Contrasting the Interaction Structure of an Email and a Telephone Corpus: A Machine Learning Approach to Annotation of Dialogue Function Units

Jun Hu
Department of Computer Science
Columbia University
New York, NY, USA
jh2740@columbia.edu

Rebecca J. Passonneau
CCLS
Columbia University
New York, NY, USA
becky@cs.columbia.edu

Owen Rambow
CCLS
Columbia University
New York, NY, USA
rambow@ccls.columbia.edu

Abstract

We present a dialogue annotation scheme for both spoken and written interaction, and use it in a telephone transaction corpus and an email corpus. We train classifiers, comparing regular SVM and structured SVM against a heuristic baseline. We provide a novel application of structured SVM to predicting relations between instance pairs.

1 Introduction

We present an annotation scheme for verbal interaction which can be applied to corpora that vary across many dimensions: modality of signal (oral, textual), medium (e.g., email, voice alone, voice over electronic channel), register (such as informal conversation versus formal legal interrogation), number of participants, immediacy (online versus offline), and so on.[1] We test it by annotating transcribed phone conversations and email threads. We then use three algorithms, two of which use machine learning (including a novel approach to using Structured SVM), to predict labels and links (a generalization of adjacency pairs) on unseen data. We conclude that we can indeed use a common annotation scheme, and that the email modality is easier to tag for dialogue acts, but that it is harder in email to find the links.

2 Related Work

Annotation for dialogue acts (DAs), inspired by Searle and Austin's work on speech acts, arose largely as a means to understand, evaluate and model human-human and human-machine communication. The need for the enterprise derives from the fact that the relationship between lexico-grammatical form (including mood, e.g., interrogative) and communicative actions cannot be enumerated; there are complex dependencies on the linguistic and situational contexts of use. Many DA schemes exist: they can be hierarchical or flat (Popescu-Belis, 2008), can comprise a large (Devillers et al., 2002; Hardy et al., 2003) or small repertoire (Komatani et al., 2005), or can be oriented towards human-human dialogue (Allen and Core, 1997; Devillers et al., 2002; Thompson et al., 1993; Traum and Heeman, 1996; Stolcke et al., 2000) or multi-party interactions (Galley et al., 2004), or human-computer interaction (Walker and Passonneau, 2001; Hardy et al., 2003), including multimodal ones (Thompson et al., 1993; Kruijff-Korbayová et al., 2006).

A major focus of the cited work is on how to recognize or generate speech acts for interactive systems, or how to classify speech acts for distributional analyses. The focus can be on a specific type of speech act (e.g., grounding and repairs (Traum and Heeman, 1996; Frampton and Lemon, 2008)), or on more general comparisons, such as the contrast between human-human and human-computer dialogues (Doran et al., 2001). While there is a large degree of overlap across schemes, the set of DA types will differ due to differences in the nature of the communicative goals; thus information-seeking versus task-oriented dialogues differ in the set of speech acts and their relative frequencies.

Our motivation in providing a new DA annotation scheme is that our focus differs from much of this prior work. We aim for a relatively abstract annotation scheme in order to make comparisons across interactions of widely differing properties. Our initial focus is less on speech act types and more on the patterns of local alternation be-

---

tween an initiating speech act and a responding one–the analog of adjacency pairs (Sacks et al., 1974). The most closely related effort is (Galley et al., 2004), which aims to automatically identify adjacency pairs in the ICSI Meeting corpus, a large corpus of 75 meetings, using a small tagset. Their maximum entropy ranking approach achieved 90% accuracy on the 4-way classification into agreement, disagreement, backchannel and other. Using the switchboard corpus, (Stolcke et al., 2000) achieved good dialogue act labeling accuracy (71% on manual transcriptions) for a set of 42 dialogue act types, and constructed probabilistic models of dialogue act sequencing in order to test the hypothesis that dialogue act sequence information could boost speech recognition performance.

There has been far less work on developing manual and automatic dialogue act annotation schemes for email. We summarize some salient recent work. Carvalho and Cohen (2006) use word n-grams (with extensive preprocrssing) to classify entire emails into a complex ontology of speech acts. However, in their experiments, they concentrate on detecting only a subset of speech acts, which is comparable in size to ours. Speech acts are assigned for entire emails, but several speech acts can be assigned to one email. Apparently, they develop separate binary classifiers for each speech act. Corston-Oliver et al. (2004) are interested in identifying tasks in email. They label each sentence in email with tags from a set which describes the type of content of the sentence (describing a task, scheduling a meeting), but are less interested in the interactive aspect of email communication (creating an obligation to respond).

There has been some work which relates to finding links, but limited to finding question-answer pairs. Shrestha and McKeown (2004) first detect questions using lexical and part-of-speech features, and then find the paragraph that answers the question. They use features related to the structure of the email thread, as well as lexical features. As do we, they find that classifying is easier than linking.

Ding et al. (2008) argue that in order to do well at finding answers to questions, one must also find the context of the question, since it often contains the information needed to identify the answer. They use a corpus of online discussion forums, and use slip-CRFs and two-dimensional

CRFs, models related to those we use. We will investigate their proposal to consider the question context in future work.

While they do not use dialogue act tagging to compare modalities, as we do, Murray and Carenini (2008) compare spoken conversation with email by comparing a common summarization architecture across both modalities. They get similar performance, but the features differ.

Table 1: DFU speech act labels

Request-Information (R-I)
Request-Action (R-A)
Inform (Inf)
Commit (Comm)
Conventional (Conv)
Perform (Perf)
Backchannel (Bch) (+/- Grounding)
Other

3 Annotation Scheme

Figure 1: Example DFU illustrating the relation of extent (segmentation) to speech act type

M1.2 I have completed the invoices for April, May and June
M1.3 and we owe Pasadena each month for a total of $3,615,910.62.
M1.4 I am waiting to hear back from Patti on May and June to make sure they are okay with her.
[**Inform(1.2-1.4)**: *status of Pasadena invoicing- completed & pending approval – versus amount due*]
Sflink(1.2-1.4)
M2.1 That's fine.
[**Inform(2.1)**: *acknowledgement of status of Pasadena invoicing*]
Blink(1.2-1.4)

The annotation scheme presented here consists of Dialogue Function Units (DFUs), which are intended to represent abstract units of interaction. The last two authors developed the annotation on three contrasting corpora: email threads, telephone conversations, and court transcripts. It builds on our previous work in intention-based segmentation (Passonneau and Litman, 1997), and on mixing a formal schema with natural language descriptions (Nenkova et al., 2007). In this

paper, we investigate the modalities of telephone two-person conversation in a library setting, and multi-party email in a workplace setting. Our initial focus is on the structure of turn-taking. By using a relatively abstract annotation scheme, we can compare and contrast this behavior across different types of interaction.

Our unit of annotation is the DFU. DFUs have an extent, a dialogue act (DA) label along with a description, and possibly one or more forward and/or backward links. We explain each component of the annotation in turn. We use the example in Figure 1; the example is drawn from actual messages, but has been modified to yield a more succinct example.

The extent of a DFU roughly corresponds to that portion of a turn (conversational turn; email message; etc.) that corresponds to a coherent communicative intention. Because we do not address automatic identification of the segmentation into DFU units in this paper, we do not discuss how annotators are instructed to identify extent.

As illustrated in Figure 1, the communicative function of a DFU is captured by a speech act type, and a natural language description. This is somewhat analogous to the natural language descriptions associated with Summary Content Units (SCUs) in pyramid annotation (Nenkova et al., 2007), or with the intention-based segmentation of (Passonneau and Litman, 1997). The purpose in all cases is to require annotators to articulate briefly but specifically the unifying intention (Passonneau and Litman, 1997), semantic content (Nenkova et al., 2007), or speech act. We use the eight dialogue act types listed in the upper left of Table 1. To accommodate discontinuous speech acts, due to the interruptions that are common to conversation, each speech act can have an operator affix such as "-Continue". We have previously shown (Passonneau and Litman, 1997) that intention-based segmentation can be done reliably by multiple annotators. For twenty narratives each segmented by the same seven annotators, using Cochran's $\mathcal{Q}$ (Cochran, 1950), we found the probabilities associated with the null hypothesis that the observed distributions could have arisen by chance to be at or below p=0.1 $\times 10^{-6}$. Partitioning $\mathcal{Q}$ by number of annotators gave significant results for all values of $\mathcal{A}$ ranging over the number of annotators apart from $\mathcal{A} = 2$. We would expect similar patterns of agreement on DFU segmen-

tation, but have not collected segmentation data from multiple annotators on the two corpora presented here.

DFU Links, or simply Links, correspond to adjacency pairs, but need not be adjacent. A forward link (Flink) is the analog of a "first pair-part" of an adjacency pair (Sacks et al., 1974), and is similarly restricted to specific speech act types. All Request-Information and Request-Action DFUs are assigned Flinks. The responses to such requests are assigned a backward link (Blink). In principle, a response can be any of the speech act types, thus it can be an answer to a question (Inform), a rejection of a Request-Action or a commitment to take the requested action (Commit), a request for clarification (Request-Information), and so on. In most but not all cases, requests are responded to, thus most Flinks and Blinks come in pairs. We refer to Flinks with no matching Blink as dangling links. If an utterance can be interpreted as a response to a preceding DFU, it will get a Blink even where the preceding DFU has no Flink. The preceding DFU taken to be the "first pair-part" of the Link will be assigned a secondary forward link (Sflink). All links except dangling links are annotated with the address of the DFU from which they originate. Figure 1 illustrates an email message (M2) containing a single sentence ("That's fine") that is a response to a DFU in a prior email (M1), where the prior email had no Flink because it only contains Inform DAs; thus M1 gets an Sflink.

4 Corpora

The Loqui corpus consists of 82 transcribed dialogues from a larger set of 175 dialogues that were recorded at New York City's Andrew Heiskell Braille and Talking Book Library during the summer of 2005. All of the transcribed dialogues pertain to one or more book requests. Forty-eight dialogues were annotated; the annotators worked from a combination of the transcription and the audio. Three annotators were trained together, annotated up to a dozen dialogues independently, then discussed, adjudicated and merged ten of them. During this phase, the annotation guidelines were refined and revised. One of the three annotators subsequently annotated 38 additional dialogues.

We also annotated 122 email threads of the Enron email corpus, consisting of email messages in the inboxes and outboxes of Enron corporation

Table 2: Distributional Characteristics of Dialogue Acts in Enron and Loqui

	Loqui		Enron	
Words	21097		17924	
DFUs	3845		1400	
Speech Act Labels				
Inform	1928	50%	853	61%
Request-Inf.	761	20%	149	11%
Request-Action	39	1%	37	3%
Commit	338	9%	3	0%
Conventional	254	7%	356	25%
Backchannel	507	13%	0	0
Other	18	0%	2	0%
Total	3845	100%	1400	100%
Links				
Paired Links	1204	63%	193	28%
Flink/Blink	702	58%	83	43%
Sflink/Blink	502	42%	110	57%
Dangling Links	90	2%	97	7%
Mutliple Blinks	4	0%	4	0%
Links by Speech Act Labels				
Inform	1003	83%	142	74%
Request-Inf.	170	14%	44	23%
Request-Action	1	0%	5	3%
Commit	13	1%	2	1%
Conventional	2	0%	0	0
Backchannel	15	1%	0	0
	1204	100%	193	100%

employees. Most of the emails are concerned with exchanging information, scheduling meetings, and solving problems, but there are also purely social emails. We used a version of the corpus with some missing messages restored from other emails in which they were quoted (Yeh and Harnly, 2006). The annotator of the majority of the Loqui corpus also annotated the Enron corpus. She received additional training and guidance based on our experience with a pilot annotator who helped us develop the initial guidelines.

Table 2 illustrates differences between the two corpora. The DFUs in the Loqui data are much shorter, with 5.5 words on average compared with 12.8 words in Enron. The distribution of DFU labels shows a similarly high proportion of Inform acts, comprising 50% of all Loqui DFUs and 61% of all Enron DFUs. Otherwise, the distributions are quite distinct. The Loqui interactions are all two party telephone dialogues where the callers

(library patrons) tend to have limited goals (requesting books). The Enron threads consist of two or more parties, and exhibit a much broader range of communicative goals. In the Loqui data, backchannels are relatively frequent (13%) but do not occur in the email corpus for obvious reasons. There are some Commits (9%), typically reflecting cases where the librarian indicates she will send requested items to the caller by mail, or place them on reserve. There are no Commits in the Enron data. Neither corpus has many Request-Actions; the Loqui corpus has many more requests for information, which includes requests made by the librarian, e.g., for the patrons' identifying information, or by the caller.

The most striking differences between the two corpora pertain to the distribution of DFU Links. In Loqui, 63% of the DFUs are the first pair-part or the second pair-part of a Link compared with 28% in Enron. In Loqui, the majority of Links are initiated by overt requests (58% of Links are Flink/Blink pairs), whereas in Enron, the majority of Links involve SFlinks (57%). There are relatively few dangling Links in either dataset, with more than three times as many in Enron (7% versus 2% in Loqui). Most of the DFU types in the second pair-part of Links are Informs and Request-Information, with a different proportion in each dataset. In Loqui, 83% of DFUs that are second pair-part of a Link are Informs compared with 74% in Enron; correspondingly, only 14% of DFUs in Links are Request-Information in Loqui versus 23% in Enron.

5 Dialogue Act Tagging and Link Prediction

There are two machine learning tasks in our problem. The first is Dialogue Act (DA) Tagging, in which we assign DAs to every Dialogue Functional Unit (DFU). The second is Link prediction, in which we predict if two DFUs form a link pair. In this paper, we assume that the DFUs are given. We propose three systems to tackle the problem. The first system is a non-strawman Baseline Heuristics system, which uses the structural characteristics of dialogue. The second is Regular SVM. The third is Structured SVM. Structured SVM is a discriminative method that can predict complex structured output. Recently, discriminative Probabilistic Graphical Models have been widely applied in structural problems (Getoor and

Taskar, 2007) such as link prediction. However, Structured SVM (Taskar et al., 2003; Tsochantaridis et al., 2005) is also a compelling method which has the potential to handle the interdependence between labeling and sequencing, due to its ability to handle dependencies among features and prediction results within the structure. sequence labeling (Tsochantaridis et al., 2005). We have adapted Structured SVM to our problem, provided a novel method for link prediction, and shown that it is superior in some aspects to Regular SVM.

5.1 Features

We have two sets of features. DFU features are associated with a particular DFU, and link features describe the relationship between two DFUs. DFU features are used in both tasks. Link features are only used in link prediction. The feature vector of a link contains two sets of DFU features and the link features that are defined over the two DFUs. Table 3 gives the features we used, which are almost identical for both corpora, so we could compare the performance.

Because a lot of Flinks are questions, we chose some features that are tailored to Question-Answer detection, such as presence of a question mark. Dialogue fillers and acceptance words affect the accuracy of Part-Of-Speech tagging. On the other hand, they are helpful indicators of disfluency or confirmation. So we hand-picked a list of filler and acceptance words, removed them from the sentence, and added features counting their occurrences.

5.2 Baseline Heuristics

Dialogue Act Tagging We use the most frequent DA as the heuristic for prediction. In both Enron and Loqui, this DA is Inform.

Link Prediction In link prediction, the heuristics for Enron and Loqui corpora are different due to structural differences. In Loqui, whenever we see a DFU with a Forward Link (DA is Request-Information or Request-Action), we predict that the target of the link is the first following DFU that is available and acceptable. "Available" means that the second DFU has not been assigned a Backward Link yet. "Acceptable" means that the second DFU has a DA that is very frequent in a Backward Link and it is of a different speaker to the first DFU. We enforce similar constraints in Enron corpus for link prediction, except that the second

Table 3: DFU features (E: Enron, L: Loqui)

Structural for DA prediction	
E,L	First three POS
E,L	Relative Position in the Dialogue
E	Existence of Question Mark
E,L	Does the first POS start with "w" or "v"
E,L	Length of the DFU
E	Head, body, tail of the Message
E,L	Dialogue Act (Only used in link prediction)
Lexical for DA prediction	
E,L	Bag of Words
E,L	Number of Content Words
L	Number of Filler Words, as "uh", "hmm"
E,L	Number of Acceptance Words, as "yes"
Structural for Link prediction	
E,L	The distance between two DFUs
Lexical for Link prediction	
E,L	Overlapping number of content words

DFU not only has to be from a different author, but also has to be in a message which is a direct descendant in the reply chain of the message that contains the first DFU. The baseline link prediction algorithm uses the DAs as predicted by the Regular SVM. If we used the baseline DA prediction, the result would be too low to make a valid comparison against other systems in terms of link prediction because all DAs would be identical.

5.3 Regular SVM

We have used the Yamcha support vector machine package (chasen.org/∼taku/software/yamcha/). The advantage of Yamcha is that it extends the traditional SVM by enabling using dynamically generated features such as preceding labels.

Dialogue Act Tagging We use the feature vector of the current DFU as well as the predicted DA of the preceding DFU as features to predict the DA of the current DFU.

Link Prediction First, in order to limit search space, we specify a certain window size to produce a space **S** of DFU pairs under consideration. For a particular DFU, we look at all succeeding DFUs and check if these two DFUs satisfy the following constraint: in Loqui, they must be of different speakers; in Email, one must be another's ancestor and they must be of different authors. We consider all valid pairs starting from the current DFU until

the number of considered valid pairs reaches the window size. Then we proceed to the next DFU and collect more DFU pairs into our consideration space.

Second, we train a link binary classifier with all DFU pairs in this consideration space along with a binary classification correct/not correct as training data. This classifier takes the feature vectors of the two DFUs as well as the link features such as the distance between these two DFUs as features.

Third, we apply a greedy algorithm to generate links in the test data with the binary classifier. The algorithm firstly uses the classifier to generate scores for all DFU pairs in the consideration space of the test data, then it scans the dialogue sequentially, checks all preceding DFUs that are allowed to link to the current DFU (i.e., the DFU pair is in the consideration space), and assigns corresponding links to the most likely DFU pair. We impose a restriction that there can be at most one Flink, one Sflink and one Blink for any given DFU.

5.4 Structured SVM

A Structured SVM is able to predict complex output instead of simply a binary result as in a regular SVM. There are several variants. We have followed the margin-rescaling approach (Tsochantaridis et al., 2005), and implemented our systems using SVM^{python}, which is a python interface to the SVM^{struct} package (svmlight.joachims.org/svm_struct.html). Generally, Structured SVM learns a discriminant function $F : \mathbf{X} \times \mathbf{Y} \to \mathbf{R}$, which estimates a score of how likely the output $\mathbf{y}$ is given the input $\mathbf{x}$. Crucially, $\mathbf{y}$ can be a complex structure. Section A in the appendix; here, we summarize the main intuitions.

Dialogue Act Tagging The input $\mathbf{x}$ is a sequence of DFUs, and $\mathbf{y}$ is the corresponding sequence of DAs to predict. Compared to Regular SVM, instead of predicting $\mathbf{y}^t$ one at a time, Structured SVM optimizes the sequence as a whole and predicts all labels simultaneously. Due to the similarity to HMM, the maximization problem is solved by the Viterbi algorithm (Tsochantaridis et al., 2005).

Link Prediction The input now contains the DFU sequence, a link consideration space, as well as a label sequence, which we get from the previous stage. The output structure chooses among the possible links in the link consideration space,

such that there is at most one Flink/SFlink or Blink for any given DFU, and that there are no crossing links. (Note that all the constraints are only enforced in training and prediction; in testing, we compare results against the complete manual annotations which do not follow these constraints.) Then the maximization problem can be solved by a straightforward dynamic programming algorithm.

Table 4: Result of DA prediction

	Baseline	Regular	Struct
Loqui	50.14%	68.30%	70.26%
Enron	60.93%	88.34%	88.71%

Note: Structured SVM parameters for Loqui are $\mathbf{C} = 300$, $\alpha = 1$; Structured SVM parameters for Enron are $\mathbf{C} = 1000$, $\alpha = 1$.

6 Experiments

We have three hypotheses for our experiments:

Hypothesis 1 Link prediction is harder than Dialogue Act prediction.
Hypothesis 2 Enron is harder than Loqui.
Hypothesis 3 Structured SVM is better than Regular SVM, and Baseline is the worst.

We have applied the algorithm described in Section 5 to both the Enron and Loqui corpora. The data set is annotated with DFUs; we focus on the DA labels and Links. As discussed before, every system is a pipeline that would preprocess the data into separate DFUs, predict the Dialogue Acts, and then feed the Dialogue Acts into the link prediction algorithm. The size of the data set is shown in Table 2. We do five-fold cross-validation.

Table 4 shows the accuracy of three systems on Enron and Loqui. Structured SVM has a clear lead to Regular SVM in Loqui; but the advantage is less clear in Enron. Tables 6 and 7 give detailed results of DA prediction.We do not show DAs that do not exist in the corpora, or that were not predicted by the algorithms. Both Regular SVM and Structured SVM performed consistently for the two corpora.

Table 5 gives Link prediction results. Note that when we compute the combined result for both types of links, we are only concerned with the Link position. The seperate results for Flink/Blink and Sflink/Blink require us to identify the types of links first, so here we not only compare the position of predicted links against the gold, but also require predicted DAs to indicate the link type (e.g., the DA of the first DFU must be Request-

Table 5: Link Prediction for Enron and Loqui

Enron	Baseline			Regular			Struct		
	R	P	F	R	P	F	R	P	R
Paired Links	16.66%	40%	23.52%	18.75%	55.38%	28.01%	31.25%	39.47%	34.88%
Flink/Blink	32.53%	33.75%	33.13%	26.50%	61.11%	36.97%	34.93%	47.54%	40.27%
Sflink/Blink	0.0%	0.0%	0.0%	11.92%	44.82%	18.83%	22.93%	27.47%	25.00%
Loqui									
Paired Links	30%	56.15%	39.11%	43.59%	60.60%	50.71%	44.15%	56.02%	49.38%
Flink/Blink	43.30%	46.47%	44.83%	40.58%	57.73%	47.66%	43.55%	60.04%	50.48%
Sflink/Blink	0.0%	0.0%	0.0%	21.76%	29.36%	25.00%	22.88%	26.24%	24.45%

Note: Structured SVM parameters for Enron are $C = 2000$, $\beta = 2.$, for Loqui $C = 1000$, $\beta = 4$.

Information or Request-Action to qualify as a Flink/Blink).

Table 6: Recall/Precision/F-measure of DA prediction for Loqui (in %)

	Regular			Struct		
	P	R	F	P	R	F
R-A	50.0	51.7	50.9	43.3	43.3	43.3
R-I	51.3	61.1	55.8	52.3	71.2	60.3
Inf	73.9	73.0	73.5	76.9	74.1	75.5
Bch	65.3	51.7	57.7	65.1	53.6	58.8
Com	5.6	33.3	9.5	5.6	33.3	9.5
Conv	81.2	84.0	82.6	83.7	83.3	83.5

Table 7: Recall/Precision/F-measure of DA prediction for Enron (in %)

	Regular			Struct		
	R	P	F	R	P	F
R-A	27.8	55.6	37.0	25.0	75.0	37.5
R-I	77.9	82.3	80.0	77.2	83.3	80.1
Inf	92.5	90.6	91.5	92.1	91.2	91.7
Conv	90.5	87.3	88.9	93.4	85.6	89.3

7 Discussion

Hypothesis 1 The result of DA prediction is drastically better than link prediction. There are usually indicators of DA types such as "thank you" for Conventional, so learning algorithms could easily capture them. But in link prediction, we frequently need to handle deep semantic inference and sometimes useful information exists in the surrounding context rather than the DFU itself. Both of these scenarios imply that in order to predict links or relationships better, we need more sophisticated features.

Hypothesis 2 This hypothesis turns out to be half-correct. The DA prediction accuracy for Enron is better than that of Loqui. The higher percentage of Inform and less diversity of DAs in Enron (See Appendix for statistics) may be part of the reason. Another possible explanation is that as a set of spoken dialogue data, Loqui is inherently more difficult to process than written form, since some common tasks such Part-Of-Speech tagging have lower accuracy for spoken data. On the other hand, the result of link prediction did confirm our hypothesis. The first reason is that there are far fewer links in Enron than in Loqui, so we have less training data. The tree structure of the reply chain in the email threads also makes prediction more difficult. And the link distance is longer, because in email, people can respond to a very early message, while in a phone conversation, people tend to respond to immediate requests.

Hypothesis 3 Both SVM models perform better than the baseline. Generally, Structured SVM performs better than Regular SVM, especially in link prediction for Enron. This confirms the advantage of using Structured SVM for output involving inter-dependencies. The only exception is the Sflink prediction in Loqui, which in turn affects the overall accuracy of link prediction.

References

James Allen and Mark Core. 1997. Damsl: Dialogue act markup in several layers. http://www.cs.rochester.edu/research/cisd/resources/damsl.

Vitor Carvalho and William Cohen. 2006. Improving "email speech acts" analysis via n-gram selection. In *Proceedings of the Analyzing Conversations in Text and Speech*.

William G. Cochran. 1950. The comparison of percentages in matched samples. *Biometrika*, 37:256–266.

Simon Corston-Oliver, Eric Ringger, Michael Gamon, and Richard Campbell. 2004. Task-focused summarization of email. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*.

Laurence Devillers, Sophie Rosset, Bonneau-Helene Maynard, and Lamel Lori. 2002. Annotations for dynamic diagnosis of the dialog state. In *LREC*.

Shilin Ding, Gao Cong, Chin-Yew Lin, and Xiaoyan Zhu. 2008. Using conditional random fields to extract contexts and answers of questions from online forums. In *Proceedings of ACL-08: HLT*, Columbus, Ohio.

Christine Doran, John Aberdeen, Laurie Damianos, and Lynette Hirschman. 2001. Comparing several aspects of human-computer and human-human dialogues. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue*.

Matthew Frampton and Oliver Lemon. 2008. Using dialogue acts to learn better repair strategies for spoken dialogue systems. In *ICASSP*.

Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. 2004. Identifying agreement and disagreement in conversational speech: use of Bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 669–676.

Lise Getoor and Ben Taskar, editors. 2007. *Introduction to Statistical Relational Learning*. The MIT Press.

Hilda Hardy, Kirk Baker, Bonneau-Helene Maynard, Laurence Devillers, Sophie Rosset, and Tomek Strzalkowski. 2003. Semantic and dialogic annotation for automated multilingual customer service. In *Eurospeech/Interspeech*.

Kazunori Komatani, Nayouki Kanda, Tetsuya Ogata, and Hiroshi G. Okuno. 2005. Contextual constraints based on dialogue models in database search task for spoken dialogue systems. In *Eurospeech*.

Ivana Kruijff-Korbayová, Tilman Becker, Nate Blaylock, Ciprian Gerstenberger, Michael Kaisser, Peter Poller, Verena Rieser, and Jan Schehl. 2006. The Sammie corpus of multimodal dialogues with an mp3 player. In *LREC*.

Gabriel Murray and Giuseppe Carenini. 2008. Summarizing spoken and written conversations. In *EMNLP*.

Ani Nenkova, Rebecca J. Passonneau, and Kathleen McKeown. 2007. The pyramid method: incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, 4(2).

Rebecca J. Passonneau and Diane J. Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1).

Andrei Popescu-Belis. 2008. Dimensionality of dialogue act tagsets: An empirical analysis of large corpora. *LREC*, 42(1).

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systemics for the organization of turn-taking for conversation. *Language*, 50(4).

Lokesh Shrestha and Kathleen McKeown. 2004. Detection of question-answer pairs in email conversations. In *COLING*.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykena, and Meteer Marie. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *International Journal of Computational Linguistics*, 26(3).

Ben Taskar, Crlos Guestrin, and Daphne Koller. 2003. Max-margin markov networks. In *NIPS*.

Henry S. Thompson, Anne H. Anderson, Ellen Gurman Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. 1993. The HCRC map task corpus: Natural dialogue for speech recognition. In *Proceedings of the DARPA Human Language Technology Workshop*.

David Traum and Peter Heeman. 1996. Utterance units and grounding in spoken dialogue. In *Interspeech/ICSLP*.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. 2005. Large margin methods for structured and interdependent output variables. *JMLR*, 6.

Marilyn A. Walker and Rebecca Passonneau. 2001. Date: A dialogue act tagging scheme for evaluation of spoken dialogue systems. In *HLT*.

Jen-Yuan Yeh and Aaron Harnly. 2006. Email thread reassembly using similarity matching. In *Conference on Email and Anti-Spam*.

A   Appendix: Structured SVM

This section provides mathematical background for Secton 5.4. The hypothesis function is given by:

$$f(\mathbf{x}, \mathbf{w}) = argmax_{\mathbf{y} \in \mathbf{Y}} F(\mathbf{x}, \mathbf{y} : \mathbf{w})$$

And in addition, we assume F to be linear to a joint feature map $\Psi(\mathbf{x}, \mathbf{y})$.

$$F(\mathbf{x}, \mathbf{y} : \mathbf{w}) = \langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{y}) \rangle$$

We also define a loss function $\Delta(\mathbf{y}, \overline{\mathbf{y}})$ which defines the deviation of the predicted output $\overline{\mathbf{y}}$ to the correct output.

As a result, given a sequence of training examples,$(\mathbf{x}_1, \mathbf{y}_1) \cdots (\mathbf{x}_n, \mathbf{y}_n) \in \mathbf{X} \times \mathbf{Y}$, the function we need to optimize becomes:

$$min_{\mathbf{w}, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{n} \sum_{i=1}^{n} \xi_i$$

$s.t. \ \forall i \forall \mathbf{y} \in \mathbf{Y} \backslash \mathbf{y}_{(i)} : \langle \mathbf{w}, \delta \Psi_i(\mathbf{y}) \rangle > \Delta(\mathbf{y}_{(i)}, \mathbf{y}) - \xi_{\mathbf{i}}$ where,

$$\langle \mathbf{w}, \delta \Psi_i(\mathbf{y}) \rangle = \left\langle \mathbf{w}, \Psi(\mathbf{x}_{(i)}, \mathbf{y}_{(i)}) - \Psi(\mathbf{x}_{(i)}, \mathbf{y}) \right\rangle$$

$\mathbf{w}$ is optimized towards maximizing the margin between the true structured output $\mathbf{y}$ and any other suboptimal configurations for all training instances.

A cutting plane optimization algorithm is implemented in SVM^{struct}. However, for any problem, we need to implement the feature map $\Psi(\mathbf{x}, \mathbf{y})$, the loss function $\Delta(\mathbf{y}, \overline{\mathbf{y}})$, and a maximization problem which enables the cutting plane optimization, i.e.

$$\overline{\mathbf{y}} = argmax_{\mathbf{y} \in \mathbf{Y}} \Delta(\mathbf{y}_{(i)}, \mathbf{y}) - \langle \mathbf{w}, \delta \psi_i(\mathbf{y}) \rangle$$

Only certain feature maps that would make solving this maximization effectively, usually by dynamic programming, could be handled this way.

For **Dialogue Act Tagging**, let $\mathbf{x} = \left(\mathbf{x}^1, \mathbf{x}^2 \ldots \mathbf{x}^T \right)$ be the sequence of DFUs, and $\mathbf{y} = \left(\mathbf{y}^1, \mathbf{y}^2 \ldots \mathbf{y}^T \right)$ the corresponding sequence of dialogue acts. $\phi(\mathbf{x}^t)$ represents the DFU features and $\phi(\mathbf{x}^t) \in R^D$. $\mathbf{y}^t \in L = \{l_1, \ldots, l_K\}$ where L contains the set of available DAs. The feature map is (Tsochantaridis et al., 2005):

$$\Psi(\mathbf{x}, \mathbf{y}) = \begin{pmatrix} \sum_{t=1}^{T} \phi(\mathbf{x}^t) \otimes \mathbf{\Lambda}(\mathbf{y}^t) \\ \mathbf{\Lambda}(\mathbf{y}^{t-1}) \otimes \mathbf{\Lambda}(\mathbf{y}^t) \end{pmatrix}$$

where $\mathbf{\Lambda}(\mathbf{y}^t) = [\lambda(l_1, \mathbf{y}), \ldots, \lambda(l_k, \mathbf{y})]$ and λ is an indicator function that returns 1 if two parameters are equal. $\otimes$-operator is defined as:

$\mathbf{R}^D \times \mathbf{R}^K \rightarrow \mathbf{R}^{D \cdot K}, \ (\mathbf{a} \otimes \mathbf{b})_{i+(j-1)D} \equiv \mathbf{a}_i \cdot \mathbf{b}_j$

In analogy to an HMM, the lower part in $\Psi(\mathbf{x}, \mathbf{y})$ encodes the histogram of adjacent DA transitions in $\mathbf{y}$; the upper part encodes the DA emissions from a specific label to one dimension in the DFU feature vector. Hence, the total number of dimensions in $\Psi(\mathbf{x}, \mathbf{y})$ is $K^2 + DK$. As a result, $F(\mathbf{x}, \mathbf{y} : \mathbf{w}) = \langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{y}) \rangle$ gives a global score based on all transitions and emissions in the sequence, which captures the dependecies among nearby labels and mimics the behaviour of an HMM. Figure 2 gives an example of how to compute the feature map.

The loss function is the sum of all zero-one losses across the sequence, i.e.

$$\Delta(\mathbf{y}, \overline{\mathbf{y}}) = \alpha \sum_{t=1}^{T} \lambda(\mathbf{y}^t, \overline{\mathbf{y}}^t)$$

α denotes a cost assigned to every DA loss.

For **Link Prediction**, the input contains the DFU sequence $\mathbf{x}$, a link consideration space $s = \{(i, j) :, \text{DFU } i \text{ and } j \text{ is being considered}\}$, as well as label sequence $\mathbf{y}$ which we get from the previous stage. $\varphi(\mathbf{x}^i, \mathbf{x}^j)$ is the link feature defined over two DFUs. Let the dimension of link feature be B. The output structure $\mathbf{u} = \left\{ \mathbf{u}^1, \mathbf{u}^2 \ldots \mathbf{u}^T \right\}$ specifies the link plan. $\mathbf{u}^t$ denotes that there is a link from DFU $t - \mathbf{u}^t$ to t with the exception that $\mathbf{u}^t$ being zero denotes there is no link pointing to t. The setup of u constraints that there can be at most one Flink/SFlink or Blink for any given DFU. In addition $\mathbf{u}$ is also subject to the constraint that all specified links must be in the link consideration space.

The discriminant function becomes $F : \mathbf{X} \times \mathbf{Y} \times \mathbf{S} \times \mathbf{U} \rightarrow \mathbf{R}$. Similar to structured DA prediction, the discriminant function should give a global evaluation as to how likely is the link plan specified by $\mathbf{U}$ with respect to all the input vectors. Our solution is to decompose the score, and correspondingly, the feature representation into two components, link emission and no-link emission; the details can be found in Figure 3 in the appendix and an example is in Figure 2.

Similarly, we could define the loss function as the sum of all zero-one losses across the sequence , i.e.

$$\Delta(\mathbf{u}, \overline{\mathbf{u}}) = \beta \sum_{t=1}^{T} \lambda(\mathbf{u}^t, \overline{\mathbf{u}}^t)$$

β denotes a cost assigned to every Link loss.

Figure 2: A full example of feature map for Structured SVM

$$
\mathbf{x}^1 = \text{"are you you sure"} \qquad \mathbf{\Psi}_{da} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 2 \\ 1 \end{pmatrix} \begin{array}{l} \text{Inform to Inform} \\ \text{Inform to Req-Info} \\ \text{Req-Info to Inform} \\ \text{Req-Info to Inform} \\ \\ \text{Inform with "are"} \\ \text{Inform with "you"} \\ \text{Inform with "sure"} \\ \\ \text{Req-Info with "are"} \\ \text{Req-Info with "you"} \\ \text{Req-Info with "sure"} \end{array}
$$

$\mathbf{x}^1 = \text{"are you you sure"}$
$\mathbf{x}^2 = \text{"sure"}$

$\mathbf{y}^1 = \text{"Req-Info"}$
$\mathbf{y}^2 = \text{"Inform"}$

$\mathbf{u}^1 = 0$
$\mathbf{u}^2 = 1$

$\phi(\mathbf{x}^1) = (1, 2, 1)$
$\phi(\mathbf{x}^2) = (0, 0, 1)$

$\varphi(\mathbf{x}^1, \mathbf{x}^2) = (1, 1)$

$$
\mathbf{\Psi}_{link} = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 0 \\ 1 \\ \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ \\ 1 \\ 1 \\ \\ 1 \\ 2 \\ 1 \\ 0 \\ 1 \end{pmatrix} \begin{array}{l} \text{1st link pair-part with "are"} \\ \text{1st link pair-part with "you"} \\ \text{1st link pair-part with "sure"} \\ \text{1st link pair-part with Inform} \\ \text{1st link pair-part with Req-Info} \\ \\ \text{2nd link pair-part with "are"} \\ \text{2nd link pair-part with "you"} \\ \text{2nd link pair-part with "sure"} \\ \text{2nd link pair-part with Inform} \\ \text{2nd link pair-part with Req-Info} \\ \\ \text{distance of link} \\ \text{overlap of link} \\ \\ \text{No-Link with "are"} \\ \text{No-Link with "you"} \\ \text{No-Link with "sure"} \\ \text{No-Link with Inform} \\ \text{No-Link with Req-Info} \end{array}
$$

Note: In this example, $\phi(\mathbf{x}^t)$ extracts the bag-of-words features from $\mathbf{x}^t$. "are","you","sure" are the 1st, 2nd and 3rd DFU feature respectively. $\varphi(\mathbf{x}^i, \mathbf{x}^j)$ extracts the distance and number of the overlap content, which are the link features, from the 1st and 2nd pair-part in a DFU link pair. There is a link from DFU 1 to DFU 2 as specified by $j - u^j = i$, but there is no link pointing to DFU 1.

Figure 3: The feature map of link prediction for the structured SVM

$$
\mathbf{\Psi}_L =
$$
$$
\begin{pmatrix}
\sum_{i=1}^{T-1} \sum_{j=i+1}^{T} \phi(\mathbf{x}^i) \lambda(i, j - \mathbf{u}^j) \\
\sum_{i=1}^{T-1} \sum_{j=i+1}^{T} \mathbf{\Lambda}(\mathbf{y}^i) \lambda(i, j - \mathbf{u}^j) \\
\sum_{i=1}^{T-1} \sum_{j=i+1}^{T} \phi(\mathbf{x}^j) \lambda(i, j - \mathbf{u}^j) \\
\sum_{i=1}^{T-1} \sum_{j=i+1}^{T} \mathbf{\Lambda}(\mathbf{y}^j) \lambda(i, j - \mathbf{u}^j) \\
\sum_{i=1}^{T-1} \sum_{j=i+1}^{T} \varphi(\mathbf{x}^i, \mathbf{x}^j) \lambda(i, j - \mathbf{u}^j)
\end{pmatrix}
$$
$$
\mathbf{\Psi}_{NL} = \begin{pmatrix} \sum_{i=1}^{T} \phi(\mathbf{x}^i) \lambda(0, \mathbf{u}^i) \\ \sum_{i=1}^{T} \mathbf{\Lambda}(\mathbf{y}^i) \lambda(0, \mathbf{u}^i) \end{pmatrix}
$$
$$
\mathbf{\Psi}(\mathbf{x}, \mathbf{y}, \mathbf{s}, \mathbf{u}) = \begin{pmatrix} \mathbf{\Psi}_L \\ \mathbf{\Psi}_{NL} \end{pmatrix}
$$

Note: $\mathbf{\Psi}_L$ and $\mathbf{\Psi}_{NL}$ correspond to the link and no-link emissions in the feature map $\mathbf{\Psi}(\mathbf{x}, \mathbf{y}, \mathbf{s}, \mathbf{u})$ respectively as shown in the equations. The total dimension of the feature map is $3D + 3K + B$.