tensorflow https://github.com/openai/generating-reviews-discovering-sentiment/tree/master
(for the pretrained mLSTM)(used as a frozen feature extractor)
https://github.com/guillitte/pytorch-sentiment-neuron
(pytorch port for it)

A Context-based Approach for Dialogue Act Recognition using Simple Recurrent Neural Networks

Chandrakant Bothe, Cornelius Weber, Sven Magg, and Stefan Wermter
Knowledge Technology, Department of Informatics,
University of Hamburg,
Vogt-Koelln-Str. 30, 22527 Hamburg, Germany
http://www.informatik.uni-hamburg.de/WTM/
{bothe,weber,magg,wermter}@informatik.uni-hamburg.de

Abstract

Dialogue act recognition is an important part of natural language understanding. We investigate the way dialogue act corpora are annotated and the learning approaches used so far. We find that the dialogue act is context-sensitive within the conversation for most of the classes. Nevertheless, previous models of dialogue act classification work on the utterance-level and only very few consider context. We propose a novel context-based learning method to classify dialogue acts using a character-level language model utterance representation, and we notice significant improvement. We evaluate this method on the Switchboard Dialogue Act corpus, and our results show that the consideration of the preceding utterances as a a context of the current utterance improves dialogue act detection.

Keywords: Dialogue Acts Detection, Recurrent Neural Networks, Context-based Learning

1. Introduction

In natural language processing research, the dialogue act (DA) concept plays an important role. Its recognition, in most cases, is considered a lexical-based or syntax-based classification at utterance-level. However, the discourse compositionality is context sensitive, meaning that the DA of an utterance can be elicited from the preceding utterances (Grosz, 1982). Hence, classifying only utterances is not enough because their DA class arises from their context. For example, the utterance containing only the lexical entry *'yeah'* might appear in several DA classes such as *Backchannel*, *Yes-Answer*, etc. For certain DA classes, the utterances are short, and most of them share similar lexical and syntactic cues (Jurafsky et al., 1998).

The aim of this article has two subgoals: first, we investigate the annotation process of DA corpora and review the modelling so far used for DA classification, and second, we present a novel model and compare its results with the state of the art. We propose to use context-based learning for the identification of the DA classes. First, we show the results without context, i.e., classifying only utterances. Including context leads to 3% higher accuracy. We use a simple recurrent neural network (RNN) for context learning of the discourse compositionality. We feed the preceding and current utterances to the RNN model to predict its DA class. The main contributions of this work are as follows:

- We provide detailed insight on the annotation and modelling of dialogue act corpora. We suggest to model discourse within the context of a conversation.
- We propose a context-based learning approach for DA identification. In our approach, we represent utterances by a character-level language model trained on domain-independent data.
- We evaluate the model on the Switchboard Dialogue Act (SwDA[1]) corpus and show how using context affects the results. For the SwDA corpus, our model achieved an accuracy of 77.3% compared to 73.9% as state of the art, where the context-based learning is used for the DA classification (Kalchbrenner and Blunsom, 2013).
- Benefits of using context arise from using only a few preceding utterances making the model suitable for dialogue system in real time, in contrast to feeding the whole conversation, which can achieve high accuracy, but includes future utterances (Liu et al., 2017; Kumar et al., 2017).

[1]Available at https://github.com/cgpotts/swda

2. Related Work

2.1. Annotation of Dialogue Act Corpora

Annotation Process and Standards: Research on dialogue acts became important with the commercial reality of spoken dialogue systems. There have been many taxonomies to it: speech act (Austin, 1962) which was later modified into five classes (Assertive, Directive, Commissive, Expressive, Declarative) (Searle, 1979), and the Dialogue Act Markup in Several Layers (DAMSL) tag set where each DA has a forward-looking function (such as Statement, Info-request, Thanking) and a backward-looking function (such as Accept, Reject, Answer) (Allen and Core, 1997). There are many such standard taxonomies and schemes to annotate conversational data, some of them follow the concept of discourse compositionality. These schemes are important for analysing dialogues or building a dialogue system (Skantze, 2007). However, there can never be a unique scheme that considers all aspects of dialogue.

Corpus Insight: We have investigated the annotation method for two corpora: Switchboard (SWBD) (Godfrey et al., 1992; Jurafsky et al., 1997) and ICSI Meeting Recorder Dialogue Act (MRDA) (Shriberg et al., 2004). They are annotated with the DAMSL tag set. The annotation includes not only the utterance-level but also the segmented-utterance labelling. The DAMSL tag set provides very fine-grained and detailed DA classes and follows the discourse compositionality. For example, the SWBD-DAMSL is the variant of DAMSL specific to the Switchboard cor-

Table 1: Example of a labeled conversation (portions) from the Switchboard Dialogue Act corpus

Speaker	Dialogue Act	Utterance
A	Backchannel	Uh-huh.
B	Statement	About twelve foot in diameter
B	Abandoned	and, there is a lot of pressure to get that much weight up in the air.
A	Backchannel	Oh, yeah.
B	Abandoned	So it's interesting, though.
		...
B	Statement-opinion	it's a very complex, uh, situation to go into space.
A	Agree/Accept	Oh, yeah,
		...
A	Yes-No Question	You never think about that do you?
B	Yes-Answer	Yeah.
A	Statement-opinion	I would think it would be harder to get up than it would be
B	Backchannel	Yeah.

pus. It distinguishes *wh-questions (qw)*, *yes-no questions (qy)*, *open-ended (qo)*, and *or-questions (qr)* classes, not just because these questions are syntactically distinct, but also because they have different forward functions (Jurafsky, 1997). A *yes-no question* is more likely to get a *"yes"* answer than a *wh-question*. This also gives an intuition that the answers follow the syntactic formulation of question which provides a context. For example *qy* is used for a question that from a discourse perspective expects a *Yes* or *No* answer.

Nature of Discourse in Conversation: The dialogue act is a context-based discourse concept that means the DA class of a current utterance can be derived from its preceding utterance. We will elaborate this argument with an example given in Table 1. Speaker *A* utters *'Oh, yeah.'* twice in the first portion, and each time it is labelled with two different DA labels. This is simply due to the context of the previously conversed utterances. If we see the last four utterances of the example, when speaker *A* utters the *'Yes-No Question'* DA, speaker *B* answers with *'yeah'* which is labelled as *'Yes-Answer'* DA. However, after the *'Statement-opinion'* from the same speaker, the same utterance *'yeah'* is labelled as *'Backchannel'* and not *'Yes-Answer'*. This gives evidence that when we process the text of a conversation, we can see the context of a current utterance in the preceding utterances.

Prosodic Cues for DA Recognition: It has also been noted that prosodic knowledge plays a major role in DA identification for certain DA types (Jurafsky et al., 1998; Stolcke et al., 2000). The main reason is that the acoustic signal of the same utterance can be very different in a different DA class. This indicates that if one wants to classify DA classes only from the text, the context must be an important aspect to consider: simply classifying single utterances might not be enough, but considering the preceding utterances as a context is important.

2.2. Modelling Approaches

Lexical, Prosodic, and Syntactic Cues: Many studies have been carried out to find out the lexical, prosodic and syntactic cues (Stolcke et al., 2000; Surendran and Levow, 2006; O'Shea et al., 2012; Yang et al., 2014). For the SwDA corpus, the state-of-the-art baseline result was 71% for more than a decade using a standard Hidden Markov Model (HMM) with language features such as words and n-grams (Stolcke et al., 2000). The inter-annotator agreement accuracy for the same corpus is 84%, and in this particular case, we are still far from achieving human accuracy. However, words like *'yeah'* appear in many classes such as *backchannel*, *yes-answer*, *agree/accept* etc. Here, the prosodic cues play a very important role in identifying the DA classes, as the same utterance can acoustically differ a lot which helps to distinguish the specific DA class (Shriberg et al., 1998). There are several approaches like traditional Naive Bayes and HMM models, which use minimal information and certainly ignore the dependency of the context within the communication (Grau et al., 2004; Tavafi et al., 2013). They achieved 66% and 74.32% respectively on the SwDA test set.

Utterance-level Classification: Perhaps most research in modelling dialogue act identification is conducted at utterance-level (Stolcke et al., 2000; Grau et al., 2004; Tavafi et al., 2013; Ji et al., ; Khanpour et al., 2016; Lee and Dernoncourt, 2016). The emergence of deep learning also gave a big push to DA classification. In a natural language conversation, most utterances are very short; hence it is also referred to as short text classification. Lee and Dernoncourt (2016) achieved 73.1% accuracy on the SwDA corpus by using advanced deep learning frameworks such as RNNs and convolutional neural networks (CNN) with word-level feature embeddings.

A Novel Approach: Context-based Learning: Classifying the DA classes at single utterance-level might fail when it comes to DA classes where the utterances share similar lexical and syntactic cues (words and phrases) like the *backchannel*, *yes-answer* and *accept/agree* classes. Some researchers proposed an utterance-dependent learning approach (Kalchbrenner and Blunsom, 2013; Ji et al., ; Kumar et al., 2017; Tran et al., 2017; Liu et al., 2017; Ortega and Vu, 2017; Meng et al., 2017). Kalchbrenner and Blunsom (2013) and Ortega and Vu (2017) have proposed context-based learning, where they represent the utterance as a compressed vector of the word embeddings using CNNs and use these utterance representations to model discourse within a conversation using RNNs. In their architecture,

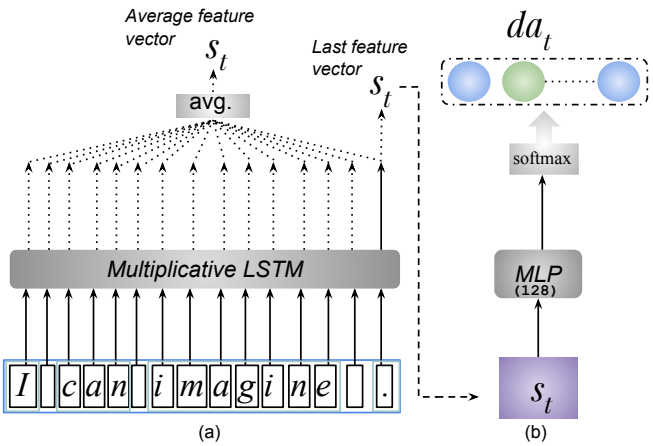

Figure 1: (a) Multiplicative LSTM (mLSTM) character-level language model to produce the sentence representation s_t. The character-level language model is pre-trained and produces the feature (hidden unit states of mLSTM at the last character) or average (average of all hidden unit states of every character) vector representation of the given utterance. (b) Utterance-level classification using a simple MLP layer with a *softmax* function (our baseline model).

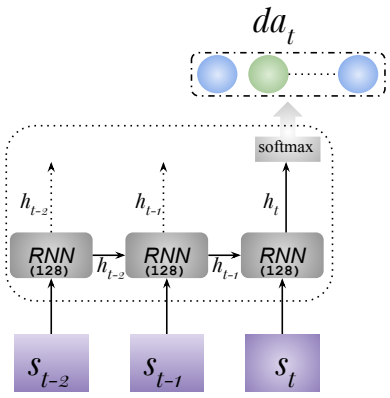

Figure 2: The RNN setup for learning the dialogue act recognition with the previous sentences as context. s_t is an utterance representation derived with a character-level language model and has a dialogue act label da_t. s_{t-1} and s_{t-2} are the preceding utterances of s_t. The RNN is trained to learn the recurrency through previous utterances s_{t-1} and s_{t-2} derived as h_{t-1} and h_{t-2} as a context to recognize the dialogue act of current utterance s_t which is represented by h_t used to detect da_t.

they also give importance to turn-taking by providing the speaker identity but do not analyse their model in this regard. This approach achieves about 73.9% accuracy on the SwDA corpus. In another line of research (Ji et al., ; Kumar et al., 2017), authors claim that their models take care of the dependency of the utterances within a conversation. Ji et al. (2016) use discourse annotation for the word-level language modelling on the SwDA corpus and also highlight a limitation that this approach is not scalable to large data. In other approaches a hierarchical convolutional and recurrent neural encoder model are used to learn utterance representation by feeding a whole conversation (Kumar et al., 2017; Liu et al., 2017). The utterance representations are further used to classify DA classes using the conditional random field (CRF) as a linear classifier. The model can see the past and future utterances at the same time within a conversation, which limits usage in a dialogue system where one can only perceive the preceding utterance as a context but does not know the upcoming utterances. Hence, we use a context-based learning approach and regard the 73.9% accuracy (Kalchbrenner and Blunsom, 2013) on the SwDA corpus as a current state of the art for this task.

3. Our Approach

Our approach takes care of discourse compositionality while recognising dialogue acts. The DA class of the current utterance is predicted using the context of the preceding utterances. We represent each utterance by the hidden state of the multiplicative recurrent neural network trained on domain-independent data using a character-level language model. We use RNNs to feed the sequence of the utterances and eventually predict the DA class of the corresponding utterance.

3.1. Utterance Representation

Character-level encoding allows processing words and whole sentences based on their smallest units and still capturing punctuation and permutation of words. We represent a character-level utterance by encoding the whole sentence with a pre-trained character language model[2]. This model consists of a single multiplicative long-short-term memory (mLSTM) network (Krause et al., 2016) layer with 4,096 hidden units. The mLSTM is composed of an LSTM and a multiplicative RNN and considers each possible input in a recurrent transition function. It is trained as a character language model on ∼80 million Amazon product reviews (Radford et al., 2017). We sequentially input the characters of an utterance to the mLSTM and take the hidden state values after the last character as shown in Figure 1 (a). The hidden vector s_t obtained after the last character is called the last feature vector, as it stores the information related to the character language model and the sentiment of the utterance. However, it was shown that the average vector over all characters in the utterance works better for emotion detection (Lakomkin et al., 2017). Hence, we extract the last feature vector and also the average feature vector representations for each utterance. We classify these representations with a multi-layer perceptron (MLP) as shown in Figure 1 (b). The results are shown in Table 2. The standard deviation (SD) is computed over ten runs. The average vector seems to carry more information related to the DA; hence we use it for future experiments. There is an advantage of using domain-independent data: it is rich regarding features being trained on big data, perhaps surpassing the limitation of scalability as mentioned in Ji et al. (2016).

[2] https://github.com/openai/generating-reviews-discovering-sentiment

Table 2: Accuracy of the dialogue act identification using the character-level language model utterance representation for 42 classes using a single MLP layer with 64 neurons.

Model input	Acc.(%)	SD
Last feature vector	71.48	0.28
Average feature vector	**73.96**	0.26
Concatenated vector	73.18	0.31

3.2. Context Learning with RNNs

We apply context learning with the help of RNNs. As shown in Figure 2, the utterances with their character-level language model representation s_t are fed to the RNN with the preceding utterances (s_{t-1}, s_{t-2}) being the context. We use the RNN, which gets the input s_t, and stores the hidden vector h_t at time t (Elman, 1990), which is calculated as:

$$h_t = f\left(W_h * h_{t-1} + I * s_t + b\right) \quad (1)$$

where $f()$ is a sigmoid function, W_h and I are recurrent and input weight matrices respectively and b is a bias vector learned during training. h_t is computed using the previous hidden vector h_{t-1} which is computed in a same way for preceding utterance s_{t-1}. The output da_t is the dialogue act label of the current utterance s_t calculated using h_t, as:

$$da_t = g\left(W_{out} * h_t\right) \quad (2)$$

where W_{out} is the output weight matrix. The weight matrices are learned using back-propagation through time. The task is to classify several classes; hence we use a $softmax$ function $g()$ on the output. The input is the sequence of the current and preceding utterances, e.g., s_t, s_{t-1}, and s_{t-2}. We reset the RNN when it sees the current utterance s_t. We also give the information related to a speaker to let the network find the change in the speaker's turn. The speaker id 'A' is represented by [1,0] and id 'B' by [0,1] and it is concatenated with the corresponding utterances s_t.

The Adam optimiser (Kingma and Ba, 2014) was used with a learning rate $1e-4$, which decays to zero during training, and clipping gradients at norm 1. Early stopping was used to avoid over-fitting of the network, 20% of training samples were used for validation. In all learning cases, we minimise the categorical cross-entropy.

3.3. Results

We follow the same data split of 1115 training and 19 test conversations as in the baseline approach (Stolcke et al., 2000; Kalchbrenner and Blunsom, 2013). Table 3 shows the results of the proposed model with several setups, first without the context, then with one, two, and so on preceding utterances in the context. We examined different values for the number of the hidden units of the RNN, empirically 64 was identified as best and used throughout the experiments. We also experimented with the various representations for the speaker id that is concatenated with the respective utterances but could find no differences. As a result, our proposed model uses minimal information for the context. The performance increases from 74% to about 77% with context. We run each experiment for ten times and take the average. The model shows robustness providing minimal variance, and using a minimum number of preceding utterances as a context can produce fair results.

Table 3: Accuracy of the dialogue act identification with the context-learning approach.

Model setup	Acc.(%)	SD
Baseline		
Most common class	31.50	
Related previous work		
Stolcke et al. (2000)	71.00	
Kalchbrenner and Blunsom (2013)	73.90	
Our work		
Our baseline (without context)	73.96	0.26
RNN (1 utt. in context w. SpeakerID)	76.48	0.33
RNN (1 utt. in context)	76.57	0.28
RNN (2 utts. in context)	76.81	0.24
RNN (3 utts. in context)	**77.34**	0.21
RNN (4 utts. in context)	77.28	0.22

4. Conclusion

In this article, we detail the annotation and modelling of dialogue act corpora, and we find that there is a difference in the way DAs are annotated and the way they are modelled. We argue to generalise the discourse modelling for conversation within the context of communication. Hence, we propose to use the context-based learning approach for the DA identification task. We used simple RNN to model the context of preceding utterances. We used the domain-independent pre-trained character language model to represent the utterances. We evaluated the proposed model on the Switchboard Dialogue Act corpus and show the results with and without context. For this corpus, our model achieved an accuracy of 77.34% with context compared to 73.96% without context. We also compare our model with Kalchbrenner and Blunsom (2013) who used the context-based learning approach achieving 73.9%. Our model uses minimal information, such as the context of a few preceding utterances which can be adapted to an online learning tool such as a spoken dialogue system where one can naturally see the preceding utterances but not the future ones. This makes our model suitable for human-robot/computer interaction which can be easily plugged into any spoken dialogue system.

5. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement number 642667 (SECURE).

6. Bibliographical References

Allen, J. and Core, M. (1997). Draft of DAMSL: Dialogue Act Markup in Several Layers.

Austin, J. L. (1962). *How to Do Things with Words*. Oxford University Press.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520.

Grau, S., Sanchis, E., Castro, M. J., and Vilar, D. (2004). Dialogue act classification using a Bayesian approach. In *9th Conference Speech and Computer SPECOM 2004*.

Grosz, B. J. (1982). Discourse Analysis. *Sublanguage. Studies of Language in Restricted Semantic Domains*, pages 138–174.

Ji, Y., Haffari, G., and Eisenstein, J.). A Latent Variable Recurrent Neural Network for Discourse Relation Language Models. In *Proc. of NAACL-HLT*, pages 332–342.

Jurafsky, D., Shriberg, E., and Biasca, D. (1997). Switchboard Dialog Act Corpus. Technical report, International Computer Science Inst. Berkeley CA.

Jurafsky, D., Shribergy, E., Fox, B., and Curl, T. (1998). Lexical, Prosodic, and Syntactic Cues for Dialog Acts. In *The ACL/COLING Workshop on Discourse Relations and Discourse Markers*.

Jurafsky, D. (1997). Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, draft 13. *Technical Report 97-01, University of Colorado Institute of Cognitive Science*, pages 225–233.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent Convolutional Neural Networks for Discourse Compositionality. In *Workshop on Continuous Vector Space Models and their Compositionality, ACL*, pages 119–126.

Khanpour, H., Guntakandla, N., and Nielsen, R. (2016). Dialogue Act Classification in Domain-Independent Conversations Using a Deep Recurrent Neural Network. In *COLING*, pages 2012–2021.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv: 1412.6980*.

Krause, B., Lu, L., Murray, I., and Renals, S. (2016). Multiplicative LSTM for sequence modelling. *Workshop track - ICLR 2016*.

Kumar, H., Agarwal, A., Dasgupta, R., Joshi, S., and Kumar, A. (2017). Dialogue Act Sequence Labeling using Hierarchical encoder with CRF. *arXiv:1709.04250v2*.

Lakomkin, E., Bothe, C., and Wermter, S. (2017). GradAscent at EmoInt-2017: Character and Word Level Recurrent Neural Network Models for Tweet Emotion Intensity Detection. In *Proc. of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis at EMNLP 2017*, pages 169–174. ACL.

Lee, J. Y. and Dernoncourt, F. (2016). Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. *arXiv preprint arXiv:1603.03827*.

Liu, Y., Han, K., Tan, Z., and Lei, Y. (2017). Using Context Information for Dialog Act Classification in DNN Framework. In *Proc. of the 2017 Conference on EMNLP*, pages 2160–2168. ACL.

Meng, Z., Mou, L., and Jin, Z. (2017). Hierarchical RNN with Static Sentence-Level Attention for Text-Based Speaker Change Detection. In *Proc. of ACM Conference on Information and Knowledge Management*, pages 2203–2206.

Ortega, D. and Vu, N. T. (2017). Neural-based Context Representation Learning for Dialog Act Classification. *Proc. of the SIGDIAL 2017 Conference*, pages 247–252.

O'Shea, J., Bandar, Z., and Crockett, K. (2012). A Multi-classifier Approach to Dialogue Act Classification Using Function Words. In *Transactions on Computational Collective Intelligence VII*, pages 119–143. Springer.

Radford, A., Jozefowicz, R., and Sutskever, I. (2017). Learning to Generate Reviews and Discovering Sentiment. *arXiv: 1704.01444*.

Searle, J. R. (1979). *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press.

Shriberg, E., Stolcke, A., Jurafsky, D., Coccaro, N., Meteer, M., Bates, R., Taylor, P., Ries, K., Martin, R., and Van Ess-Dykema, C. (1998). Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? *Language and Speech*, 41(3-4):443–492.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvey, H. (2004). The ICSI meeting recorder dialog act (MRDA) corpus. Technical report, International Computer Science Inst. Berkeley CA.

Skantze, G. (2007). Error Handling in Spoken Dialogue Systems-Managing Uncertainty, Grounding and Miscommunication: Chapter 2, Spoken Dialogue Systems. *KTH Computer Science and Communication*.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., and Meteer, M. (2000). Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339–373.

Surendran, D. and Levow, G.-A. (2006). Dialog act tagging with support vector machines and hidden Markov models. In *Interspeech – ICSLP*.

Tavafi, M., Mehdad, Y., Joty, S. R., Carenini, G., and Ng, R. T. (2013). Dialogue Act Recognition in Synchronous and Asynchronous Conversations. In *SIGDIAL Conference*, pages 117–121. ACL.

Tran, Q. H., Zukerman, I., and Haffari, G. (2017). Preserving Distributional Information in Dialogue Act Classification. In *Proc. of Conference on EMNLP*, pages 2141–2146. ACL.

Yang, X., Liu, J., Chen, Z., and Wu, W. (2014). Semi-supervised Learning of Dialogue Acts Using Sentence Similarity Based on Word Embeddings. In *Proc. of International Conference on Audio, Language and Image Processing*, pages 882–886.

Appendix: Analysis of the state of the RNN

We also analyze the internal state h_t of the RNNs for a two-utterance setup. We plot them on a 2D graph with the t-SNE algorithm for the first 2,000 utterances of the SwDA test set. Figure 3 shows the clusters of all the DA classes. The classes which do not share any information are grouped without any interference such as *Non-verbal*, and *Abandoned*. Figure 4 shows some particular classes with utterances in their vector spaces, the (1) current utterance and (2) a preceding utterance in the context.

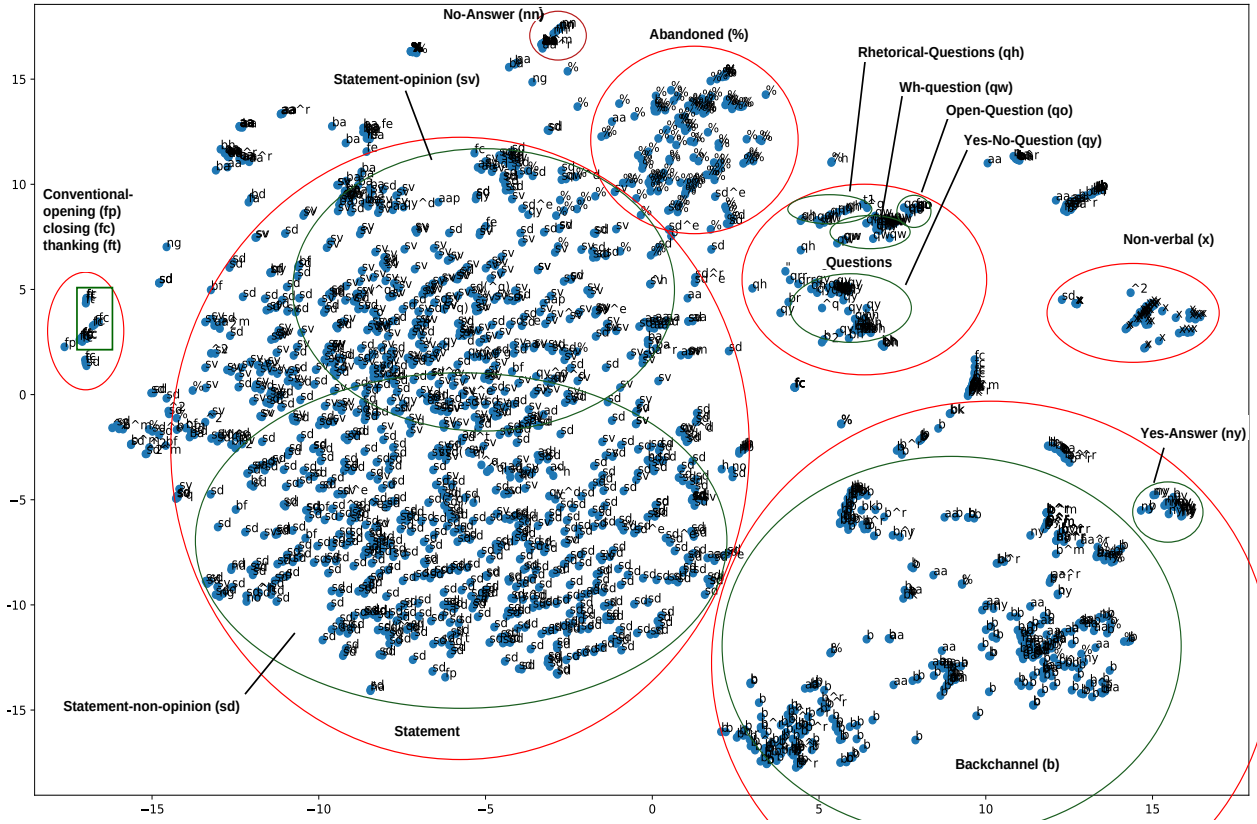

Figure 3: Clusters of all classes. Big clusters belong to the dominating *Statement* classes, *sv* and *sd*. The *Question* classes, *qy*, *qw*, *qh* and *qo* are clustered within the big class. The classes *Backchannel*, *Yes-answers*, and *Agree/Accept* share a lot of syntactic information hence they are clustered together, and our approach makes those classes separable within the cluster.

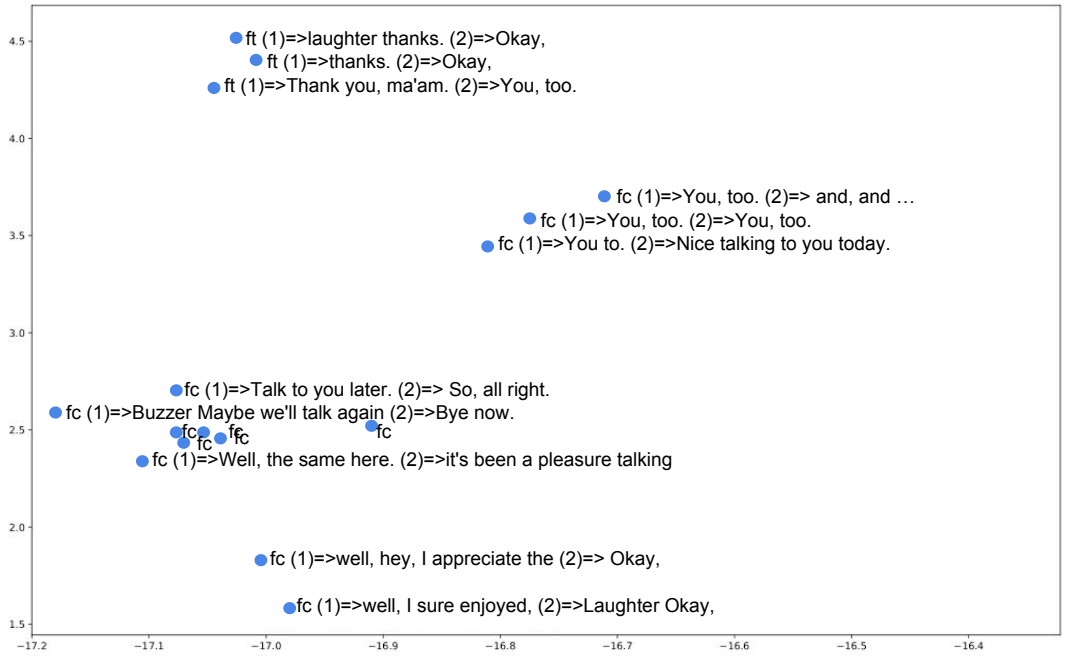

Figure 4: A blowup of the rectangle in Figure 3 from the *Conventional closing (fc)* and *thanking (ft) function* classes with their utterances. For readability, some utterances have been omitted and we show only the labels. These are examples of the context-sensitive dialogues, where we can see one cluster of the *ft* class and three groups of the *fc* class.