the three feature sets are ablated. Among the positive feature sets, the post structural features improves the performance while post linguistic features pull down the model performance the most.

Additionally for the post level accuracy, bag of words and structural features (on both the post and sentence level) are important in achieving good performance. For thread level accuracy, the post level structural feature set improves the performance significantly while post level linguistic feature set is the only feature set that has a negative influence on the final performance.

Overall, the performance of models with different feature combinations are consistent on the five metrics. The structural features at both post and sentence level have a positive effect on the model performance. This is mainly because the classification of the CRF model mainly relies on the sequence of instances and the structural features provide more information on the structure and position of the sentences in the thread. In contrast, the linguistic features, especially the post linguistic features, pull down performance. The reason is that the post linguistic features contain the User Profile feature set which was also proven to have negative effect on the performance on the post level research (Wang *et al.*, 2011).

	Post	Thread	Sentence
Post Level	0.619/0.665	0.484/0.524	-
Sentence Level	0.491/0.539	0.291/0.33	0.527/0.594

Table 4.2: All features/ component wise (best result) comparison

	Sentence level	Post level
Multi-headedness ratio	0.02%	5%

Table 4.3: Multi-headedness ratio comparison between sentence and post level discourse structure

4.2.1 Comparison with post level research

Comparison on prediction performance

In Wang *et al.* (2011), CRF models were also used in predicting labels at the post level. We directly compare our post-level results in Table 4.2. Here we mainly compare between the all features model and the component wise model, which achieves the best performance on post, thread and sentence levels.

In the post level results, the metrics are post level F1-score and thread level accuracy. In the sentence level results, the metrics are sentence level F1-score, post level F1-score (majority of sentence labels as post label) and thread level accuracy (all posts are correct). As shown in Table 4.2, we cannot compare on the sentence level as the post level research did not deep into the sentence level. But on the corresponding metrics on the post and thread level, we can see that the post level research results are much better than the sentence level results.

As the sentence level research is on a more granular level, it is more demanding to predict the label for each sentences. But this also indicates that sentence level discourse research can reveal more information. So a possible future direction is to combine the post labels predicted by post level models with sentence level features to see if it can achieve better performance than pure post level or pure sentence level models.

Analysis of multi-headedness issue

According to the statistical analysis in Section 3.2.4, post level multi-headedness is reduced substantially in shifting our analysis to the sentence level as shown in Table 4.3. What's more, we can find which sentence within the post leads to the post-level multi-headedness based on the sentence level annotation.

Label	Count
1+Answer-answer	2841
0+Question-question	2423
Other	527
1+Resolution	379
2+Answer-answer	252
1+Question-add	204
1+Answer-add	176
2+Question-add	116
1+Question-confirmation	55
3+Answer-answer	12
1+Answer-confirmation	4
5+Answer-answer	1
4+Question-information	1

Table 4.4: Distribution of predicted labels

4.3 Error Analysis

To have a clearer look at what kind of sentences can be classified by our model and what kind can not be classified, we performed error analysis. We applied three methods, confusion matrix analysis, correlation analysis and manual analysis.

4.3.1 Confusion matrix

Putting the predicted results of each copy of 10 fold cross validation together, we construct an integrated confusion matrix. We select the top 10 labels among the raw labels to analyse. The other labels are merged into the category of Misc. Compared with the 51 labels in gold standard data set, the predicted labels only have 13 labels. 1+Answer-answer, 0+Question-question, Other and 1+Resolution make up the greatest proportion of predicted labels.

From the confusion matrix Table 4.5 and predicted labels distribution in Table 4.4, we can see that 0+Question-question and 1+Answer-answer account for the vast proportion of the predicted labels. 1+Resolution and 1+Question-add are also identified but a big proportion of these two labels are mislabelled as each other. Apart from these, the other labels are mostly mislabeled as 1+Answer-answer. To improve this model, one possible future work would be to optimise the classification on the labels which are misclassified as 1+Answer-answer.

	0QQ	1AAn	O	2AAd	1AAd	1Res	1QAd	3AAn	2QAd	Misc
0QQ	2035	8	27	1	0	4	0	0	0	2
1AAn	14	1271	25	65	57	8	6	6	17	12
O	47	57	427	5	0	13	3	0	2	1
2AAn	5	350	3	23	25	4	0	0	0	0
1AAd	0	319	2	27	28	3	2	0	1	2
1Res	19	22	22	1	0	184	70	0	6	0
1QAd	98	29	8	6	3	76	37	0	44	2
3AAn	0	128	1	40	18	0	0	0	0	2
2QAd	46	31	5	0	0	26	42	0	27	4
Misc	159	626	7	84	45	61	44	6	19	36

Table 4.5: Confusion matrix of predicted labels

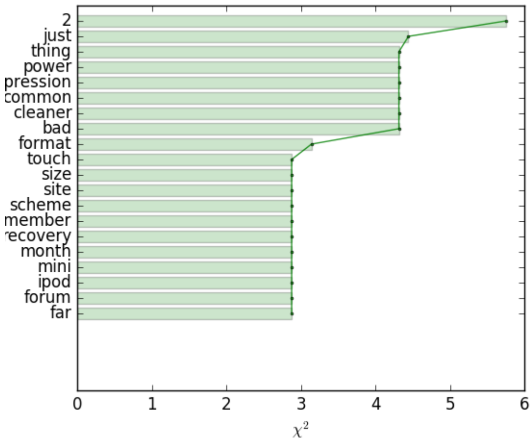


Figure 4.1: Chi square ranking for the tokens

4.3.2 Correlation analysis

In this part, we randomly select 100 sentences to see the what kind of characteristics that they have. We mainly compare the relationship between the feature and the prediction correctness. The prediction correctness is a binary value (1 for correctly classified while 1 for wrongly classified) indicating whether the sentence is correctly predicted. Among the 100 sentences, 59 sentences are correct while the others are wrong.

Linguistic characteristics

As the bag of words are all binary value (either 0 or 1) and whether prediction is equal to the gold standard is also binary value, we use the Chi-Square analysis to see

	Head-Correctness	Body-Correctness	Tail-Correctness
Entropy	3.714	2.079	2.303

Table 4.6: Prediction correctness entropy comparison between sentence positions

the correlation between the tokens and the prediction correctness which is also either 0 (correct prediction) or 1 (right prediction). Given $\alpha = 0.05$ and $\chi^2 > 3.84$, we can say this feature is related to predicted correctness.

As shown in Figure 4.1, the top ranking tokens are *2*, *just*, *thing*, *power*, *pression*, *common*, *cleaner* and *bad* whose $\chi^2 > 3.84$. So these tokens can be used to see if the sentences can be predicted correctly. Besides, the initiator feature of this sentence also has $\chi^2 = 7.96$ so the *initiator* feature of the sentence is also related to correctness of prediction.

Structural characteristics

As the prediction correctness is either 0 (wrong prediction) or 1 (right prediction), we can use `Shannon Entropy` to evaluate the effect of position has on the prediction correctness. The greater entropy is, the more significantly relevant the feature is to the prediction correctness. From Table 4.6, we can see that sentences located at the start of the post achieve highest entropy, while the tail sentences have the lowest entropy.

4.3.3 Manual analysis

In the manual analysis, we randomly select 100 *misclassified* instances and analyse the what characteristics they have. We mainly analyse the instances from the following two types of special instances:

Sequential exception: if the instance has a label different from its surrounding instances (i.e. the instance has different labels with its previous instance and its next one), this instance can be regarded as a sequential exception.

Lexical exception: one instance can be regarded as a lexical exception if it has at least one *unique token*, (the tokens not belonging to any other instance in the training dataset). For example, if a instance has a token *"well"* but the *"well"* tokens can not be found in any other instance in the training dataset, this instance can be regarded as a lexical exception.

Among the 100 misclassified instances, we use the ratio of sequential exceptions and lexical exceptions to respectively indicate the lexical and sequential characteristics. The results are shown in Table 4.7. We can see that sequential exception have

	Lexical exception	Sequential exception
Ratio	38%	56%

Table 4.7: Lexically and sequentially exceptional instances ratio

a greater effect on the wrongly-classified instances as there are 56 wrongly classified instances are caused by difference with its surrounding instances. 38% of the 100 are wrongly classified as a result of being a lexical exception.

As we have defined, one instance can be regarded as a lexical exception only if there is one unique token in this sentence. However, the length of sentences varies, so we future define another metric to evaluate the extend to which current instance overlaps with the training dataset.

Overlap degree: can be defined as Equation 4.1. It is equal to the number of unique tokens divided by the number of all tokens in this instance.

$$Overlap\ degree = \frac{Number\ of\ unique\ tokens}{Number\ of\ all\ tokens} \tag{4.1}$$

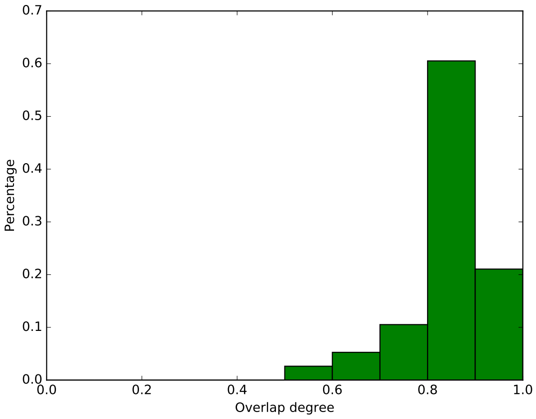


Figure 4.2: Lexical exceptions overlap degree frequency distribution histogram

Here we select the 38 lexical exceptions from the 100 randomly selected instances. From the histogram shown in Figure 4.2, we can see that more than 60% of the 38 lexical exceptions have an overlap degree between 0.8 and 0.9 and more than 20% of the lexical exceptions have an overlap degree more than 0.9. So no more than 20% of the lexical exceptions have an overlap degree less than 0.8. No lexical exceptions have an overlap degree less than 0.5 (0.5 can be roughly regraded as half of the tokens from this instance are unique with the training dataset).

Finally, we manually observe the unique tokens. Generally, most of the Url tokens are unique tokens. But the other unique tokens varies from instance to instance. Some are regular words, such as *flame* and *overheat*. Some are numbers, such as *1320* and *378*. It is hard to find a pattern among the other unique tokens.

Chapter 5

Conclusion And Future Work

5.1 Conclusion

In this research, we explore the discourse structure at sub post level based on CNET dataset. As the CNET dataset is not annotated and only has post level discourse structure, our first task is to segment the posts into sentences. Here we applied the combination method of automatic post segmentation and manual correction of errors. We used `tokenizer` sentence boundary detection tools to segment posts into sentences and then correct errors manually.

After sentence segmentation, we use the dialogue act tag set in previous research (Kim *et al.*, 2010b) to annotate the sentences making it consistent with the post level models. However, two exceptions need to be added into the original dialogue act tag set. One is that sentences such as author signature have no linking information so they are simply labelled as Other. The other one is annotation of non troubleshooting-oriented threads. For these threads, because only Question-information and Other can be used to annotate sentences within such threads, we simply label the sentences in the initial post as 0+Question-information while others are labeled as and hop counts + Other in which *hop counts* mean the distance to the initial post.

As the sentence level model are more granular, we introduce 5 metrics to evaluate the models, including sentence level F1 score, post level F1 score, post accuracy, thread sentence accuracy and thread post accuracy. To compare with post level model, we use the majority of sentence level labels as corresponding parent post labels and get the thread post accuracy within which all the posts in the thread correct.

After annotation, we come to find that the graph structure of sentence level discourse structure is different from the post level model - the links starts from a sentence but end at a post in the sentence level model while the links starts from a post but end at another post in the post level model. That is to say, the annotated sentence level discourse structure can not be constructed as a typical tree or rooted directed acyclic graph. So the dependency parser can not approach the sentence level model.

What's more, CRFs have been proven to achieve better performance on classification of post labels in Wang *et al.* (2011). For the above two reasons, CRFs are used as the only feasible classifier in our experiment.

In the experimental result analysis, heuristic model performs better than the dummy models. All features model gets better performance than both heuristic and dummy models. In the component wise experiments, we can find that post level structural feature set have the greatest positive effect on the final results while the post level linguistic feature set pull down the performance most. In comparison with post level model, sentence level model does not achieve as good performance as post level models in predicting the post labels .

In the resolvent of inherent issues in higher level models. As shown in the experimental results, we can see that the sentence level model can resolve the multi-handedness issue, which reduce the ratio of multi-headedness from 5% to 0.02%. The performance of corresponding prediction of post labels and thread labels by sentence level model is not as good as the post level models. As sentence level model is at a more granular level, it is demanding to predict the higher level label by the model at lower level.

In the error analysis, we perform three kinds of error analysis, confusion matrix analysis, correlation analysis and manual analysis. In confusion matrix analysis, we can see that the CRFs model mainly predicts the labels into 0+Question-question and 1+Answer-answer. Among other labels, 1+Resolution and 1+Question-add are also identified but a big proportion of these two labels are mislabelled as each other. Apart from these, the other labels are mostly mislabeled as 1+Answer-answer. In correlation analysis, we randomly select 100 instances to see the characteristics of the sentences. We investigate the relationship between the binary value features and the prediction correctness. The results shows that some tokens, such as *2*, *just*, *thing*, *power* and so forth, are relevant to the prediction correctness while sentences which are the head of post tends to be more easily classified correctly than on other positions. Finally, we perform manual error analysis. We randomly select 100 misclassified instances, define 2 characteristics which may lead to misclassificaion (sequentially exceptional and lexically exceptional). Among the 100 misclassified instances, 56% are cased by the difference with previous instance and next instance and 38% has unique tokens. Among the 38 lexically exceptional instances, more than 80% have an overlap degree above 0.8. No instance have overlap degree below 0.5.

5.2 Future Directions

This research just focuses on one perspective of predicating the sentence labels by CRF model. To our best knowledge, not too much previous work has been done on the sentence level discourse structure analysis. So there are many topics can be investigated within this area. Here we proposed three future research topics that are closely related to our work and can improve our model performance potentially.

Combination with results of post level models As the sentence level model does not achieve as good predication performance on thread and post level as post level models. So if we want to get a model with better prediction performance in post labels. One research direction is to get labels predicted by post level models, i.e. use the joint prediction by `CRFSGD` in Wang *et al.* (2011) as a feature in the setence level model. This technique can be understood as a partial *stacking* technique, we combine prediction results with other features to see if the other features or what features can improve the prediction performance.

Analysis of features weight In our research, we did not focus on the weight of features. The optimisation technique we used is to perform component-wise experiments to see the effects of features on the final results. Analysis of the weight of features can also potentially improve the model performance. As the number and sparsity of features used in this research varies greatly. For example, bag of words have more than 8000 features which makes up more than 90% of the feature but at the same time they are also sparser than all the other features. So it is necessary and possible to investigate what effect the weights of features will have on the model results.

Classification of wrongly classified labels As stated in the error analysis, the CRF model of sentence level classifies most of the labels into 0+Question-question and 1+Answer-answer. The labels with small amount, such as 1+Question-add, 3+Answer-answer and so forth, are wrongly classified into 1+Answer-answer. So another future research is to explore better techniques to classify the wrongly-classified instances, which may greatly improve the prediction models.

Bibliography

Ang, Jeremy, Yang Liu, and Elizabeth Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *ICASSP (1)*, 1061–1064.

Austin, John Langshaw. 1975. *How to do things with words*. Oxford university press.

Baldwin, Timothy, David Martinez, and Richard B Penman. 2007. Automatic thread classification for linux user forum information access. In *Proceedings of the Twelfth Australasian Document Computing Symposium (ADCS 2007)*, 72–79.

Bhatia, Sumit, Prakhar Biyani, and Prasenjit Mitra. 2012. Classifying user messages for managing web forum data.

——, and Prasenjit Mitra. 2010. Adopting inference networks for online thread retrieval. In *AAAI*, volume 10, 1300–1305.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. " O'Reilly Media, Inc.".

Cong, Gao, Long Wang, Chin-Yew Lin, Young-In Song, and Yueheng Sun. 2008. Finding question-answer pairs from online forums. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 467–474. ACM.

Dhillon, Rajdip, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. 2004. Meeting recorder project: Dialog act labeling guide. Technical report, DTIC Document.

Ding, Shilin, Gao Cong, Chin-Yew Lin, and Xiaoyan Zhu. 2008. Using conditional random fields to extract contexts and answers of questions from online forums. In *ACL*, volume 8, 710–718. Citeseer.

Dridan, Rebecca, and Stephan Oepen. 2012. Tokenization: returning to a long solved problem a survey, contrastive experiment, recommendations, and toolkit. In *Proceedings of the 50th Annual Meeting of the Association for Computational*

Linguistics: Short Papers-Volume 2, 378–382. Association for Computational Linguistics.

Feng, Donghui, Erin Shaw, Jihie Kim, and Eduard Hovy. 2006. Learning to detect conversation focus of threaded discussions. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 208–215. Association for Computational Linguistics.

Finke, Michael, Maria Lapata, Alon Lavie, Lori Levin, Laura Mayfield Tomokiyo, Thomas Polzin, Klaus Ries, Alex Waibel, and Klaus Zechner. 1998. Clarity: Inferring discourse structure from speech. In *Applying Machine Learning to Discourse Processing. Papers from the 1998 AAAI Spring Symposium. Technical Report SS-98-01*, 25–32.

Fortuna, Blaz, Eduarda Mendes Rodrigues, and Natasa Milic-Frayling. 2007. Improving the classification of newsgroup messages through social network analysis. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 877–880. ACM.

Francis, W, and Henry Kucera. 1982. Frequency analysis of english usage.

Jeong, Minwoo, Chin-Yew Lin, and Gary Geunbae Lee. 2009. Semi-supervised speech act recognition in emails and forums. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, 1250–1259. Association for Computational Linguistics.

Kim, J-D, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. 2003. Genia corpusa semantically annotated corpus for bio-textmining. *Bioinformatics* 19.i180–i182.

Kim, Jihie, Grace Chern, Donghui Feng, Erin Shaw, and Eduard Hovy. 2006. Mining and assessing discussions on the web through speech act analysis.

Kim, Su Nam, Lawrence Cavedon, and Timothy Baldwin. 2010a. Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 862–871. Association for Computational Linguistics.

——, Li Wang, and Timothy Baldwin. 2010b. Tagging and linking web forum posts. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, 192–202. Association for Computational Linguistics.

Lui, Marco, and Timothy Baldwin. 2009. You are what you post: User-level features in threaded discourse. In *Proceedings of the 14th Australasian Document Computing Symposium*, 98–105.

Mike, Thelwall, Buckley Kevan, Paltoglou Georgios, and Cai Di. 2010. Sentiment in short strength detection informal text. *JASIST* 61.2544–2558.

Morante, Roser, and Eduardo Blanco. 2012. * sem 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 265–274. Association for Computational Linguistics.

Okazaki, Naoaki, 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Read, Jonathon, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg. 2012. Sentence boundary detection: A long solved problem? In *COLING (Posters)*, 985–994.

Řehůřek, Radim, and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Searle, John R. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. cambridge university press.

Shriberg, Elizabeth, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech communication* 32.127–154.

Soricut, Radu, and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 149–156. Association for Computational Linguistics.

Wanas, Nayer, Motaz El-Saban, Heba Ashour, and Waleed Ammar. 2008. Automatic scoring of online discussion posts. In *Proceedings of the 2Nd ACM Workshop on Information Credibility on the Web*, 19–26. ACM.

Wang, Li, Marco Lui, Su Nam Kim, Joakim Nivre, and Timothy Baldwin. 2011. Predicting thread discourse structure over technical web forums. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 13–25. Association for Computational Linguistics.