IDN Dialogue Act Classification With Conditional Random Field and Recurrent Neural Network

Christopher Koenig
Stanford University
koenig97@stanford.edu

Homero Roman Roman
Stanford University
homero@stanford.edu

Amanda Lim
Stanford University
aplim@stanford.edu

Abstract

Recent research in the field of dialogue act classification has made significant progress through integrating discourse-level context dependencies with deep learning approaches. This study seeks to translate the success of such recent work, conducted on the Switchboard Dialogue Act Corpus dataset, to the Interaction Dynamics Notation dataset developed at the Stanford Center for Design Research. We explore the use of a Conditional Random Field for this task in addition to a LSTM RNN. Ultimately, we find the CRF achieves best performance, setting a new standard of accuracy on the IDN dataset.

1 Introduction

Designing as a team is one of the most critical tasks for companies and organizations, yet it is also a very difficult process to understand given the multidimensional facets of a design meeting. Just a few of these facets that have been studied in the past include gestures (Tang, 1991), questions (Eris, 2003), emotions (Leifer and Steinert, 2011), sketching (Van der Lugt, 2005), and team composition (Kim and Kang, 2003). Continuing efforts to better understand the complex design process, the Stanford Center for Design Research has developed the Interaction Dynamics Notation (IDN) as a way of capturing the moment to moment interpersonal team dynamics in a design meeting. The labels for IDN are also known as dialogue acts (DA), such as those developed for the SWDA and MRDA datasets. Because IDN labels provide a tangible method of tracking team interaction dynamics during a meeting, they can be used to study patterns of interaction that distinguish high performing teams from low performing teams.

However, labeling of IDN transcripts is currently conducted manually and is therefore a long and labor-intensive process. In this study, we seek to build a classifier for IDN dialogue acts given the transcripts from the IDN corpus. Automatic classification of IDN dialogue acts could enable the development of systems that analyze and provide feedback on design team interactions in order to facilitate more effective teams. Inspired by the recent performance of discourse-level context dependent models on the SWDA dataset, we explore two approaches, one based on a Conditional Random Field (CRF) and one based on a Long Short-Term Memory (LSTM) RNN. These models hold great promise for utilizing discourse-level context dependencies due to their ability to incorporate sequential information into their DA predictions.

2 Related Work

While a great deal of past work has been conducted on the MRDA and SWDA datasets, almost none of this work has involved approaches with CRFs. We first review the only pre-existing dialogue act classification study with a CRF approach, proceed to cover the recent research gains made through discourse-level context dependent deep learning models, and close by examining past performance achieved on the IDN corpus.

2.1 Segmentation and Classification of Dialog Acts Using Conditional Random Fields

Zimmermann (2009) investigates the use of CRFs for the joint segmentation and classification of dialogue acts. He exclusively bases the joint segmentation and classification on features directly produced from an available speech to text system and ran his experiments using the MRDA corpus. The results from the study are not directly applicable to comparisons with other dialogue act classification systems as the metrics incorporate combined performance on segmentation and classification (an instance is only considered correctly classified if it is both correctly segmented and assigned the correct dialogue act). However, the conceptual simplicity of the model combined with its power to capture contextual information between words yielded better performance than previous joint segmentation and classification approaches.

2.2 Discourse-Level Context Dependent Deep Learning Models

Kalchbrenner and Blunsom (2013) engineered a new approach to dialogue act classification through the implementation of two models working in conjunction to model the two levels of compositionality in a dialogue - the individual sentence level and the discourse overall. Specifically, sentences in a DA dataset were passed through a convolutional neural network (CNN) that output semantic vector representations. These representations were then fed into the discourse model - a RNN conditioned on both the current sentence and the speaker - which created an overview of the entire dialogue and was therefore able to take into account a variable number of previous sentences as well as interpersonal dynamics between speakers in the dialogue. Training this model on the SWDA dataset with a depth of 2 (predictions were conditioned on the previous two sentences), Kalchbrenner and Blunsom (2013) achieved an accuracy of 73.9%.

The success achieved by Kalchbrenner and Blunsom (2013) inspired a host of other teams to experiment with similar architectures. One such approach (Lee and Dernoncourt, 2016) compared the semantic vector representations produced by sentence-level CNNs and RNNs for use in feeding into a two-layer artificial neural network (ANN), along with the representations and classifications of a variable number of preceding sentences. With the use of 200-dimensional GLoVe embeddings trained on Twitter, the best model achieved an accuracy of 73.1% on the SWDA dataset. CNNs were found to be more effective than RNNs, as accuracy fell to 69.6% with RNN representations.

Another such approach (Ji et al., 2016) involved the combination of a sentence-level LSTM neural network and a discourse-level latent variable model (LVM). The incorporation of the LVM allowed the model to treat the relationships between adjacent sentences as latent variables, while the RNN learned distributed representations for each sentence. Because the LVM focused the model on shallow discourse relations - relations between sentences that are adjacent - the model does not possess the flexibility to expand the number of previous sentences that it incorporates into predictions, but it was still able to achieve 77% accuracy on the SWDA dataset. This surpassed the 73.9% mark set by Kalchbrenner and Blunsom (2013).

2.3 Previous Work on the IDN Corpus

One of the first DA classification models for the IDN corpus (Chan et al., 2015) leveraged a RNN with Gated Recurrent Units (GRU). In this model, pre-trained GLoVe embeddings were fed into the GRU cells, the final hidden states of which were then used as sentence representations to be fed into a softmax classifier for multi-class classification. The model achieved an accuracy of 64.2% on the top five most common labels in the IDN dataset. An important difference to note between this approach and the approaches detailed in the previous section is that this model assumes the independence of sentences, resulting in the loss of discourse-level context information. As will be discussed in the next section, discourse-level contextual information is critical to distinguish between labels in the IDN dataset, and thus the incorporation of such information was a major area of opportunity to expand on this work.

The next model implemented for the IDN corpus (Roman, 2017) explored the opportunity detailed above through a bidirectional multi-layer LSTM. The LSTM operated at the sentence level, taking in information word-by-word for a given sentence to make a prediction, but its design also allowed it to capture discourse-level information as well. Additionally, the approach augmented the relatively small IDN dataset with the larger SWDA dataset by building two classifiers; one classifier was trained on SWDA data for the standard 42 labels, and the second classifier was trained on IDN with the features extracted on IDN using the SWDA classifier. In other words, a model was first pre-trained on SWDA and then leveraged to produce output features for training a second model on IDN. This approach enabled the model to overcome some of the challenges posed to deep learning models by the smaller size of the IDN dataset and achieve a 73% accuracy on the corpus.

3 Dataset

3.1 Corpus Composition

The IDN dataset was collected by the Stanford Center for Design Research (CDR). To gather this data, the CDR conducted experiments in which small groups of two to five people (mostly Stanford students) were assigned tasks requiring design team collaboration. These tasks ranged from developing a system to retrieve water from an underground water supply to improving massive open online (MOOC) education. The design team sessions were recorded, transcribed, segmented by sentence turns, and labeled with IDN dialogue act types. The dataset is still under development, but as of this writing it consists of conversations from 23 distinct teams for a total of 7230 sentence turns.

3.2 Notation Description

The IDN notation consists of 12 symbols that map the flow of conversation in a manner similar to the way notes of a music staff map the flow a song. The 12 symbols are as follows: move (M), support(SU), yesand(Y-A), question(Q), humour(H), silence(SI), block(BL), overcoming(OV), deflection(DE), block-support(B-S), yesandquestion(Y-Q), ignored(IG). Past research (Sonalkar et al., 2013) has shown these symbols to be effective in revealing patterns of design team interaction.

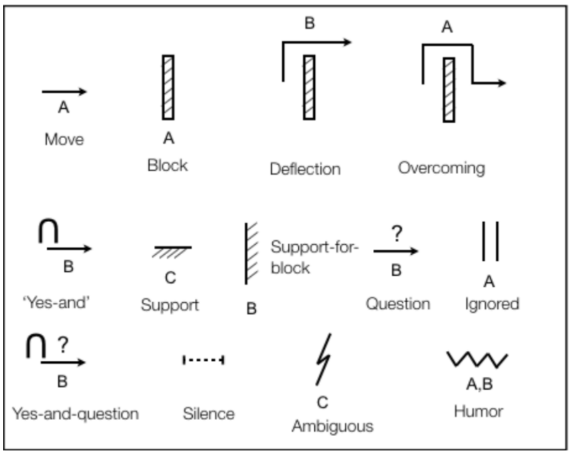


Figure 1: IDN dialogue acts and their symbols

While some of these dialogue acts are intuitive, several of the labels are not at first clear or possess significant nuance. In particular, Move is an expression that seeks to move the conversation in a given direction. In addition, Deflection is a move in a very specific context; it must be in response to a block, and has to present an alternative direction from the previous move. An important takeaway from this specification is that the IDN dialogue acts are explicitly context dependent at a discourse-level; that is, the criteria for many of the DAs do not stand on their own but are intelligible only in reference to the DAs of the preceding sentences. It is also important to note that because these DAs are tailored to the context of design interactions, it is difficult to compare results with existing literature on the SWDA or MRDA datasets.

3.3 Data Preprocessing

No data preprocessing is conducted on the IDN data for the CRF model. For the LSTM model, we represent each word in a sentence with its 300-dimensional word embedding, pre-trained using Facebook's fastText approach on a large corpus of Wikipedia files. Principal component analysis is then conducted on the embeddings to reduce dimensionality from 300 to 150. Finally, the dataset is partitioned with a 70% training, 15% validation and 15% testing split.

3.4 Biases and Noise

As can be seen in figure three, the dialogue act distribution of the IDN dataset is heavily skewed, with the top five labels accounting for nearly all of the samples. This of course poses challenges in that only these five most prevalent classes will be learned by a model as there are simply too few instances of the remaining labels. In addition, because the dataset is transcribed by multiple professionals, there are slight variations in the transcriptions of the speech into text. For instance, silence may be transcribed as an empty sentence or with contextual denotation in brackets; this decision is at the discretion of the transcriber and sometimes varies.

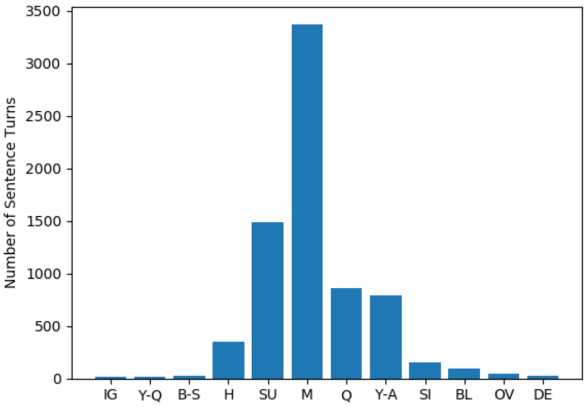


Figure 2: Frequency of IDN dialogue acts

4 Model

As noted in section 2, one of the most critical themes across successful models on both the IDN and SWDA dataset has been the incorporation of information on context dependencies at the discourse level. This theme was at the core of the innovative model developed by Kalchbrenner and Blunsom (2013), which introduced the notion of an architecture with double sequencing in which a model at the sentence level feeds sentence representations into a model at the discourse level. This theme was also critical in producing the best performance thus far on the IDN dataset (Roman, 2017). Intuitively, this finding makes sense, as a sentence in a conversation can only be fully understand in the context of the dialogue as a whole. This is particularly true with regard to design team collaborations, as team members must work together at a high level to achieve success, and IDN labels are explicitly dependent upon the labeling of preceding sentences. We thus develop CRF and LSTM models due to the ability of these models to effectively capture sequential information.

4.1 Conditional Random Field

A CRF is a Markov network over variables X∪Y which specifies a conditional distribution

$$P(y \mid x) = \frac{1}{Z(x)} \prod_{c \in C} \phi_c(x_c, y_c)$$

where phi is a factor that describes the joint probabilities between X and Y. In turn, a Markov field is a probability distribution p over variables x_1,..., x_n defined by an undirected graph G in which a node corresponds to variable x_i. The probability p has the form

$$p(x_1, .., x_n) = \frac{1}{Z} \prod_{c \in C} \phi_c(x_c)$$

where C denotes the set of fully connected subgraphs of G, and Z is a normalizing constant that ensures that probability distributions add up to 1.

In the case of CRFs for dialogue act classification tasks, the feature representation of a sentence is the evidence which then enables the CRF model to perform probabilistic inference to calculate the argmax(label, evidence) and output the label with the highest probability as the DA prediction.

Feature Engineering

Despite having access to the IDN corpus videos, we use only textual features in our model due to previous findings (Chan et al., 2015) on the IDN dataset that the use of audio features decreases model accuracy. The representation of each sentence thus consists of the following 10 features:

- First two words, filtering stop words.
- Last two words, filtering stop words.
- Last two words of the preceding sentence.
- First two words of the succeeding sentence
- Sentence length.
- ”Long” sentence.
- Punctuation of current sentence.
- Punctuation of preceding sentence.
- Punctuation of succeeding sentence.
- Latent summary of sentence.

The first five features build on findings of useful features in previous machine learning DA classification approaches (Ang et al., 2005). Before obtaining the first two and last two words of the sentence, we filter out stop words so as to increase the likelihood of capturing words with a greater impact upon the sentence meaning. For each of the first four features, sentences with less than two words are padded with the empty string. The final feature calculates sentence length by word count.

Closely related to the length of the sentence is a binary feature indicating whether or not a sentence is considered ”long”. This feature was crafted to help distinguish labels that tended to correspond to sentences consistently longer than other sentences in the transcripts, such as the ”yesand” label. The current, preceding, and succeeding punctuation are included for the primary purpose of distinguishing questions from other dialogue acts.

Finally, we incorporate the ”latent summary” of a sentence. The latent summary of a sentence is the single word from the sentence with a word embedding closest to the mean embedding of the sentence overall, obtained by calculating the mean of the sum the embeddings of every word in the sentence. The IDN id of this word is then included in the feature representation. By incorporating latent summaries, we hoped to leverage the intuition captured in word embeddings for the CRF model by providing a feature that captures which word most reflects the overall meaning of a sentence.

4.2 Discourse Level LSTM

Input

In addition to the CRF model, we implement a single-layer LSTM at the discourse level. Unlike the double-sequencing architecture developed by Kalchbrenner and Blunsom (2013) and further explored in additional studies (Lee and Dernoncourt, 2016; Ji et al., 2016) in which a sentence level deep learning model feeds sentence representations into a separate discourse-level model for prediction, this model takes as input vector representations of sentences derived from the summation of the 150-dimensional fastText embeddings (detailed in section 3.3) for the words in that sentence. The decision to use word embedding sums rather than to learn sentence representations was made because it was determined that it would be difficult, given the size of the IDN dataset, to learn parameters for both sentence representations and DA classifications. Finally, our model makes predictions utilizing a sentence depth of two, meaning representations for the preceding two sentences are incorporated into classifications.

Architecture

We define the LSTM cell at time step t to be a set of vectors in R^d. The formal definition of the cell is specified in the equations of figure three.

$$i_t = \sigma\left(W^{(i)}X_t + U^{(i)}h_{t-1} + b^{(i)}\right)$$

$$f_t = \sigma\left(W^{(f)}X_t + U^{(f)}h_{t-1} + b^{(f)}\right)$$

$$o_t = \sigma\left(W^{(o)}X_t + U^{(o)}h_{t-1} + b^{(o)}\right)$$

$$u_t = \tanh\left(W^{(u)}X_t + U^{(u)}h_{t-1} + b^{(u)}\right)$$

$$c_t = i_t \bigodot u_t + f_{(t)} \bigodot c_{t-1}$$

$$h_t = o_t \bigodot \tanh(c_t)$$

Figure 3: LSTM definition

In these equations, X_t is the d dimensional vector input at time t, W^{hh} and W^{hx} are weight matrices and σ represents the sigmoid function. Conceptually, i_t is the input gate, f_t is the forget gate, o_t is the output gate, c_t is the memory cell, and h_t is the hidden state. The interaction between these gates in the LSTM cell structure is demonstrated in figure 4.

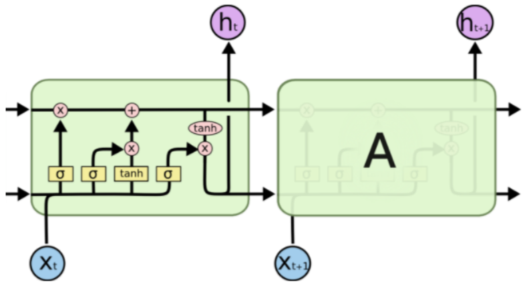


Figure 4: LSTM cell structure and parameters

We utilize a simple single-layer LSTM rather than a multi-layer LSTM as implemented in previous studies on the IDN dataset (Roman, 2017) because multiple LSTM layers are primarily used to capture longer dependencies between input items. However, previous work (Kalchbrenner and Blunsom, 2013; Ji et al., 2016) on the development of discourse-level context dependent models has found optimum performance achieved in these models when sequential information from only the previous one or two sentences is incorporated into predictions. This finding was reflected in performance testing on the development set for our LSTM model as well, as performance with the incorporation of information beyond two sentences previous to the current sentence. Therefore we maintain only a single layer LSTM since the range of dependencies we seek to capture is fairly short.

Hyper-parameters

The primary hyper-parameters tuned for our model was the number of hidden dimensions across the following range: {250, 500, 750, 1000}. We also interrogated the use of three activation functions: rectified linear activation, scaled exponential linear activation, and hyperbolic tangent. The best combination of these parameters selected via grid search was utilized for experiments. Training procedure was monitored by validation set performance. We found around 200 epochs were needed to achieve sufficient learning.

5 Results

5.1 Baseline

Given our motivation of incorporating sequential information at the discourse level in DA prediction on the IDN dataset, we choose as our baseline the non-sequential version of the CRF, logistic regression. Logistic regression is a special case of a CRF where the sequences are of length 1.

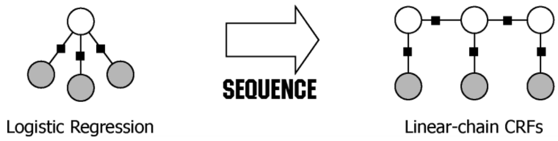


Figure 5: Graphical representation of logistic regression and conditional random field models

Input to the logistic regression model consists of a bag of words approach. Every word in the IDN vocabulary is assigned an index from zero to the vocabulary size. A sentence is then represented by a sparse vector of the same size as the vocabulary in which the i_{th} index represents the frequency at which the i_{th} word in the vocabulary appears.

5.2 Evaluation

Accuracy has served as the standard metric for performance throughout dialogue act classification literature on both the SWDA and IDN datasets. Because of this established standard, accuracy is the primary metric provided in our study as well. However, given that accuracy does not provide per-class performance granularity nor control for size imbalances in the classes, accuracy seems a poor metric for this task. These drawbacks are especially relevant given the unbalanced nature of the dataset detailed in section 3.4.

Therefore we provide the weighted F_1 score as well. We choose the weighted F_1 score over the macro-averaged F_1 score because, as depicted in figure two, several of the classes (ignored, deflection, etc.) have so few counts as to be nearly impossible for the model to learn. The low F_1 scores for these labels would then distort the overall macro-averaged score. One approach to this difficulty in past work (Chan et al., 2015) has been to only report performance on the top five labels.

	Accuracy (%)	Weighted F_1 Score
Majority Baseline	53.9	0.377
Logistic Regression Baseline	68.9	0.680
Unidirectional Multi-Layer LSTM (Roman, 2017)	70.0	N/A
Bidirectional Multi-Layer LSTM (Roman, 2017)	71.0	N/A
Context-Dependent Bidirectional Multi-Layer LSTM (Roman, 2017)	73.0	N/A
Discourse-Level Single-Layer LSTM	73.3	0.720
Conditional Random Field	**79.0**	**0.766**

Table 1: Model Accuracies and F_1 Scores

	Precision	**Recall**	**F1-Score**	**Support**
move	0.782	0.869	0.823	579
support	0.777	0.800	0.788	270
question	0.856	0.820	0.838	167
yesand	0.250	0.017	0.032	58
block	0.000	0.000	0.000	6
overcoming	0.000	0.000	0.000	3
deflection	0.000	0.000	0.000	1
yesandquestion	0.000	0.000	0.000	1
ignored	0.000	0.000	0.000	0
silence	0.000	0.000	0.000	0
block-support	0.000	0.000	0.000	0
humour	0.000	0.000	0.000	0
Average/Total:	**0.756**	**0.790**	**0.766**	**1085**

Table 2: CRF per-label performance scores

	Precision	**Recall**	**F1-Score**	**Support**
move	0.750	0.810	0.780	579
support	0.670	0.760	0.720	270
question	0.850	0.690	0.770	167
yesand	0.270	0.050	0.090	58
block	0.000	0.000	0.000	6
overcoming	0.000	0.000	0.000	3
deflection	0.000	0.000	0.000	1
yesandquestion	0.000	0.000	0.000	1
ignored	0.000	0.000	0.000	0
silence	0.000	0.000	0.000	0
block-support	0.000	0.000	0.000	0
humour	0.000	0.000	0.000	0
Average/Total:	**0.710**	**0.730**	**0.720**	**1085**

Table 3: LSTM per-label performance scores

6 Analysis

6.1 Model Comparison

As seen in table one, the CRF performs the best of all models developed for the IDN dataset so far by a significant margin with a 79.0% test accuracy, constituting an absolute gain of 10.1% over the logistic regression baseline. Perhaps more informative, however, is the 0.766 F_1 score of the CRF, which represents a 100% relative improvement over the majority baseline and a 10% relative improvement over the logistic regression baseline. The discourse-level LSTM just surpasses the performance of the context-dependent bidirectional multi-layer LSTM, but still lags significantly behind the CRF. The gap between the CRF and the LSTM is reinforced by the superior precision and recall performance of the CRF on every label other than ”yesand” in the IDN dataset, as displayed in tables two and three.

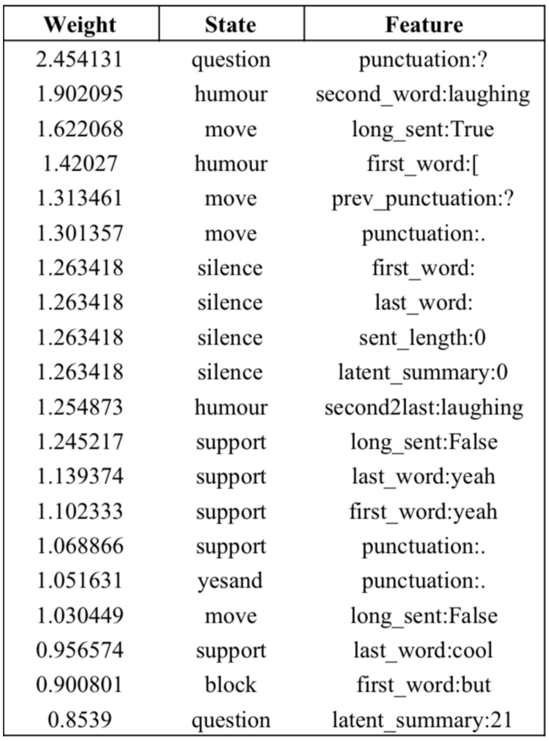
6.2 Comparison to Previous Work

In the same study in which Roman (2017) implements the bidirectional LSTM against which comparisons are made above, he also details the implementation of unidirectional and bidirectional multi-layer LSTMs that do not take discourse-level context dependencies into account. Rather, these models make a prediction for a sentence by processing that sentence word-by-word without the incorporation of information at the discourse (i.e. inter-sentence) level. As demonstrated in table 1, both the LSTM developed by Roman and the LSTM developed in this study that incorporated sequential information at the discourse-level outperformed the LSTM models that did not. This trend reinforces the primary theme drawn across previous work in the literature review; namely, that the incorporation of discourse-level sequential information is critical for the development of effective dialogue act classifiers.

However, even the LSTMs that incorporate discourse-level sequential information fail to obtain maximum performance within even 5% of the performance of the CRF. The 8% relative improvement of the CRF in comparison to the best of the LSTM models provides supporting evidence for a driving hypothesis behind this study; namely, that the smaller size of the IDN dataset would place particular pressure on deep learning approaches and would be best handled by a non-deep machine learning model. Results indicate that even as the field of dialogue act classification has been swept with the ”deep-learning tsunami” (Manning, 2015) throughout NLP, there are still critical areas in which traditional machine learning approaches are capable of superior performance.

6.3 Feature Analysis

Many of the top features learned by the CRF models are to be expected, particularly cues regarding question punctuation and transcript denotation for humour and silence. More interesting to note are the states missing from the table - overcoming, deflection, block-support, yesandquestion, and ignored are absent, while block and yesand each make a single appearance. This is likely due to two factors: the most clear factor is the dataset imbalance discussed previously, in which the five most common IDN labels account for over 97% of all sentences in the corpus. However, ”yesand” is the fourth most common label in the dataset with over 10% of the sentence share, and thus the model should have had sufficient opportunity to learn features for this label. One potential reason it did not is that sentences labeled ”yesand” lack the type of distinctive notation or characteristics that distinguish other labels such as question, humour, and support. Thus although the yesand label accounts for over 10% of sentences in the corpus, it could be the case that the CRF fails to distinguish between sentences labeled ”yesand” and more common sentences labeled ”move”, which dominates the dataset with nearly 50% of the share of the total number of sentences.



Weight	State	Feature
2.454131	question	punctuation:?
1.902095	humour	second_word:laughing
1.622068	move	long_sent:True
1.42027	humour	first_word:[
1.313461	move	prev_punctuation:?
1.301357	move	punctuation:.
1.263418	silence	first_word:
1.263418	silence	last_word:
1.263418	silence	sent_length:0
1.263418	silence	latent_summary:0
1.254873	humour	second2last:laughing
1.245217	support	long_sent:False
1.139374	support	last_word:yeah
1.102333	support	first_word:yeah
1.068866	support	punctuation:.
1.051631	yesand	punctuation:.
1.030449	move	long_sent:False
0.956574	support	last_word:cool
0.900801	block	first_word:but
0.8539	question	latent_summary:21

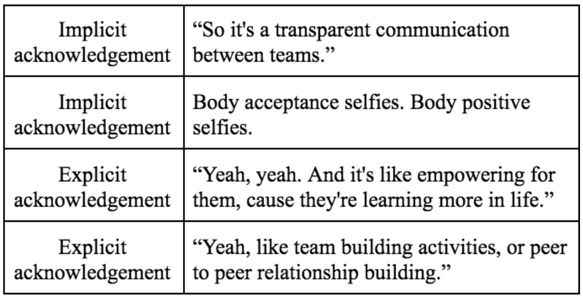
Table 4: CRF top 20 positive features

6.4 Error Analysis

The hypothesis articulated above is borne out by error analysis. As can be seen in tables two and three, both the CRF and LSTM models struggled with predicting the yesand label, earning F_1 scores of .03 and .09, respectively, on instances of those labels in the test set. The CRF, in particular, achieves only a 1.7% recall of yesand labels in the test set, indicating the yesand prediction is almost never made. When analyzing what predictions the CRF makes instead, we found that 84.2% of yesand misclassifications were incorrectly labeled as move, with the support and question labels accounting for the remaining misclassifications.

It is thus clear that the models are mostly incapable of distinguishing yesand labels from move labels. This limitation makes sense in context of the close relationship between these two labels specified in the IDN coding manual: ”A ’yes and’ response...derives from the external [group space], is modified and extended in the internal [mental space] and is again expressed in the external. The ’yes and’ response thus builds on the previous move that was expressed.” (Sonalkar et al., 2013) The complex internal-external processing the yesand label seeks to capture, in essence, specifies a sentence that accepts the content of the previous move and adds on to it.

Because a yesand sentence adds on to a previous move, there are often no clear dialogue markers or recurring words or phrases that distinguish a yesand sentence from a move sentence. Consider the misclassified yesand sentences in figure six.



Implicit acknowledgement	“So it's a transparent communication between teams.”
Implicit acknowledgement	Body acceptance selfies. Body positive selfies.
Explicit acknowledgement	“Yeah, yeah. And it's like empowering for them, cause they're learning more in life.”
Explicit acknowledgement	“Yeah, like team building activities, or peer to peer relationship building.”

Figure 6: Yesand label misclassification examples

The last two examples showcase yesand examples that begin by explicitly acknowledging the idea from the previous sentence (a ”move” sentence) before proceeding to expand on that idea in some way. This pattern is common in the dataset. However, more common are the types of sentences depicted in the first two examples in which no explicit acknowledgement of the previous idea is provided. Rather, the sentence’s relationship to the previous idea is understood to be implicit, resulting in no explicit textual cues that the yesand sentence builds on the ideas of the previous move.

This model limitation currently does not impede strong performance from the CRF, as the model still outperforms all LSTM approaches and matches state-of-the-art accuracies achieved on the SWDA and MRDA datasets as well. However, it raises questions about the ability of the model to maintain its performance as the IDN dataset continues to grow and labels such as block, block-support, deflection, and overcoming become more common in the corpus. As representation of these labels increases, the model will face similar challenges with label pairs such as question and block, support and block-support, and more pairs that cannot be known until the corpus expands.

7 Conclusion

In this article we have presented a CRF and LSTM approach to sequential dialogue act classification on the IDN dataset. We demonstrate that the incorporation of sequential information at the discourse level yields superior performance on the IDN dataset as has been found in previous work on the SWDA and MRDA datasets. Our LSTM model matches the performance of the best preexisting LSTM for IDN, while our CRF model achieves state-of-the-art results on the IDN dataset by a margin of more than 5% over other models.

7.1 Future Work

There are several opportunities for expansion on the current work. The first and most direct path of exploration would be to utilize GLoVe word embeddings rather than fastText for sentence representations in the LSTM model and compare the results to those achieved in this study, as GLoVe embeddings could potentially provide richer inputs for the LSTM model that have a significantly positive impact upon LSTM performance.

Another area of expansion could involve exploring methods of augmenting the smaller IDN dataset with with other, larger dialogue act datasets - such as the SWDA or MRDA corpora - through some sort of distant supervision. Roman (2017) begins to explore this direction in his 2017 study in which one of the components of input for his LSTM models consists of 42-length vectors output by a secondary SWDA-trained LSTM for each sentence in the IDN corpus. This approach can be used as a starting point to innovate methods of integrating the IDN dataset with other corpora.

Finally, as discussed in error analysis, developing a set of features capable of helping the CRF model detect and track implicit inter-sentence relationships (i.e. relationships that aren’t cued by specific words or phrases) would be critical.

Acknowledgements

We would like to thank Christopher Potts and Bill MacCartney for their guidance, feedback, and recommendation of resources throughout this study.

References

J. Ang, Yang Liu, and E. Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*. volume 1, pages 1061–1064. https://doi.org/10.1109/ICASSP.2005.1415300.

Ethan Chan, Aaron Loh, and Connie Zeng. 2015. Dialogue acts in design conversations .

Ozgur Eris. 2003. Perceiving, comprehending, and measuring design activity through the questions asked while designing. .

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. *arXiv preprint arXiv:1603.01913* .

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. *arXiv preprint arXiv:1306.3584* .

Yong Se Kim and Byung Gu Kang. 2003. Personal characteristics and design-related performances in a creative engineering design course. In *The 6th Asian Design Conference, Tsukuba, Japan*.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:1603.03827* .

Larry J Leifer and Martin Steinert. 2011. Dancing with ambiguity: Causality behavior, design thinking, and triple-loop-learning. *Information Knowledge Systems Management* 10(1-4):151–173.

Christopher D Manning. 2015. Computational linguistics and deep learning. *Computational Linguistics* 41(4):701–707.

Roman. 2017. Automatic idn dialog act tagging of design-team conversations .

Neeraj Sonalkar, Ade Mabogunje, and Larry Leifer. 2013. Developing a visual representation to characterize moment-to-moment concept generation in design teams. *International Journal of Design Creativity and Innovation* 1(2):93–108.

John C Tang. 1991. Findings from observational studies of collaborative work. *International Journal of Man-machine studies* 34(2):143–160.

Remko Van der Lugt. 2005. How sketching can affect the idea generation process in design group meetings. *Design studies* 26(2):101–122.

Matthias Zimmermann. 2009. Joint segmentation and classification of dialog acts using conditional random fields. In *Tenth Annual Conference of the International Speech Communication Association*.